Improving Worker Performance with Human-Centered Data Science

by

Teng Ye

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2021

Doctoral Committee:

Professor Qiaozhu Mei, Chair
Professor Yan Chen
Professor Rayid Ghani
Professor Jieping Ye

Teng Ye

tengye@umich.edu

ORCID iD: 0000-0002-5402-6412

To my parents and friends.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Qiaozhu Mei. Without his open-minded research philosophy, his incisive options about every problem, and his tremendous help and guidance, I would not have been able to complete this interdisciplinary dissertation. From him, I learned how to tackle meaningful real-world problems while keeping high scientific standards, how to think independently while understanding and supporting others in collaborations, and how to provide constructive guidance to students while empowering them with trust and freedom in their research. I could not have wished for a better advisor! I am grateful for learning such a great deal from Qiaozhu. His insights, personality, and philosophy will positively influence me for a lifetime. I hope I can carry these forward to my own students and become a great advisor like him.

I would like to sincerely thank other members of my committee: Prof. Yan Chen, Prof. Rayid Ghani, and Prof. Jieping Ye, for sharing their broad vision, for providing insightful feedback, and for spending their valuable time on my dissertation.

I am especially indebted to Prof. Yan Chen. I learned so much from her high standards, great passion, and strong dedication to research. Her expertise in causal inference, experimental design, and theories in behavioral economics helped a lot with this dissertation. She has definitely been a role model in both research and life for junior female scholars like me. I want to express my extreme appreciation for her continuous support in my studies, career, and life.

I want to specially thank Prof. Rayid Ghani, who introduced me to the realm of

data science for social good when I was a fellow of the DSSG at the University of Chicago. I benefited a lot from his unique vision of data science, from his sincere care for ordinary people, and from his high commitment to approaching social problems with technical solutions. Rayid has given me a lot of encouragement and great advice. It is my privilege to have him as my mentor and collaborator.

I worked with Prof. Jieping Ye for about two years at Didi Chuxing, during which time I acquired the skills to connect machine learning and causal inference, making this dissertation possible. I deeply thank Jieping! I have learned so much from his insightful perspective on machine learning, his acute sense of applications in business, and his super supportive and inspiring personality. His vision in bridging the frontier techniques with real-world applications and products has influenced me a lot!

It is my great honor to be part of the Foreseer Group, where the discussions and feedback have quite often sparked my ideas. I would like to thank Yang Liu, Danny Tzu-Yu Wu, Zhe Zhao, Yue Wang, Cheng Li, Sam Carton, Wei Ai, Shiyan Yan, Tera Reynold, Jiaqi Ma, Cristina Garbacea, Huoran Li, Xuedong Li, Yutong Xie, Yachuan Liu, Sui Li, V.G. Vinod Vydiswaran, Jian Tang, Xuan Lu, Jie Gui, Kai Xiong, Yumin Zhang, Jingnan Zheng, and Junhui Jin. This sweet, inclusive, and intellectual home has been a source of continuous inspiration and support.

At the School of Information, University of Michigan, I have received help and support from interacting with numerous collaborators, colleagues, and friends. I am especially grateful for Prof. Thomas Finholt, Prof. Paul Resnick, and Prof. Nicole Ellison, all of whom have provided insightful comments and helpful guidance towards my study and career, helping me arrive where I am now. I also want to thank Prof. Lionel Robert, Prof. Katharina Reinecke, and Dr. Sangseok You for introducing me to the theories and mysterious frontiers of human-computer interaction. I thank the organizers and participants of the CS/DSS seminar and the BEE lab meetings, especially Prof. Daniel Romero, Prof. Kevyn Collins-Thompson, Prof. Paramveer

Dhillon, Prof. David Jurgens, Prof. Alain Cohn, and Prof. Tanya Rosenblat, for their helpful discussions. I am so appreciative to have many friends at UMSI with whom I can share tears and joys, including Youyang Hou, Fangzhou Zhang, Shiqing He, Ming Jiang, Xuan Zhao, Carrie Xu, Yingzhi Liang, Tao Dong, Zhuofeng Wu, Xinyan Zhao, Linfeng Li, Yixin Zou, Hao Peng, Yulin Yu, Yan Chen, Ruihan Wang, Chanda Phelan, Ryan Burton, Xin Rong, Daphne Chang, Jasmine Jones, Fengmin Hu, Yichi Zhang, Lei Zhang, Tawfiq Ammari, Charles Senteio, Hariharan Subramonyam, Carol Moser, Cindy Lin, Jean Hardy, Joey Hsiao, Rasha Alahmad, and Priyank Chandra.

I am also extremely grateful for the help and care I have received across the campus. My special thanks go to Prof. Yan Huang, who was a committee member of my field prelim and who has always shared constructive advice and feedback towards my research and career. I am privileged to have her guidance and friendship. I sincerely appreciate the endless support from my good friends at U-M, Yan Chang, Tong Guo, Lai Wei, Shiya Song, Ziyong Lin, and Luhong Wang. They have witnessed every step of my PhD journey.

I deeply thank my colleagues during my fellowship and internship, especially Joe Walsh and Rebecca Johnson at DSSG; and Lingyu Zhang, Ning Luo, Lulu Zhang, Miao Liang, Tao Song, Quanjiang Wan, Lin Zeng, Hui Wang, Junling Zhang, Yan Liu, Hongtu Zhu, and Dan Li at DiDi. Without their support, I could not have finished this dissertation. My gratitude goes to DSSG and DiDi (via MIDAS) for their generous financial support. I thank the participants of the UChicago Rising Star in Data Science Workshop, especially Prof. Chenhao Tan, for their great suggestions.

In addition, I sincerely thank Prof. Huizhang Shen, Prof. Lichuan Han, and Prof. Pengzhu Zhang for their advisorship, help, and continuous support since when I was a student at Shanghai Jiao Tong University.

Finally, I would like to thank my parents for their deepest love. I would not have been anywhere close to where I am right now without their unfailing support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Advances in information technologies not only provide novel tools to support work in the traditional sectors; they also create additional employment opportunities in the modern workforce where work contexts have been largely changed. All these changes call for new efforts to study worker performance. Indeed, information technologies, especially data science techniques, render unprecedented large-scale rich data and sophisticated analytic tools to investigate worker performance. However, it remains unclear how we can combine the strengths of big data analytics in data science and our existing knowledge in social science to enhance worker performance.

In this dissertation, we propose a human-centered data science framework that integrates machine learning, causal inference, field experiments, and social science theories: First, machine learning (with counterfactual reasoning) enables the prediction (and explanation) of human behavior in work practice via large-scale data analysis. Existing insights from social theories can further enhance its predictive power by informing feature construction, model architecture, and model explanation. Field experiments can help to evaluate the effectiveness of these models in real-world practices. Second, field experiments perform precise interventions and establish causality with randomized controlled trials. Yet, the experimental analysis mainly supports the understanding of treatment effects at aggregate levels, such as average treatment effect. Machine learning empowers more sophisticated analyses of experimental data by revealing heterogeneous effects at a finer granularity, such as individual treatment effects. Third, while these data-driven discoveries complement social science theories

and provide rich insights for describing, explaining, and predicting human behavior, they require rigorous analytic tools, such as experiments and machine learning, to validate or disconfirm their applicability in specific contexts. In addition to testing theories, causal insights derived from field experiments and counterfactual machine learning models could support the development of new theories that better reflect reality.

To exemplify the various applications of this framework in both traditional sectors and in the modern workforce, we present three empirical studies: developing machine learning models to improve the outreach performance for government specialists, leveraging a field experiment to enhance the performance of the gig economy workers, and using counterfactual machine learning to unpack individual treatment effects of field experiments on worker performance in the gig economy. These studies illustrate that the framework of human-centered data science is effective and flexible in increasing worker performance.

CHAPTER I

Introduction

The last three decades, from 1990 to 2019, have seen a great increase in the labor force participation from 2.3 billion (43.5% of the world population) to 3.5 billion (45.6% of the world population) people across the world [79, 80]. As of 2019, the United States' labor force alone has involved 167 million workers, which represents about 50% of the total American population ([30, 79]). Investigating the performance of workers has been increasingly critical for the society. Given its importance, worker performance has been studied for centuries, establishing a great understanding of related predictors, mechanisms, and interventions. However, the vast changes in work contexts and work-support tools brought by modern information technologies call for new examinations and insights on worker performance.

As highlighted by the World Bank, advances in information technologies have been changing the nature of work ([12]). On one hand, the development of new technologies is reshaping the existing work in the traditional sectors by providing more tools to support work. For example, Zoom and other computer-mediated tools have been widely used to support distributed collaboration while we have worked at home during the period of COVID-19. As another example, machine-learning algorithms have been leveraged to help doctors with medical diagnosis [86]. These are just two examples among the booming variety of technology-supported tools for

work. Nonetheless, we are far from fully understanding how to facilitate work with information technologies, especially data science, and how to do this better.

On the other hand, information technologies create new work contexts. For example, the fast growing gig economy platforms, such as Uber, Fiverr, and TaskRabbit, have facilitated the birth of a modern work force. Compared to the jobs in traditional companies and organizations, the gig-economy jobs are typically characterized by low barriers to join, high flexibility in time and location, and high autonomy. These new work contexts blur the boundaries between full-time employees and casual labor, and even between work and leisure [126], challenging the traditional definition of work. Therefore, it is largely unknown whether what we have learned about work performance holds in the new contexts and whether there would be novel insights for the modern work force. As of 2019, gig economy platforms had attracted 57 million workers in the United States alone, and the jobs on such platforms have been frequently referred to as the future of work (e.g., [99]). The broad participation and wide recognition further increase the significance of investigating worker performance in the modern era.

Therefore, in this dissertation, we unpack such mysteries about worker performance. Specifically, we explore how to improve worker performance.

In approaching this question, we must understand that big data, online platforms, and the advanced data science techniques have provided us unprecedented opportunities. First, modern technologies allow for the digitization of comprehensive records regarding workers and customers, organizations, and societal contexts. This large-scale documentation of data at a fine granularity provides new possibility to more accurately capture behavioral and contextual characteristics, facilitating the investigation of worker performance. Second, the online platforms not only collect big data, they also very importantly render easy access to precise and sophisticated behavioral interventions. The ability to control interactions between users and the platform at

both individual and session interaction levels affords exciting options for designing new interventions in randomized field trials, and field trials can add valuable experimental records to the general big data documentation. Third, to make sense of the vast documentation of data and discover actionable insights, data science provides advanced analytic tools, such as machine learning.

These emerging opportunities brought by big data, online platforms, and data science techniques complement our existing knowledge and methods to enhance worker performance. In this dissertation, we take such opportunities by proposing a human-centered data science framework to improve worker performance.

1.1 A Human-Centered Data Science Framework

The special challenge of human-centered data science framework lies in that the goal of this framework is to describe, predict, understand, and ultimately promote human behaviors, which are highly complex and heterogeneous. Therefore, this framework must be able to incorporate our existing knowledge about human factors — commonly reflected in social science theories — in addition to data science techniques and behavioral interventions. Specifically, as shown in Figure 1.1[1], this framework consists of three major components centering around human behavior (i.e., worker performance in our context): machine learning, field experiment, and social science theories.

1.1.1 Key Components

Machine Learning

Machine learning enables the prediction of human behavior via large-scale data analysis. While some machine learning models make it easy to explain the predictive

[1]Note that research potentials represented by the dashed lines are not covered by the empirical studies in this dissertation.

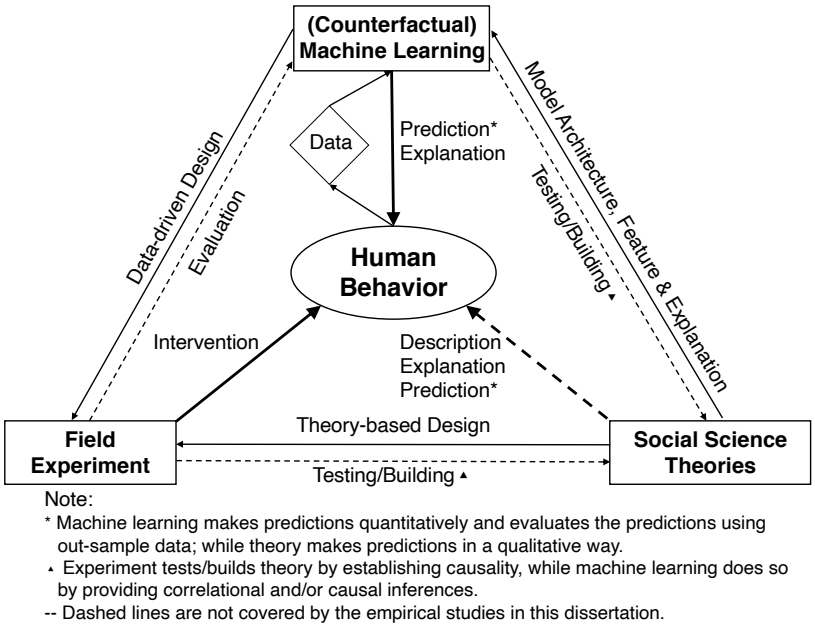


Figure 1.1: A human-centered data science framework

power of features, counterfactual reasoning further strengthens the explainability of machine learning models. Instead of predicting what will happen as general machine learning models, counterfactual machine learning predicts what will happen if something is changed, making the predicted results more interpretable.

Social Science Theories

Social science theories provide rich insights in describing, explaining, and predicting human behavior. These theories generally refer to theories in social sciences, such as psychology, economics, organization, marketing, and management.

Field Experiment

Field experiments are randomized control trials that are carried out in real-world settings [131]. Field experiments establish causal relationships between manipulated factors (i.e., treatment) and target outcomes (i.e., effect) by performing precise behavioral interventions on the randomized treatment and control groups. In contrast

to lab experiments that usually involve hypothetical experimental settings, field experiments demonstrate causality with data from the real-world context and alleviate the difficulties of external validity.

1.1.2 Challenges of Traditional Mono-Methods and Opportunities

While these three components each provide powerful tools to investigate worker performance, traditional mono-method approaches cannot tackle the challenges rendered by large and complex data in the modern era.

Machine learning models are characterized by scalability and can handle high-dimensional complex data. However, ways to fully incorporate existing knowledge into the data-driven methods and to discover insights that are more explainable and actionable remain largely mysterious.

Field experiments are commonly used to establish causality; yet, it requires insights to design effective interventions in the first place and data analytics to discover reliable and useful results. The latter part is especially challenging given that field experiments may involve high-dimensional and large-scale datasets, such as those involving thousands of human and contextual factors and millions of records, where traditional experimental analysis is commonly insufficient.

Social science theories embed our existing knowledge about worker performance in the traditional job context. However, sophisticated operationalization and analytical methods are required to examine whether a theory is applicable to new contexts and to build new theories.

Taken together, these mono-methods either cannot handle complex relationships among human and contextual factors, or cannot fully incorporate existing knowledge, or lack the scalability to deal with high-dimensional large-scale data. These challenges call for interdisciplinary solutions leveraging both the advanced analytic skills of data science and the deep insights of social sciences.

1.1.3 Human-Centered Data Science Synthesizing Social Science Theories, Field Experiment, and Machine Learning

To approach these challenges, we complementarily bridge machine learning, field experiments, and social science theories into a human-centered data science framework as demonstrated in Figure 1.1. First, we root the design of field experiments in theoretical insights, and the causal insights derived from field experiments can then support the testing and development of social science theories. Second, after collecting large-scale and complex observational data, especially those from field experiments, we leverage machine learning and counterfactual reasoning to analyze the data to identify data-driven insights. While traditional experimental analysis mainly supports the understanding of treatment effects at aggregate levels, such as average treatment effect, machine learning empowers more sophisticated analyses of experimental data by revealing heterogeneous effects at a finer granularity, such as individual treatment effects. These data-driven discoveries complement theory-based insights to inspire better experimental designs, and we can in turn evaluate the effectiveness of these models and insights in real-world practices via field experiments. Third, existing knowledge from social science theories can further enhance the predictive power of machine learning models by informing feature construction, model architecture, and model explanation. The patterns identified by machine learning models can in turn support the testing of the applicability of existing knowledge and the development of new theories that better reflect reality. Together, social science theories, field experiments, and machine learning strengthen one another to better study human behaviors.

We would like to point out that human-centered data science is different from the general data science. While human-centered data science leverages general data science techniques, the goal is mainly to describe, predict, and understand human behavior by combining data analytic techniques and social theories. Human-centered

data science emphasizes the interpretation of models through the perspective of human behaviors. In contrast, general data science either focuses on data analysis without considering social science theories or separately investigates the three components.

In addition, while this framework connects social science theories, field experiments, and machine learning into an organic integration, applications of this framework are very flexible. One does not have to apply the entire framework in a single study, and there is a rich exploration space for every component and the interaction between any pair of these components. We illustrate this flexibility with three empirical studies in the next section (as represented by the solid lines in Figure 1.1).

1.1.4 Empirical Applications to Improve Worker Performance

In this dissertation, we present three research projects applying human-centered data science to improve worker performance.

In the first project, we use machine learning models to help improve the performance of the outreach specialists in New York City's Tenant Support Unit (a traditional sector). By providing the predicted probability that tenants in a given building would need the specialists' assistance, our models are able to inform the outreach priorities and facilitate better planning, enhancing the effectiveness and efficiency of outreach. While an ideal natural next step is to conduct randomized field trials to verify the performance of the model, we have not finished these trials due to constraints in reality. Alternatively, we illustrate the deployment of field trials in another scenario – the ride-sharing economy. In the second project, we conduct a field experiment and show the effectiveness of theory-informed interventions in enhancing worker performance on a ride-sharing platform (Didi Chuxing). To further understand the heterogeneous treatment effect of the field experiment on Didi Chuxing, in the third project, we deploy counterfactual machine learning to analyze hundreds of

large-scale field experiments, the results of which are shown to be directly actionable to increase worker performance in the ride-sharing economy.

In the following section, we present the outline of this dissertation with a summary of the three projects.

1.2 Dissertation Outline

This dissertation proceeds with several chapters. In Chapter II, we review the preliminaries on field experiments, counterfactual machine learning, and worker performance. We first introduce field experiments. Field experiments are rooted in the theme of causal inference, which also connects to counterfactual reasoning. Therefore, here we discuss the preliminaries of causal inference as well as the pros and cons of field experiments in establishing causality. Next, we present the background knowledge about counterfactual machine learning, with a comparison between counterfactual machine learning and machine learning. Last, within the huge body of worker performance literature, we briefly introduce the research studying how to improve worker performance through predictions and interventions, including studies applying machine learning and field experiments. In this chapter, we prepare the readers with the fundamentals of field experiments, counterfactual machine learning, and domain knowledge about worker performance.

Chapters III to V present three research projects that deploy human-centered data science in real-world applications. The workers who benefit from these applications span the traditional sectors and the rising gig-economy work force. They each reflect different parts of the human-centered data science framework, exemplifying the flexibility of applying this framework to real-world practices.

In Chapter III, we present a study that applies machine learning to improve the performance of workers in a traditional sector — the New York City government. To ensure the existence of affordable housing for low-income tenants, New York City

has implemented rent-stabilization policies to restrict the maximum annual rate at which the rent of certain units can be increased. However, some landlords try to circumvent the laws and "destabilize" these units by illegally forcing the tenants out so that they can greatly increase the rent. Therefore, the New York City Public Engagement Unit (PEU) conducts proactive outreach to identify tenants vulnerable to such rental harassment and assist them in exercising tenants' rights. However, the current system that PEU uses only provides a map of residential buildings. In this project, we help to increase PEU's work productivity in identifying harassment cases by providing additional information about the harassment risk levels associated with these buildings. This study sits in the framework as an application of machine learning in predicting human behaviors (i.e., rental harassment), the results of which can further inform intervention (i.e., the outreach to tenants). The analyses based on historical data show promising results. We note that field trials are the golden standard to evaluate the effectiveness of such machine learning models, but we have not had luck to deploy field trials for this project so far. In the next chapter, we will illustrate the execution of field experiments in the gig economy.

Chapter IV describes the design and implementation of a field experiment to enhance worker performance in a modern work force of the gig economy — Didi Chuxing (DiDi). The gig economy provides flexible and low-barrier jobs for millions of workers globally. However, a lack of both organization identity and social bonds contributes to the high attrition rate experienced by gig platforms [112]. To help engage workers, we propose to enhance worker performance by engaging them in virtual teams and team contests, an intervention inspired by social identity theory and contest theory. Through a large-scale field experiment with 27,790 drivers, we show that virtual teams are able to enhance worker performance on DiDi and that the treated workers continue to work longer hours on the platform even three months after the end of the experiment. This study demonstrates the effectiveness of informing

field experiments with theories and using experimental evidence to support theories. Meanwhile, we note that the experimental analysis of this study stays at the aggregate levels, leaving rich opportunities to further unpack the treatment effect at a finer granularity.

In Chapter V, we show that experimental analysis via counterfactual machine learning enables sophisticated evaluation and optimization of interventions, further improving worker performance in the gig economy. While virtual teams and team contests have been shown to effectively enhance worker performance, huge variation in treatment effects is observed across individuals, teams, and contests. To uncover the decisive factors behind the treatment effects and identify better team formation strategies and contest designs, we combine machine learning and counterfactual reasoning to answer these questions: (1) What *will* be the treatment effect on worker performance if this worker participates in this contest with this team? and (2) What will be the treatment effect if we *change* the contest design and team formation? In addition, to better capture worker behavior, demographics, and contextual factors, we employ insights from theories in virtual teams, social psychology, and behavioral economics to design features. The results show the effectiveness and promising potential of deploying counterfactual machine learning to predict treatment effects at the individual level and to derive data-driven insights for future experimental intervention designs. This study complementarily bridges machine learning, field experiments, and social science theories in promoting worker performance.

Taken together, this dissertation proposes a human-centered data science framework to improve worker performance and presents three studies that illustrate its flexible applications to solve real-world problems. We summarize the dissertation in Chapter VI, with a discussion of future research directions.

CHAPTER II

Preliminaries

To facilitate the understanding of the following chapters, in this chapter, we introduce the key methodologies — field experiments and counterfactual machine learning, as well as the application domain — worker performance. We first present the concept of field experiments, typical experimental analysis embedded in the general theme of causal inference, and its relative advantages and disadvantages in comparison to lab experiments. Next, we discuss counterfactual machine learning and we point out its distinguishing characteristics from machine learning in general. At the end of this chapter, we further prepare the readers with two streams of worker performance research — prediction and intervention. More relevant studies associated with each project are introduced in the following chapters.

2.1 Field Experiment

2.1.1 Field Experiment and Causal Questions

Humans have a natural interest in discovering causation. In childhood, many of us have probably been curious about why we have two eyes and one nose, and why we see the sun in the day and the moon in the night. Our interests in causation extend to the social and behavioral world as we grow up. What contributes to a good relationship between parents and children? What are the decisive factors of individual financial

success? How can we promote social equity among groups of divergent socio-economic levels? Will employees put more effort into work if they receive additional incentives to do so? These causal questions span fields of psychology, economics, organization studies, management, sociology, etc., and all of them fall into key interests of social science researchers.

Experimentation is one approach to establish causality. By performing different interventions on the randomly assigned treatment group and control group, randomized experiments are able to claim causal relationships between treatment variable(s) and outcome(s). Experiments can be carried out in the lab and in real-world settings, and the latter approach is referred to as a field experiment or A/B testing.

To better understand the establishment of causality via field experiments and experimental analysis, we introduce the field experiments under the umbrella of a causal inference framework.

2.1.2 The Potential Outcome Framework of Causal Inference

Generally speaking, causal inference answers the question of to what level the outcome changes if the treatment changes. Neyman [103] and Rubin [117] propose to estimate the causal effect under a framework of potential outcomes (hereafter referred to as the Neyman-Rubin framework).

2.1.2.1 Key Components

There are three key elements in this framework: treatment, unit, and outcome. To make it simple, we illustrate this with binary treatment status in the following sections. In other words, there is one treatment condition (i.e., treated) and one control condition (i.e., not treated).

Treatment. Treatment refers to the intervention or action applied to the experimental subject. For example, if we want to understand whether providing additional

financial incentives increases the salesman's revenue, the financial incentive would be the treatment. Treated salesmen would receive the financial incentive, while non-treated salesmen would not. We denote treatment as $W \in (0, 1)$, with $W = 1$ representing treated and $W = 0$ representing not treated.

Unit. A unit is an object that receives the treatment (or not). In the salesman example, a unit is an individual salesman. However, we would like to point out that a unit is usually, but not always, an individual participant. For example, in the study of racial discrimination in the job market [22], the researchers sent out resumes with the applicant name assigned as African American sounding or White sounding to job ads and compared the difference in interview callback rates. In this case, the unit is not a person, but rather an application/resume.

Potential Outcomes. For each unit, the potential outcome under a given treatment is the outcome if the unit receives the given treatment. The outcome is denoted as Y. For example, when the treatment variable is binary, there are two potential outcomes for a unit i: $Y_i(W = 1)$ if i is receiving the treatment and $Y_i(W = 0)$ if i is receiving no treatment.

2.1.2.2 Treatment Effects Estimation

The treatment, unit, and potential outcomes defined in the Neyman-Rubin framework make a good foundation to define treatment effects.

Individual Treatment Effect (ITE). Individual treatment effect refers to the causal effect of the treatment on a unit. The individual treatment effect of unit i is:

$$ITE_i = Y_i(W = 1) - Y_i(W = 0). \tag{2.1}$$

Average Treatment Effect (ATE). Average treatment effect aggregates the individual treatment effects at the population level:

$$ATE = \mathbb{E}[Y(W=1) - Y(W=0)], \tag{2.2}$$

where $Y(W=1)$ ($Y(W=0)$) refers to the average potential treated (non-treated) outcome of the entire population.

Average Treatment Effect on the Treated Group (ATT). Because the treated group might have different characteristics and potential outcomes from the entire population, social scientists also define the Average Treatment Effect on the Treated Group as:

$$ATT = \mathbb{E}[Y(W=1)|W=1] - \mathbb{E}[Y(W=0)|W=1], \tag{2.3}$$

where $Y(W=1)|W=1$ ($Y(W=0)|W=1$) represents the average potential treated (non-treated) outcome of the treated units.

Conditional Average Treatment Effect (CATE). Focusing on the treatment effect of a subgroup, the Conditional Average Treatment Effect is represented as:

$$CATE = \mathbb{E}[Y(W=1)|X=x] - \mathbb{E}[Y(W=0)|X=x], \tag{2.4}$$

where $Y(W=1)|X=x$ ($Y(W=0)|X=x$) represents the average potential treated (non-treated) outcome of a subgroup of units with $X=x$.

Difference-in-Differences (DID). The DID model considers not only the comparison between the treated and non-treated units, but also the within-unit difference between the pre-treatment and post-treatment periods. This model is most commonly used in natural experiments where the treatment groups and control groups are naturally assigned by policy changes or natural events, and it is also frequently applied in the analyses of randomized experiments. Although this is not in perfect parallel with other treatment effects presented above, we introduce the DID model here because it is applied in Chapter IV and Chapter V, and this fits best for the

reading purpose.

The average treatment effect under the DID model can be represented as:

$$\begin{aligned} ATE_{DID} = &\mathbb{E}\left[\big(Y(W=1,T=T_1) - Y(W=1,T=T_0)\big)\right. \\ &\left. - \big(Y(W=0,T=T_1) - Y(W=0,T=T_0)\big)\right] \end{aligned} \tag{2.5}$$

where T_0 (T_1) is denoted as the pre-treatment (post-treatment) period.

The regression equation of a DID model generally follows the form of:

$$Y_{iT} = \beta_0 \cdot W_i + \beta_1 \cdot I_T + \beta_2 \cdot W_i \cdot I_T + \epsilon_{iT}, \tag{2.6}$$

where Y_{iT} represents the potential outcome of unit i at time T; $W_i \in \{0, 1\}$ shows the treatment assigned to unit i; I_T is a binary variable indicating the period, with $I_T = 0$ if $T = T_0$ and $I_T = 1$ if $T = T_1$; ϵ_{iT} refers to the residual term of unit i at time T; and β_2 presents the treatment effect. Equivalently, it can take the form of:

$$\Delta Y_i = \theta + \beta \cdot W_i + \epsilon_{iT}, \tag{2.7}$$

with ΔY_i ($= Y_{i,T=T_1} - Y_{i,T=T_0}$) indicating the within-individual outcome change and β representing the treatment effect.

2.1.3 Use of Field Experiments to Establish Causality

As suggested by Box et al. [26] that "to find out what happens when you change something, it is necessary to change it," randomized (field) experiments, with treatment interventions implemented, have been commonly recognized as the golden stan-

dard to establish causality.

This is because of the fundamental challenge of causal inference in identifying the counterfactual outcome [78]. For a unit i, we can only observe $Y_i(W = 1)$ if i gets the treatment or $Y_i(W = 0)$ if i is not treated, but not both at the same time. Estimating the counterfactual outcome — the outcome of treatment that i does not receive — is thus the key challenge of causal inference.

Randomized experiments help to provide a good approximation of the counterfactual outcome of the treated (control) groups at the aggregate level. The randomized treatment assignment leads to the independence between treatment assignment W and the potential outcomes Y (sometimes given X), i.e., $W \perp\!\!\!\perp Y(W = 1), Y(W = 0)|X$ [1]. If we denote the observed outcome as Y^{Obs} , this gives us:

$$\mathbb{E}[Y(W = w)|X = x] = \mathbb{E}[Y(W = w)|W = w, X = x]. \tag{2.8}$$

Therefore, we will have:

$$\begin{aligned}\mathbb{E}[Y(W = 1)|X = x] &= \mathbb{E}[Y(W = 1)|W = 1, X = x] \\ &= \mathbb{E}[Y^{Obs}|W = 1, X = x],\end{aligned} \tag{2.9}$$

and

$$\begin{aligned}\mathbb{E}[Y(W = 0)|X = x] &= \mathbb{E}[Y(W = 0)|W = 0, X = x] \\ &= \mathbb{E}[Y^{Obs}|W = 0, X = x],\end{aligned} \tag{2.10}$$

which leads to:

[1] This is also commonly referred to as ignorability or unconfoundedness.

$$
\begin{aligned}
ATE &= \mathbb{E}[Y(W=1) - Y(W=0)] \\
&= \mathbb{E}_x \left[\mathbb{E}[Y^{Obs}|W=1, X=x] - \mathbb{E}[Y^{Obs}|W=0, X=x] \right] \\
&= \mathbb{E}[Y^{Obs}|W=1] - \mathbb{E}[Y^{Obs}|W=0]. \qquad (2.11)
\end{aligned}
$$

However, we also note that not all randomized assignments successfully satisfy the randomization threshold. Treated groups and control groups may still be significantly different on key characteristics. A randomization check helps to rule out some of these violations.

2.1.4 Field Experiments Versus Lab Experiments

Randomized experiments can be carried on both in the lab (i.e., lab experiments) and in the field (i.e., field experiments). Compared to the lab experiments that mostly involve college students as subjects, the results of field experiments rely on the data in real-world settings, participants of which are more representative of the target population. This greatly enhances the *external validity* of experiments.

Despite the rising favor toward field experiments, critics raise concerns about real-world contexts being too complicated. For example, the messy real-world practice may challenge the assumptions of causal inference [16], and some assumptions, like randomization, are hardly verifiable in a large-scale experiment. Moreover, experiments on the online platforms that support social networking, such as social media platforms, can go against the assumption of Stable Unit-Treatment-Value Assumption (SUTVA) because participants in different treatment groups may have interactions with one another. In addition, concurrently launching multiple A/B testing, a type of field experiments, may expose a participant to several experiments at the same time, also undermining the reliability of the experimental results. Researchers have

started to explore methods to approach such problems (e.g., [141]).

Experimentation is commonly compared with other causal inference methods for observational data, such as propensity score matching, regression discontinuity, instrumental variables, and structural causal models. While this is outside the scope of this dissertation, we refer the readers to related reviews for more details, such as [106], [134], and [143].

2.2 Counterfactual Machine Learning

2.2.1 Counterfactual Machine Learning

In the prior section, we discuss that to address counterfactual questions, a field experiment (i.e., A/B testing) is commonly regarded as the golden standard. However, field experiments are both financially costly and time-consuming, and only a limited number of experiments, if not none, are feasible to launch due to real-world constraints. Therefore, it is important to estimate the effectiveness of a new treatment before it has been deployed online. This is referred to as counterfactual estimation (also called off-policy evaluation and estimation of treatment effects). Counterfactual estimation enables prediction and evaluation of the effects of a new treatment by learning from the existing data without collecting new data. This further allows for policy/treatment optimization. Counterfactual estimation and policy optimization (also referred to as counterfactual learning), together make up the key components of counterfactual machine learning (CML) [83].

2.2.1.1 Counterfactual Evaluation and Counterfactual Learning

If we denote Y as the targeted outcome variable, w as the historical treatment(s), and X as the vector representing individual and contextual factors, CML leverages the existing data about treatment W to learn a regression of

$$Y(X, w) = f(X, w) \tag{2.12}$$

and predict $Y(X, w')$ for a new treatment w'.

The most intuitive estimation of the average new treatment outcome is to directly take the outcome mean of every unit under this treatment. This direct method is also referred to as the "model-the-world" approach [83]. In other words, the outcome of the new treatment w' can be represented by the outcome of unit $i \in 1, ..., n$ treated with w' as:

$$Y(W = w') = \frac{1}{n} \sum_{i} Y_i(x_i, w') \tag{2.13}$$

However, potential selection bias can undermine the validity of this estimation. To achieve unbiased results, literature has incorporated inverse propensity score (IPS) [116]. This is referred to as importance sampling [62], or the "model the bias" [83] approach. Compared to the direct method, this approach is unbiased but shows higher variance. The doubly robust estimator combines the direct estimation and the importance sampling to achieve both low bias and low variance at the same time [55, 81]. Self-normalized IPS model also reduces the variance by incorporating normalization to address propensity overfitting [127].

Once $f(X, W)$ is realized by counterfactual estimation, counterfactual learning can be leveraged to inform the optimized treatment by solving

$$W^* = \underset{W}{\operatorname{argmax}}\, f(X, W) \tag{2.14}$$

Similarly, this learning approach might suffer from selection bias as well as model bias. To address such problems, more sophisticated unbiased counterfactual learning models have been developed by incorporating weighting (e.g., [148]).

2.2.1.2 Comparing counterfactual machine learning and machine learning

While both machine learning and counterfactual machine learning make predictions, note that they are different in the objectives of their predictions. Machine learning focuses on predicting what is going to happen given the features, i.e., predicting the outcome variable Y given $X = x$. In contrast, counterfactual machine learning predicts what will happen if the treatment changes, i.e., predicting Y given $W = w'$ and $X = x$.[2] The underlying goal of counterfactual machine learning is to optimize the treatment and derive new treatments through predicting what is going to happen given the data of historical treatment(s) on some population under some context. In other words, the goal of machine learning is to make predictions, while the objective of counterfactual machine learning is to evaluate and optimize through prediction.

2.2.2 Applications and Evaluations of Counterfactual Machine Learning

The last decade has observed increasing applications of counterfactual machine learning in industry, which focus on improving user-engagement metrics. Specifically, most research in this area has been focused on increasing the click-through rate (CTR) of recommender systems, such as ad placement recommenders [25], ad format recommenders [129], news article recommenders [95], and search engine result recommenders [108], except that a few studies have gone beyond CTR to investigate other user engagement metrics that might be more robust with a longer horizon, such as advertiser retention, brand-search behavior, and brand-sites navigation in online ad campaigns (e.g., [31, 92]). For example, Li and his colleagues analyze the search log data with the IPS counterfactual estimator to evaluate and optimize the click-through metrics of the spelling correction recommendation in a commercial search

[2]More generally, counterfactual machine learning is focused on predicting what will happen if there is some change in predictors, no matter the change is in treatment and/or other feature(s) in X, i.e. predicting Y given $W = w'$ and/or $X = x'$.

engine [94]. To evaluate the effectiveness of this counterfactual model, they conduct an A/B testing online and show that the model significantly improves the existing approach.

Our work in Chapter V extends the application domains of counterfactual machine learning in the industry from CTR to worker performance. Enhancing worker performance has inherent distinctions from improving CTR. While CTR is more about accepting and consuming the recommendation of information, service, and goods consumption, worker performance is more about the supply of labor. This difference may reflect the divergent motivations and thus mechanisms behind the click-through and working behaviors. In addition, the decisive factors of worker performance, such as skills, task assignment, organizational context, social relationships with co-workers, and contextual constraints could all be different by nature from the predictors of CTR, such as the position and the layout of ads. In the next section, we will have an overview of worker performance research, with a focus on improving worker performance through predictions and interventions.

2.3 Worker Performance

2.3.1 The Definition and Measurement of Worker Performance

In this dissertation, worker performance is referred to as the aggregated value of the set of behaviors that a worker contributes to organizational goals both directly and indirectly [28]. The measurements of worker performance vary across contexts and disciplines. For example, the work performance of a customer service specialist is commonly measured by the percentage of tickets successfully solved, whereas the performance of an Uber driver can be reflected by revenue or the number of rides.

In the empirical studies in this dissertation, we operationalize worker performance according to organization practices and contexts. In Chapter III, we consider worker

performance as the number of buildings (or rental units) with rental harassment that a worker is able to identify. This metric is selected to reflect one of the main goals of the organization (i.e., NYC's tenant support unit): to identify as many tenant rental harassment cases as possible in order to assist the tenants involved. In the context of the ride-sharing economy (see Chapters IV and V), worker performance is represented by driver revenue, which is one of the most commonly used metrics in DiDi, the ride-sharing platform that we collaborate with.

2.3.2 Improving Worker Performance: Prediction and Intervention

Examining established literature on worker performance that broadly spans over management, organization, economics, human-computer interaction, and data mining, we find that prior studies have leveraged predictions and interventions to improve worker performance.

Prediction. Predicted results can improve worker performance by informing better decisions and interventions. For example, machine learning has been used to promote city inspector performance by informing inspector allocation (e.g., [84]), to reduce potential adverse events in the police department by helping with police assignment (e.g., [29]), and to facilitate educational proactive intervention by forecasting student grades and dropout risk (e.g., [87, 91]). Machine learning has also been widely used to support clinical care management and disease diagnosis (e.g., [86, 111]). These studies illustrate the power of machine-learning predictions: predicted results from even off-the-shelf machine-learning algorithms are able to effectively enhance worker performance. Recent work has also leveraged more advanced prediction techniques, such as deep learning (e.g., [137]), transfer learning (e.g., [88]), and federal learning (e.g., [43, 142]), to better incorporate large heterogeneous data, from multiple sources, organizations, and domains. Our work in Chapter III contributes to this area by identifying an opportunity to help tenant support specialists: we use machine learning to

predict the risk of rental harassment, informing more effective specialists' outreach to tenants.

Intervention. To enhance worker performance, prior researchers have performed interventions in lab and field experiments. For example, prior literature has examined the effect of for-profit versus non-profit motivation designs, performance feedback incentives, and financial incentives on crowdsourcing worker performance (e.g., [98, 113, 122, 146]). As another example, workers, especially who are geographically distributed from their co-workers, have been grouped into virtual teams to enhance their performance; experimental studies have shown that the performance of virtual teams can depend on various factors related to team design, inputs, and processes, such as group size, group diversity, anonymity, availability of performance feedback, group history, and task features (e.g., [14, 45, 65, 107, 147]). A recent study in the ride-sharing economy shows that grouping drivers into virtual teams and engaging the teams into cash-rewarded contests are effective in promoting worker performance [3]. However, it remains unknown whether the effect of virtual teams holds without financial incentives. Our work in Chapters IV and V approaches this problem with a large-scale field experiment and counterfactual machine learning in a ride-sharing platform.

In this dissertation, we contribute to improving worker performance via predictions (i.e., machine learning) and interventions (i.e., field experiments), leveraging the framework of human-centered data science.

CHAPTER III

Using Machine Learning to Improve the Performance of Government Specialists in New York City

Technology is reshaping the nature of work, including that in the traditional sectors [12]. Given the many ways that technology can affect traditional jobs, data-science approaches have delved into the possibility of improving work effectiveness and efficiency by using predictions to inform better decisions and interventions. This is exactly where we as human-centered data scientists hope to contribute. In this chapter, we present an example of such intervention-oriented machine learning: we use machine learning to predict the work outcomes by analyzing historical data of work results and contextual factors, the results of which help to improve work planning and thus worker performance. This can be conceptually mapped to the link from machine learning to human behavior in the framework (Figure 1.1).

Specifically, in this project, we collaborate with the New York City's government and help its outreach specialists to improve their performance by informing work planning with machine learning predictions. Through broad data collection and careful data analytics, we show that even simple machine learning models have the potential to increase work performance by as much as 59% for workers in traditional sectors.

We also discuss the importance of getting the empirical problem appropriately formulated by comparing the results of two different formations. These promising findings illustrate the effectiveness of applying data science to promote worker performance in traditional sectors.

3.1 Introduction

In New York City (NYC), one of the world's most populous and dense cities, housing availability and affordability is a major concern for residents and city government. From 2009 to 2017, rents rose at twice the rate of wages [136], making it more difficult for New York City tenants to afford housing.

To help ensure the long-term existence of affordable housing, the New York State and New York City governments have implemented housing policies, such as rent stabilization, which restricts yearly rent increases, and a voucher program, which subsidizes rent for low-income households. Currently, the city has more than 1 million rent-stabilized housing units [52, 53].

However, the landlords of rent-stabilized units often want to "destabilize" these units [140] by forcing tenants out: that is, they want tenants to move out, voluntarily or through an eviction, to force a larger allowable rent increase that eventually places the unit beyond the purview of rent-stabilization policies. While the overall number of housing units has increased, the number of rent-controlled and rent-stabilized apartments in New York City has decreased by 146,902 units since 1991 [52, 53]. Some of this turnover is the result of landlord harassment, which can take the form of refusal to make essential repairs, illegally locking tenants out of units they have a right to live in, and other tactics aimed at inducing tenant turnover [120].

To help vulnerable tenants handle these tactics, in 2015, New York City's Mayor's Office established a Tenant Support Unit (TSU), a team of outreach specialists from the Mayor's Public Engagement Unit (PEU). TSU specialists proactively canvass

door-to-door throughout the city and hold events with local community partners to find tenants in need of assistance with housing challenges. Once they identify a case of harassment or other serious housing challenges, specialists further case-manage tenants to help them access a range of city services, such as emergency repairs, vouchers and free legal assistance.

Canvassing to find tenants in need is a time-sensitive process — TSU's goal is to reach tenants before their problems progress to more serious cases of eviction or other forms of displacement. Currently, TSU identifies buildings that have rent stabilized units in 20 ZIP codes prioritized as part of anti-harassment protection legislation. To locate the buildings, TSU uses an internal address database and canvasses every apartment unit in these buildings. PEU team leads in each borough send specialists to each area until all apartment units have been attempted. Once an area is completed, canvassing begins again in an adjacent area. There are about 150,000 rental units in the 20 ZIP codes where funding is available for TSU to help tenants in need, but TSU specialists only have the resources to knock on an average of 5,000 units a month. Our work is focused on helping TSU prioritize locations where tenants face a high risk of harassment to help TSU specialists better plan their outreach, increase their work performance, and serve more tenants in need proactively. We note that, in this study, work performance is operationalized as the number of harassment cases identified in a unit of time period.

In collaboration with TSU, we[1] deployed machine learning models to help predict which buildings house tenants who face a high risk of harassment by their landlords. By analyzing historical outreach results and building and neighborhood characteristics, we showed that a Gradient Boosting model successfully outperformed the current outreach practice. Specifically, our model increased the precision relative to our baseline — the unit's expert-driven success rate — by 59%, helping TSU better allocate its

[1]This work was done at the Data Science for Social Good Fellowship program at University of Chicago.

outreach resources to people most in need and improving its efficiency at helping vulnerable tenants. In addition, we also provided analyses of feature importance, helping the team understand which attributes of buildings and neighborhoods contribute to the likelihood of rental tenant harassment.

In summary, this paper provides the following contributions:

1. This paper contributes to the prediction of landlord harassment risk by deploying various machine learning models with a direct measure of landlord harassment and well-defined evaluation metrics.

2. Our model shows significant improvement at identifying buildings at high harassment risk over TSU's current approach.

3. In addition to yielding risk scores for tenant harassment, this paper also highlights features that can potentially be used as "early warning signs" of future harassment or proxy markers for the presence of harassment.

3.2 Related Work

3.2.1 Housing Assistance for Low-income Renters

Social science research documents the negative consequences of housing instability and shows the mixed effects of rent-stabilization and other rental assistance policies on combating this instability. On one hand, such assistance may reduce homelessness [139] and rental burden, increasing financial security (such as to afford health care) among low-income households [100]. On the other hand, suppressing a unit's rent at a level below the rate it would receive on the open market can result in lower-quality housing [72, 133] and creates incentives for landlords to use legal loopholes, such as those that allow landlords to increase the rent each time a tenant moves out, to eventually convert the units to market rate [13].

Thus, policymakers face a dilemma: how can they use policies such as rent stabilization (which sets an upper limit on the rate at which the rent can be increased annually) to promote *access* to affordable housing, while also ensuring that tenants renting in these affordable units live in habitable conditions and do not face landlord harassment aimed at getting them to move out? The bulk of existing research focuses on the former part of the dilemma (the effect of policies on housing access). Less research investigates strategies to ameliorate potential byproducts of rent regulation policies.

Our work, by predicting where tenants in affordable units are likely to experience landlord harassment, fills an important gap. The Mayor's Office of Data Analytics ([104]), referred to as MODA hereafter, also has studied data-driven protection from landlord harassment, and our project builds upon their efforts in several ways. First, through this paper we had a more direct measure of landlord harassment. While MODA ([104]) defined harassment using a proxy variable (i.e., the number of rent-stabilized units a building lost during a particular time period), TSU's historical canvass data allowed us to use harassment cases tenants reported during outreach. Second, we estimated many different models and evaluated model performance with well-defined metrics. Finally, the different machine learning models we estimated allow us to use significantly more features and to learn complex relationships — i.e., both linear and nonlinear relationships— between these features and a building's observed harassment risk.

3.2.2 Machine learning for Social Good

In recent years, machine learning has been widely applied to problems of social good and to inform public policies. For example, it has been introduced to forecast issues of criminal justice [20], detect online rumors on social media [151], identify political bias in text [70], map wealth and poverty in given areas [23, 67] and even

facilitate medical diagnoses [86].

In particular, government agencies have used machine learning to inform better allocation of resources and work outcome. For example, random forest and logistic regression have been used to identify students at risk of not graduating, so that school districts can prioritize their limited intervention resources to help these students [91]. Machine learning models can also help government inspectors prioritize inspections to high-risk units. These efforts include using Yelp reviews to help a government agency target hygiene inspections [84, 66] and predicting which buildings face a high fire risk to help the New York City Fire Department narrow its inspection focus [9, 115].

However, far less work has been done to explore how machine learning can inform housing policies and facilitate housing workers in the public sector, except for making policy recommendations to reduce home abandonment in Mexico [1] and detecting home locations by real life photos on social media [152] or by tweets [130], as well as MODA's study mentioned in the previous section [104]. In this paper, we highlight a new application by deploying machine learning methods to predict which buildings house tenant(s) facing a high risk of harassment by their landlords.

3.3 Problem Formulation

We formulate the tenant harassment risk prediction as a binary classification problem. For each building, our model produces a risk score for whether there will be at least one harassment case identified if the TSU specialists canvass the building in the next month. Our model answers the question: *Will there be any cases of harassment in a given building in the next month?*

This formulation leads to two further decisions: 1) what time horizon to predict for (e.g., a harassment case within the next week, next month, or next year) and 2) the unit of prediction (e.g., modeling which residential unit faces a high harassment risk versus modeling which buildings contain tenants who face a high harassment

Table 3.1: Data sources summary

Dataset	Records #	Time Window
(Internal) Knock attempts	100K	2016.4 - 2018.2
(Internal) Case records	8K	2015.6 - 2018.2
(Internal) Case issues	30K	2015.6 - 2018.2
(Internal) Building address	1M	N/A
(External) ACS (tract-level)	2000	2013 to 2016
(External) PLUTO buildings	1M	till 2018.1
(External) HPD violations	4M	till 2018.6
(External) Hous. Court litigation	150K	till 2018.6
(External) Subsidized housing	16K	till 2016

risk). Both of these questions need to be answered reflecting the operational and policy constraints of our partner, the Tenant Support Unit at NYC.

For 1), we use a month as the time horizon for our prediction because TSU specialists typically plan their work at the beginning of each month. Monthly prediction thus matches their outreach planning process.

For 2), we focus on each building rather than each tenant for two reasons. First, TSU conducts a building-level outreach process. Out of concern for equity among tenants, TSU specialists believe they should knock on every single unit in a building once they enter. Second, the majority of information in both TSU internal databases and public available datasets describes buildings rather than units. Therefore, it's both more feasible and more important to know the building-level risk of harassment.

3.4 Data

To explore variables that can help us predict which buildings may be at risk of harassment, we combined data from multiple sources. Table 3.1 summarizes the information presented in the data. Details are described in the following sections.

3.4.1 TSU (Internal) Data

3.4.1.1 Building address

To locate residential units for canvassing, TSU uses an internal database (which was built using a publicly available dataset) that contains addresses for all the residential buildings in NYC. For each building, the database records the number of units and location information such as address, building identifier number and the tract it belongs to, making it convenient to join with data from other spatial sources.

3.4.1.2 Knock attempts and case records

During canvassing, TSU specialists knock on every apartment unit in the targeted building(s). If a tenant answers the door, they talk to the person about whether he or she is facing harassment. These activities are recorded at the unit level in *knock attempts* and *case records*, respectively. Each of the records describes the location of the unit, the date it was canvassed, the specialist team that did the canvassing, and the result of the attempt (i.e., knocked, answered, and case identified). The case database also records the source of the case, allowing us to know which cases came from canvassing as opposed to other sources, such as referrals. *Case records* contain information about our outcome variable — whether or not there was at least one case of harassment identified in the building.

3.4.1.3 Case issues.

Once a harassment case is identified, such as a landlord refusing to do essential repairs, the specialists will follow up with the case and separately record each issue related to the housing unit in the *case issues* database. The specialists can then connect the tenants to relevant assistance resources, such as city services or legal support.

Figure 3.1 shows the TSU specialists' canvassing process and our definition for having a case identified (i.e., the label).

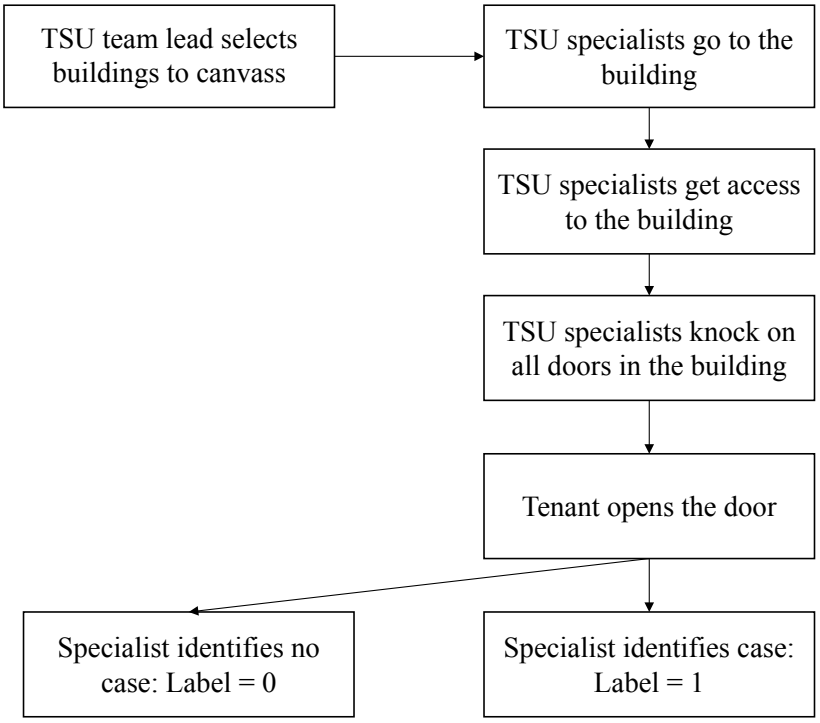


Figure 3.1: Canvassing process with our definition of the outcome label.

3.4.2 Public (External) Data

While the internal canvassing records are critical for understanding where harassment occurs, external data are also important to capture information about buildings not canvassed by TSU yet or longer term historical data before TSU began their outreach activities. TSU's records focus on violations the agency finds based on outreach that began in 2015. External data provides both an expanded time window — which buildings face high rates of landlord issues documented by agencies that predate TSU's existence? — and a lens into the characteristics of buildings and neighborhoods where TSU has historically detected cases.

3.4.2.1 American Community Survey (ACS)

To gain insight into the demographics of tenants whom TSU specialists conduct outreach to, we collected American Community Survey 5-year estimates from 2013 to 2016 at the census tract level. The ACS data contain demographic information such as racial composition, average income, work hours, age distributions and other demographics of the census tract in which a building is located.

3.4.2.2 Primary Land Use and Tax Lot Output (PLUTO)

The PLUTO records describe attributes of each building, such as its renovation history, its building class (e.g., is it a high-rise or a walk-up apartment?), the number of floors, and its recorded owner. We introduced PLUTO data into our model because we believed building information could shed light upon tenant harassment. For example, landlords often own multiple buildings — if TSU canvassing finds harassment at one of a landlord's buildings, that same landlord might be engaging in harassment in other buildings he or she owns. In addition, if a building has been recently renovated, this could be a signal that the landlord is hoping to displace current tenants and lease the building's units to higher-paying tenants. Therefore, we believe that PLUTO features should improve our predictions of harassment.

3.4.2.3 Department of Housing Preservation and Development (HPD) violations

The HPD issues violations when, after sending inspectors to a unit in response to a complaint, they find evidence of a Housing Code violation. This database contains recorded housing violations, which range from more minor, non-hazardous violations to severe, immediately hazardous violations (e.g., no heat or hot water, a rodent infestation, lead paint). These housing violations could be indicators of rental harassment since some reflect extreme landlord neglect of living conditions. Mr. Sidibe, a New

York resident, is a recent example reported in *The New York Times*. He was first hurt by a broken hot water tap and then was improperly evicted while he was recovering in the hospital [13]. Therefore, we hope to use the HPD violation records to improve the predicted harassment risk of a given building.

3.4.2.4 Housing court litigation.

Similar to the HPD violations, housing court litigation can help the model by integrating historical violations. It shows the cases that city agencies levy against an owner when he or she fails to properly address a violation, such as a case legally compelling an owner to fix the heat and hot water in a unit.

3.4.2.5 Subsidized housing.

This database contains building-level information of 53 different subsidy programs a building might participate in, such as the low-income affordable marketplace program and the HPD mixed income program. The subsidy data complement other building-specific characteristics in the databases described here.

3.5 Methods

To predict which *buildings* are likely to house tenants susceptible to experiencing harassment *in the next month*, we experimented with Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), and Gradient Boosting (GB), all of which are implemented with *scikit-learn*. We used the data described above to extract numerous features of buildings: in total we used 92 original features (before further processing such as reformatting to one-hot vectors) generated from our data sources.

3.5.1 Feature Generation

We generated features based on our discussion with experts at the PEU as well as past research on landlord–tenant issues.

3.5.1.1 Building-level features

Building-level features mainly included *dynamic features* of what harassment-related behaviors have occurred before and *static features* of basic building characteristics. For *dynamic features*, we first generated behavioral features by aggregating the *canvassing* activities and the results at the building level. To predict harassment risk in the upcoming (*next*) month, for example, we counted the number of knocks, doors opened and case identifications in the current (*this*) month in a given building. We also calculated the number of issues associated with these cases for each type (e.g., repair, legal) separately. Apart from the count, we created binary variables that indicate whether there were any knocks, doors opened, or case identifications in the current (*this*) month. In addition to recording activity in *this* month, we aggregated all the prior historical records (until *this* month) to assess the predictive utility of aggregate measures.

Similarly, we created the *HPD violations* and the *housing court litigation* features. The records are aggregated to indicate the number or existence of violations and litigation, both in *this* month and all the months until now. To further break down the type of violations, we included features that describe the number of violations for each severity class. We also grouped housing court litigation by litigation type (such as heat and hot water litigation versus tenant actions against owners).

For *static features*, ZIP codes and borough information were generated from the internal *building address* database. We also included dummy variables describing each canvassing team to account for potential variation between the individual specialists responsible for given buildings or areas.

We further extracted basic building characteristics from *PLUTO*, such as ownership features like owner name and owner type, as well as building renovation features including the year of each renovation. We also considered the size of the building (indicated by the number of floors and number of residential units), the class of buildings (identifying whether the building was made of brick and whether it has an elevator), and the assessed total value of the building.

Additionally from the *subsidized housing* database, we generated a feature to describe whether the building is included in a subsidy program or not.

3.5.1.2 Tract-level features

At the tract level, we generated demographic features by extracting records from the *American Community Survey* database. PEU managers suggested local areas with a certain demographic composition of tenants might contain buildings with more harassment. For example, tracts with a higher percentage of low-income tenants might be more likely to have both a higher concentration of tenants living in rent-stabilized units and a higher concentration of tenants who, due to a lack of awareness of city resources, have unmet needs for help with landlord issues. Our features contain measures of racial demographics, measures of when residents work outside of the home (which affects the tenants' ability to answer the door during the main TSU canvassing hours), and measures of income insecurity, such as receipt of public assistance like Supplemental Security Income (SSI).

We cleaned (i.e., preprocessed such as removing duplicate records) all the data mentioned above to generate the features and match data from different sources by location indicators. We used extrapolation to impute missing data in the features (not the label), such as imputing missing records in 2018.2 with data from 2018.1. We further used the min-max scaler in scikit-learn to normalize continuous features, especially for use in regularized logistic regression models.

3.5.2 Splitting Data into Training and Testing Sets

To evaluate models with temporal cross-validation, we followed the rule of time-dependent knowledge restriction to temporally split the data into training and testing sets. We needed to ensure that the knowledge in the *future* (i.e., the testing set) does not inform predictions in the *past* (i.e., the training set). For example, in one data split, if we wanted to use data until end of March 2017 (i.e., testing features) to predict the risk of harassment during April 2017 (i.e., testing label), the training set should contain features only until end of February 2017. The training label would then be generated using cases from records during March 2017. Figure 3.2 shows an example of these training and testing splits, with each row representing one split.

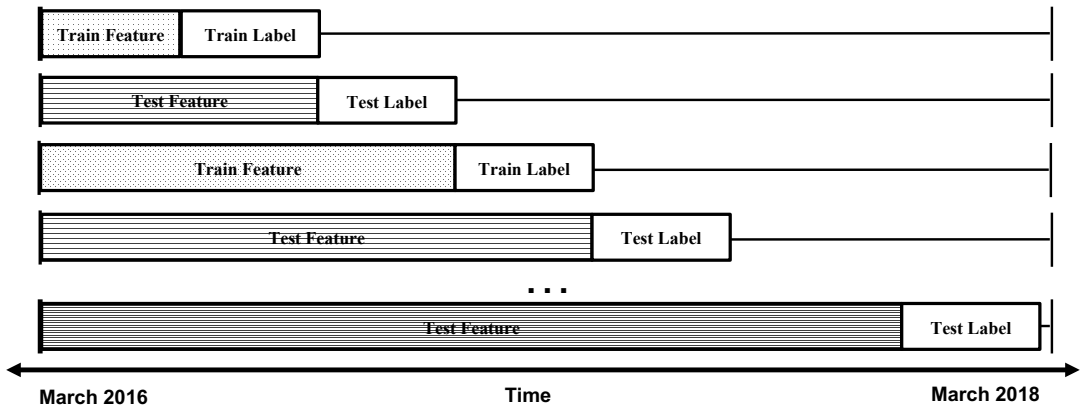


Figure 3.2: An example illustrating training and testing splits.

3.5.3 Model Evaluation

3.5.3.1 Metrics

We used variations of standard metrics to evaluate the model performance: precision and recall at highest predicted risk buildings with a total of k residential units based on the outreach capacity. We select evaluation metrics that have enough flexibility when applied to labels with missing values since many of the buildings we

predict at risk will not have been canvassed historically (since the goal of this project is to suggest new buildings to canvass) and we need to evaluate our models in that setting.

To help TSU plan their outreach, at the beginning of each month, we will use the prediction model to recommend a list of buildings with the highest predicted risk of harassment, adding up to k residential units (hereafter denoted as top k, with k limited to TSU's monthly outreach capacity — the number of units they are able to knock on for outreach in a given month).

We want to evaluate the performance of the model according to the true labels of buildings in this prioritized list. Our test set (that we predict on) contains three of types of buildings:

1. buildings with true positive labels, where TSU knocked and identified case(s)

2. buildings with true negative labels, where tenant(s) opened the doors when TSU canvassed, but no cases were identified

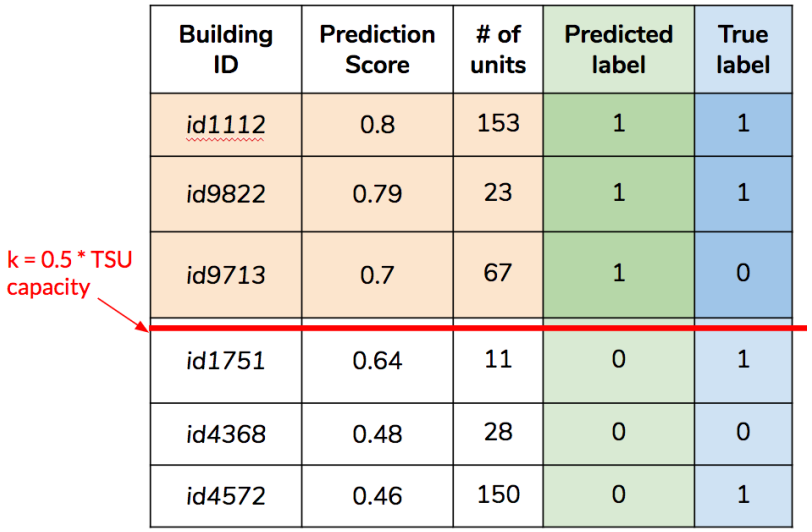
3. buildings missing labels, where (i) TSU specialists did not go to the building (no knocks) or (ii) no doors were opened when TSU canvassed the building (knocks but no opens). Traditional precision and recall metrics are not very informative in this case when the true labels of buildings predicted as positive might be missing.

We built upon previous literature [91] focusing on resource allocation in scarce resource settings and used precision and recall at top k as the evaluation metrics. We denote $N_{k,all}$ as total number of buildings in the top k building list, $N_{k,lp}$ as the number of buildings labeled as positive in the top-k list and $N_{k,ln}$ as the number of buildings labeled as negative in the top-k list. $N_{k,u}$ refers to the number of unlabeled buildings. Obviously, $N_{k,all} = N_{k,lp} + N_{k,ln} + N_{k,u}$. As shown by Equation 3.1, precision at the top k is the proportion of buildings that are labeled as positive (i.e., resulted in true cases)

in the top k building list. Recall at the top k represents the proportion of buildings with true positive labels (i.e., with cases identified) that the model captures in the top k list (as shown by Equation 3.2). While precision measures the efficiency of the model, recall measures model coverage. Figure 3.3 shows an example of calculating precision and recall at top k.

$$\begin{aligned} precision\ at\ top\ k &= \frac{\#\ of\ true\ positive\ labels\ in\ top\ k}{\#\ of\ total\ labels\ in\ top\ k} \\ &= \frac{N_{k,lp}}{N_{k,lp} + N_{k,ln}} \end{aligned} \tag{3.1}$$

$$\begin{aligned} recall\ at\ top\ k &= \frac{\#\ of\ true\ positive\ labels\ in\ top\ k}{\#\ of\ true\ positive\ labels\ in\ testing\ set} \\ &= \frac{N_{k,lp}}{\#\ of\ true\ positive\ labels\ in\ testing\ set} \end{aligned} \tag{3.2}$$



Building ID	Prediction Score	# of units	Predicted label	True label
id1112	0.8	153	1	1
id9822	0.79	23	1	1
id9713	0.7	67	1	0
id1751	0.64	11	0	1
id4368	0.48	28	0	0
id4572	0.46	150	0	1

Figure 3.3: Example of metrics calculation — if a half of TSU capacity $k = 200$, then precision $= \frac{2}{3}$ and recall $= \frac{2}{4}$.

3.5.3.2 Choices in determining the top k list

First, to determine k, TSU indicated that they would like to keep half of the capacity to their own expert-selected buildings so that TSU specialists could also help

residents who lived in buildings outside the top k list. Therefore, each month, we set k as half of TSU's canvassing capacity in a given month (k = 3,000, approximately).

[Top-k list for TSU to canvass]. Second, to suggest a list of the buildings for TSU to canvass, we first rank *all* residential buildings by predicted risk scores and then take the top ones that add up to contain k (apartment) units since the TSU capacity is based on the number of units and we are predicting at the level of buildings. Note that if k is in between two buildings in our list, we include the entire building with at least one unit in the top-k list.

[Top-k list for model performance evaluation]. Third, to evaluate model performance, we generated the top-k list of buildings by only including the *labeled* buildings. We ranked labeled buildings by the predicted risk of harassment, and marked the top-k-units buildings as positive. We didn't deploy the k cut-off on *all* buildings since the top-k list of *all* buildings did not contain enough labeled data to make the precision scores reliable. On average, TSU canvasses about 300 buildings per month out of a total of 6,437 in their outreach area, which covers $< 5\%$ of all buildings. It was highly likely that most, if not all, of the (previously canvassed) 300 buildings fell out of the top-k list, leading to few labeled data in top-k list. In fact, about 20% of the top-k lists generated by each model in each test month contained no labeled building, with the rest 80% of models only include a few labeled data. For example, a Random Forest model proposed 19 buildings in the top-k building list, with only one of them observed by TSU. The precision would be 1 if TSU identified case(s) in this building and 0 otherwise. This challenges our confidence in using these precision and recall metrics to represent the model performance. We thus chose to use the labeled data in determining the top k list for model performance evaluation. This is typical in problems with missing labels and we recommend to conduct a field trial with proactive canvassing on the previously not canvassed buildings to further validate the model on both labeled and unlabeled data.

3.6 Results

3.6.1 Predictive Performance

3.6.1.1 Baseline: TSU's current outreach method

TSU currently uses a simple approach to plan its outreach in the targeted 20-ZIP-codes areas. TSU specialists systematically go block by block attempting to enter every building where there is at least one rent-stabilized unit. A list of buildings to attempt is assigned via a custom-built canvassing app loaded on an iPad.

3.6.1.2 The performance of our models

Our final models were trained on data from July 2016 to December 2017 and were tested on outreach records from January 2018. We further split the training data into 17 folds as illustrated in the previous section to conduct temporal validation.

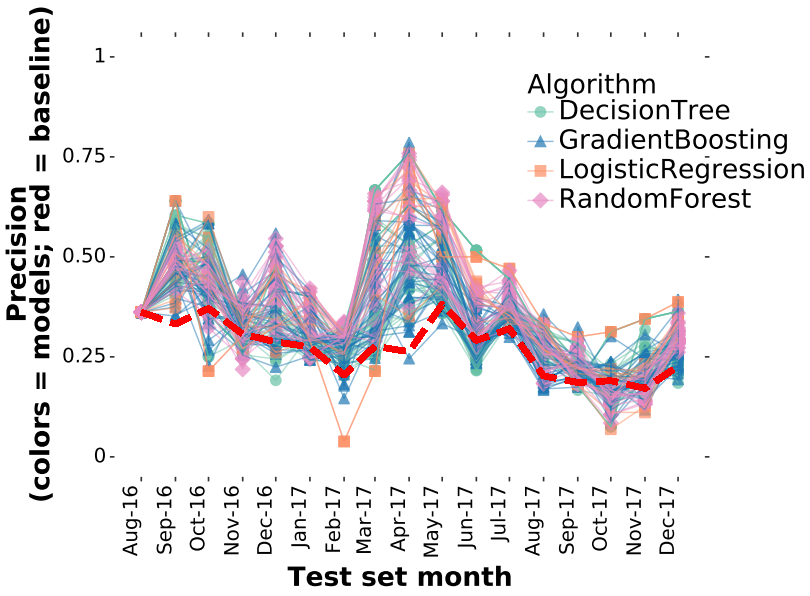


Figure 3.4: Model performance over time (in training stage). X axis represents the time period (month) and Y axis represents the precision scores of models in the month. The figure shows that the baseline has been varying across months and our models generally performed better than the baseline.

Figure 3.4 shows the performance of every model on each data split during the

training stage. The TSU baseline is represented by the red dashed line. The machine learning models performed better than the baseline by 36% on average.

The figure also shows that the effectiveness of outreach efforts by TSU in terms of found cases of tenant harassment varies over time as well. Therefore, to better interpret how much better our model performed than the baseline in each data split, we calculated the ratio of model precision to baseline precision (hereafter named as precision ratio).

To select the best performing model, we first took the average precision ratio score of all data splits and narrowed down to models that had precision ratios ranked in the top 10. Because we want the model that TSU uses to not only exhibit high *average* precision but also exhibit high *stability* in performance, we incorporated the standard error of the precision ratio scores [64] into the evaluation of a model's performance by calculating:

$$\text{precision standard error} = \frac{\text{precision mean}}{\text{precision std}/\sqrt{\#\text{ of precisions}}}$$

The best model to predict whether there will be at least one case in a building next month was a Gradient Boosting classifier with 100 estimators. In our test month (February, 2018), TSU was able to inspect 312 buildings of 7,374 residential units, covering about 4.85% of all buildings. Therefore, we set $k = 3,687$ units to generate the top k list for evaluation. Table 3.2 shows how our model performed in terms of false positives, false negatives, true positives and true negatives. Our model was able to identify about 59% more high-risk buildings than the baseline (with a precision score of 0.25 in the test month).

Table 3.2: Confusion matrix of the best performing model.

	Actual True	Actual False
Predicted True	33	50
Predicted False	46	183

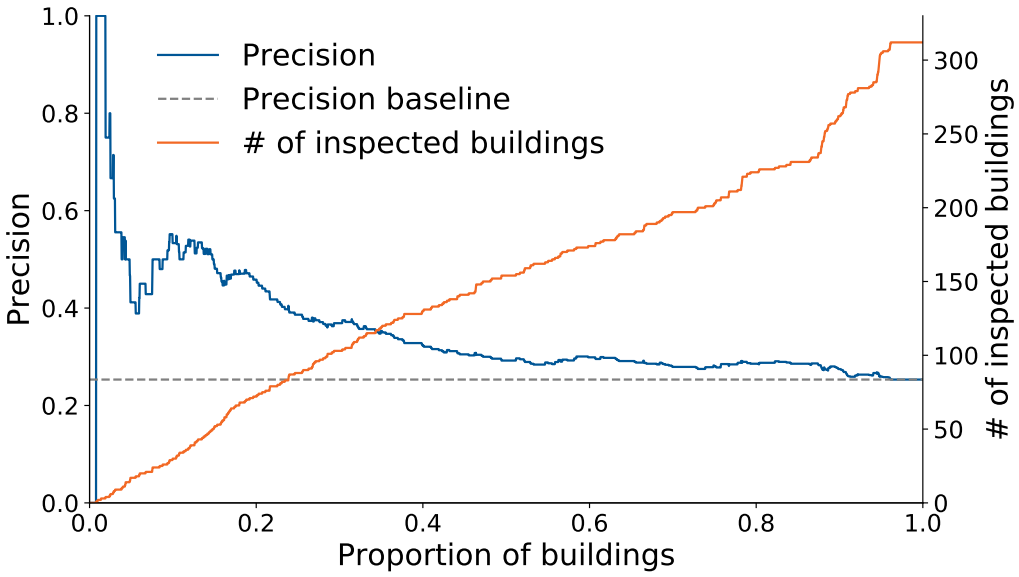


Figure 3.5: Precision and number of labeled data at each k proportion for the Gradient Boosting model.

In Figure 3.5, the precision scores of the building-level prediction at different levels of k is represented by the blue line, with X axis representing the proportion of buildings at k (i.e., $N_{k,all}/\textit{Total number of buildings}$). We also plotted an orange line to visualize the number of labeled data at each k (i.e., $N_{k,lp} + N_{k,ln}$), which shows the number of (labeled or successfully canvassed) buildings supporting the precision calculations. Since TSU only inspected about 300 buildings per month and left most buildings unlabeled, this supporting number at k helps us assess our confidence in the precision score at k.

In addition to precision, we calculated two measures of recall: recall of the *total* count of cases across buildings and recall of buildings with *any* case. Figure 3.6 shows that recall of cases (represented by the yellow-green line) was in general higher than the recall of buildings with any case (represented by the orange line), indicating that our model was good at predicting buildings with a larger number of cases rather than buildings with only one case.

Since both precision and recall measures are relying on labeled data, we wanted

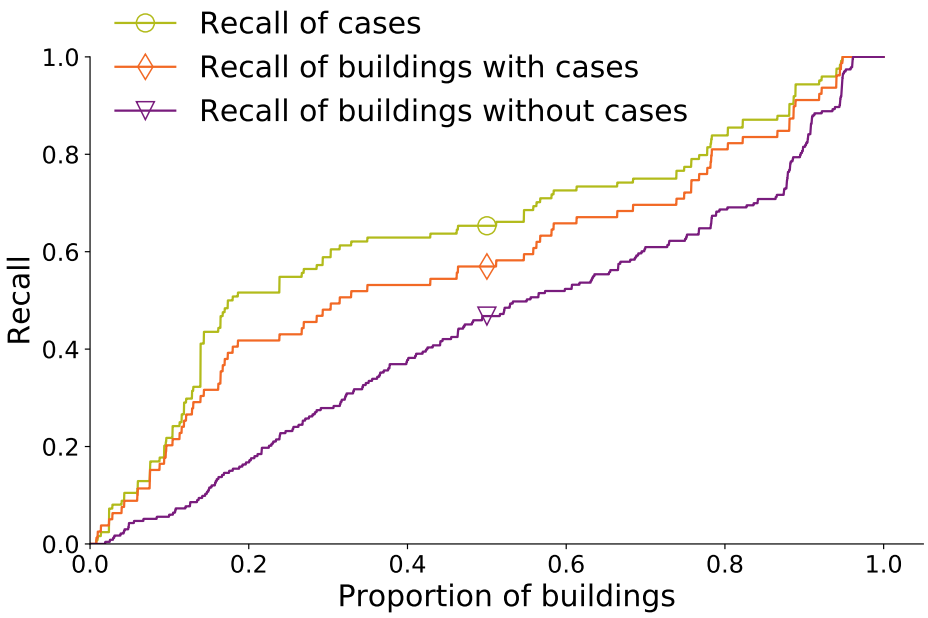


Figure 3.6: Recall curves at each k proportion for the Gradient Boosting model.

to understand if our overall list of high risk buildings was good at ranking high risk buildings above low risk buildings. In addition to recall on the labeled positive examples (orange line), we calculated recall on the labeled negative examples (buildings with no cases) using all buildings as the denominator (purple line). The intuition is that a good ranked list will have more (labeled) positive examples than negative examples at the top of the list and vice versa at the bottom of the list. The gap between recall on positive examples and recall on negative examples in Figure 3.6 allows us to see this was actually the case. The orange line goes up steeply at the beginning (more positive examples) and the purple line goes up steeply at the bottom of the list (more negative examples) giving us confidence in the ranking performance of our model.

3.6.2 Interpreting the Models: Feature Interpretation

Figure 3.7 shows the top 20 features that have the highest feature importances in the best performing Gradient Boosting model.

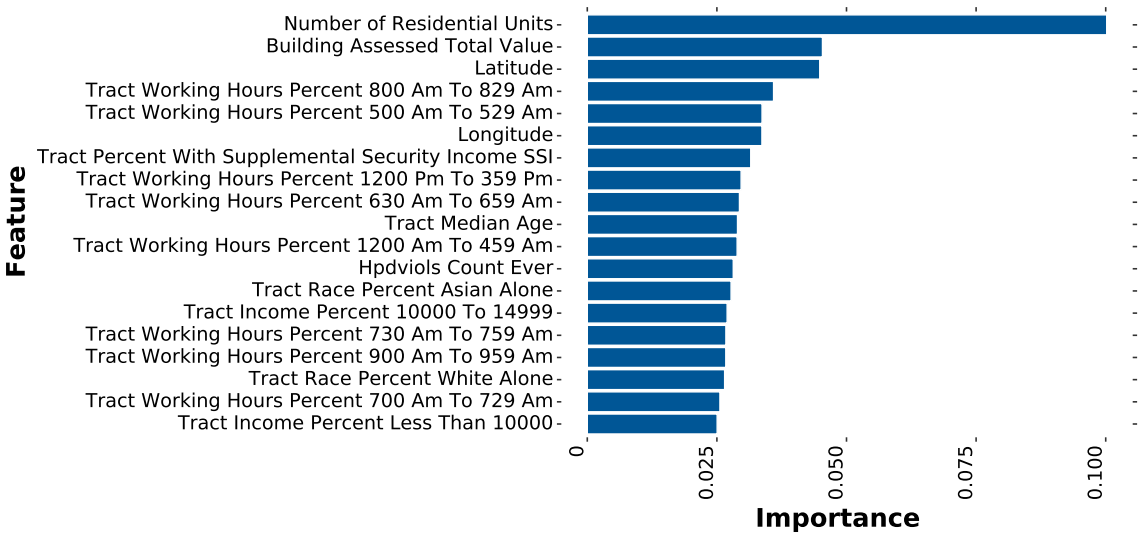


Figure 3.7: Feature importance from the gradient boosting model. We plot the 20 most important features to understand the top predictors that help us identify buildings of high risks.

3.6.2.1 Tract-level demographic features

From the figure, we see that most features in the top 20 feature list were generated from American Community Survey data, which reflects the demographic characteristics of residents in the tract in which a building is located. For example, measures of income insecurity were important in predicting harassment — these included the proportion of people receiving Supplemental Security Income (SSI) in a given tract (*Tract Percent With Supplemental Security Income SSI* in figure 3.7) and the percentage of households earning less than $10,000 per year (*Tract Income Percent Less Than 10000* in figure 3.7). These features might be important because they may reflect unmet need — that is, areas where people are both particularly vulnerable to illegal tactics by landlords and where they also may, prior to TSU's visit, have the most difficulty navigating city services that can help. In addition to the income variables and, more interestingly, 8 of the 20 top features were indicators for the hours that a tract's residents work outside the home. For example, the feature, *Tract Working Hours Percent 800Am to 829Am*, represented the proportion of people who

usually leave their apartment to work between 8:00AM and 8:29AM. These features could be important for two reasons — first, they might serve as additional indicators of socioeconomic status (e.g., lower-income individuals might face less standard work schedules); second, they might reflect which tenants are home to answer the door when TSU specialists go canvassing on weekdays and weekends.

3.6.2.2 Building history and value features

The figure also shows that building-level indicators, such as the total number of HPD violations in a building up until the given month and the total monetary value of a building (generated from PLUTO dataset) are informative in predicting harassment risks. These observations provide support to the idea that external information, including a building's history of violations as well as the physical and economic attributes of a building (i.e., how much is a building worth?), is valuable in predicting whether there will be at least one case of harassment in the building next month.

3.6.2.3 Building location features

In addition, the model identified the longitude and latitude of a building as important features. This indicates that high-risk buildings are perhaps clustered in specific locations. To highlight this clustering feature, we predicted the risk of each building with our best-performing model. We further separated buildings into different levels of risk, with high risk representing buildings with the highest 33.33% risk scores, low risk representing buildings with the lowest 33.33% risk scores and medium risk representing the rest. We plotted each building according to its point location, with high-risk buildings in red, medium-risk buildings in yellow and low-risk buildings in green (see Figure 3.8). The map highlights clusters of high-risk buildings in Manhattan and the Bronx, with low-risk buildings dispersed throughout Brooklyn, Queens and Staten Island. This finding suggests that in order to balance canvassing efforts

across boroughs, we would need to separately rank buildings and provide a high-risk building list for each borough when deploying the model in practice.

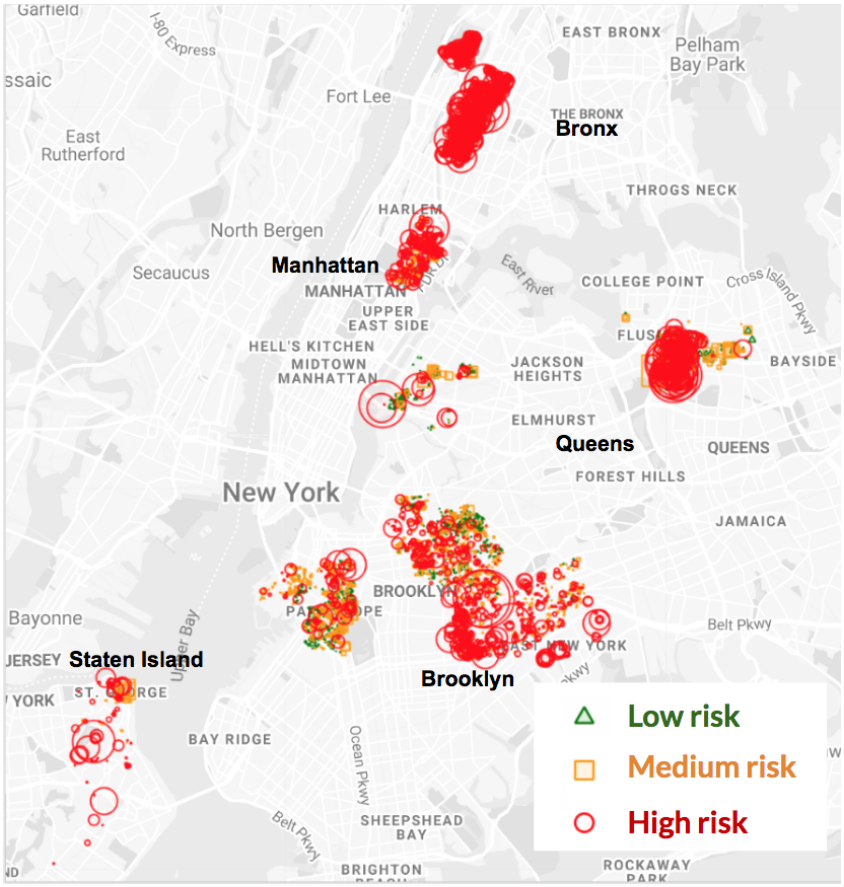


Figure 3.8: Map of buildings predicted as different risk levels. Each point represents a building: high risk (red), medium risk (yellow), low risk (green). Manhattan and the Bronx had most of the high-risk buildings. Low- and medium- risk ones were mainly spread out among Brooklyn, Queens and Staten Island. The marker (i.e., circle, triangle, square) size reflects the # of units in the building.

3.6.2.4 Building size

While these features mentioned above indicate that the model took advantage of information in the data, the high importance of the *Number of Residential Units* shows that our problem formulation — predicting *any* case in a building — leads us to identify buildings with many residential units. These buildings have a higher "denominator" of tenants at risk of harassment to generate the label of a single case in

a particular month. For the test month (Feb, 2018) in particular,buildings predicted to have high risk of harassment on average contained 70 units per building, which was about 3 times as many as the average size of all buildings in the targeted area. Figure 3.8 further highlights the correlation between a building being larger and a building being identified as higher risk: buildings with larger numbers of units (indicated by marker size) were more likely to be predicted as buildings of high rental harassment risk (indicated by color of red). Therefore we defined another problem formulation to try and standardize a building's count of cases by the number of tenants who might have a case.

3.6.3 Reformulation: Predicting Case per Unit Ratio above a Threshold

Our reformulated problem uses the label — hereafter called the *any-case* label — defined as follows: $Y \in \{1 = \text{any case}, 0 = \text{no case}\}$ in building i in month m. What we call the threshold label constructed a binary label using a two-step procedure: first, we calculated the ratio of cases in a building per number of units; second, we constructed the binary label as follows: $Y \in \{1 = \text{ratio} \geq \text{threshold}, 0 = \text{ratio} < \text{threshold}\}$ in building i in month m. The results we present focus on buildings with a ratio in the top 10% of the training set.

We used the same procedure as in section 6.1 to select the best performing model for the threshold label. The best-performing model (a Gradient Boosting classifier) identified 14% more buildings of high case-per-unit ratio than the baseline (with *precision* = 0.15 in the test month). Figure 3.9 plots its precision scores and the supporting number of buildings at each level of k. We found that the threshold model successfully prioritized buildings with a higher proportion of cases than the any-case model (with case-per-unit ratio = 0.94 and 0.30, respectively), which means about 213% more cases could be identified by the threshold model than the any-case model, holding the number of units canvassed constant.

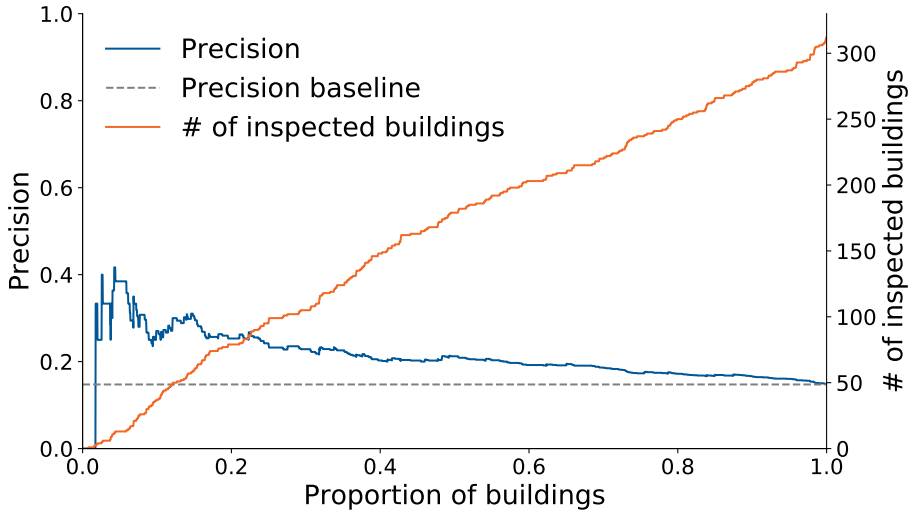


Figure 3.9: Precision and number of labeled data at each proportion of buildings for the Gradient Boosting model using the threshold label.

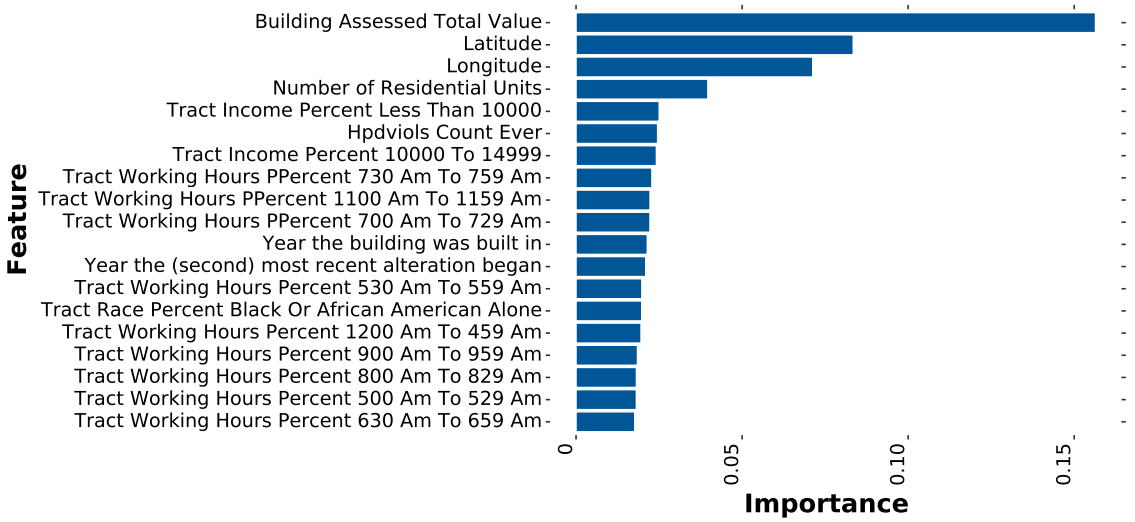


Figure 3.10: Feature importance from the best-performing threshold model. We plot the 20 most important features to understand the top predictors that help us identify buildings of high risks.

Figure 3.10 plots the 20 most important features from this model. Comparing to the model using the any-case label, the best-performing model using the threshold label put more weights on features such as the assessed building value, the year the building was built in, the year the building was recently renovated (i.e., *Year the (second) most recent alteration began*), and number of violations HPD had ever

recorded, while it was less informed by features such as percentage of households receiving SSI in the tract and number of units.

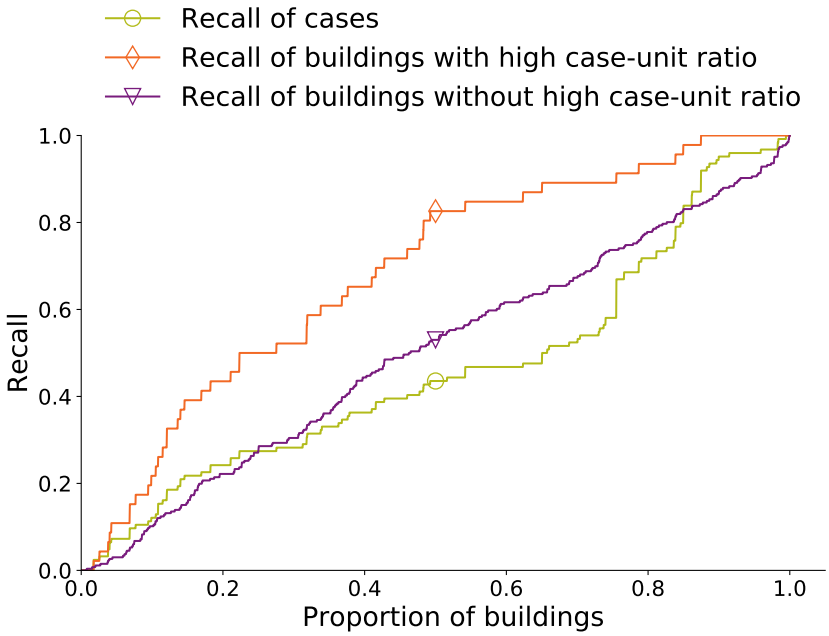


Figure 3.11: Recall curves at each proportion of buildings for the Gradient Boosting model using the threshold label.

In fact, the model using the *threshold label* prioritized buildings with smaller sizes (average number of units = 11) than the model using the *any-case label* (average number of units = 70). This may also account for the phenomenon in Figure 3.11: Recall of buildings with high case-per-unit ratio was higher than average ($slope > 1$); recall of cases was not as high.

To further understand how model of any-case label over optimized large buildings, we additionally compared the two models in two ways. First, we ranked all 6,437 buildings according to the predicted risk score using both models and found that they were somewhat uncorrelated. Second, for each model, we plotted the buildings in the top-k list the model suggested TSU to canvass, respectively (see Figure 3.12). Each point refers to a building with the size of the point representing the number of units in the building. This map shows that predictions using the threshold label (represented

by orange square) top ranked more small-size and geographically distributed buildings than predictions using the any-case label (represented by green circle).

These findings support our assertion that if we only predict whether there would be any case next month, the model would be more likely to provide a list of large buildings as opposed to buildings of high case-per-unit ratio. Depending on their goals, policymakers and canvassing specialists might prefer one or the other — larger buildings might allow for more efficient canvassing to knock on doors that are more geographically co-located (supporting the *any-case label*). On the other hand, the threshold label gives tenants living in smaller buildings more of an opportunity to receive outreach and results in a possibly more equitable outreach process.

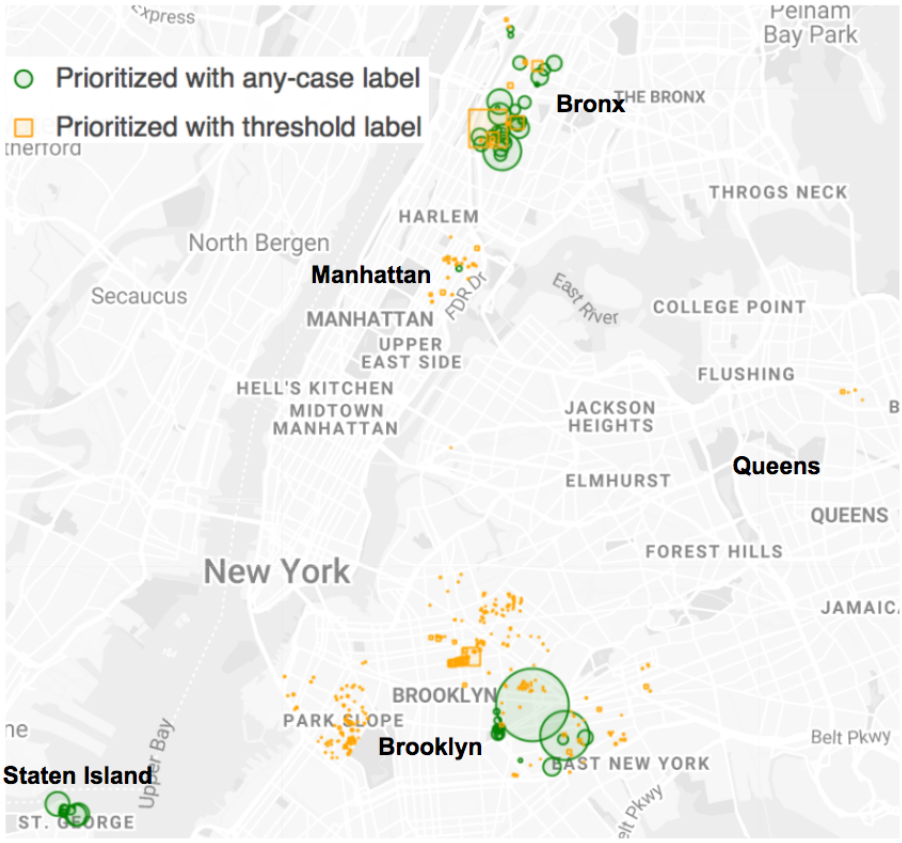


Figure 3.12: Comparing of any-case label and threshold label suggestions. Each point is a building with size representing the # of units. Predictions using any-case label prioritize large size buildings and were more geographically clustered.

3.7 Discussion: Practical Implications and Next Steps Prior to Implementation

The Tenant Support Unit hopes to efficiently find more individuals in need of their help with fewer outreach knocks by generating a list of buildings where tenants are most likely to experience harassment. At the beginning of each month (when TSU team leads typically decide which areas to visit next in the upcoming month), the model will generate a list of buildings for each borough where tenants face high risks of harassment.

A field trial. Prior to the results being used to inform TSU's process, the agency should conduct a field trial to validate the predictions of our model. This field trial can better inform whether buildings that the model flags as high risk are more likely to yield cases than buildings that the model flags as low risk.

Selection bias in the labels. One area of future work we want to explore is to deal with selection bias in our labels and actively collect new labels, with which randomized field trials may also help. Since we only have labels from buildings canvassed by TSU, and there is some bias in how they select buildings to canvass, our model is trained only on that data and will most likely be only confident on predictions made on similar buildings. We want to use the field trials to understand this bias and use the TSU team to also help improve the model by canvassing new buildings to provide more representative labels to train our model.

Clustering buildings of high risks. If the model is able to successfully differentiate buildings, next steps should include efforts to use the list more efficiently — that is, to not waste time travelling across the city to canvass buildings in exact order of high to low risk. TSU could determine clusters of buildings that have a high enough density of units in high-risk buildings to canvass in one or over multiple days (see Figure 3.13). Once these cluster areas are created, specialists can canvass ev-

ery target building in the cluster area without unnecessary travel among exclusively high-risk buildings across boroughs or neighborhoods.

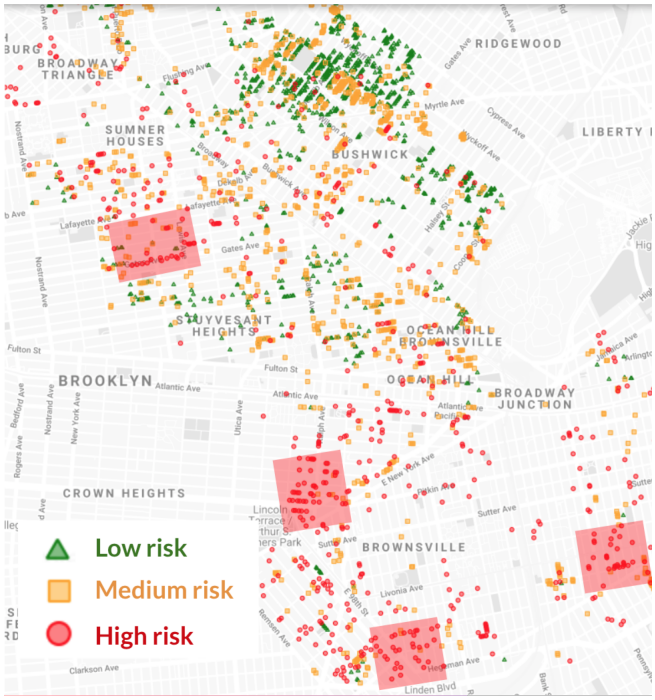


Figure 3.13: Example of post-model implementation with high-risk buildings clustered.

Ethics and fairness. We also note that prioritizing buildings with high predicted harassment risks might raise inequality across different areas and sub-population groups. This inequality in where to canvass (i.e., labeled data) may lead to further inequality in the understanding of and assistance for different groups. Such potential inequality, together with the biases in canvassing data collection and modelling phases, all call for a thorough bias and fairness analysis.

3.8 Conclusion and Take Away

In this study, we use a machine learning approach to help NYC outreach specialists identify buildings where tenants might face landlord harassment. Our model significantly outperforms the current outreach method. The predicted risk scores can

help the agency more accurately prioritize areas of high rental harassment and better allocate building canvassing specialists to help more tenants in need in an equitable manner.

In addition, our model provides insights into the important correlates of harassment that might be useful for researchers without access to agency data confirming harassment, complementing efforts to look at harassment using proxies like a loss in rent-stabilized units [104]. Our results on feature importance not only find the relevance of building-specific attributes — for instance, the building's history of code violations — but also the utility of local demographic and behavioral data to highlight where tenants might face housing issues.

By comparing different formulations of the prediction problem, with different prediction labels, we also show that although formulating the harassment prediction as a binary classification — whether there would be any case next month — significantly increased the precision, it might be biased towards buildings with many units. Finally, we discuss how the model can better facilitate canvass planning and resource allocation by clustering the high-risk buildings for efficient deployment and outreach.

This project shows that even simple machine-learning models have the potential to largely increase work productivity (59% in this case) and that it is critical to formulate a practical problem in the right way. Note, too, that in order to verify the practical effectiveness of the interventions proposed by machine learning models, randomized field trials in the wild are highly suggested. While the constraints in TSU's practice have limited the opportunities for such field trials, in the next chapter, we discuss how to leverage field experiments to examine and establish the causal relationship between interventions and work performance .

CHAPTER IV

Improving Worker Performance in a Gig Economy with a Field Experiment

In this chapter, we describe the design and deployment of a field experiment to examine the effectiveness of behavioral interventions in the wild. The study in Chapter III shows that machine learning models can provide data-driven evidence supporting interventions for outreach specialists (i.e., informing where to go). To improve worker performance, one can also come up with interventions by leveraging insights from social science theories and domain knowledge, as suggested in the framework (Figure 1.1).

However, questions at large are whether these interventions are effective, and to what extent. To answer such causal questions, randomized field experiments have been commonly regarded as a gold standard.

In Chapter IV, we illustrate the design and deployment of field experiments with an application in a gig economy. Specifically, we design a randomized field experiment to examine the effect of virtual teams on worker performance on a leading gig-economy platform. In collaboration with a ride-sharing platform, Didi Chuxing (DiDi), we conduct a large-scale field experiment with 27,790 drivers to organize drivers into teams that are randomly assigned to one of three experimental conditions. Treated drivers receive either their team ranking or their individual ranking within

their team, whereas those in the control condition receive individual performance information without social comparison. We find that treated drivers generate 2% higher revenue than those in the control condition. Further, drivers in the team leaderboard treatment continue to work longer hours on the platform three months after the end of the experiment. Lastly, we show that drivers with below-median revenue prior to the experiment benefit the most from a team contest. This study speaks to the powerful interaction of field experiments and social science theories in this framework (Figure 1.1): one can leverage field experiments to examine the effectiveness of theory-informed interventions in order to promote worker performance.

4.1 Introduction

The gig economy provides workers with the benefits of autonomy and flexibility [36], but it does so at the expense of work-related identity and co-worker bonds. Indeed, many gig platforms have experienced low engagement and high attrition rates among their workers, who note that they typically work alone with "no interaction or relationship with other colleagues," on jobs "that don't lead to anything" [75, 112]. The 2020 Covid-19 pandemic has created a work structure that has placed exponentially more workers in a work-from-home scenario that is susceptible to the same issues related to the lack of in-person interaction as those in a gig economy. By August 2020, 42 percent of the U.S. labor force was identified as working from home full-time [15]. Given that we expect at least some portion of this remote work to remain post-pandemic, an important question is how organizations can help their workers create and maintain positive work-related social connections while working remotely.

To answer this question, we conduct a large-scale natural field experiment using a global ride-sharing platform. Specifically, we form drivers into virtual teams and engage the teams in contests to strengthen team identity. We then evaluate the effects

of these virtual teams on worker productivity and retention.

Our research applies insights from the social identity research from psychology [128, 27] as well as studies in behavioral economics [4, 5]. In a lab setting, this research shows that, when people feel a stronger sense of common identity with a group using either induced [56, 35, 38] or natural identities [68, 41], they exert higher effort and make more contributions to improve group outcomes. Field experiments show a similar positive effect of identity-based teams in increasing pro-social behavior in fruit harvesting [58] and online peer-to-peer pro-social lending [2, 34]. By contrast, other field experiments have found that when workers are paid by piece rate, providing team ranking information might reduce average worker productivity for teams that are not randomly assigned [11].

To estimate the causal effects of team incentives on productivity and retention, we randomly assign teams into different experimental conditions using DiDi Chuxing (DiDi), which is the largest ride-sharing platform in China. We then examine the effect of team contests on individual driver behavior. To our knowledge, this is the first natural field experiment to examine the effect of virtual teams in a large-scale field setting. To design our contests, we draw on insights obtained from an earlier field experiment conducted in the southern Chinese city of Dongguan in August 2017. In this earlier experiment, we randomly assigned 2,100 drivers drivers into seven-person teams to compete for a cash prize across a five-day period. Team compositions are determined either randomly or based on homophily in age, hometown location, or productivity. The results from this earlier experiment show that, compared to those in the control condition, treatment drivers work longer hours and earn 12% higher revenue during the contest period. We find that teams formed based on age similarity are more productive two weeks post-contest than randomly-formed teams [3].

Encouraged by the results of this first field experiment, in 2018, DiDi conducted 1,548 team contests across 180 cities in China, involving over two million drivers

placed into teams based on hometown or age similarity. These contests, typically one week in duration, helped the platform meet the high tourist demands during national holidays, and increased both driver income and retention [145]. A common feature among the 1,548 team contests DiDi ran in 2018 is that they were all one-week contests with cash incentives, and the teams existed only for the duration of the contest. As a result of this latter factor, the DiDi contest initiative did not provide an opportunity to study the long-term effects of team membership on organization identity and teammate bonds.

Our study investigates the long-term effects of team formation on the same platform in the context of contests, but without a monetary incentive to participate in the contest. Specifically, in October 2018, we conducted a natural field experiment on the DiDi platform involving 27,790 drivers across three cities: Beijing, Kunming, and Taiyuan. The experiment ran from October 22 to December 3, 2018. To evaluate our treatment effect on driver retention, we continued to collect data for three months after our experiment ended.

In our experiment design, we vary whether teams receive social comparison information through the provision of a leaderboard that indicates team ranking or individual ranking within a team, or whether they receive only individual performance information (control). In the Team Leaderboard treatment, drivers are provided with access to both team and individual leaderboards. We send a daily reminder to these drivers to check the rankings of the same five teams within their leaderboard, as well as individual teammate rankings within their team. In the Individual Leaderboard treatment, drivers are provided with access only to the individual leaderboard within their team. Again, we send a daily reminder to drivers to check their individual rankings. Finally, in the control condition, drivers receive no leaderboard access. However, to maintain the same communication frequency across experimental conditions, these drivers receive a daily reminder that they are able to access their own income statis-

tics in the app. With the exception of the normal piece rate, there is no monetary incentive in any of the experimental conditions.

Across the three-week contest intervention, we find that drivers in the team and individual leaderboard treatments generated 1.7% higher revenue than those in the control condition. Investigating the two treatments separately, we find that drivers under the team (individual) leaderboard treatment generated 1.8% (2%) higher revenue than those in the control condition. Examining the city level results, the team leaderboard treatment in Taiyuan (individual leaderobard in Beijing) leads to a 5.3% (2.3%) increase in driver revenue compared to the control condition, whereas neither treatment has a significant effect in Kunming. Our observed city-level difference is likely due to the fact that both Beijing and Taiyuan had a respective 90% passenger order fulfillment rate pre-intervention, compared to Kunming, which was already meeting 98% of passenger orders pre-intervention. Three months after the experiment ended, we find that drivers in the team leaderboard treatment continue to work longer hours on the platform. Within the teams themselves, we find that those identified as "laggards" benefit the most from team contests.

Overall, our results show that platform designers can leverage team identity and team contests to increase revenue and worker engagement in a gig economy. More broadly, our research demonstrates the value of a social-relational approach [110] which puts teams and social relationships into the gig economy.

4.2 Related Work

4.2.1 The Gig Economy

Our research contributes to the rapidly growing literature on the gig economy and the future of work more broadly. To inform better design of the gig-economy work practices, a growing literature investigates the operation and estimates the benefits

and risks of the gig economy for individuals and society (e.g., [54, 59, 90, 93, 109]). Specifically, in the ride-sharing economy, researchers have explored the socio-economic effects on and consequences of ride-sharing platforms, such as Uber and Lyft [49, 54, 82, 96, 150]. Another stream of literature focuses on motivation and incentives for participation. Inspired by the findings that economic gains positively influence people's intention to participate [74], research has quantified the positive effect of dynamic pricing [37], subsidy [61] on improving supply-demand efficiency, the gender wage gap in ride-sharing [50], the value of flexible work [36], the determinants of tipping [32, 33], the effects of apologies for late trips [73], and the value of passenger waiting time [69]. Our study adds to this literature by investigating the effectiveness of team identity and social information on driver participation in a ride-sharing economy.

4.2.2 Team Contest and Team Identity

Team contests have been widespread among humans in the real world, such as the sports teams who compete for the victory of games [118]. A recent survey study summarizing findings in more than a hundred studies shows a common phenomenon that people expend more effort when participating in the team contests [124]. While this finding is robust in both theoretical prediction and laboratory experiments, there is seldom an opportunity to test in the field.

With the rise of modern technologies, team competitions have been increasingly applied in online communities, such as in crowdsourcing [114], education [119], online games [44], charitable giving [39, 2], and ride-sharing service provision [3]. These empirical studies have shown that team competitions are effective in promoting participation and work quality (e.g., [105, 119, 3]). In particular, a crowdsourcing field experimental study examining the effect of team contest (versus individual and individual contest) and the bonus allocation strategy (rewarded according to team performance or individual performance) on crowdsourcing productivity [114] shows that

workers are the most productive when they are in teams with both team-based bonus and individual-based bonus. Recent field work in the ride-sharing context shows that cash-rewarded team contests are effective in promoting driver participation during the contests [3].

4.3 Experiment Design

As mentioned, we conduct a natural field experiment on the DiDi platform involving 27,790 drivers across Beijing, Kunming, and Taiyuan, three cities chosen to exemplify diversity in demographics, location, and the number of team contests hosted on DiDi prior to our experiment (see Table 4.1 for more details). Our experiment is approved by the University of Michigan IRB (HUM00153090), and pre-registered at the AEA RCT Registry [144].

Table 4.1: City Characteristics

City	Location	# of historical contests	Order fulfillment rate	Population (million)	# of drivers registered the experiment	# of participants
Beijing	Northern	17	0.90	21.54	21,126	18,900
Taiyuan	Central	14	0.90	4.42	4,648	3,815
Kunming	Southwest	5	0.98	6.85	5,776	5,075

Note: Order-fulfillment rate is calculated by data of two weeks before the experiment (i.e., 2018/10/08-2018/10/21).

Of the three cities where we implemented our experiment, Beijing is the capital of China, located in northern China, with over 21.54 million residents. Taiyuan is the capital of Shanxi province, located in central China, with a population of 4.42 million. Kunming is the capital of Yunnan province, located in southwest China, with a population of 6.85 million.[1]

[1]Population data at the end of the year of 2018 are retrieved from:
the National Bureau of Statistics of China (http://www.stats.gov.cn/tjsj/ndsj/2019/indexch.htm),
the Province Bureau of Statistics of Yunnan (http://stats.yn.gov.cn/tjsj/tjnj/201912/t20191202_908222.html),
and the City Bureau of Statistics of Taiyuan (http://stats.taiyuan.gov.cn/doc/2019/05/14/845586.shtml).

The experiment was conducted from October 22, 2018 to December 3, 2018. To evaluate our treatment effect on driver retention, we continue to collect data for three months after our experiment, until March 15, 2019. In addition to our recruitment and team formation stage, our experiment is organized into pre-intervention, intervention, and post-intervention stages (see Figure 4.1 for the experimental process).

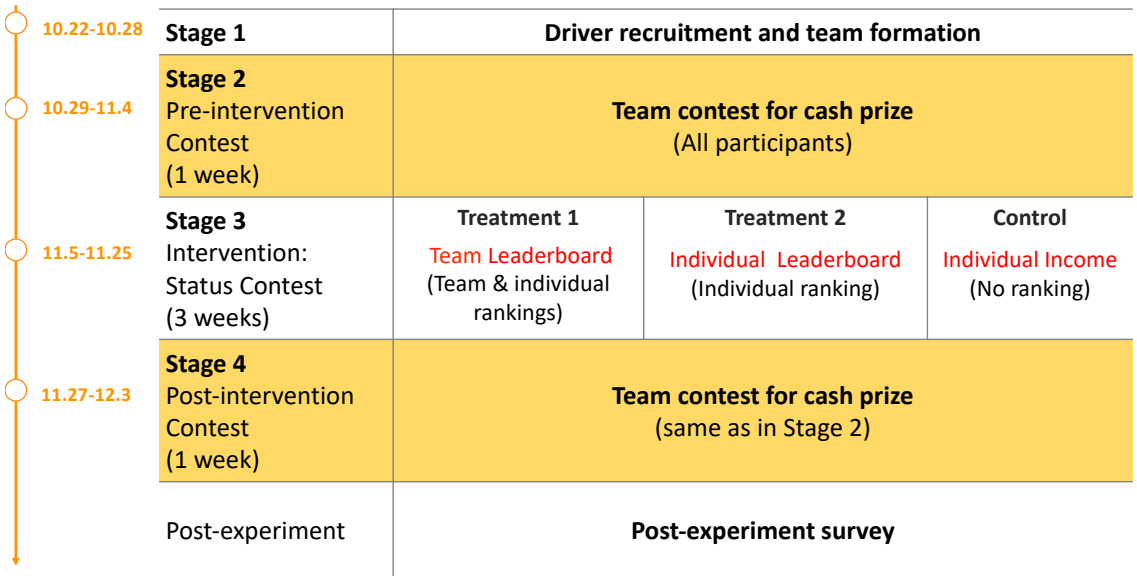


Figure 4.1: Experiment process

Driver recruitment and team formation. For the driver recruitment and team formation stage, the platform used its built-in process which was informed by our earlier experiment [3]. To obtain our participants, we ask DiDi to send an invitation on October 22, 2018 to all drivers in our three cities to participate in a week-long team contest for a monetary prize. To do so, our collaborators at DiDi send out both text messages and in-app push notifications[2] to inform all drivers in the experimental cities about the team contest. The English translation of the call for participation to the drivers reads as follows:

The DiDi driver team contest is about to start soon! Say goodbye to the

[2]Text message refers to the normal message sent out by DiDi. In-app push notification refers to the message popping up within the DiDi app.

> lonely driving work on your own. Get to know new driver friends and compete for rewards with your teammates! Click here to register for the contest. Please keep up your good service and drive safely.

Interested drivers are invited to sign up for the contest and start forming teams. Drivers can create a new team as a captain, invite others to join their team, or join an existing team if invited to do so.

While teams are designed to have seven members, 36% of our teams achieved the desired size during the team formation period. Those that reached the desired size during the team formation period are referred to as *self-formed* teams. At the end of recruitment stage, the system then randomly selects 90% of the drivers in under-sized teams and groups them into full-sized teams, which we refer to as *system-formed* teams. The system-formed teams are based on either hometown or age similarity, two of the most successful team formation algorithms from our earlier experiment [3]. The remaining 10% are not assigned to any team and do not participate in the contest. These drivers are referred to as *solo drivers*. In our analysis, we control for whether a team is self-formed.

Finally, we sort teams into contest groups. To assign the teams into contest groups, we first sort teams within each city decreasingly based on their prior revenue (the sum of individual team members' revenue in the two weeks prior to the beginning of the experiment). We then partition every five adjacent teams into a contest group, also referred to as a *leaderboard*. Teams compete only with other teams in the same leaderboard. Our grouping method ensures that teams in the same leaderboard have similar prior productivity. We now describe the three stages of the team contest.

The pre-intervention contest. Following previous studies that find that inter-group competition is among the most successful methods used for creating a strong sense of group identity [56], we conduct a pre-intervention best-of-five team contest.

In this contest, within each leaderboard, the team with the highest cumulative team revenue during the contest week wins a cash prize, whereas the other four teams receive no prize. Following DiDi's current contest practice, we exclude the lowest driver revenue in a given team in each day when calculating the team's daily cumulative team revenue. This allows one driver on a team to take a day off without affecting team performance.[3] The cash prize is 1,000 RMB (per winning team) for Beijing, and 650 for Taiyuan and Kunming, respectively, adjusted by the drivers' average hourly revenue in each city. The prize is allocated to members of the winning team proportional to their contributions to the cumulative team revenue, an allocation shown to incentivize group members in laboratory contests [124], and is credited to their driver accounts immediately after the contest.

During this stage, all drivers participating in the contest can use the DiDi app to access both a team leaderboard and an individual leaderboard for social information, as illustrated in Figure 4.2. The team leaderboard shows the cumulative revenue of each of the five teams in the contest group in descending order (top left panel of Figure 4.2). The top three teams are highlighted with badges. The individual leaderboard shows individual members' daily revenue in descending order for those within a given team (top right panel of Figure 4.2). In addition, we mark the average performance of that team with a line on the individual leaderboard to enhance the effect of ranking [40, 42]. The team ranking is updated every hour while individual revenue is updated in real time. We send each driver a daily reminder of the contest and the leaderboards at the end of each day. The reminder is sent by both text message and in-app push notification as follows.

> The driver team contest has become more intense! Want to know your team's ranking? Want to check your teammates' performance? Want to know your competitors' performance? Click this link and you can access

[3] In some cities, such as Beijing, to reduce air pollution, each license plate must be off the street on a designated day of the week, typically determined by the last digit of the plate number.

all the above information. Please keep up your good service and drive safely.

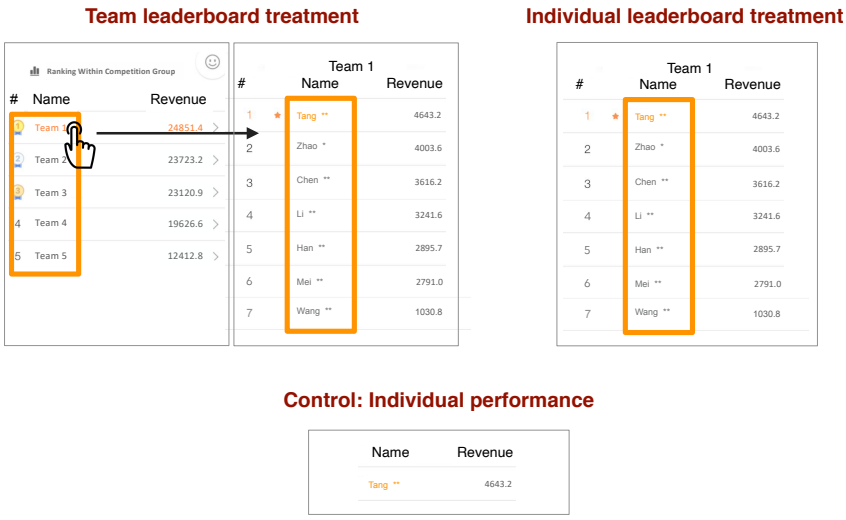


Figure 4.2: APP interfaces (mock-up) of team leaderboard, individual leaderboard, and control group

The intervention: A status contest. Immediately after the pre-intervention contest, we randomly assign each leaderboard to one of three experimental conditions and conduct a three-week status contest between November 5-25 to examine the effect of team identity on driver revenue and retention.

- *Team Leaderboard.* In this treatment, drivers continue to have access to both the team and individual leaderboards as in the short-term contest. We send out a daily reminder to these drivers to check the rankings of the same five teams within their leaderboard, as well as individual member rankings within their team.

- *Individual Leaderboard.* In this treatment, drivers have access to only the individual leaderboard within their team. Again, we send out a daily reminder to drivers to check their individual rankings.

Table 4.2: Randomization check and summary of statistics

	Beijing			Taiyuan			Kunming		
	Team	Individual	Control	Team	Individual	Control	Team	Individual	Control
Daily Revenue	381.97	381.71	381.70	171.64	180.54	176.09	212.71	214.21	218.03
before experiment	(215.35)	(216.15)	(213.83)	(126.25)	(129.99)	(125.88)	(144.30)	(143.99)	(144.56)
Age	36.82	36.91	37.35	36.53	36.63	36.86	36.49	35.91	37.02
	(8.12)	(8.09)	(8.28)	(8.26)	(8.22)	(8.34)	(8.50)	(8.58)	(8.81)
Male	0.97	0.97	0.97	0.97	0.95	0.96	0.93	0.92	0.93
	(0.17)	(0.16)	(0.17)	(0.18)	(0.22)	(0.19)	(0.26)	(0.26)	(0.26)
DiDi age (month)	24.69	24.97	24.86	24.08	23.88	24.55	15.06	14.70	14.51
	(13.19)	(13.11)	(13.01)	(11.13)	(11.62)	(11.04)	(11.04)	(11.01)	(10.98)
Self-formed teams	0.38	0.37	0.37	0.36	0.38	0.26	0.31	0.31	0.30
	(0.49)	(0.48)	(0.48)	(0.48)	(0.49)	(0.44)	(0.46)	(0.46)	(0.47)
Hometown distance	451.93	465.04	463.66	114.50	121.68	109.06	249.72	293.28	289.13
to the contest city	(281.53)	(283.85)	(285.47)	(117.96)	(135.84)	(100.40)	(219.29)	(326.88)	(343.39)
# of leaderboards	180	180	180	37	36	36	49	48	48
# of teams	900	900	900	185	180	180	245	240	240
# of drivers	6,300	6,300	6,300	1,295	1,260	1,260	1,715	1,680	1,680

Standard deviation in parentheses
Pairwise Kolmogorov-Smirnov tests for every variable across the three conditions are not significant ($p > 0.10$) for each of Beijing, Taiyuan, and Kunming.

- *Control.* In the control condition, drivers cannot access either leaderboard. However, to keep the same communication frequency, drivers continue to receive a daily reminder that they can access their own income statistics in the app.

While drivers continue to earn piece rate, we do not provide additional monetary incentives for the status contest.

The randomization is stratified based on the average productivity of a given leaderboard prior to the experiment. Kolmogorov-Smirnov tests show that the distribution of pre-experiment revenue, age, gender, length of time with DiDi, team formation approach, and hometown distance to the contest city is not significantly different in pairwise comparisons across the three conditions ($p > 0.10$, see Table 4.2). Table 4.2 also reveals interesting facts about our drivers: more than 95% of them are male, with an average age of 37. Looking at their hometown distance to the contest city, we conclude that Taiyuan drivers are predominantly local, whereas Beijing and Kunming drivers are mostly migrants. In China, DiDi drivers comprise of workers laid off from their traditional jobs, veterans, migrant workers from rural areas, and commuters

who offer rides during their daily commute.

We carefully manipulate the communications provided to drivers during this stage. The corresponding notification and reminder for the three experiment conditions in the long-term status competition include:

1. **Team leaderboard treatment.**

   At the beginning of this stage, drivers in the team leaderboard condition are notified by text message that:

   > The team contest is over. The ranking information will continue to be updated during November. Please pay attention to the performance of your team and your teammates. DiDi is amazing because of you!

   The following reminder is sent by text message and in-app push notification once a day during the evening:

   > Latest performance just came out! Want to know your team's and teammates' performance? Click this link and you can access all the information. Please keep up your good service and drive safely.

2. **Individual leaderboard treatment.** At the beginning of this stage, we send the following text message to notify drivers in the individual leaderboard condition that:

   > The team contest is over. The ranking information will continue to be updated during November. Please keep your attention on the performance of your teammates. DiDi is amazing because of you!

   The following individual performance reminders are sent every evening by both text message and in-app push notification:

> Latest performance just came out! Want to know your teammates' performance? Click this link and you can access all the information. Please keep up your good service and drive safely.

3. **Control.**

   At the beginning of this stage, we send the following text message to drivers in the control group that:

   > The team contest is over. Please pay your attention to your performance. DiDi is amazing because of you!

   An individual performance update reminder is sent every evening by text message and in-app push notification as follows:

   > Latest performance just came out! Want to know the your outcome? Click this link and you can go to the your revenue page. Please keep up your good service and drive safely.

The post-intervention contest. On November 26, we send each driver a message announcing a one-week contest for a cash prize from November 27 to December 3 under the same leaderboard groups, and prize parameters as the pre-intervention contest. This post-intervention contest is designed to evaluate the treatment spillover effects on individual driver productivity immediately after the intervention. The following text message announcement is sent to all drivers in all three conditions on the day before the pre-intervention contest to notify them of the contest:

> Here comes the driver team contest again (from 2018.11.27 to 2018.12.3)! You don't need to form a team again. Team members and competitor teams will remain the same as in the last contest. Please contact your team members and get ready to compete for the cash prize!

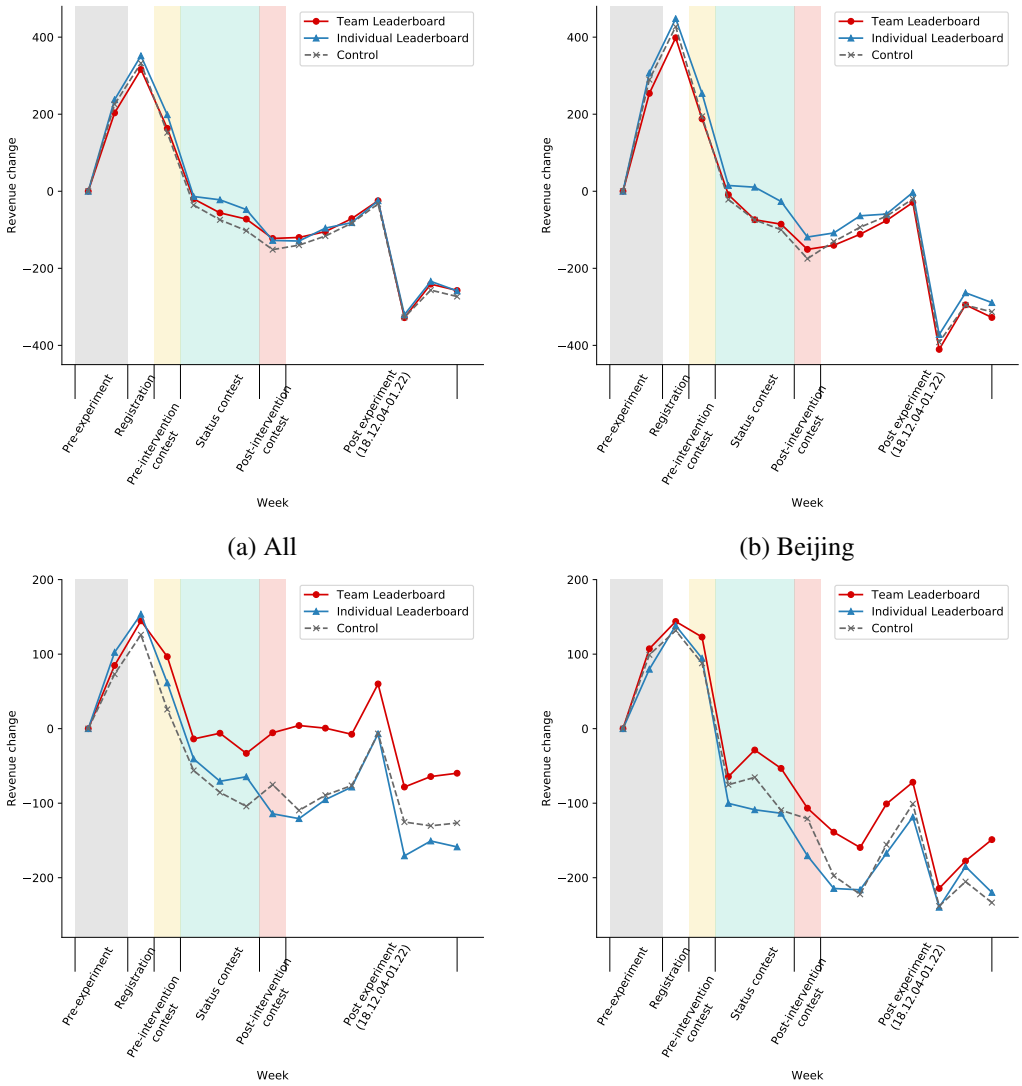
The post-experiment survey. After the post-intervention contest, all drivers receive a survey which evaluates their sense of belonging related to their team as well as to the organization (DiDi). The survey questions and responses are included in Section 4.6.9 of the Extended Materials.

4.4 Results

Our experiment yields findings related to the immediate and long-term effects of virtual teams on driver productivity and retention, both overall [144] and at the city level. On the DiDi platform, drivers receive 81% of the revenue they generate and give the remaining 19% to the platform. Therefore, using revenue as one of our outcome variables is equivalent to using driver earning or platform profit.

We first examine the average treatment effect on driver revenue during the experiment period. In Figure 4.3, we plot the weekly average driver revenue for each experimental group. To better compare the treatments, we realign the lines based on revenue earned during the pre-experiment period. The y-axis presents the revenue difference between a given week and the baseline week in the pre-experiment period. Note that the three lines coincide up to the start of the pre-intervention contest period. However, during the status contest intervention, since drivers in different treatment conditions receive different social information, the lines in Figure 4.3 (a) start to diverge. Pooling all three cities, we observe that our treatment drivers are more productive on average than those in the control condition both during and after the intervention.

In our first pre-registered hypothesis, based on prior laboratory experiments on social identity and team competition [56], we predict that drivers in our treatment conditions will generate higher revenue than those in the control condition as their exposure to a leaderboard should facilitate a team identity. The comparison between team leaderboard and individual leaderboard is motivated by laboratory experiments



(a) All

(b) Beijing

(c) Taiyuan

(d) Kunming

Figure 4.3: Average weekly driver revenue under each experimental condition. To better visualize the changes over time, we re-scale the revenue within each experimental condition with reference to its pre-experiment average weekly revenue from the week of October 8-14, i.e., two weeks before the start of the experiment. For example, each point represents the weekly average revenue per driver under that experimental condition minus the pre-experiment weekly average revenue per driver under the same experimental condition.

in group contests [46, 47].

Hypothesis 1 (Status contest)**.** *(a) Treated drivers are more productive than those in the control condition; and (b) drivers in the team leaderboard condition are more productive than those in the individual leaderboard condition during the status contest phase.*

To quantify the average treatment effects on outcome, Y, we construct the following difference-in-differences model:

$$\Delta \mathrm{Y}_{i,t} = \beta_0 + \beta_1 \mathrm{Treated} + \alpha_c + \epsilon_{i,t}, \tag{4.1}$$

where $\Delta \mathrm{Y}_{i,t}$ represents the outcome change in the t-th week in the current period compared to the corresponding pre-contest week(s),[4] and α_c captures city fixed effects. Hypothesis 1(a) implies that $\beta_1 > 0$ in Equation (4.1).

We report the main results in Tables 4.3 to 4.6 in the main text, and the results of our robustness checks in the Extended Materials (EM). To correct for multiple hypothesis testing, we report the false discovery rate adjusted q-values in square brackets [18, 7]. To claim significance, we use a 5% (10%) cutoff for our p-values (q-values) [57].

The results in column 1 of Table 4.3 show that our treatment conditions increase driver revenue by 34.53 RMB, or 1.66% of the average weekly revenue per driver, during the three-week intervention ($p < .05$). Therefore, we reject the null hypothesis in favor of Hypothesis 1(a). We further find a significant treatment effect for drivers in Beijing (41.67 RMB, $p < .05$, 1.69% of average weekly revenue), but not in Taiyuan or Kunming. Our findings are strengthened (39.08 RMB, or 1.88% of average weekly

[4]For the pre-experiment baseline week(s), we use the week before the experiment (Oct. 15-21, 2018) for one-week target periods, i.e., the pre-intervention contest, the post-intervention contest, and retention. For the status contest, we use the two weeks before the experiment (Oct. 8-21, 2018) as our baseline, as the week of October 1-7, 2018 was a national holiday with drastically different demand and supply for ride-sharing.

Table 4.3: Average and heterogeneous treatment effects on weekly revenue during the intervention (status contest): Difference-in-differences panel regressions.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Treated	34.53**	41.67**	33.99	8.25	39.08**	45.82**	38.40	14.53
(In a virtual team)	(15.37)	(21.01)	(23.86)	(24.97)	(15.31)	(20.93)	(23.69)	(24.81)
		[0.17]	[0.18]	[0.33]		[0.09]	[0.12]	[0.23]
Age					6.98***	7.47***	1.90	8.39***
(Year)					(0.83)	(1.17)	(1.37)	(1.27)
DiDi age					32.16***	40.85***	3.64	3.43
(Year)					(7.47)	(9.59)	(11.53)	(13.39)
Hometown distance					-0.02	-0.01	-0.12**	-0.03
to contest city (km)					(0.02)	(0.02)	(0.05)	(0.02)
Self-formed team					-45.25***	-60.09***	-24.18	-4.10
					(16.09)	(21.59)	(27.49)	(26.90)
City fixed effect	Yes	-	-	-	Yes	-	-	-
# of clusters	11,890	8,100	1,625	2,165	11,890	8,100	1,625	2,165
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team (individual) level for treated (control) drivers. False Discovery Rate adjusted q-values calculated separately for individual cities (2-4) & (6-8) are reported in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

revenue) when we control for demographics and self versus system formation of teams (columns 5-8). Consistent with social identity theories focusing on the effects of social status and social distance on individual identification with social groups [21, 123], a driver's hometown distance from the contest city is negatively correlated with their productivity ($p < 0.05$, column 7). Interestingly, self-formed teams generate lower revenue compared to system-formed teams using hometown or age similarity ($p < 0.01$, columns 5-6). We also note that older drivers and those who have joined the platform earlier generate higher revenue.

Investigating the two types of interventions separately (Hypothesis 1(b)), we fur-

ther expect that drivers in the team leaderboard treatment will generate higher revenue than those in the individual leaderboard treatment, who in turn will generate higher revenue than those in the control group during our intervention period. This hypothesis implies that $\beta_1 > 0$, $\beta_2 > 0$, and that $\beta_1 > \beta_2$ in Equation (4.2) below.

$$\Delta \mathrm{Y}_{i,t} = \beta_0 + \beta_1 \text{Team Leaderboard} + \beta_2 \text{Individual Leaderboard} + \alpha_c + \epsilon_{i,t}, \quad (4.2)$$

The results in column 1 of Table 4.4, show that drivers in the team leaderboard treatment generate 32.12 RMB marginally higher weekly revenue compared with the control group ($p < .10$), while those in the individual leaderboard condition generate 36.96 RMB higher revenue compared with the control group ($p < .05$). After controlling for demographics and self versus system formation of teams, we see from column 5 that the team (individual) leaderboard generates 36.7 RMB (41.47 RMB) more weekly revenue, equivalent to a 1.76% (1.99%) increase ($p < .05$ in each case), although the difference between the two treatments is not significant ($p > .10$).

We next examine our city-level results. From Table 4.4, we see that only the individual leaderboard treatment has a significant effect on revenue (56.32 RMB per week, or 2.29% of the weekly revenue of the control group, $p < .05$) for drivers in Beijing (columns 2 and 6), whereas in Taiyuan (columns 3 and 7), only the team leaderboard treatment has a significant effect on revenue compared to the control condition (58.49 RMB per week, $p < .05$). By contrast, neither treatment has a significant effect on weekly revenue for drivers in Kunming. As shown in Table 4.1, passenger order fulfillment rate was already quite high (98%) in Kunming before our experiment; thus, there was little room for a substantial improvement in revenue. In comparison, 90% of the orders were fulfilled in Beijing and Taiyuan during the same time period. After controlling for demographics and team formation methods

Table 4.4: Average and heterogeneous treatment effects on weekly revenue during the intervention (status contest): Difference-in-differences panel regressions investigating the two treatments separately.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	32.12*	27.03	58.49**	30.54	36.70**	32.40	62.31**	33.81
(β_1)	(17.97)	(24.61)	(26.60)	(29.91)	(17.90)	(24.50)	(26.57)	(29.69)
	[0.08]	[0.44]	[0.09]	[0.44]	[0.04]	[0.33]	[0.06]	[0.34]
Individual leaderboard	36.96**	56.32**	8.81	-14.50	41.47**	59.24**	13.68	-5.18
(β_2)	(17.90)	(24.49)	(28.76)	(28.03)	(17.82)	(24.37)	(28.43)	(27.86)
	[0.08]	[0.09]	[0.86]	[0.86]	[0.04]	[0.06]	[0.61]	[0.62]
Age					6.98***	7.47***	1.91	8.35***
(Year)					(0.83)	(1.17)	(1.37)	(1.28)
DiDi age					32.15***	40.77***	3.57	3.29
(Year)					(7.46)	(9.59)	(11.57)	(13.39)
Hometown distance					-0.02	-0.01	-0.12**	-0.03
to contest city (km)					(0.02)	(0.02)	(0.05)	(0.02)
Self-formed team					-45.22***	-59.76***	-23.62	-3.96
					(16.10)	(21.59)	(27.40)	(26.91)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.79	0.25	0.08*	0.13	0.80	0.29	0.08*	0.18
# of clusters	11,890	8,100	1,625	2,165	11,890	8,100	1,625	2,165
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team (individual) level for treatment (control) conditions. False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(columns 6-8 of Table 4.4), the city-level treatment effects remain statistically and economically significant, with the size of the individual and team leaderboard effect equal to 59.24 RMB in Beijing and 62.31 RMB in Taiyuan ($p < .05$ in each case). Furthermore, the difference between the two treatments in Taiyuan is in the direction

we hypothesize (Hypothesis 1b), albeit marginally significant ($p < .10$, columns 3 and 7). This leads to our first main result.

Result 1 (Virtual teams and productivity). During the status contest intervention, (1) drivers in virtual teams generate 1.9% higher revenue than those in the control condition; (2) drivers in the team (individual) leaderboard treatment generate 1.8% (2%) higher revenue than those in the control condition; and (3) at the city level, the team (individual) leaderboard treatment leads to a 5.3% (2.3%) increase in driver revenue in Taiyuan (Beijing) compared to the control group, whereas neither treatment has a significant effect in Kunming.

Note that our status contest belongs to the class of information provision experiments. The effect sizes reported in Result 1 are largely consistent with the meta-analysis results using 126 randomized control trials covering 23 million individuals [51].

We are also interested in the question of whether our team effect persists over time. To evaluate the short-term spillover effect of our intervention, we implement a one-week best-of-five contest with a monetary reward immediately after the intervention. We expect that the treatment effects will persist during this post-intervention contest (pre-registered Hypothesis 2).

Hypothesis 2 (Treatment Persistence)**.** *Drivers in the team leaderboard condition are more productive than those in the individual leaderboard condition, who in turn are more productive than those in the control condition during the post-intervention contest.*

The results in column 1 of Table 4.5 show that drivers in the team leaderboard treatment generate 49.91 RMB higher weekly revenue during our post-intervention contest, or a 2.49% increase, compared to those in the control group ($p < .05$). By contrast, drivers in the individual leaderboard treatment do not differ significantly

Table 4.5: Average and heterogeneous treatment effects on weekly revenue in the post-intervention contest: Difference-in-differences panel regressions.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	49.91**	59.89*	58.03	6.05	55.75**	67.20**	59.78	11.14
(β_1)	(23.80)	(32.49)	(37.50)	(39.57)	(23.44)	(31.92)	(36.92)	(39.00)
	[0.08]	[0.32]	[0.32]	[0.56]	[0.04]	[0.27]	[0.27]	[0.39]
Individual leaderboard	11.75	38.98	-68.26*	-30.36	17.55	42.82	-65.75*	-18.30
(β_2)	(24.30)	(33.12)	(39.25)	(39.52)	(23.84)	(32.42)	(38.27)	(39.01)
	[0.46]	[0.32]	[0.32]	[0.36]	[0.30]	[0.27]	[0.27]	[0.39]
Age					10.56***	11.31***	4.72***	11.57***
(Year)					(1.07)	(1.50)	(1.70)	(1.68)
DiDi age					84.14***	97.94***	38.20**	38.55**
(Year)					(9.62)	(12.33)	(15.49)	(17.20)
Hometown distance					-0.03	-0.04	-0.16**	0.02
to contest city (km)					(0.02)	(0.03)	(0.06)	(0.03)
Self-formed team					-20.55	-39.15	23.93	28.60
					(21.57)	(28.73)	(38.61)	(37.24)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.11	0.53	0.00***	0.33	0.11	0.45	0.00***	0.42
# of clusters	3,970	2,700	545	725	3,970	2,700	545	725
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team level. False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

from those in the control group ($p > .10$). Although the coefficient for the team leaderboard dummy is greater than that for the individual leaderboard, this difference is not significant at the aggregate level ($p = .11$, column 1). Our results are strengthened when we control for demographics as well as self versus system formation of teams (column 5).

At the city level, Beijing drivers in the team leaderboard treatment generate 59.89 RMB marginally higher revenue during our post-intervention contest than those in

the control group ($p < .10$, column 2), and significantly higher when we control for demographics and team formation methods (67.20 RMB, $p < .05$, column 6). By contrast, we find no persistent effect of the individual leaderboard treatment for Beijing drivers.

For drivers in Taiyuan, those in the team leaderboard treatment do not differ significantly from those in the control group ($\beta_1 = 58.03$ RMB, $p = .12$), but do generate significantly higher revenue during our post-intervention contest than those in the individual leaderboard treatment ($\beta_1 \neq \beta_2$, $p < .01$ in columns 3 and 7). It is worth noting that those in the individual leaderboard treatment exhibit a marginally significant reduction in average weekly revenue during the post-intervention contest compared to the control group (-68.26 RMB, $p < .10$, columns 3 and 7). Again, we observe no treatment effect for drivers in Kunming (columns 4 and 8). Based on a theoretical model of individual status contests [101], depending on the properties of the ability distribution function, the aggregate revenue under an individual leaderboard can be lower than that under the control condition, as we observe in Taiyuan. We state our results related to the persistence of our treatment effect below.

Result 2 (Treatment Persistence). During the one-week post-intervention contest, drivers in the team leaderboard treatment continue to generate 2.49% more weekly revenue compared to those in the control group, whereas the individual leaderboard treatment no longer has an effect.

In addition to testing whether teams incentivize individual drivers to generate more revenue, we are interested in whether these individuals are more likely to continue working as drivers. Driver retention is a key challenge for ride-sharing platforms across the globe. As such, an important goal for our intervention is to evaluate the effects of virtual teams on driver retention. Specifically, we hypothesize that drivers who are part of a virtual team are more likely to continue as drivers than those in

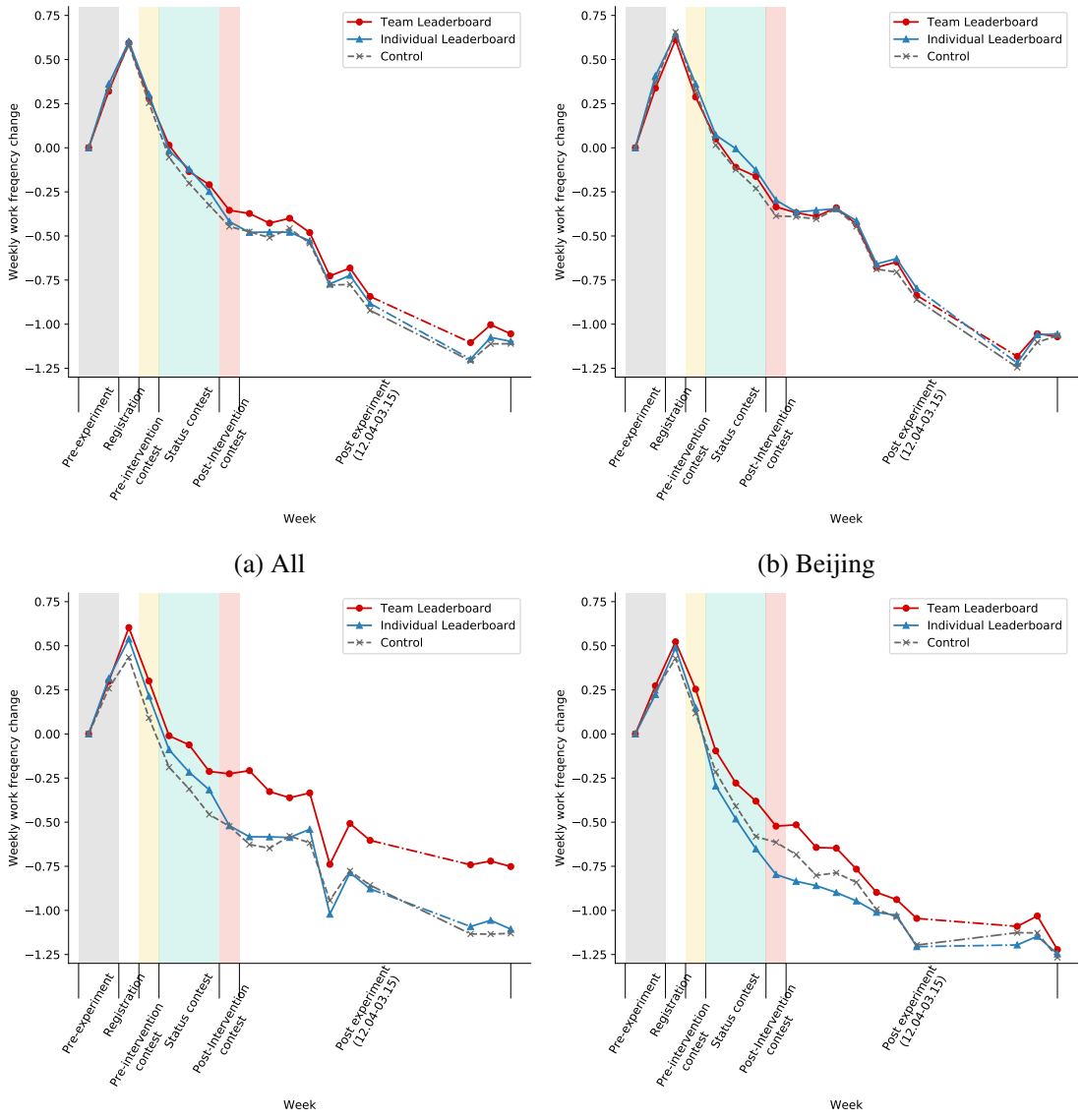
the control group (our pre-registered Hypothesis 3).

Hypothesis 3 (Retention)**.** *Drivers in the team and individual leaderboard conditions are more likely to stay in DiDi than those in the control condition both during and post our experiment.*

To examine the effect of team membership on driver retention, we measure driver retention one week, one month, and three months after the end of our experiment. Unlike workers in traditional sectors whose departure is unambiguous, gig workers who quit typically do not delete their app. Furthermore, it is possible those who have quit driving may still log into the app. Therefore, we use whether they drive for the platform in a given day rather than app login as our retention measure. Specifically, we measure retention as the number of days that a driver provides at least one ride and separately analyze retention during the week immediately (Table 4.11), one month (Table 4.13), and three months (Table 4.6) after the post-intervention contest.

As shown in Figure 4.4, drivers in the team leaderboard treatment consistently exhibit higher retention than those from either of the other experimental conditions. From Table 4.6, we see that drivers in the team leaderboard treatment on average work 0.1 days (or an hour) more than those in the control group in the week three months after the experiment ended ($p < 0.01$, columns 1 and 5). Furthermore, we find that drivers in the team leaderboard treatment also outperform those in the individual leaderboard treatment ($p = 0.02$, columns 1 and 5). The effect is robust and the effect size is stable across different time windows (Tables 4.11 and 4.13 in SM). Finally, we observe no significant difference in retention across any of the periods between those in the individual leaderboard treatment and those in the control group. These results are robust after controlling for demographic covariates and team characteristics, such as whether a team has won the post-intervention contest.

Examining our city-level results, columns 2 to 4 in Table 4.6 show significant differences in driver retention across cities. Indeed, only in Taiyuan do we see a



(a) All

(b) Beijing

(c) Taiyuan

(d) Kunming

Figure 4.4: Average work frequency of each condition over a week (all cities). To better visualize the change over time, we scale each condition by taking a difference of the average weekly days of driving during the week before the experiment. For example, each point in the treatment line equals the weekly average working days per driver of treatment group minus the mean of the pre-experiment weekly average working days per driver of the treatment group. The month of Spring Festival is omitted where the temporary retention (compared to that of the week before the experiment) ranges from -3.59 to -0.95 across different conditions.

Table 4.6: Average and heterogeneous treatment effects on weekly number of working days during the second week of March (March 4-10, 2019), about three months after the experiment ended: Difference-in-differences panel regressions.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	0.10**	0.06	0.33***	0.05	0.11**	0.08	0.33***	0.06
(β_1)	(0.05)	(0.06)	(0.12)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.06]	[1.00]	[0.03]	[1.00]	[0.02]	[0.50]	[0.02]	[1.00]
Individual leaderboard	-0.01	-0.01	-0.02	0.01	0.01	-0.00	-0.01	0.04
(β_2)	(0.05)	(0.06)	(0.12)	(0.10)	(0.04)	(0.05)	(0.12)	(0.10)
	[0.70]	[1.00]	[1.00]	[1.00]	[0.77]	[1.00]	[1.00]	[1.00]
Age					0.03***	0.03***	0.02***	0.03***
(Year)					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.22***	0.24***	0.08	0.18***
(Year)					(0.02)	(0.02)	(0.05)	(0.05)
Hometown distance					-0.00***	-0.00***	-0.00**	-0.00
to contest city (km)					(0.00)	(0.00)	(0.00)	(0.00)
Self-formed team					-0.07*	-0.16***	0.10	0.16*
					(0.04)	(0.05)	(0.10)	(0.09)
Team won in post-					0.66***	0.68***	0.63***	0.61***
intervention contest					(0.05)	(0.06)	(0.12)	(0.11)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.02**	0.17	0.00***	0.66	0.02**	0.12	0.00***	0.85
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets. The results hold if we alternatively control for the number of wins in the two short contests instead of the team that wins the post-intervention contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

consistent positive effect of the team leaderboard treatment on retention (0.33 days, $p < 0.01$), with a similar significant effect between the team and individual leaderboard treatments. In Kunming, we find a positive albeit insignificant effect of the team leaderboard treatment on retention only during the one-week window compared to the control group (0.22 days, $p < 0.05$, Table 4.11). We observe no significant difference between treatments for drivers in Beijing. We summarize the results of our driver retention analysis below.

Result 3 (Virtual Teams and Retention). For up to three months after the experiment, drivers in the team leaderboard treatment work an average of 0.1 days longer per week than those in the control group. At the city level, Taiyuan drivers in the team leaderboard treatment work 0.3 days longer per week, whereas treated drivers in Beijing and Kunming do not behave differently from those in their respective control groups.

To better understand driver incentives within each team, we conduct analyses on driver preferences to be a team captain based on our last pre-registered hypothesis.

Hypothesis 4 (Leadership)**.** *Drivers with higher productivity prior to our experiment, a longer tenure on the platform, and previous contest captain positions will be more likely to volunteer to be a team captain.*

We use a Logistic regression model (eq. 4.3) to understand how past experience on DiDi affects a driver's choice to be a team captain. The results (Table 4.16), which reject the null in favor of Hypothesis 4, show that drivers with higher revenue prior to the experiment and who have served as captains before are significantly more likely to volunteer to be a captain overall and at the city level.

Result 4 (Leadership). Drivers with a higher revenue, a longer tenure on the platform, and previous contest captain positions prior to our experiment are more likely to volunteer to be team captains.

To rule out the possibility that captains are the main drivers of our treatment effects, we re-run all analyses excluding team captains, and find that our results are robust to this specification (Tables 4.9 for Hypothesis 1, 4.10 for Hypothesis 2, and 4.12, 4.14, and 4.15 for Hypothesis 3). This indicates that captains are not the only people benefiting from the team contests.

In addition, to understand who benefits more from virtual team contests, we partition the drivers into two subgroups by whether their pre-experiment revenue was above or below the median in their respective city. As shown in Fig. 4.5, drivers whose revenue falls in the lower half in their city consistently generated a larger revenue increase than their above-median counterparts in the pre-intervention, status, and post-intervention contests. More specifically,in the pre-intervention contest (Table 4.17), below-median drivers generate a 628.50 RMB revenue increase compared to their above-median counterparts ($p < .01$). This asymmetric effect has been observed in other information intervention field experiments [40, 42], and could be attributed to any combinations of social information, team identity, and monetary rewards. When the latter is removed in the three-week status contest, social information and team identity remain present among treated drivers, whereas none of the three channels is available to drivers in the control condition, although we cannot rule out the possibility that drivers in the control condition continue to use the social information from the pre-intervention contest as a reference point. The fact that below-median drivers in the control condition continue to outperform their more productive counterparts during the three week intervention indicates that social information alone could sustain better performance for workers who used to be lagging behind.

At the end of our experiment, we sent out a survey to all drivers (EM Section

4.6.9). While the survey response rate is only 15%, feedback from the 4,295 drivers who completed the survey yields insights on how drivers benefit from virtual teams. More than 82% of the drivers like the contests (Q1), citing team belonging (Q17), making friends (Q2, Q6), and identification with the organization (Q18) as benefits. We also find evidence of peer information exchange, learning and skill improvement among team members (Q4e), providing empirical evidence for information sharing in teams [19].

4.5 Main Conclusion

Our study examines the effect of virtual teams on worker productivity, retention, and well-being on an online ride-sharing platform. Hailed as the future of work, the gig economy provides flexible, low-barrier jobs for millions of workers globally. However, a lack of both organization identity and social bonds contributes to the high attrition rate experienced by gig platforms [112]. In this paper, we investigate the efficacy of virtual teams on worker productivity and retention in a global ride-sharing platform. Using a large-scale natural field experiment with 27,790 drivers, we organize drivers into virtual teams via self and system formation. We then randomly assign teams to one of three conditions: team leaderboard, individual leaderboard, and no leaderboard/social comparison information (control). We find that treated drivers are significantly more productive in terms of revenue generated than those in the control group. Three months after the experiment ended, we find that drivers in the team leaderboard treatment continue to work longer hours on the platform, indicating that virtual teams have the potential to increase team identity and facilitate bonds with co-workers, which in turn increases productivity and worker retention.

4.6 Extended Materials

4.6.1 Power Analysis

We use a subset of the experimental data from our 2017 field experiment conducted among DiDi drivers in the city of Dongguan to generate an estimated effect size and variance parameters for our power analysis and sample size calculation. For our experiment, we would like to have a sample size large enough to obtain 90% power.

In the 2017 experiment, drivers are randomized into treatment and control conditions. Among the treated drivers, we deem teams for which the captain submits a pre-contest questionnaire as responsive and those who do not submit a questionnaire as unresponsive. In our power analysis, we use the responsive teams as our treatment condition and the unresponsive teams as our control condition since the 2017 placebo control drivers are not formed into teams. We use the five contest days as five periods. With this setting, we run the following fixed effects panel regression:

Table 4.7: Panel analysis with 2017 experiment data by fixed-effects (within-subject) regression

	Δ of Daily Orders
Game day	-1.35**
	(0.29)
Responsive	2.81 **
	(0.37)
# of observations = 17,500; # of groups = 250;	
$\sigma_u = 4.01$; $\sigma_e = 12.10$; $\rho = 0.10$;	
Standard errors in parentheses.	
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$	

According to the results in Table 4.7, we use the PowerBBK package [17] to compute the power of the new design, assuming similar behavioral responses as in the 2017 experiments.

The parameters are determined based on the following considerations (see Table 4.7 for statistics):

- budget = 125 teams per condition × 2 experimental conditions × 5 contest periods = 1250.
- beta = $(15.24, 2.8)$ since (1) $15.24 = 16.582 - 1.347$ is the daily number of trips of the unresponsive teams during the contest, (2) whereas 2.8 is the treatment effect of responsiveness.
- muvar = $\sigma_u^2 = 16$.
- espva = $\sigma_e^2 = 144$.
- panel allocation = 0.4 since 40% of the teams were unresponsive.

This command yields a power of 0.896. As we have three experimental conditions in our main analyses, we need 375 teams.

Increasing the budget by 1.5 (from 250 to 375 teams in two conditions) would give us a power of 0.982. In this case, having 564 teams (4,000 drivers) would be sufficient for our analysis. The caveat is that we do not know the potential treatment effect in the leader board phase, and therefore, cannot account for this effect in our power calculation.

4.6.2 Prize Determination across Cities

To make the experiment in each city most comparable, we determine the bonus volume for the winner team by keeping the rate of the bonus above the city-specific drivers' hourly earnings. Specifically, we first calculate the average hourly pay using the 30-day data from DiDi prior to the experiment. We carefully exclude the national holiday period (2018/10/01 - 2018/10/07) from our calculations to obtain a better indication of the average hourly earnings. As a result, we measure the average hourly pay based on data from 2018/09/10 - 2018/09/29 and 2018/10/08 - 2018/10/17. The details of the financial reward for each city are reported in Table 4.8.

Table 4.8: Details of prize in each city (money in CNY)

City	Calculated team prize	Rounded team prize	Team leader extra prize
Beijing	1,000	1,000	10
Taiyuan	654.21	650	10
Kunming	663.02	650	10

4.6.3 Robustness Checks: Treatment Effects on Driver Revenue after Excluding Team Captains

Table 4.9: Average and heterogeneous treatment effects on weekly revenue during the intervention (status contest) after excluding team captains: Difference-in-differences panel regressions investigating the two treatments separately.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	35.90*	26.81	69.93**	43.43	41.13**	32.82	73.10**	47.72
(β_1)	(19.30)	(26.54)	(28.34)	(30.87)	(19.24)	(26.47)	(28.36)	(30.58)
	[0.03]	[0.46]	[0.04]	[0.27]	[0.02]	[0.27]	[0.03]	[0.19]
Individual leaderboard	48.39**	65.11**	25.95	2.55	52.92***	68.21***	30.54	11.61
(β_2)	(19.12)	(26.14)	(31.12)	(29.97)	(19.07)	(26.06)	(30.96)	(29.89)
	[0.02]	[0.04]	[0.47]	[0.87]	[0.01]	[0.03]	[0.31]	[0.48]
Age					6.54***	7.13***	1.31	7.75***
(Year)					(0.90)	(1.26)	(1.54)	(1.37)
DiDi age					30.20***	38.62***	7.69	-1.25
(Year)					(8.05)	(10.32)	(13.22)	(14.06)
Hometown distance					-0.01	-0.00	-0.13**	-0.02
to contest city (km)					(0.02)	(0.03)	(0.05)	(0.03)
Self-formed team					-43.83**	-54.71**	-19.99	-21.97
					(17.19)	(23.11)	(29.20)	(28.05)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.52	0.16	0.14	0.18	0.55	0.19	0.15	0.23
# of clusters	10,570	7,200	1,445	1,925	10,570	7,200	1,445	1,925
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Notes: Standard errors in parentheses are clustered at the team (individual) level for ranking (control) conditions. False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.10: Average and heterogeneous treatment effects on weekly revenue in the post-intervention contest after excluding team captains: Difference-in-differences panel regressions.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	56.43**	61.42*	72.75*	24.86	64.07***	70.28**	74.40**	32.40
(β_1)	(24.94)	(34.15)	(38.10)	(40.85)	(24.58)	(33.61)	(37.54)	(40.17)
	[0.05]	[0.28]	[0.28]	[0.35]	[0.02]	[0.17]	[0.17]	[0.34]
Individual leaderboard	21.20	47.30	-55.35	-19.28	27.71	51.81	-52.66	-7.36
(β_2)	(25.36)	(34.63)	(39.99)	(40.84)	(24.95)	(34.02)	(39.08)	(40.50)
	[0.25]	[0.28]	[0.28]	[0.35]	[0.15]	[0.21]	[0.22]	[0.40]
Age					10.60***	11.55***	4.08**	11.27***
(Year)					(1.14)	(1.60)	(1.77)	(1.79)
DiDi age					81.98***	95.10***	29.31*	45.30**
(Year)					(10.34)	(13.22)	(16.78)	(18.50)
Hometown distance					-0.02	-0.03	-0.15**	0.04
to contest city (km)					(0.02)	(0.03)	(0.07)	(0.03)
Self-formed team					-28.48	-45.97	18.55	9.83
					(22.57)	(30.18)	(39.59)	(38.17)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.16	0.68	0.00***	0.26	0.14	0.59	0.00***	0.30
# of clusters	3,970	2,700	545	725	3,970	2,700	545	725
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Notes: Standard errors in parentheses are clustered at the team level. False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.6.4 Robustness Checks: Treatment Effects on Driver Retention after Excluding Team Captains or Using Different Time Windows

Table 4.11: Average and heterogeneous treatment effects on weekly number of working days during the week after the experiment ended (December 5-11, 2018): Difference-in-differences panel regressions.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	0.11***	0.05	0.39***	0.14	0.12***	0.06	0.41***	0.15
(β_1)	(0.04)	(0.05)	(0.11)	(0.09)	(0.04)	(0.05)	(0.11)	(0.09)
	[0.01]	[0.61]	[0.002]	[0.46]	[0.004]	[0.43]	[0.001]	[0.34]
Individual leaderboard	-0.03	-0.01	-0.01	-0.12	-0.02	-0.00	0.02	-0.09
(β_2)	(0.04)	(0.05)	(0.11)	(0.09)	(0.04)	(0.05)	(0.11)	(0.09)
	[0.30]	[0.90]	[0.90]	[0.55]	[0.48]	[0.86]	[0.86]	[0.51]
Age					0.02***	0.02***	0.02***	0.03***
(Year)					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.14***	0.15***	0.01	0.15***
(Year)					(0.02)	(0.02)	(0.05)	(0.04)
Hometown distance					-0.00	-0.00	-0.00	0.00
to contest city (km)					(0.00)	(0.00)	(0.00)	(0.00)
Self-formed team					-0.08**	-0.13***	-0.16*	0.19**
					(0.03)	(0.04)	(0.09)	(0.08)
Team won in post-					0.86***	0.91***	0.77***	0.71***
intervention contest					(0.04)	(0.05)	(0.11)	(0.10)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.00***	0.24	0.00***	0.01***	0.00***	0.18	0.00***	0.01**
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets. The results hold if we alternatively control for the number of wins in the two short contests instead of the winning team in the post-intervention contest.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.12: Average and heterogeneous treatment effects on weekly number of working days during the week after the contest (December 5-11, 2018) after excluding team captains: Difference-in-differences panel regressions.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	0.13***	0.04	0.47***	0.22**	0.15***	0.05	0.49***	0.23**
(β_1)	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.01]	[0.94]	[0.001]	[0.09]	[0.002]	[0.59]	[0.001]	[0.06]
Individual leaderboard	-0.00	-0.01	0.09	-0.05	0.01	0.00	0.10	-0.02
(β_2)	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.95]	[1.00]	[0.94]	[1.00]	[0.66]	[1.00]	[0.59]	[1.00]
Age					0.02***	0.02***	0.02***	0.03***
(Year)					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.13***	0.15***	-0.02	0.15***
(Year)					(0.02)	(0.02)	(0.05)	(0.05)
Hometown distance					-0.00	-0.00	-0.00*	0.00
to contest city (km)					(0.00)	(0.00)	(0.00)	(0.00)
Self-formed team					-0.07*	-0.13***	-0.09	0.16*
					(0.04)	(0.04)	(0.10)	(0.09)
Team won in post-					0.87***	0.95***	0.74***	0.71***
intervention contest					(0.04)	(0.05)	(0.12)	(0.10)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.00***	0.41	0.00***	0.01***	0.00***	0.32	0.00***	0.01**
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Notes: False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets. The results hold if we alternatively control for the number of wins in the two short contests instead of the winning team in the post-intervention contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.13: Average and heterogeneous treatment effects on weekly number of working days during the week of January 12-18, 2019, about one month after the experiment ended: Difference-in-differences panel regressions.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	0.11**	0.08	0.24**	0.11	0.12***	0.10*	0.26**	0.12
(β_1)	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.02]	[0.43]	[0.18]	[0.56]	[0.01]	[0.18]	[0.11]	[0.39]
Individual leaderboard	0.03	0.05	-0.08	0.03	0.04	0.06	-0.05	0.06
(β_2)	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.36]	[0.57]	[0.63]	[0.98]	[0.20]	[0.39]	[0.71]	[0.71]
Age					0.03***	0.03***	0.01***	0.03***
(Year)					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.19***	0.22***	0.03	0.18***
(Year)					(0.02)	(0.02)	(0.05)	(0.04)
Hometown distance					-0.00***	-0.00**	-0.00**	-0.00
to contest city (km)					(0.00)	(0.00)	(0.00)	(0.00)
Self-formed team					-0.05	-0.10**	-0.09	0.18**
					(0.04)	(0.05)	(0.10)	(0.09)
Team won in post-					0.69***	0.73***	0.58***	0.60***
intervention contest					(0.04)	(0.05)	(0.11)	(0.10)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.05*	0.52	0.00***	0.37	0.05*	0.43	0.00***	0.53
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets. The results hold if we alternatively control for the number of wins in the two short contests instead of the winning team in the post-intervention contest.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.14: Average and heterogeneous treatment effects on weekly number of working days during the week of January 12-18, 2019, about one month after the experiment ended, after excluding team captains: Difference-in-differences panel regressions.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	0.12**	0.07	0.31***	0.14	0.14***	0.09*	0.32***	0.16
(β_1)	(0.05)	(0.06)	(0.12)	(0.11)	(0.05)	(0.06)	(0.12)	(0.11)
	[0.02]	[0.49]	[0.05]	[0.49]	[0.01]	[0.30]	[0.05]	[0.30]
Individual leaderboard	0.04	0.06	-0.01	0.02	0.06	0.07	0.00	0.05
(β_2)	(0.05)	(0.06)	(0.12)	(0.11)	(0.05)	(0.06)	(0.12)	(0.11)
	[0.24]	[0.65]	[0.88]	[0.88]	[0.12]	[0.38]	[0.53]	[0.53]
Age					0.03***	0.03***	0.01**	0.04***
(Year)					(0.00)	(0.00)	(0.01)	(0.01)
DiDi age					0.19***	0.22***	0.02	0.16***
(Year)					(0.02)	(0.02)	(0.05)	(0.05)
Hometown distance					-0.00***	-0.00**	-0.00**	-0.00
to contest city (km)					(0.00)	(0.00)	(0.00)	(0.00)
Self-formed team					-0.05	-0.12**	-0.00	0.19**
					(0.04)	(0.05)	(0.11)	(0.09)
Team won in					0.69***	0.76***	0.54***	0.58***
surprise short contest					(0.05)	(0.06)	(0.12)	(0.11)
City fixed effect	yes	-	-	-	yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.09*	0.77	0.01***	0.24	0.09*	0.65	0.01***	0.32
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Notes: False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets. The results hold if we alternatively control for the number of wins in the two short contests instead of the winning team in the post-intervention contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.15: Average and heterogeneous treatment effects on weekly number of working days during the second week of March (March 4-10, 2019), about three months after the experiment ended, after excluding team captains: Difference-in-differences panel regressions.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team leaderboard	0.09*	0.04	0.38***	0.06	0.11**	0.06	0.37***	0.07
(β_1)	(0.05)	(0.06)	(0.12)	(0.11)	(0.05)	(0.06)	(0.12)	(0.11)
	[0.15]	[1.00]	[0.01]	[1.00]	[0.06]	[1.00]	[0.02]	[1.00]
Individual leaderboard	-0.03	-0.05	0.04	0.00	-0.01	-0.03	0.03	0.03
(β_2)	(0.05)	(0.06)	(0.13)	(0.11)	(0.05)	(0.06)	(0.13)	(0.11)
	[0.42]	[1.00]	[1.00]	[1.00]	[0.78]	[1.00]	[1.00]	[1.00]
Age					0.03***	0.03***	0.02***	0.03***
(Year)					(0.00)	(0.00)	(0.01)	(0.01)
DiDi age					0.21***	0.24***	0.08	0.16***
(Year)					(0.02)	(0.02)	(0.06)	(0.05)
Hometown distance					-0.00***	-0.00***	-0.00**	-0.00
to contest city (km)					(0.00)	(0.00)	(0.00)	(0.00)
Self-formed team					-0.07	-0.17***	0.19*	0.15
					(0.04)	(0.05)	(0.11)	(0.10)
Team won in post-					0.64***	0.68***	0.56***	0.54***
intervention contest					(0.05)	(0.06)	(0.13)	(0.11)
City fixed effect	Yes	-	-	-	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.02**	0.12	0.01***	0.61	0.02**	0.11	0.01***	0.74
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Notes: False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets. The results hold if we alternatively control for the number of wins in the two short contests instead of the winning team in the post-intervention contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.6.5 Preference for Being a Captain

To better understand driver incentives, we conduct an additional analysis on driver preferences to be a team captain. We use a Logistic regression model (eq. 4.3) to

Table 4.16: Results of preference for being team captains: Logistic regression with all participants.

	Outcome: Whether drivers volunteer to be captains			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Pre Experiment Revenue	0.04***	0.04***	0.07**	0.08***
(in 10,000 RMB)	(0.01)	(0.01)	(0.03)	(0.02)
	[0.00]	[0.00]	[0.01]	[0.00]
Served as captain before	0.22***	0.22***	0.23***	0.22***
(Binary indicator)	(0.00)	(0.00)	(0.02)	(0.01)
	[0.00]	[0.00]	[0.00]	[0.00]
DiDi age	0.01***	0.02***	-0.01**	0.00
(Year)	(0.00)	(0.00)	(0.01)	(0.01)
	[0.00]	[0.00]	[0.01]	[0.10]
City fixed effect	Yes	-	-	-
# of drivers	27,790	18,900	3,815	5,075

Notes: Average marginal effect with *delta-method* SE in parentheses.False Discovery Rate adjusted q-values are calculated separately for all cities (1) and for individual cities (2-4) and are reported in square brackets.
$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

understand how past experience on DiDi affects a driver's choice to be a team captain (H4), where V refers to the indicator function which equals 1 if a driver volunteers to be a team captain, and *Pre-Experiment Productivity* is operationalized as driver revenue in the two weeks before our experiment. *Served as Captain Before* is a binary variable that shows whether the driver had been a captain before he participated in the current team contest. We include γ_c to control for city-specific characteristics.

$$Pr(V = 1) = \Phi(\beta_0 + \beta_1 \text{Pre-Experiment Productivity} + \beta_2 \text{Served as Captain before} + \beta_3 \text{Didi Age} + \gamma_c) \tag{4.3}$$

The results (Table 4.16) show that drivers with higher performance prior to the experiment and who have served as captains before are significantly more likely to volunteer to be a captain overall and at the city level. However, the effects of DiDi age are more complicated. DiDi age is positively correlated with captain preference

overall and in Beijing (with $\beta = 0.01$, $p < .01$ and $\beta = 0.02$, $p < .01$, respectively), while it is negatively related to captain preference in Taiyuan (with $\beta = -0.01$, $p < .05$) and has no significant relationship with captain preference in Kunming.

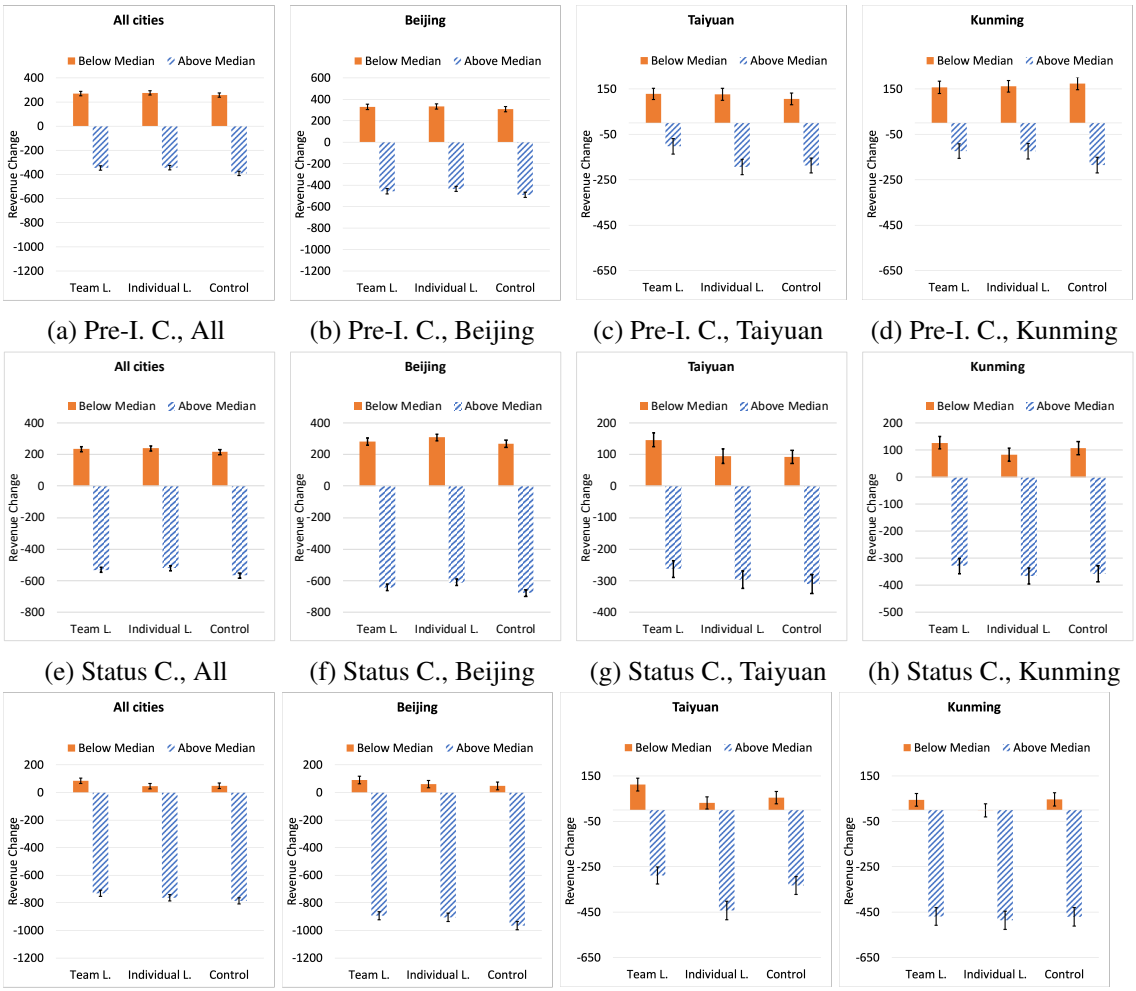
4.6.6 Who Benefits More from Team Contests? Below- versus Above-median Drivers

In this analysis, we examine who benefits more from the team identity and social information by testing the heterogeneous treatment effects on drivers with different levels of pre-experiment revenue. We differentiate drivers by whether their pre-experiment revenue is below the city median. As shown in Fig. 4.5, drivers whose revenue falls in the lower half in their city consistently exhibit a greater revenue increase than their counterparts, in the status contest across all cities.

Specifically during the longer-term contest, pooling drivers in all cities (table 4.18 (1)), we find that drivers whose pre-experiment revenue is below the city median generate 782.07 Yuan more than drivers whose pre-experiment revenue is above the city median ($p < .01$), accounting for about 37.53% of the average weekly revenue. This pattern is consistent in each of the three cities, with a revenue increase of 943.36 Yuan in Beijing (38.32% of Beijing average weekly revenue, $p < .01$), 401.99 Yuan in Taiyuan (36.08% of Taiyuan average weekly revenue, $p < .01$), and 462.37 Yuan in Kunming (33.19% of Kunming average weekly revenue, $p < .01$). No interaction effect is identified across cities and treatments. Additional tests show that drivers with below-median revenue in the team (H0: $\beta_3 + \beta_4 = 0$) and individual (H0: $\beta_3 + \beta_5 = 0$) leaderboard conditions exhibit a greater revenue increase during the status competition overall and in each of the three cities.

From Table 4.19, we see that drivers with below-median revenue also benefit more in the rewarded post-intervention contest: they generate a higher revenue of 837.31 Yuan ($p < .01$) than the above-median drivers overall, which accounts for 41.80%

of the average weekly revenue of all drivers in the control groups in the three cities. Among these drivers, below-median drivers in Beijing exhibit a higher increase of 1013.26 Yuan (43.08% of Beijing average weekly revenue, $p < .01$), while drivers in Taiyuan and Kunming generate 386.66 Yuan (34.50% of Taiyuan average weekly revenue, $p < .01$) and 517.18 Yuan (38.15% of Kunming average weekly revenue, $p < .01$), respectively, compared to the above-median drivers. Results of additional tests (H0: $\beta_3 + \beta_4 = 0$ and H0: $\beta_3 + \beta_5 = 0$) confirm that the below-median drivers in both the team and individual leaderboard conditions exhibit a greater revenue increase during the post-intervention contest period overall and in each of the three cities.



(a) Pre-I. C., All (b) Pre-I. C., Beijing (c) Pre-I. C., Taiyuan (d) Pre-I. C., Kunming

(e) Status C., All (f) Status C., Beijing (g) Status C., Taiyuan (h) Status C., Kunming

(i) Post-I. C., All (j) Post-I. C., Beijing (k) Post-I. C., Taiyuan (l) Post-I. C., Kunming

Figure 4.5: The effect of team and individual leaderboards for drivers with below and above median pre-contest revenue with standard error as error bars. (Pre-I. C.: Pre-intervention contest; Status C.: Status contest; Post-I. C.: Post-intervention contest.)

Table 4.17: Below- versus above-median drivers: Difference-in-differences regressions during the pre-intervention contest.

	Outcome: Δ of Weekly Revenue (CNY)			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Below median	628.50***	784.24***	281.89***	308.65***
	(15.95)	(20.65)	(26.39)	(26.39)
City fixed effect	Yes	-	-	-
# of drivers	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.18: Below- versus above-median drivers: Difference-in-differences regressions during the intervention.

	Outcome: Δ of Weekly Revenue (CNY)			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Team leaderboard	35.43	34.86	47.50	28.68
(β_1)	(23.86)	(31.53)	(40.82)	(43.01)
	[0.11]	[0.81]	[0.81]	[1.00]
Individual leaderboard	46.50**	67.87**	14.69	-8.29
(β_2)	(23.59)	(31.28)	(42.13)	(42.99)
	[0.11]	[0.22]	[1.00]	[1.00]
Below median	782.07***	943.36***	401.99***	462.37***
(β_3)	(23.15)	(31.32)	(36.21)	(38.36)
Team leaderboard * Below median	-17.03	-20.60	6.94	-7.68
(β_4)	(34.34)	(46.24)	(51.32)	(53.54)
Individual leaderboard * Below median	-22.48	-27.47	-13.00	-14.47
(β_5)	(34.15)	(45.60)	(52.96)	(54.99)
City fixed effect	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.65	0.31	0.43	0.40
H_0: $\beta_3 + \beta_4 = 0$ (p-value)	0.00***	0.00***	0.00***	0.00***
H_0: $\beta_3 + \beta_5 = 0$ (p-value)	0.00***	0.00***	0.00***	0.00***
H_0: $\beta_1 + \beta_4 = 0$ (p-value)	0.46	0.68	0.09*	0.56
H_0: $\beta_2 + \beta_5 = 0$ (p-value)	0.33	0.23	0.96	0.50
H_0: $\beta_1 + \beta_4 = \beta_2 + \beta_5$ (p-value)	0.83	0.46	0.13	0.21
# of drivers	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team (individual) level for the leaderboard (control) conditions. False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.19: Below- versus above-median drivers: Difference-in-differences regressions during the post-intervention contest.

	Outcome: Δ of Weekly Revenue (CNY)			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Team leaderboard	55.11	72.12	43.92	1.54
(β_1)	(34.27)	(45.54)	(59.38)	(59.27)
	[0.28]	[0.52]	[0.61]	[0.96]
Individual leaderboard	22.57	60.18	-109.98*	-15.16
(β_2)	(34.60)	(45.88)	(62.99)	(61.09)
	[0.35]	[0.52]	[0.52]	[0.93]
Below median	837.31***	1013.26***	386.66***	517.18***
(β_3)	(32.26)	(42.87)	(52.94)	(49.32)
Team leaderboard * Below median	-21.54	-29.80	13.82	-3.69
(β_4)	(45.47)	(60.95)	(71.06)	(68.99)
Individual leaderboard * Below median	-25.31	-47.23	86.81	-33.15
(β_5)	(45.43)	(60.55)	(74.45)	(70.54)
City fixed effect	Yes	-	-	-
H_0: $\beta_1 = \beta_2$ (p-value)	0.35	0.80	0.01**	0.77
H_0: $\beta_3 + \beta_4 = 0$ (p-value)	0.00***	0.00***	0.00***	0.00***
H_0: $\beta_3 + \beta_5 = 0$ (p-value)	0.00***	0.00***	0.00***	0.00***
H_0: $\beta_1 + \beta_4 = 0$ (p-value)	0.27	0.31	0.17	0.96
H_0: $\beta_2 + \beta_5 = 0$ (p-value)	0.93	0.76	0.56	0.25
H_0: $\beta_1 + \beta_4 = \beta_2 + \beta_5$ (p-value)	0.23	0.48	0.05**	0.28
# of drivers	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team level. False Discovery Rate adjusted q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.6.7 The Effect of Being Treated on Driver Revenue Change

To examine the general effect of having a leaderboard, we code the binary variable treated as 0 if the driver is in the control group and as 1 if the driver is in the team or individual leaderboard condition. We use models represented by Equation 4.1 to capture the effect.

We have discussed the effects of being treated int the main text and Table 4.3. Here we examine the persistent effects of being treated during the post-intervention contest. We see from the results in Table 4.20 that the treatment of having a leaderboard marginally significantly improves drivers revenue by 49.44 RMB ($p < .10$, 2.10% of average weekly revenue) in Beijing, but has no significant effect overall, or in Taiyuan or Kunming. Controlling for individual heterogeneity, we find that having a leaderboard improves drivers revenue by 36.72 RMB (1.83% of average weekly revenue) with marginal significance ($p < .10$) and by 55.00 RMB (2.34% of average weekly revenue) with significance ($p < .05$) overall.

Table 4.20: Average and heterogeneous treatment effects on weekly revenue during the post-intervention contest: Difference-in-differences panel regressions.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Treated	30.90	49.44*	-4.25	-11.97	36.72*	55.00**	-1.94	-3.42
(In a virtual team)	(20.81)	(28.32)	(33.04)	(34.82)	(20.45)	(27.77)	(32.22)	(34.38)
		[0.32]	[1.00]	[1.00]		[0.17]	[1.00]	[1.00]
Age					10.57***	11.31***	4.67***	11.61***
(Year)					(1.07)	(1.50)	(1.70)	(1.68)
DiDi age					84.07***	97.86***	38.37**	38.65**
(Year)					(9.62)	(12.33)	(15.43)	(17.18)
Hometown distance					-0.03	-0.04	-0.16**	0.02
to contest city (km)					(0.02)	(0.03)	(0.06)	(0.03)
Self-formed team					-20.27	-38.85	22.50	28.50
					(21.57)	(28.71)	(39.11)	(37.24)
City fixed effect	Yes	-	-	-	Yes	-	-	-
# of clusters	3,970	2,700	545	725	3,970	2,700	545	725
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Notes: Standard errors in parentheses are clustered at the team level. False Discovery Rate adjusted q-values are calculated separately for individual cities (2-4) & (6-8) and are reported in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.6.8 Does Virtual Team Contests Encourage Risky Driving?

To understand whether virtual team contests have an adverse effect on driving safety, we additionally analyze the driver safety score provided by DiDi. This is a comprehensive indicator representing driver's overall risk of accidents and transportation violations. Specifically, to calculate the safety score, DiDi incorporates multiple driving-behavior indicators, including over-speed driving, distracted driving, fatigue driving, harsh braking, and acceleration. DiDi has an algorithm to periodically update the safety score using the latest driving data. Therefore, the safety score is a good proxy for safe driving. Figure 4.6 shows the average safety score for each experimental condition over time.

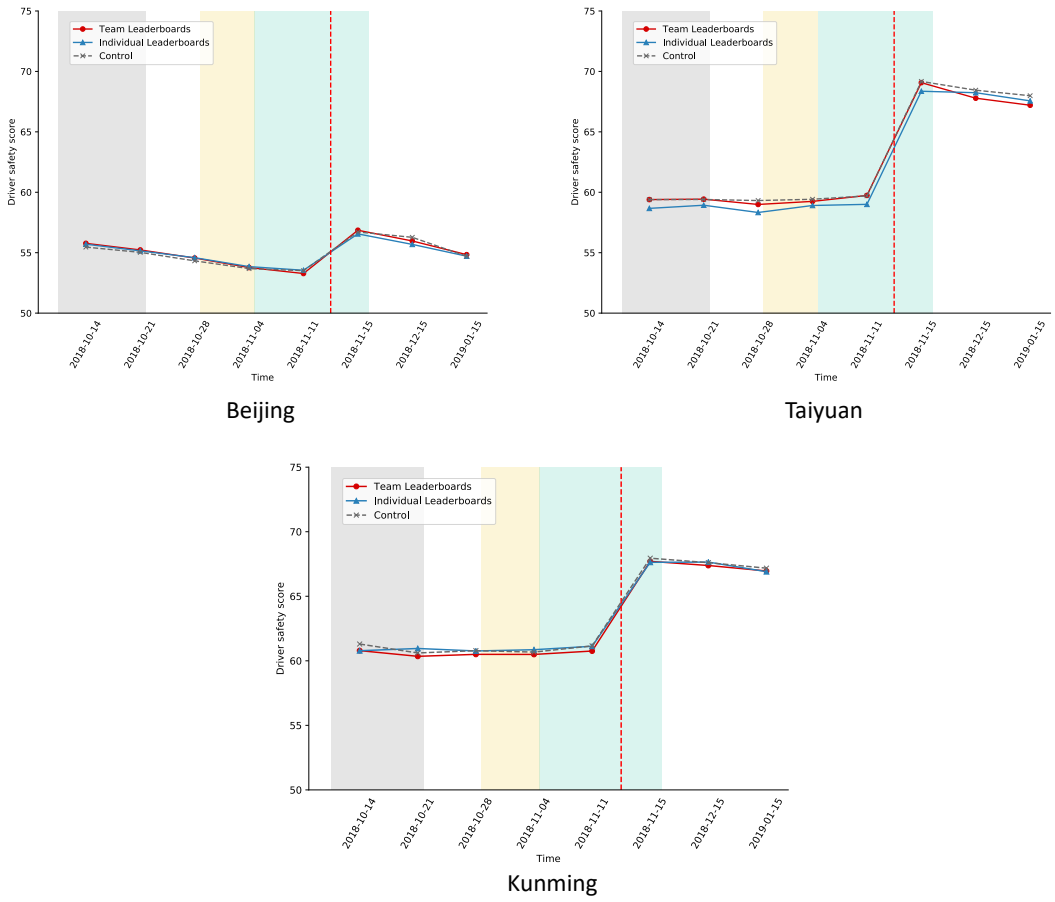


Figure 4.6: Average driver safety score of each condition over week. The red dashed lines separate the new and old safety-score formulas.

We conduct pairwise Kolmogorov-Smirnov tests across the three experiment conditions.[5] To capture the safety score pre-experiment, during contests, and post-experiment, we conduct separate tests on 2018.10.21 (the day before the experiment), 2018.11.04 (the last day of the pre-intervention contest), 2018.11.15 (during the status contest), and 2018.12.15 (about 10 days after the end of the experiment) for each pair of conditions. No significant difference is identified at the .05 significance level. Therefore, we conclude that drivers in each condition had no significantly different safety score before, during, and after the experiment.

4.6.9 Survey and Results

After the experiment, we sent a survey to every teamed-up driver in the contest; 4,295 drivers completed our survey in Beijing, Taiyuan and Kunming together, which covered about 15.46% out of 27,790 teamed drivers.

To examine the tendency of drivers to complete the survey, we conduct logistic regression analysis with results shown in Table 4.21.

[5]Since during the experiment period (on 2018.11.15), DiDi implemented a new safety score formula during the experiment period (on 2018.11.15) and the score has been updated monthly since then, we choose not to conduct a pre-experiment and post-experiment difference-in-differences model to avoid possible confounds. This is also why there is a similar score increase for each condition on 2018.11.15 as shown in Figure 4.6.

Table 4.21: Logistic regression results of driver tendency to complete the survey.

	Outcome: Whether driver completes survey			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Is captain	0.08***	0.09***	0.06***	0.06***
(Binary)	(0.01)	(0.01)	(0.02)	(0.01)
Team won in post-intervention contest	0.12***	0.11***	0.13***	0.13***
(Binary)	(0.00)	(0.01)	(0.01)	(0.01)
Pre-contest average daily revenue	0.07***	0.00	0.22***	0.21***
(in 1000 Yuan)	(0.01)	(0.01)	(0.05)	(0.04)
Male	0.06***	0.03*	0.11***	0.10***
	(0.01)	(0.02)	(0.04)	(0.02)
Hometown distance to contest city	-0.03***	-0.02***	-0.01	-0.02
(in 1000 km)	(0.01)	(0.01)	(0.03)	(0.01)
Age	0.03***	0.03***	0.03***	0.02***
(in 10 years)	(0.00)	(0.00)	(0.01)	(0.01)
DiDi age	0.01***	0.01**	-0.01	0.00
(Year)	(0.00)	(0.00)	(0.01)	(0.01)
# of drivers	34,335	18,900	3,815	5,075

Notes: Average marginal effect with *delta-method* SE in parentheses. The results hold if we alternatively control for the number of wins of the two short-term contests.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

1. To what extent do you like the recent team contest from October 29, 2018 to December 3, 2018? Please rate on a scale between 0 (I don't like it at all) and 6 (I like very much). Depending on your answer, choose either question #2 or #3.

   0 - I don't like it at all. (*201 out of 4,295, 4.68%*)

   1 - I don't like it a moderate amount . (*53 out of 4,295, 1.23%*)

   2 - I don't like it a little. (*75 out of 4,295, 1.75%*)

   3 - Neither like nor dislike. (*196 out of 4,295, 4.56%*)

   4 - I like it a little. (*152 out of 4,295, 3.53%*)

   5 - I like it a moderate amount. (*245 out of 4,295, 5.70%*

   6 - I like it very much. (*3,373 out of 4,295, 78.53%*)

   [Branch: for those who choose like]

2. Why do you like this team contest? (Please check all that apply.)

   (a) Because I like the sense of team belonging. (*2,601 out of 3,966, 65.58%*)

   (b) Because I like the fun and excitement of the contest. (*2,025 out of 3,966, 51.06%*)

   (c) Because I got to know more friends during the contest. (*2,025 out of 3,966, 51.06%*)

   (d) Because winning the contest gave me a sense of honor. (*2,417 out of 3,966, 60.94%*)

   (e) Because I won the monetary bonus. (*2,196 out of 3,966, 55.37%*)

   (f) Other reasons. Please specify ___.

[Branch: for those who choose dislike]

3. Why do you dislike this team contest? (Please check all that apply.)

   (a) Because my team members were not collaborative or united enough. (*118 out of 330, 35.76%*)

   (b) Because my team was not active enough to justify its existence. (*121 out of 330, 36.67%*)

   (c) Because the captain did not have good leadership or management skills. (*83 out of 330, 25.15%*)

   (d) Because the contest rules were too complicated to understand. (*77 out of 330, 23.33%*)

   (e) Because the contest rules were unfair. (*106 out of 330, 32.12%*)

   (f) Because the financial bonus was not large enough to attract me. (*172 out of 330, 52.12%*)

   (g) Other reasons. Please specify ___.

4. As a team member, what did you get from this team contest? (Please check all that apply.)

   (a) I got to know more friends. (*2,749 out of 4,295, 64.00%*)

   (b) I improved my leadership skills. (*1,443 out of 4,295, 33.60%*)

   (c) I improved my communication skills. (*2,067 out of 4,295, 48.13%*)

   (d) I improved my collaboration skills with other drivers. (*2,541 out of 4,295, 59.16%*)

   (e) I became more experienced and skillful about taking DiDi orders. (*2,452 out of 4,295, 57.09%*)

(f) I received emotional support from my teammates when I was down. (*1,516 out of 4,295, 35.30%*)

(g) Other reasons. Please specify ___.

5. During this event, which option best describes how your team members got along with each other?

   (a) Our team shared commonalities and common interests. (*586 out of 4,295, 13.64%*)

   (b) Although team members each had our own personalities, we got along well. (*683 out of 4,295, 15.90%*)

   (c) Everyone contributed for our team honor during the contest. (*2,377 out of 4,295, 55.34%*)

   (d) Inactive team members influenced others' enthusiasm for the contest. (*649 out of 4,295, 15.11%*)

   (e) Other reasons. Please specify ___. (0)

6. To what extent do you agree that you have developed deep friendship with your teammates? (from 0 being strongly disagree to 6 being strongly agree)

   0 - Strongly disagree. (*288 out of 4,295, 6.71%*)

   1 - Disagree. (*49 out of 4,295, 1.14%*)

   2 - Somewhat disagree. (*100 out of 4,295, 2.33%*)

   3 - Neither agree nor disagree. (*268 out of 4,295, 6.24%*)

   4 - Somewhat agree. (*203 out of 4,295, 4.73%*)

   5 - Agree. (*264 out of 4,295, 6.15%*)

   6 - Strongly agree. (*3,123 out of 4,295, 72.71%*)

7. (A reverse coding question) To what extent do you not believe that you are a part of your team? (from 0 being not agree at all to 6 being agree very much)

    0 - Strongly disagree. (*1,481 out of 4,295, 34.48%*)

    1 - Disagree. (*312 out of 4,295, 7.26%*)

    2 - Somewhat disagree. (*236 out of 4,295, 5.49%*)

    3 - Neither agree nor disagree. (*255 out of 4,295, 5.94%*)

    4 - Somewhat agree. (*177 out of 4,295, 4.12%*)

    5 - Agree. (*94 out of 4,295, 2.19%*)

    6 - Strongly agree. (*1,740 out of 4,295, 40.51%*)

8. Which option do you prefer if you participate in a team contest again?

    (a) I prefer to be a team captain. (*2,648 out of 4,295, 61.65%*)

    (b) I prefer to be a team member. (*1,647 out of 4,295, 38.35%*)

    [Branch: if choose team member]

9. Why did you choose NOT to be a team captain? (Please check all boxes that apply.)

    (a) I don't want to initiate communications with strangers. (*146 out of 1,647, 8.86%*)

    (b) I don't know how to lead a team. (*519 out of 1,647, 31.51%*)

    (c) The extra bonus for a captain was not enough. (*196 out of 1,647, 11.90%*)

    (d) I was concerned that being a captain would entail a lot of extra work. (*257 out of 1,647, 15.60%*)

(e) I was inexperienced with team leadership and needed more practice in the first place. (*1,053 out of 1,647, 63.93%*)

(f) Other reasons. Please specify ___.

[Branch: if choose team captain]

10. What do you think a team captain should do? (Please check all boxes that apply.)

(a) A captain should be a good example for other teammates. (*2,351 out of 2,648, 88.78%*)

(b) A captain should be positive and energetic. (*2,093 out of 2,648, 79.04%*)

(c) A captain should help his teammates to become more active. (*2,108 out of 2,648, 79.61%*)

(d) A captain should help his team win the contest. (*1,940 out of 2,648, 73.26%*)

(e) A captain should provide feedback and suggestions to the DiDi platform on behalf of team members. (*1,621 out of 2,648, 61.22%*)

(f) Other. Please specify ___.

11. Through which approach do you prefer to build your team?

(a) I prefer to wait for others' phone calls to invite me to join a team. (*480 out of 4,295, 11.18%*)

(b) I prefer to call other people and ask if I can join their team. (*2,983 out of 4,295, 69.45%*)

(c) I prefer to join a team without prior communication and then contact teammates online. (*832 out of 4,295, 19.37%*)

(d) Other. Please specify ___.

12. What do you hope would happen to your team?

    (a) I hope it was a temporary team and I might be able to join a different team next time. (*3,457 out of 4,295, 80.49%*)

    (b) I hope it is a long-lasting team and team members will keep in touch after the contest. (*838 out of 4,295, 19.51%*)

13. How do you communicate with your teammates during the contests?

    (a) WeChat (*3,372 out of 4,295, 78.51%*)

    (b) phone calls (*2,300 out of 4,295, 53.55%*)

    (c) text messages (*1,363 out of 4,295, 31.73%*)

    (d) face-to-face (*966 out of 4,295, 22.49%*)

14. How often do you communicate with your teammates during the first-week contest? During the three weeks in between the contests and during the last contest?

    (a) Never (*First short term: 712 out of 4295, 16.58%; Longer-term: 717 out of 4,295, 16.69%; Post-intervention contest: 755 out of 4,295, 17.58%*)

    (b) Once a week (*First short term: 725 out of 4295, 16.88%; Longer-term: 796 out of 4,295, 18.53%; Post-intervention contest: 757 out of 4,295, 17.63%*)

    (c) Multiple times a week, but not every day (*First short term: 1,142 out of 4295, 26.59%; Longer-term: 1,153 out of 4,295, 26.85%; Post-intervention contest: 1,097 out of 4,295, 25.54%*)

(d) At least once per day (*First short term: 1,716 out of 4295, 39.95%; Longer-term: 1,629 out of 4,295, 37.93%; Post-intervention contest: 1,686 out of 4,295, 39.25%*)

15. (Treated drivers only.) During the three-week contest (November 5-25, 2018), do you hope to see your team ranking on top? (from 0 being not at all to 6 being very much so)

    0 - Not hope so at all (*47 out of 2,824, 1.66%*)

    1 - Not hope so (*10 out of 2,824, 0.35%*)

    2 - Somewhat not hope so (*28 out of 2,824, 0.99%*)

    3 - Neither hope nor not hope (*73 out of 2,824, 2.58%*)

    4 - Somewhat hope so (*50 out of 2,824, 1.77%*)

    5 - Hope so (*73 out of 2,824, 2.58%*)

    6 - Hope so very much (*2,543 out of 2,824, 90.05%*)

16. (Treated drivers only.) During the three-week contest (November 5-25, 2018), which statement(s) about the leaderboard do you agree with? Please check all that apply.

    (a) Although there was no team bonus, keeping the team relationship makes me feel not lonely anymore. (*1,813 out of 2,824, 64.20%*)

    (b) Although there was no team bonus, I was curious about my ranking within my team members. (*1,459 out of 2,824, 51.66%*)

    (c) (Team-leaderboard drivers only.) Although there was no team bonus, I was curious about my team ranking among our competitor teams. (*694 out of 1,390, 49.93%*)

(d) The ranking was meaningless since there was no monetary bonus, so I didn't care about the ranking and team. (*561 out of 2,824, 19.87%*)

17. On a scale of 0 to 6, 0 being not at all, and 6 being very much so, how would you evaluate your sense of belonging to your team?

    0 - Very not strong (*207 out of 4,295, 4.82%*)

    1 - Not strong (*70 out of 4,295, 1.63%*)

    2 - Somewhat not strong (*92 out of 4,295, 2.14%*)

    3 - Moderate (*257 out of 4,295, 5.98%*)

    4 - Somewhat strong (*205 out of 4,295, 4.77%*)

    5 - Strong (*296 out of 4,295, 6.89%*)

    6 - Very strong (*3,168 out of 4,295, 73.76%*)

18. On a scale of 0 to 6, 0 being not at all, and 6 being very much so, how would you evaluate your sense of belonging to DiDi?

    0 - Very not strong (*237 out of 4,295, 5.52%*)

    1 - Not strong (*74 out of 4,295, 1.72%*)

    2 - Somewhat not strong (*91 out of 4,295, 2.12%*)

    3 - Moderate (*237 out of 4,295, 5.52%*)

    4 - Somewhat strong (*187 out of 4,295, 4.35%*)

    5 - Strong (*256 out of 4,295, 5.96%*)

    6 - Very strong (*3,213 out of 4,295, 74.81%*)

19. To what extent do you believe that your DiDi income is the primary source of income for your household?

(a) Yes, it's the only source of income for our household. (*2,076 out of 4,295, 48.34%*)

(b) It's the primary source of income, but not the only one. (*1,110 out of 4,295, 25.84%*)

(c) It's a good amount of income, but not the primary income of the household. (*660 out of 4,295, 15.37%*)

(d) It's just an additional source of income. We don't depend on DiDi's income to live a life at all. (*449 out of 4,295, 10.45%*)

20. Why do you want to be a DiDi driver?

(a) I would like to be a full-time DiDi driver for a long time. (*3,188 out of 4,295, 74.23%*)

(b) I am and will be a full-time DiDi driver until I find the next job. (*406 out of 4,295, 9.45%*)

(c) I have another job. I regard DiDi revenue as my extra pocket money in addition to my job. (*375 out of 4,295, 8.73%*)

(d) I want to kill time by driving. It doesn't matter too much for me whether I make money from it. (*77 out of 4,295, 1.79%*)

(e) Simply driving is my habit. I like driving. (*249 out of 4,295, 5.80%*)

21. What suggestions do you have for future team activities?

22. Please fill out the phone number which you use to log into the DiDi driver APP: ___.

4.7 Discussion and Take Away

This chapter details a field experiment that is designed to examine whether virtual-team interventions are effective in improving worker performance in a ride-sharing platform. Results show that virtual teams are able to increase driver performance during a bonus-free status contest and enhance driver retention even three months after the experiment ended.

Mapping to the framework of human-centered data science (Figure 1.1), this chapter illustrates that social science theories are able to inform intervention design, the effectiveness of which can be examined by a field experiment.

While this study mainly examines the effect on driver revenue, we note that it is important to examine how virtual teams and contests affect other critical metrics, such as driver happiness, driver health, and rider satisfaction. From an ethical point of view, it would be important to understand, for instance, whether longer working hours are negatively associated with driver health; while there is no evidence to confirm a negative relationship, we believe such ethical question is critical, which both we and the platform designer have been caring about. Future work could help to further understand how virtual teams and team contests affect other aspects of workers and riders besides driver revenue.

In addition, while the causal findings of this experiment are promising in general, there are many open questions related to the effects and optimal design of the interventions. For example, we observe variations of the treatment effects across different cities. Why does the same intervention work in one city but not in another? Will this intervention work in a new city? Who can benefit more from the contest? In addition, most of the contest design options, such as bonus and contest group size, are set up according to theoretical evidence or domain expert suggestions. Yet, what is the optimal design in the ride-sharing context? Furthermore, would the optimal design be different for another city or different driver participants?

These questions are challenging to answer because they require sophisticated analysis of heterogeneous treatment effects at a fine granularity on large-scale high-dimensional data, which traditional experimental analysis can hardly support. To approach such problems, we provide data-driven insights by adopting counterfactual machine learning, as discussed in the next chapter.

CHAPTER V

Predicting Individual Treatment Effects of Field Experiments with Counterfactual Machine Learning

To improve worker performance in the gig economy, in this chapter, we further integrate machine learning, field experiments, and social science theories. As described in Chapter IV, participants may experience different behavioral changes even if they engage in the same experimental intervention. We therefore ask, (1) Who benefits more from the intervention? and (2) How should we design optimal interventions for different participants subgroups and divergent contexts? These questions require a comprehensive understanding of heterogeneous treatment effects at a finer granularity, such as at the individual level.

In industry, the common practice of launching a series of field experiments with some variation in intervention provides rich data and unprecedented opportunity to answer such questions. For example, DiDi has launched thousands of team-contest experiments, which vary in terms of participating drivers, teams, cities, and contest designs. These team-contest experiments have generated large-scale and high-dimensional data to unpack heterogeneous treatment effects.

However, traditional experimental analyses are not enough to take full advantage

of such rich data. First, traditional experimental analysis usually focuses on the treatment effect at the aggregate levels and can hardly approach individual treatment effect analysis. Second, traditional experimental analyses have constraints on dealing with high-dimensional data that might involve hundreds of features.

To approach these questions, this chapter uses counterfactual machine learning to predict the effect of team contests at the individual level and discovers actionable insights by interpreting the predictive models. Our best-performing models are able to reduce the prediction error from the baseline by more than 24%. By interpreting the model, we identify findings that are directly actionable to inform team formation and contest design for future experimental interventions. Further counterfactual analysis via simulation shows that our findings have the potential to increase the treatment effect of a real contest by as much as 26%. We also highlight that theoretical insights from social sciences are additionally adopted in the feature generation process so that we can leverage existing knowledge about human behavior to improve the predictions.

This study demonstrates that integrating machine learning, field experiments, and social science theories could expand our understanding of human behavior at a finer granularity, which enables precise intervention and data-driven design for the follow-up experiments, as shown in the framework (Figure 1.1).

5.1 Introduction

The rise of the sharing economy has brought dramatic changes to work and life in modern society. The financial benefits and work schedule flexibility offered by online ride-sharing platforms, such as Uber, Lyft, and Didi Chuxing, have attracted tens of millions of drivers to serve as ride providers. While the drivers enjoy all the values of the ride-sharing economy [36], they commonly complain about new barriers to job satisfaction and retention, such as working alone, having few bonds with colleagues, no clear career paths, and a lack of a sense of achievement (e.g., [75]). How to retain

and incentivize service providers to better cover the dynamics of demand has also been a critical problem for the platforms.

Team contests, practices rooted in social identity theory [4] and contest theories [135], have been recognized as a potential cure for the pain on both sides. Through competing as teams, drivers are able to (1) build team identity and social bonds with teammates; (2) create a sense of achievement by winning a contest; and (3) increase their satisfaction and performance at work [3]. The increase in driver productivity often outweighs the cost of organizing and providing financial incentives for these contests, which creates a win-win situation for both the drivers and the platform.

Indeed, Didi Chuxing (DiDi), one of the world's leading ride-sharing companies, has launched recommender systems to help their drivers form teams and has organized many financially rewarded team contests to enhance their satisfaction and productivity [149]. In 2018 alone, more than 1,400 team contests were successfully held across 180 cities, which together involved more than 1 million drivers, who provided 130 million rides. These contests have yielded promising outcomes overall: the average return on investment is larger than 5, indicating that the increased platform revenue through these contests is five times the cost.

Behind the overall success, however, plenty of unknowns, pitfalls, and challenges remain. There is huge heterogeneity among the cities, the contests, the teams, and the drivers. Such heterogeneity produces variation in outcomes (or the treatment effects of these experiments): *What types of drivers and teams* benefit more from team contest? *What contest designs* better increase driver performance? *In what context* is a contest more likely to be effective? Why does a design work *in one city but not in another*? Understanding how these factors predict the outcomes of individual drivers would not only help the platforms find the optimal design of team contests for different populations of drivers, but would also help them generalize the success to new contexts.

Addressing these questions is challenging not only for human operational practitioners but also for data mining algorithms. First, it is intrinsically difficult to measure the causal effect of experiments, which requires a careful definition of individual outcome measures and targets of prediction. Second, the variable space to capture driver, team, contest, and context characteristics is high-dimensional, with complex relationships among them. Identifying the potential predictive factors calls for sophistication in both domain knowledge and data analytics. Third, the large-scale data involve millions of drivers and transactions and many real-world contexts, requiring the prediction algorithms to be scalable and interpretable.

In this paper, we take a systematic approach to address these challenges. We formulate the problem as a task to predict the treatment effects of a team contest on *individual drivers*, to which we apply both linear and non-linear machine learning models. Combining insights from both business practice and literature on virtual teams and team contest, we construct a large variety of features and train the prediction model using the data of hundreds of large contests and half a million drivers. The objective of this study is not to prove the causal effect of team contest but to predict individual driver's performance in out-of-sample/future contests. The former is analyzed in an earlier study based on a rigorously randomized field experiment (with no self-selection or pre-participation) using formal econometric analysis [3].

Evaluated on out-sample contests, the best-performing model is able to reduce the prediction error from the baseline by 24.50%. A careful interpretation of the models reveals intriguing predictive power of many factors (for individual treatment effects): some are intuitive, such as team homophily, social influence, supply-demand ratio, and weather conditions; some are rather surprising, such as team diversity, pre-contest activities, and the design of monetary incentives; and many of them have never been reported in the literature. Some of the factors are directly actionable in business practice, and a simulation analysis demonstrates that by simply varying

several contest design options, one is expected to increase the average treatment effect of a contest by as much as 26%.

To summarize, we make the following major contributions:

- We present the first study of individual treatment effects of team contests in a sharing economy. While existing work measures the average effect of an experiment, we analyze heterogeneous, per-driver outcomes across many experiments.
- We define a robust estimation of individual treatment effects and formulate a novel approach to predicting individual treatment effects through machine learning.
- We train effective machine learning models on large-scale data collected from hundreds of historical experiments, which combine a comprehensive set of features of individual drivers, teams, contest designs, and experimental environments, and we evaluate the models on out-sample experiments.
- We reveal the predictive power of a variety of factors for the outcome of individual drivers, most of which are novel.
- We identify actionable implications for business practice and demonstrate significant potential improvements in experimental outcomes by varying several contest design options.

5.2 Related Work

This study is related to the following lines of literature:

Sharing economy. A growing literature investigates the socio-economic effects on and consequences of ride-sharing platforms, such as Uber and Lyft [150]. Inspired by the findings in [74] that economic gains positively influence people's intention to participate, a stream of work quantifies the positive effect of financial incentives,

such as subsidy [61], on improving supply-demand efficiency. Our study adds to this literature by investigating the effect of rewarded team contests on service provision in a ride-sharing economy.

Team contest. Team competitions and team contests have been increasingly applied in online communities, such as crowdsourcing [114], education [119], online games [44], and charitable giving [39]. It has been shown that team contests are effective in improving key metrics, such as participation [119]. Data-mining researchers have developed team matching algorithms to ensure team formation of high efficiency, effectiveness, and fairness, taking into account factors such as demographics, social networks, and tasks (e.g., [2, 149, 6]).

Most of these studies demonstrate the effect of team contests through either field experiments or analyzing observational data. The former usually estimate the treatment effect at the experiment level, averaged over all treated teams and participants (e.g., [39, 119, 114]). Studies of the latter have examined team-level properties and their influences on team performance in online games, such as the positive factor of diverse team composition [44]. To the best of our knowledge, few have aimed to analyze and predict the heterogeneous effect of team contests on individual team members, especially in the context of the sharing economy.

Individual treatment effects & counterfactual analysis. Recent work in causal inference and machine learning has focused on a finer granularity – individual treatment effect (ITE) estimation, citing its potential in precision medicine [60] and online platforms [97]. Estimating ITE has been done with random forests [10] and deep neural networks [121], and it has taken into account hidden confounders from network information [71]. We base our analysis on a collection of online controlled experiments [85]. We are able to estimate ITE with difference-in-differences (DID), as the team contests already include randomly selected control groups. We thus focus on the prediction of ITE.

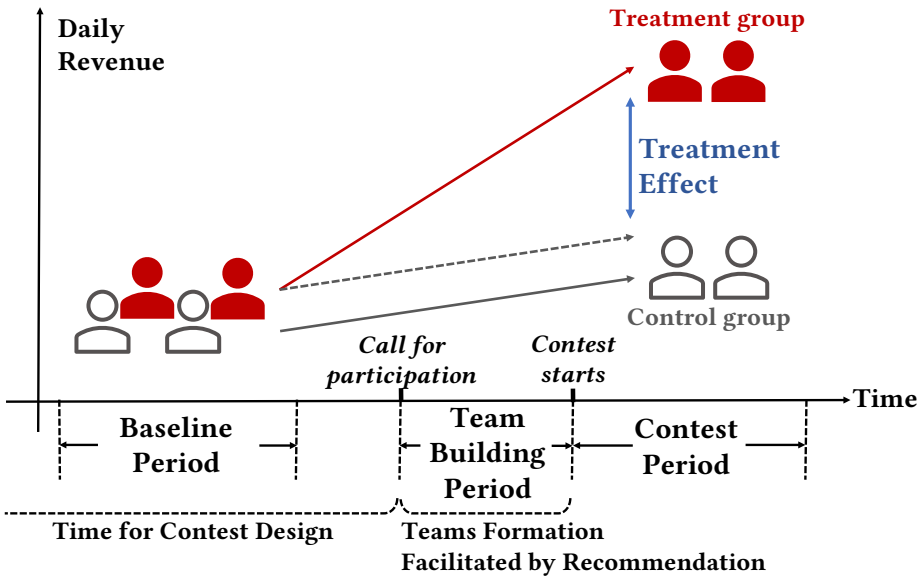


Figure 5.1: Workflow and treatment effect of a team contest

Another related stream of literature is counterfactual learning, where the focus is to learn what policies maximize some rewards, such as engagement or conversion in online advertising [25, 127]. The counterfactual estimators are typically based on importance sampling. Our paper also examines how policy (which is the contest design in our setting) predicts ITE, but we study the predictors of ITE in a much more complex socio-economic setting.

5.3 Problem Setup

5.3.1 Team Contests on DiDi

Since 2017, team contests have been widely introduced as driver incentive campaigns in DiDi [3]. A typical team contest is held in one city and consists of two periods: a *team building period* and a *contest period* (see Figure 5.1).

Team building period. The team building period starts with a call for participation and usually lasts 3-7 days. During this period, interested drivers sign up for the contest and start teaming up. Drivers can create a new team as captain or join

an existing team by invitation, and they can invite other drivers into a team either manually or assisted by a recommender system [149]. All participating teams in one contest have the same size: one captain and 2 to 7 other regular members.

About half of the teams achieve the desired size during the team building period; these are referred to as *self-formed* teams. At the end of the team building period, the system *randomly* selects 90% of the unteamed drivers and groups them into full-sized teams, which we refer to as *system-formed* teams. The other 10% are not assigned to any team and will not participate in the contest; they are referred to as *solo drivers*. These solo drivers have the same motivation level (to participate in the contest), productivity, and other demographic properties compared to the drivers who are assigned to teams, and they form a nice group for comparison. The system keeps track of the solo drivers for control.

Contest period. Both self-formed and system-formed teams will compete during the contest period. The teams are further partitioned into smaller contest groups. Each contest group contains the same number of (usually 5) teams of comparable competitiveness, measured by their productivity prior to the contest. A team only competes with other teams within the contest group and will win a cash reward according to its standing in that group. The performance of a team is calculated by summing the productivity of team members, measured by their daily revenue, number of rides, or a combination of both. During the contest period, a driver can check the performance of their team members and competitor teams through a real-time leaderboard. Under these general constraints, every city can choose among finer-grained design options (such as incentive structures). We will summarize these contest design options in Section 5.4.1.

These team contests have been quite successful in general. During a contest, a driver's daily revenue on average increases by 22%, and the revenue over investment

(ROI, which measures revenue of the platform over cost) is over 5. While the average treatment effect provides an overall picture of the effectiveness of team contests, it is critical to understand the treatment effect on individual drivers to untangle the complex interplays among participants, teams, contest design, and experimental environments. Only through this can the platform optimize their recommender systems and contest designs, provide targeted interventions for different population of drivers, and to generalize the success to new contests, cities, and countries.

5.3.2 Estimating the Individual Treatment Effect

We need to first estimate the individual treatment effects before analyzing and predicting them. Estimating the individual treatment effect by itself can be challenging in natural experiments and observational data [10, 97, 121]. In our scenario it is easier, as all the contests followed a rigorous experimental design.

The individual treatment effect (ITE) refers to the effect of a single team contest on the revenue of an individual driver. In other words, the effect measures how much additional revenue a driver generates by participating in a team contest as opposed to otherwise. Given the contest setting, we estimate the individual treatment effect using a standard difference-in-differences (DID) approach [8] in causal inference. The intuition of DID is to first compute the difference in revenue before and during the contest for each driver, aggregate such within-driver differences by treatment status (treatment vs. control), and compare the differences between the two conditions. In our case, the control group is clear - the solo drivers (drivers who are not teamed). We have two possible definitions of the treatment group: (1) drivers in both system-formed and self-formed teams; (2) drivers in system-formed teams only. Ideally, drivers in system-formed teams are the most comparable to solo drivers, as self-formed teams might differ in motivation or pre-contest history, which introduces a potential selection bias. In business operation, however, we do care about making predictions

for all drivers. We therefore separately analyze the two scenarios: using "all teams" and using "system-formed teams" as treatment group. If the results are consistent, that means the estimation of ITE can generalize from system-formed teams to all teams.

Formally, we define $R_{j,T}$ as the average daily revenue generated by driver j in the time period T. $T = T_1$ indicates the contest period while $T = T_0$ indicates a *baseline period* before contest starts. T_0 is selected as the most recent days prior to the call for participation, conditioned on matching the length and the day(s)-of-the-week of T_1. The choice of T_0 rules out day-of-the-week confounds on revenue (see Figure 5.1 for illustration).

The within-driver difference in revenue between the contest period and the baseline period can thus be calculated as

$$\Delta R_j = R_{j,T_1} - R_{j,T_0}. \tag{5.1}$$

We then aggregate the revenue change in the control group as

$$\Delta R_{\text{control}} = \frac{1}{|\text{control}|} \sum_{i \in \text{control}} \Delta R_i. \tag{5.2}$$

Finally, we can obtain the individual treatment effect as

$$\Delta R_j^{\text{ITE}} = \Delta R_j - \Delta R_{\text{control}}, \tag{5.3}$$

for every driver j in a team. If we calculate the average value of the ITE of a given contest, we will get the *average treatment effect (ATE)* of that contest. More precisely, since we can only obtain the ITE of treated drivers (participating in the team contest), the aggregated ITE represents the *average treatment effect on the treated (ATET)*.

5.3.3 Predicting the Individual Treatment Effect

We collect a dataset from all contests held between January 1, 2018 and August 23, 2018. Contests that did not hold out the 10% solo drivers are excluded, as we lack

Table 5.1: Summary of statistics

Item	Number	Item	Number
# of Cities	143	# of Unique Drivers	520,611
# of Contests	520	# of Cumulative Participation	887,842

the control condition to calculate ITE. We also exclude the contests conducted during the lunar new year, as the supply and demand pattern in that period is irregular. For all selected contests, we collect the demographics and historical activities of all drivers who sign up for the contests, regardless of whether they are in the treatment or control group. Table 5.1 presents the summary statistics of the contests included.

Based on this dataset, given every contest C_k, we are able to represent it with a list of *driver-independent* features (such as information about the city and the contest design), $\mathcal{F}_{C_k}$. For every treated driver j in C_k, we are able to estimate the treatment effect of C_k on j, $\Delta R^{\text{ITE}}_{C_k,j}$. Let the start time of the team contest period of C_k be t_k; we represent a driver j with a set of features about their demographics or activities that are observed *before* t_k, denoted as $\mathcal{F}_{j,t_k}$. We are also able to represent the team that j joins, team(j), with a set of features $\mathcal{F}_{\text{team}(j)}$. Note that $\mathcal{F}_{\text{team}(j)}$ could contain aggregated features of its members, or $\mathcal{F}_{\text{team}(j)} \sim g(\mathcal{F}_{i,t_k} | i \in \text{team}(j))$.

Given these notations, we define the problem of predicting the individual treatment effect as finding a function $f(\cdot)$ that maps the feature representations of the contest C_k, a driver j, and the team team(j) to the treatment effect of C_k on j, that is,

$$\Delta R^{\text{ITE}}_{C_k,j} = f(\mathcal{F}_{C_k}, \mathcal{F}_{j,t_k}, \mathcal{F}_{\text{team}(j)}). \tag{5.4}$$

The prediction problem as defined is intrinsically challenging. First, predicting human behavior is hard given the great complexity in cognition and decision making [125]. Second, $\Delta R^{\text{ITE}}_{C_k,j}$ as defined is essentially a "change" in behavior, which is harder to predict than the behavior itself. Moreover, the huge heterogeneity among drivers, teams, contests, time, and environments results in a wide variation in the ITE. These challenges call for a careful selection of features and predictors. In the

following sections, we show how to extract the feature representations of $\mathcal{F}_{C_k}$, $\mathcal{F}_{j,t_k}$, and $\mathcal{F}_{\text{team}(j)}$, and how to find the function $f(\cdot)$ through a machine learning approach.

5.4 Predictive Features

Our comprehensive dataset presents unprecedented opportunities to measure a wide portfolio of conditions related to the driver, the team, the contest, and the experimental environment. In this section, we characterize these conditions as informative features, generated based on the theoretical insights from the literature on contest theory, social identity theory, and virtual teams, as well as the domain knowledge from the operational practitioners at DiDi.

5.4.1 Contest Design

We start with contest design features, such as the winning condition and the prize structure. This set of features determine the utility function of the participants and directly affect their motivation and efforts devoted. Currently, the platform relies on their intuitions to decide contest designs. They are eager for actionable insights and guidance on how to optimize these designs. Apart from execution options such as team size, contest-group size, and timing, we build upon the theoretical inferences in contest theory or social identity theory to describe the incentive mechanisms in contest design.

For example, how to allocate the prizes in a contest group? Give them all to the best-performing team or split over several placements? Although this question has been analyzed in contest theory: under certain assumptions, rewarding the best in the contest group is the optimal strategy [101], it is seldom tested in field. We code the team bonus for each of the top 5 teams in a contest group.

5.4.2 Driver Properties

This set of features capture the demographics and behavioral patterns of a driver before the contest, which we assume would affect the outcomes. To depict driver behavioral patterns before contests, we retrieve drivers' daily revenue, daily number of rides, and daily hours on the platform, each in three periods: the *baseline* period (see Section 5.3.2 and Figure 5.1), 7 days before the contest starts, and 30 days before the contest starts. These features are designed to capture the most comparable activities to the contest period, the most recent activities before contest, and the longer-term work habits. We also collect driver demographics, such as age, gender, and number of months on platform (i.e., DiDi age).

5.4.3 Team Properties

This set of features are related to team-level characteristics that may significantly influence the behaviors of a member. Apart from basic team characteristics (e.g., size), we investigate *team diversity*, *team history*, *team competitiveness*, and *the influence of team on individual driver*, drawing upon previous literature [138, 2, 107, 102].

For example, we capture team diversity from three aspects: *age diversity*, *hometown diversity*, and *diversity in activity region*. As illustrated by Figure 5.2a, age diversity is shown to be a potential strong predictor of ITE. For another example, to depict *team history*, we calculate the average number of times that any two teammates have been in the same team before *this* contest. While literature has reported both the positive and the negative effect of team history on team performance [107], Figure 5.2b shows that the relationship between team history and ITE follows an inverse-U shape: no history and too much history could be equally harmful! Teams perform the best when on average half of the pairs of drivers have been teammates before, or translated to roughly 70% old members and 30% new members if a team is built on a previous team.

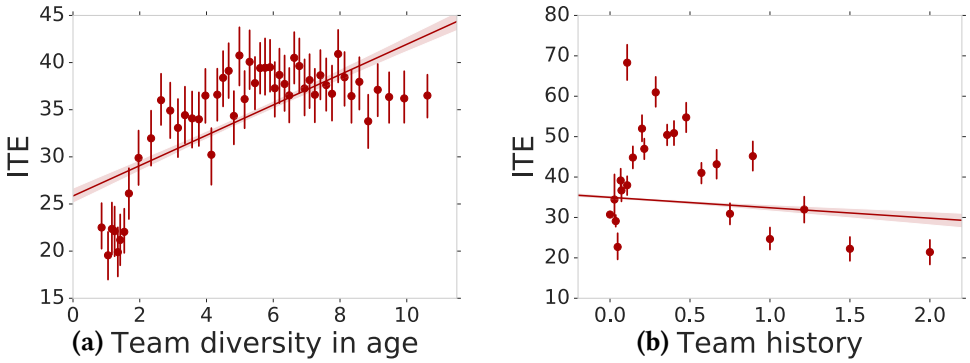


Figure 5.2: Relationship between features and ITE

5.4.4 City Properties

We also consider the environments where a contest is held, which may influence the motivation and outcome of the contest. We describe the status quo of DiDi in the contest city with the number of historical team contests, the number of DiDi drivers, and their average hourly pay. Moreover, we consider general demographics of the city, such as its development level and the province it belongs to. We also retrieve the weather reports of the city during a contest.

A more comprehensive list of features can be found in Table 5.2. Preliminary analysis has identified many correlations between these features and the ITE, although we only show two of them due to the space limit, promising the feasibility of predicting the ITE.

5.5 Predicting ITE

To what extent can the combination of the factors in Section 5.4 jointly predict the ITE? Practically, it is also valuable to know how these predictions generalize to out-sample, new contests. Building machine learning predictors is a desirable solution for both aspects.

Table 5.2: Examples of features with detailed description

Contest Design		
Bonus	Fixed Team Bonus	Whether there is a fixed-amount team bonus for the 1st-rank team in each competition group [Total fixed-amount bonus] & [Individual expectated bonus amount] for the 1st-rank team in each competition group Daily Avg of total fixed-amount bonus for the 1st-rank team in each competition group
	Pooling Team Bonus	Whether the competition uses pooling (vs fixed) bonus for the 1st-rank teams in each competition group [Total amount] & [Individual expectated bonus amount] of bonus pool for the 1st-rank teams in each competition group
	Threshold	Whether there is a minimal-performance requirement to get a team bonus
	Captain bonus	Whether there is an extra bonus for the captain of the 1st-rank team in each competition group Total extra bonus for the captain of the 1st-rank team in each competition group
	Other Bonus	Whether there is an extra individual goal-setting bonus: one can get a reward as long as him/herself satisfies the requirement
	Evaluation Metrics	Whether the worst individual performance counts towards team performance and bonus allocation
Other	Team size	Number of teams in a competition group
	Day of week	Number of workdays in the competition
Driver Properties		
Demographics	Basics	Age and gender of driver
	DiDi Specific	DiDi age (time after joining DiDi) of driver
Behavioral	Productivity	Daily revenue Avg. & Std. of the driver [in the baseline period] & [in 7 days before the contest] & [in 30 days before the contest]
	Contest History	Number of historical competitions a driver has participated in before ITE of the driver in last competition participated
Team Properties		
Diversity	Age	Std. of driver age in a team
	Hometown	Avg. pairwise geographical distance of hometowns
	Region of Activity	Avg. pairwise distance of the center locations of driving activities Avg. pairwise cosine similarity of the vectors representing number of rides taken in each sub-area
History	Team History	Avg. & Std. of pairwise number of times competing in the same team before this competition Avg. & Std. of number of times any half of the team competing in the same team before this competition
Competitiveness	Absolute	Avg. of team daily revenue [in the baseline period] & [in 7 days before the contest] & [in 30 days before the contest]
	Relative	Difference of team Avg daily revenue between this team and [the mean of all teams] & [the best team] in the competition groups
Social Influence	Team-driver	Difference of driver Avg daily revenue between this driver and the mean of all drivers in the team
	Best Team-driver	Difference of driver Avg daily revenue between this driver and the mean of the best team in the competition group
Other	Team Size	Number of drivers in the team
	Formulation	System-formed versus self-formed
City Properties		
DiDi Related	Hourly Pay	Avg. of hourly pay of all drivers in the city [in the baseline period] & [in 7 days before the contest] & [in 30 days before the contest]
	Supply-demand	Avg. of city-level daily supply-demand rate [in the baseline period] & [in 7 days before the contest] & [in 30 days before the contest]
	# of Drivers	Number of drivers in the city worked [in the baseline period] & [in 7 days before the contest] & [in 30 days before the contest]
	Rewarded Activity	Number of days that the city has other city-level rewarded activities events during the competition
Demographics	Region	Administrative (Province) and geographical region of the city
	City Classification	Tier of the city which comprehensively represents the development, population, economics, etc. of the city

5.5.1 Model Training and Evaluation

We expand the feature exploration and craft 555 features to represent factors of contests, drivers, teams, and cities (see Table 5.2).

We follow the standard practice and split the contests in our analysis into training, validation, and test sets based on the time of the contests. Contests that ended on or before June 30, 2018 are used for training and contests that fell entirely in July are used as validation set.

To determine the hyperparameters, we conduct grid-search using the training and validation set. Apart from the model specific hyperparameters, we also select the best configuration of feature scales (i.e., original, Min-Max, standardization). We apply Min-Max and standardization for Lasso and Ridge, finding standardization performing the best. For GBRT models, the data of the original scale out of all three scaling methods derives the best performance.

Finally, we use all contests that ended on or before July 31, 2018 to retrain the model and report its performance on the test set, which contains the contests starting in August 2018.

The performance of a machine learning predictor can be measured with *RMSE*:

$$\text{RMSE} = \sqrt{\sum_{k,j} \left(\Delta R^{\text{ITE}}_{C_k,j} - \Delta \hat{R}^{\text{ITE}}_{C_k,j} \right)^2 / \sum_{k} N(C_k)}, \tag{5.5}$$

where $N(C_k)$ is the number of drivers participating in contest C_k.

There are many machine learning models that can be used for building the predictors. Our main goal is not to optimize the prediction accuracy but rather to understand the effect of individual predictive factors on the target – the ITE. Therefore, we consider two objectives in selecting the machine learning algorithms: (1) they should be able to capture the linear and non-linear effects of features and their interactions; (2) they should provide an easy mechanism to interpret the predictive power of individual features. We select two standard and commonly used algorithms.

One is Lasso [132]. As a linear model, the learned coefficients provide a natural interpretation of the predictive power of features. The other is Gradient Boosted Regression Tree (GBRT) [63], which can capture the non-linearity and interactions of the features. The feature importances reported by GBRT can help interpret the contributions of different features in predicting ITE.[1] We also train Ridge models [77] to verify the robustness of linear models to different regularization. We did not choose neural networks as it is harder to interpret the importance of individual features with a deep neural network.

5.5.2 The Prediction Performance

We tune the hyperparameters of the machine learning models rigorously based on validation RMSE and report the performance of the models on test set (contests starting in August) in Table 5.3. We construct two baseline predictors for comparison. The uniform baseline predicts all ITE as the mean ITE in the training set, while the random baseline draws from a Gaussian distribution estimated from the ITEs in the training set. We separately train the models in two settings, one with drivers in all teams and one with system-formed teams only. From Table 5.3, GBRT, Lasso, and Ridge all achieve similar performance, reducing RMSE from the better baseline (Uniform) by up to 24.50% ($p < .01$) on all teamed drivers and 24.77% ($p < .01$) on drivers in the system-formed teams only. The consistency between the two settings suggests that the estimation of ITE can generalize from the system-formed teams to all teams.

Note that both GBRT and Lasso are "selecting" features during the training process. By examining the non-zero coefficients in Lasso and the positive feature importances in the GBRT, we can know which salient factors the two models rely on to make predictions. As we can see from Table 5.3, the numbers of features selected

[1]We use glmnet 3.0-2 package (https://cran.r-project.org/web/packages/glmnet/index.html) for Lasso, Ridge; scikit-learn 0.20.0 package (https://scikit-learn.org/stable/) for GBRT.

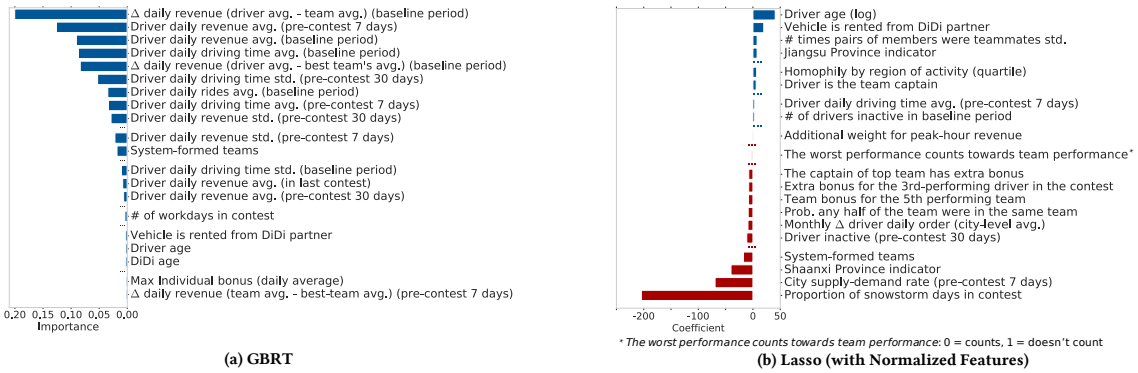


Figure 5.3: Importance scores of selected features from the best-performing GBRT and Lasso model for all teamed drivers

by the different models are quite different (251 vs. 119). In other words, the two models achieve similar performance based on different sets of features, due to the different model structures.

Table 5.3: Model performance, evaluated by RMSE

	All-teams Drivers			System-formed-teams Drivers		
Model	Val. R.	Test R.	# Ftr.	Val. R.	Test R.	# Ftr.
GBRT	139.19	147.96	251	125.00	139.67	248
Lasso	141.75	148.46	119	137.25	141.40	116
Ridge	142.16	150.55	555	136.26	143.65	552
Uniform	-	195.97	-	-	185.66	-
Random	-	266.34	-	-	250.63	-

5.6 Analyzing Prediction Results

5.6.1 Which Features Predict Treatment Effects?

We examine the most predictive features nominated from both models. Figure 5.3a and 5.3b each show 20 selected features from the best-performing GBRT and Lasso models with all-teams dataset. Both all-teams and system-formed-teams datasets produce similar results, and we choose to report the former since we do care about making predictions for everyone when deployed in the operations.

5.6.1.1 Contest Environment

We start with a set of factors about the environment of the contest.

Weather. The largest (negative) factor by Lasso for the individual treatment effect is the proportion of snowstorm days during a contest ($p < .01$). This is easy to understand as severe weather conditions would limit travel activities and driving efficiency.

Location. We observe clear heterogeneity of ITE in different locations. Contests held in certain provinces or cities have significantly higher/lower effects. Basic demographics of the city (such as population) do not appear to be predictive. The geographical heterogeneity may attribute to other properties of the locations.

Supply-Demand Rate. Surprisingly, the second largest negative factor identified by Lasso is the supply-demand ratio of the city where a contest is held. Team contests are more effective in cities of greater supply shortage ($p < .01$). This makes sense, as when supply can't meet demand, more effort of a driver ensures more profit. When supply exceeds demand, even if drivers intend to work harder, they are unlikely to receive more orders. This finding is directly actionable: sharing economy platforms should prioritize incentive-based experiments in areas of a greater supply shortage.

5.6.1.2 Driver Demographics

Young means high? No! The sharing economy has been commonly perceived as a "young people's business." [2] However, we find that middle-aged drivers and those who have joined the platform earlier experience greater treatment effects. In both GBRT and Lasso, age of driver is one of the most predictive features. Indeed, in Figure 5.4a, we observe that the treatment effect of team contests increases with age, tops among drivers in their 40's, and decreases when drivers are over 50. One possible

[2] *https://www.forbes.com/sites/homaycotte/2015/05/05/millennials-are-driving-the-sharing-economy-and-so-is-big-data/*, retrieved in October, 2019.

interpretation may be the high economic pressure on the middle-aged group. ITE also increases with a driver's age on platform. A longer lifespan on the platform indicates more experience and a greater motivation to stay in the business. From Figure 5.4b, veterans (on DiDi for over a year) have higher ITE ($p < .05$), and the trend does not drop down.

Rental Cars. Results show that drivers are more productive in contests when they don't own their vehicles but have rented from a DiDi partner ($p < .01$). One possible reason is that these drivers are more motivated to earn extra rewards to cover the rental cost, or simply the rental vehicles are in better conditions.

5.6.1.3 Pre-contest Activities

The pre-contest activities of a driver show strong predictive power.

Productivity in Previous contest. Results (see Figure 5.3a) suggest that the individual treatment effect of *this* contest depends on the revenue the same driver received in the previous contest they participated in ($p < .01$). Not surprisingly, if a treatment was effective on someone, the same thing would likely work again.

Productivity Variation. One of the most surprising factors is the variance in pre-contest activity levels. Results show that the *standard deviation* of a driver's daily revenue in the 30-day period before the contest has a positive correlation to the treatment effect ($p < .01$) from Figure 5.4c. Similar effects are observed when productivity is measured by work time or number of orders. When a driver's work habits are irregular, inner-team coordination may drag their behavior towards the social norm. Drivers of a high variation are also likely to be working part-time, and they have more room to improve through the contest.

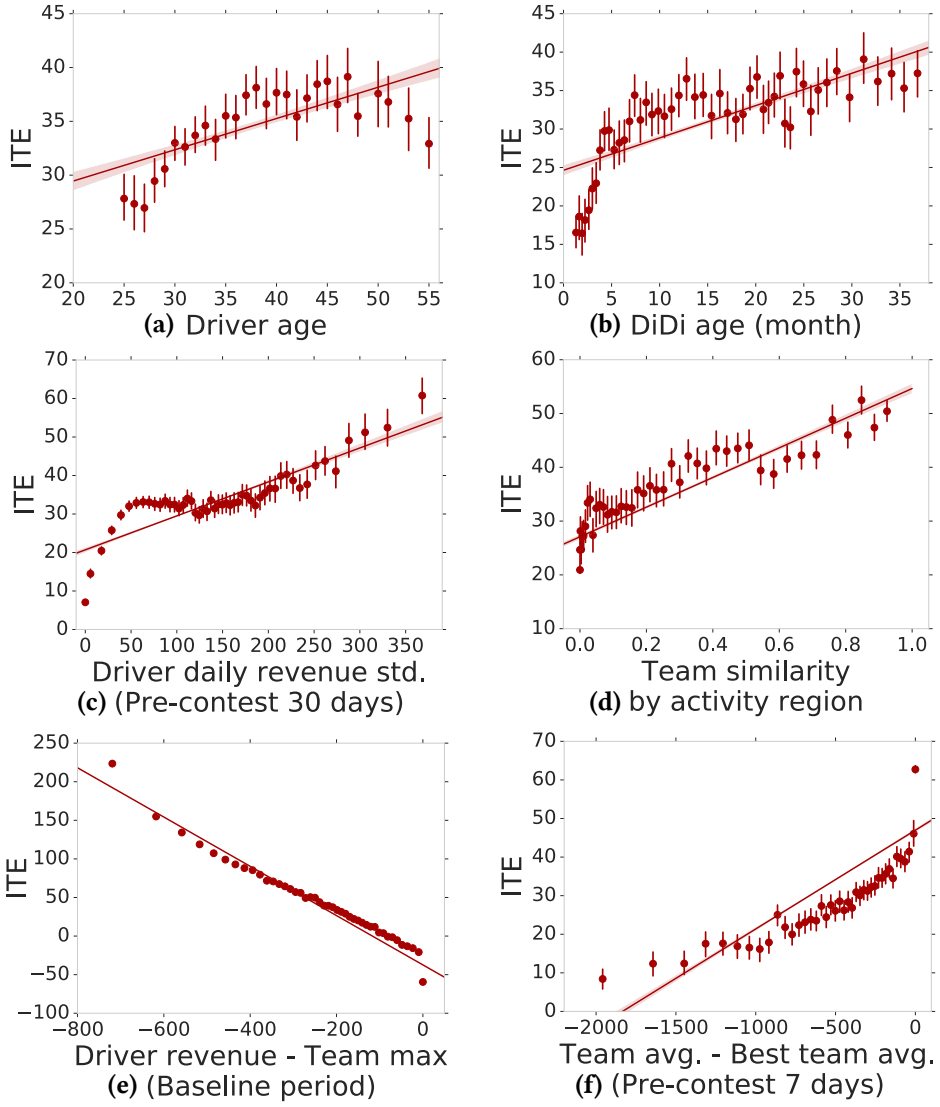


Figure 5.4: Relationships between features and ITE

5.6.1.4 Team Properties

Team structures and interrelations between members are also predictive. In social network and organization literature, there are theoretical and empirical discussions about how structural properties affect the functionality of a team or community (e.g., [44, 24]). Our analysis provides empirical evidence (in the context of the sharing economy) to these theories while also reveals novel findings.

Homophily. From Figure 5.3b, we observe that *homophily (similarity of team members) by region of activities* is a strong predictor of the treatment effect. This

effect is positive and almost linear (see Figure 5.4d). Previous literature suggests that physical distance (the inverse of homophily) negatively influences the performance of virtual work teams as it reduces shared contextual knowledge, emotional attachment, and non-verbal cues in collaborations among team members [89]. Our result extends previous work by finding that physical distance is harmful ($p < .01$) even when it requires little coordination and communication to complete the team tasks.

System-formed Teams. The method of team formation is an important predictor in both models. Teams filled by the system on average yield a smaller treatment effect than teams fully formed by drivers ($p < .01$). We note an apparent confound that drivers who form teams without the help of the platform already knew each other: they may be acquaintances in real life (related to homophily) or they may have been teammates in previous contests.

Role of captain. We find that *team captains* generally have higher ITE than other team members during contests ($p < .01$, see Figure 5.3b). This is intuitive, as drivers who volunteer to be captains are likely to be more dedicated. Another possible explanation is that the captains are "leading by example" [76].

Social influence. A rather intriguing finding by the GBRT is that social influence, rather than individual behaviors, is a strong predictor of ITE. As shown by Figure 5.3a, the top feature measures the difference between the pre-contest productivity of a driver and the average pre-contest productivity of the team. The lower a driver's pre-contest productivity is than the team average, the higher their productivity increases through the team contest ($p < .01$). This desirable outcome may be attributed to how a team functions, as social influence drags the inactive or inexperienced drivers towards the norm [48]. Note that for drivers who were already significantly more productive than their team average, the team may have also dragged them backwards towards the norm. Do these drivers constitute a large proportion in each team? By calculating the difference between the pre-contest productivity of

individual drivers and the most productive team member instead of the team average (Figure 5.4e), we see that most drivers receive a positive social influence, unless they are (or are close to) the most productive ones in their teams (with this difference close to zero).

In contrast, drivers are more motivated when the pre-contest productivity of their team is closer to that of their competitors. As shown in Figure 5.4f, ITE is higher when the pre-contest productivity of a team is closer to that of the winning team in its contest group.

These findings provide novel insights for both team formation and contest design: it is desirable to mix drivers with different activity levels, so that the more productive/experienced drivers may help the others and improve team performance. However, such a service role may hinder the motivation of the top drivers and slow down their own productivity, so it is important to provide additional incentives to the helpers. It is also important to match the competitors so that all the teams are competitive in the group.

5.6.1.5 Contest Design

More is Less! Contrary to common sense, our results show that providing more bonuses does not necessarily lead to a better outcome. Specifically, the Lasso model suggests that while in general drivers work harder for high financial rewards, an ill-designed extra bonus could inhibit the treatment effect. For example, when the 5th-performing team (the bottom team in most contests) in a contest group is rewarded, drivers become less motivated as everyone is guaranteed some reward ($p < .01$). In addition, if team captains receive an extra bonus, drivers in general become less productive ($p < .01$). The inequality between captains and members might have shifted the team goal from fighting for team identity to fighting for the captain, reducing the motivation of others.

Inner-team Competition. Adding enforced within-team competition might hurt the treatment effect: drivers are less productive if the revenue of the worst-performing driver is excluded from calculating team performance and bonus allocation ($p < .01$). Note that without this arbitrary mechanism, there is also implicit, natural competition among team members, as in most contests, the rewards are allocated based on the contributions of members.

In general, the above findings are directly actionable by contest organizers, to improve the outcomes of team contests by simply altering a few design options, at an even lower cost. We will show the potential of these opportunities with more details in Section 5.7.1.

5.6.2 Which Cases are Harder to Predict?

While the the best-performing models have already improved the baseline by 24% and generated lots of insights, the accuracy numbers do not look perfect. Indeed, individual treatment effect is perhaps one of the hardest targets for a prediction task [60]. We conduct an error analysis of the best-performing GBRT model, trying to obtain insights into what have been the harder/easier cases.

We calculate both the prediction error ($\Delta\hat{R}^{\text{ITE}}_{C_k,j} - \Delta R^{\text{ITE}}_{C_k,j}$) and its absolute value for each driver in the test set and examine their correlations with the features, using Pearson's correlation coefficient r for continuous and Student's t-test score for dummy features.

We find that the GBRT is less accurate when drivers have a high variance in pre-contest revenue, ($r = 0.41$, $p < .01$). This is intuitive: when the activities of a driver are irregular, their future activities are also hard to predict. This again highlights that predicting individual treatment effect is intrinsically challenging, especially in our context due to the huge heterogeneity of drivers. It is harder to predict for team captains than for team members ($t = 12.74$, $p < .01$), and for drivers in self-formed

teams than for those in system-formed teams ($t = 23.07$, $p < .01$). Our model also tends to underestimate the ITEs when the average hometown distance between a driver and their teammates is larger ($r = -0.02$, $p < .01$).

In addition, the absolute prediction error is significantly higher when there are more teams in one contest group ($t = 18.93$, $p < .01$ comparing groups of 3 vs. 5 teams) and when drivers' average hourly income of the city is higher ($r = 0.23$ and $p < .01$).

Overall, these factors that correlate with prediction errors are not hard to understand. Although we did not observe concerning biases, it is important to consider these patterns when applying the prediction models to different driver groups and new contexts.

5.7 Design Implications

We have obtained promising and actionable implications for the future design of team contests, which could affect the current practice of two aspects: *contest design* and *team recommender systems*.

5.7.1 Contest Design

Many findings about better contest design are immediately actionable. They are mostly about how to design incentives to balance the intensity and fairness of inter- and intra-team competitions. For example, (1) providing an extra bonus for the captain of the top team creates an inequality between captains and team members, which has a negative impact on the individual treatment effect; (2) excluding the lowest-performing driver from bonus allocation also results in unfair treatment within the team, which hurts the team performance in general. In a contest group, however, it is important to make sure that all teams have comparable levels of performance, so that no one loses the motivation to win. (3), it is also harmful to give awards to every

Table 5.4: Performance of three prototype contests under the original design and simulated new designs

	Period	C1	C2	C3	Design	True ROI	Best-design ROI (with 95% C.I.)	Worst-design ROI (with 95% C.I.)
A	Train	Y	Y	Y	Worst	2.86	4.43 (4.09, 4.76)	2.86 (2.58, 3.13)
B	Test	Y	N	Y	Bad	10.61	13.50 (12.68, 14.30)	10.50 (9.61, 11.34)
C	Train	N	N	N	Best	2.58	2.58 (2.21, 2.94)	0.71 (0.40, 0.99)

C1: Has captain bonus for top team; C2: Has team bonus for 5th team in group; C3: Worst individual score included in team performance and bonus allocation.

team, as free lunch hinders the motivation of active competitors. These design options can all be easily reversed in future contest. To demonstrate the potential benefit of such changes, we conduct a simple counter-factual analysis through simulation.

First, we select three real contests with different choices on the three dimensions above. We hypothetically vary these design choices with everything else kept the same (such as participants, team structures, etc.), and we simulate the "expected" outcome through predicting the ITE of every driver in the three contests under each new design. The benefit of changing a design option can be measured by the difference between the simulated outcome under the new contest design and the outcome of the true design. Table 5.4 lists the original design choices of the selected contests.

Through simulations using the trained Lasso model, we can compare the expected outcome of the best and the worst possible configurations and the configuration in reality. Because the trained predictor is not perfect, we further adjust the prediction results by adding Gaussian noise following (1) the prediction error distribution of the training period or the test period (depending on which period the simulated contest fell into) and (2) the prediction error distribution of the original contest (with the unchanged design). Intuitively, because all other factors are controlled, we anticipate that the expected prediction error for the simulated contest shouldn't diverge much from that of the original contest.

For each simulated contest, we bootstrap 1,000 times by sampling the number

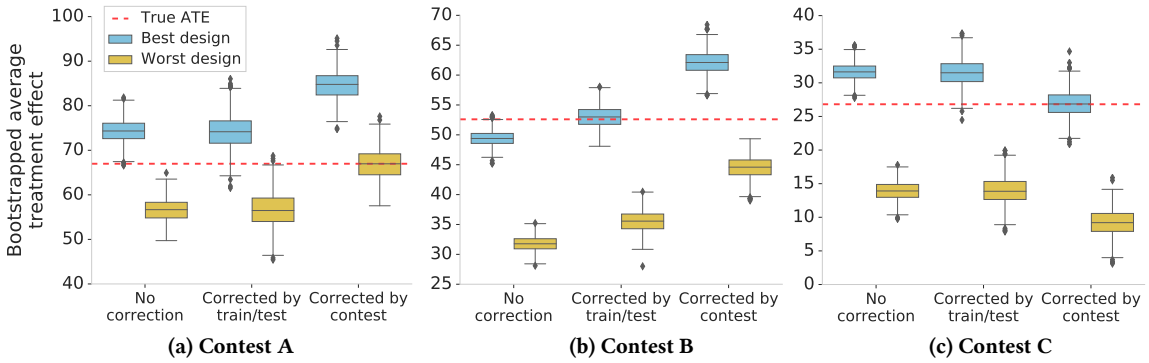


Figure 5.5: Simulated ATE of three prototype contests under the best design and the worst design

of treated drivers in the contest with replacement. Bootstrapping helps us estimate the confidence intervals of the expected average treatment effect. In Figure 5.5, we report the bootstrapped average treatment effect of different simulated designs for the 3 prototype contests, including the best, the worst, and the original designs. We report the simulated ATE with prediction error uncorrected, corrected using period-level distribution, and corrected using contest-level distribution. Clearly, there is a significant difference in average treatment effect between the best and the worst design choices (26%, 39%, and 191% improvement over the worst design respectively for contest A, B, and C). In contest A and B, the expected ATE (prediction error corrected at the contest-level) of the optimal design significantly outperforms the ATE of the original contest (using the actual design), with an improvement of as much as 26 percent. The expected ATE does not outperform the true ATE in contest C, as the original design is already the best. Moreover, the design choices may also affect the ROI (Revenue-over-Investment) of the contests. As shown in Table 5.4, the ROI can increase by as much as 55% from the original to the best design in simulation.

The results above are promising. They demonstrate that by simply varying a few design options, both the drivers and the platform can benefit significantly. Many other design options could be improved based on the analysis in Section 5.6.1.5, although it's harder to demonstrate them through a simple counter-factual simulation.

5.7.2 Team Recommendation

Our findings also shed light on how to better design team recommender systems. For example, it is better to first team up friends and former teammates, and then introduce new drivers to the team. It is beneficial to combine low-performing and newer drivers with high-performing and experienced drivers in one team. Teaming people who are from the same hometown and who work in similar areas can also boost performance.

5.8 Limitations and Future Opportunities

First, this study focuses on exploring predictive factors that explain the variance of ITE across individuals, teams, contests, and cities. Although the estimation of the ITE follows the standard practice of causal inference, the prediction model does not guarantee that relations discovered between the features and the ITE are causal. Future studies are needed to establish causal relationships between the predictors identified and the ITEs.

Second, the benefit of optimized contest-design options is estimated based on simulations. While the three design options are carefully selected to be independent as possible from other factors (so we can control the confounds), it is not impossible that changing these options would result in a change in others. For example, there is a probability that dropping the bonus for the 5th team might result in less participation. As in the literature [94], we note that a field A/B testing is highly suggested to examine the effectiveness of the counterfactual model in the wild.

In addition, all analyses and findings are based on field experiments and data collected from one ride-sharing platform in one country. Our conclusions may be generalized to other platforms, countries, and domains with caution.

Finally, we acknowledge that future research should analyze bias and fairness

across drivers and understand the effects on other outcomes besides revenue. Biases can appear in any phases of data collection, model development, and results implementation. A bias and fairness analysis should ideally be performed before widely applying these results. In addition, while this paper focuses on promoting driver revenue, note that the design of contests and recommender system may lead to other results beyond increasing driver revenue. For example, would drivers on average be less happy or more happy if the last individual performance were invalid for team performance? As another example, would removing the bonus for the top-performing captain affect the captain hierarchy design and career path development? More research is expected to examine the effect of team contests beyond driver revenue.

5.9 Conclusion and Take Away

This is the first predictive analysis of individual treatment effects of team contests in DiDi, a leading platform of the ride-sharing economy. The analysis investigates hundreds of large-scale team contests in 143 cities, involving half a million drivers, tens of millions of rides, and a comprehensive set of features of the drivers, teams, contest design, and experimental conditions. Through linear and nonlinear machine learning algorithms, these features demonstrate decent predictive power of individual outcomes in team contests. Our findings present many new insights and useful implications for the research and business practices of team contest, the sharing economy, and online field experiments in general. Some of the findings are immediately actionable in optimizing the design of upcoming team contests. Future directions of the work include testing these insights with field experiments, investigating the causal links between the heterogeneous factors and the ITE, and generalizing the procedure to other sharing economy platforms.

Connecting to the human-centered data science framework (see Figure 5.4), we show that counterfactual machine learning helps to analyze experimental data and

that theory can inform feature generation to enhance the predictive power of machine learning models. These results also suggest that counterfactual machine learning is able to uncover data-driven insights to optimize interventions.

CHAPTER VI

Conclusion

This dissertation proposes a framework of human-centered data science combining data science techniques and social science theories to investigate worker performance. In this chapter, we summarize the dissertation and conclude with a discussion of future directions.

6.1 Summary

The development of information technologies is reshaping the work of the labor force. These technologies foster millions of jobs in the novel contexts of gig economies and provide new tools to support work in traditional sectors. The recent outbreak of COVID-19 has added to the changes by greatly shifting work towards the virtual end of the spectrum, bringing tremendous challenges for workers across the world. All these new work contexts and work-support tools bring into question our existing knowledge and methods toward improving worker performance, prompting a need to study worker performance in the modern era.

To approach this problem, online platforms, the big data documentation describing workers, organizations, and societal contexts, as well as data science techniques, have provided unprecedented opportunities: online platforms allow fast, precise, and large-scale interventions in the wild, big data contains rich information, and data

science techniques provide advanced tools to discover data-driven insights. These large-scale field interventions and advanced data analytics complement our existing knowledge about human behavior that is commonly embedded in social science theories. Therefore, in this dissertation, we propose a human-centered data science framework that synthesizes the strengths of machine learning, field experiments, and social science theories to promote worker performance.

While each of machine learning, field experiments, and social science theories have each been separately examined to study worker performance, traditional monomethod approaches cannot tackle the challenges rendered by large and complex data in the modern era: they either cannot handle complex relationships among human and contextual factors, or cannot fully incorporate existing knowledge, or lack the scalability to deal with high-dimensional large-scale data. These challenges call for interdisciplinary solutions leveraging both the advanced analytic skills of data science and the deep insights of social science.

To approach this question, our human-centered data science framework complementarily connects machine learning, field experiments, and social science theories to study human behavior. This framework is rooted in our recognition of the advantages of each component – machine learning is featured by an ability to handle large and complex data to predict human behavior; field experiments perform precise interventions and establish the causality by real-world observations; and social science theories present rich existing insights in describing, predicting, and explaining human behavior. As such, this framework emphasizes the strengthened power of interactions among these components. To empower (counterfactual) machine learning, we would like to leverage existing insights from social theories by informing feature construction, model architecture, and model explanation and deploy field experiments to help evaluate the effectiveness of these models in real-world practices. Moreover, to discover nuanced data-driven insights from field experiments, we incorporate ma-

chine learning to conduct more sophisticated analyses of experimental data by revealing heterogeneous effects at a finer granularity, such as individual treatment effects. These data-driven discoveries complement theory-based insights to inspire better experimental designs. In addition, causal insights derived from field experiments and counterfactual machine learning models could support the testing of existing theories and the development of new theories that better reflect reality.

How can we apply this framework in real-world work practices? In this dissertation, we present three studies in both traditional jobs and the modern workforce to exemplify the flexibility and effectiveness of applying the framework.

First, we show that the predictions generated by machine learning models could help the tenant support specialists with more informed and effective decisions, increasing work performance. To help detect landlord harassment and provide follow-up assistance, New York City's government has established a Public Engagement Unit (PEU). The PEU's outreach specialists go across the city and knock door-by-door with little idea of whether there will be a harassment case behind a door, which significantly limits the number of harassment cases they are able to identify. By analyzing their historical canvassing records and contextual data about local areas, we are able to use machine learning to help predict the harassment risk level associated with every residential building in the city. Our best-performing model has the potential to increase their work performance by 59%.

While these results are promising, they are based on historical data. Would the new outreach strategy proposed by our model be effective in reality? Or, more generally speaking, would a new algorithm or intervention strategy have effects in real-world practices? A gold standard to address this question is to conduct field experiment, which we illustrate in our second study.

In the second application, we design a new intervention and demonstrate its effectiveness in improving worker performance with a large-scale field experiment. Specif-

ically, in collaboration with DiDi, a ride-sharing economy platform, we leverage social science theories and propose a team-based solution by first organizing workers into virtual teams and then engaging teams into team contests. The results of the field experiment show that virtual teams and team contests significantly increase worker performance (reflected by revenue) even when there is no financial incentive during the contest. The treated drivers continue to work longer on the platform even three months after the end of the experiment.

The virtual team solution accomplishes great success in general. Yet, we observe a huge variation of the treatment effect across individuals, teams, and cities. Can human-centered data science help us to understand which drivers and teams benefit more from team contests, which contest design generates the highest effect for a given sample, and why a contest design works for one city but not another? We deploy the next study to address these questions.

In the third study, we deploy counterfactual machine learning models to predict the individual treatment effects of more than 500 large-scale field experiments. Leveraging features inspired by social science theories, our models make more precise predictions than the baseline. By interpreting the model, we are able to identify insights that are directly actionable to customize contest design and team formation. Further counterfactual analysis illustrates the potential effectiveness of the new experimental designs in increasing the treatment effects. This study presents the power of integrating counterfactual machine learning, field experiments, and social science theories in enhancing work practices.

Taken together, these three studies effectively apply the framework of human-centered data science in improving worker performance. They together show the flexibility and power of integrating machine learning, field experiment, and social science theories in addressing practical behavioral problems in various work contexts.

6.2 Discussion and Implications

Overall, we believe this framework of human-centered data science speaks to a diverse audience. For machine learning researchers, this framework highlights the great potential of applying machine learning to approach causal questions and social science problems, such as improving worker performance. For social scientists, we point out the possibilities of incorporating advanced data analytics in causal research, as well as the informative power of theoretical insights on data science techniques. In addition, this framework presents an organic integration of data science techniques and social science theories to industry and domain experts, who can benefit from flexible and effective applications combining machine learning, field experiments, and theories.

The empirical studies in this dissertation present promising results in improving worker performance; yet, we acknowledge potential ethical concerns especially when the results of machine learning models are directly and widely applied without field trials and bias/fairness analysis. First, biases in data collection may lead to biases in prediction results, raising ethical concerns. For example, in Chapter III, since our models are trained only on the data of buildings canvassed by TSU, and there is some bias in how TSU selects buildings to canvass, the accuracy of predictions might be higher on similar buildings than unlike ones, resulting in inequality in assisting tenants. Second, machine learning models may prioritize certain subgroups, leading to unfair prediction-based interventions across the population. The original problem formulation in Chapter III shows an example: the model prioritizes the outreach to large buildings because of the flawed problem formulation, reducing the probability of being assisted for tenants living in small buildings. Moreover, one should pay special attention to model fairness when demographic factors are included in features, such as gender and race. Third, the data-driven interventions suggested by machine learning predictions may lead to unwanted effects on the aspects outside the scope of machine

learning features and outcomes. Taken the study in Chapter V as an example, it is unknown whether the contest designs proposed by the machine learning models have additional effects on driver happiness and driver health beyond their outcome variable — driver performance. Machine learning models typically consider only a few, if not one, outcome variables in the optimization objective. However, real-world problems related to human behavior are commonly much more complicated. We suggest that human-centered applications of machine learning models should consider the effects of model implementation on other factors besides the outcome variable(s) of the models.

In addition, several interactions among machine learning, field experiments, and social science theories in the framework have not been deeply investigated by our three empirical studies, on which we expect more solid future work. For example, the framework suggests that counterfactual machine learning can be deployed to test existing theories or build new theories. While we observe preliminary evidence from the team-contest studies that drivers in the gig economy might undermine the common assumption of independent and identical distribution in economics theories, we suggest more effort in discovering theoretical insights via machine learning.

We also would like to apply this framework in other human behaviors beyond the two presented work contexts and beyond worker performance. The focus of this dissertation is to improve worker performance; yet we expect this framework to be more generalizable to other application domains, such as health, education, and the pro-social behaviors of the general public.

Third, we hope to strengthen this framework by incorporating more methods. The current framework focuses on analyzing intervention data, particularly the data generated by experiments. Therefore, it has limitations in discovering actionable insights from other types of data, such as observational data. To enrich the data and problems that the framework is able to handle, future research should bring in more methods, such as more causal inference methods.

In all, this dissertation proposes a framework of human-centered data science to improve worker performance. We suggest more efforts in addressing potential ethical issues and encourage further research that deploys the framework to more application domains and that strengthens the framework with additional methods. Our work opens many research directions for further exploration, and we conclude the dissertation by outlining a few specific promising opportunities.

6.3 Future Direction

Promoting Worker Performance in Traditional Sectors. First we hope to provide predictions at a finer granularity to better support decision making. For example, in the NYC project, we predict whether there will be cases in a given building if the specialists visit in the next month. However, the outreach outcomes might change depending on the visit day of the week or time of the day. We would like to adapt the predicted units and involve more factors to inform more precise interventions, such as by reformulating the problem to predict whether there will be cases in a given building if the specialists visit in the morning time during the next month. Second, we would like to incorporate fairness and bias analysis in the implementation of the model. It is ethically important to understand the heterogeneous effects of applying the models for different subgroups before we can deploy the models comfortably and confidently. This indeed also echoes the special challenge and value of human-centered data science. For example, our models prioritize the residential buildings only according to the predicted risk, which may (or may not) lead to inequity or biases across different subareas and subgroups. We would like to investigate the actual and perceived fairness/bias of these algorithms and propose methods to balance fairness and efficiency in the implementation of the models.

Enhancing Worker Performance in the Modern Workforce. In the DiDi

project, we discover many insights regarding the optimal contest design and team formation strategy by off-policy evaluation. In addition to the evaluation, we would like to deploy counterfactual learning to optimize the contest design and team recommender for given participants and contexts. Another natural follow-up of this study is to examine the effectiveness of these new strategies via randomized field trials, closing the loop between machine learning and field experiments. Moreover, we would like to investigate new interventions to strengthen team identity. In the field-experiment study, we use a rewarded team contest to enhance team identity, which represents the route of inter-team competition. Next, we would like to explore interventions to improve team identity by within-team coordination leveraging theoretical insights in social psychology and behavioral economics.

Comparing Worker Performance in Traditional and Modern Workforces. In the future, we hope to understand the differences and similarities in applying advanced technologies in different contexts. For example, we hope to compare the dynamics of teams and team contests between modern housing agents and the gig-economy drivers. The work context of modern housing agents is different from that of the ride-sharing work in several ways. First, housing-agent teams involve both intensive online and hardcore offline activities, while the driver teams on DiDi are geographically distributed. Second, drivers in the same city rarely directly compete with one another in picking up rides, but housing agents are highly competitive in accessing housing sources and customers. And one customer might have several agents helping at the same time. Third, housing agent work requires more skills, training, and collaboration than driving on DiDi. Given such differences, would the same team contests be equally effective for housing agents and DiDi drivers? Would there be any novel predictors of treatment effect or any new interventions to facilitate worker performance among housing agent teams? It would be interesting to compare the

treatment effects of the same interventions in different contexts and to design novel interventions for further worker performance improvement.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] K. Ackermann, E. Blancas Reyes, S. He, T. Anderson Keller, P. van der Boor, R. Khan, R. Ghani, and J. C. González. Designing policy recommendations to reduce home abandonment in mexico. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 13–20. ACM, 2016.

[2] W. Ai, R. Chen, Y. Chen, Q. Mei, and W. Phillips. Recommending teams promotes prosocial lending in online microfinance. *Proceedings of the National Academy of Sciences*, 113(52):14944–14948, 2016.

[3] W. Ai, Y. Chen, Q. Mei, J. Ye, and L. Zhang. Putting teams into the gig economy: A field experiment at a ride-sharing platform. Under revision for resubmission to Management Science, 2019.

[4] G. A. Akerlof and R. E. Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753, August 2000.

[5] G. A. Akerlof and R. E. Kranton. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton University Press, Princeton, New Jersey, 2010.

[6] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 839–848. ACM, 2012.

[7] M. L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.

[8] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

[9] S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[10] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[11] O. Bandiera, I. Barankay, and I. Rasul. Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114, 10 2013.

[12] W. Bank. World development report 2019: The changing nature of work, 2018.

[13] K. Barker, J. Silver-Greenberg, G. Ashford, and C. Sarah. The eviction machine churning through New York City. *The New York Times*, 2018.

[14] H. Barki and A. Pinsonneault. Small group brainstorming and idea quality: Is electronic brainstorming the most effective approach? *Small Group Research*, 32(2):158–205, 2001.

[15] J. M. Barrero, N. Bloom, and S. J. Davis. Covid-19 is also a reallocation shock. Working paper available at SSRN, 2020.

[16] C. B. Barrett and M. R. Carter. Retreat from radical skepticism: Rebalancing theory, observational data and randomization in development economics. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, pages 58–77, 2014.

[17] C. Bellemare, L. Bissonnette, and S. Kröger. Simulating power of economic experiments: the powerbbk package. *Journal of the Economic Science Association*, 2(2):157–168, Nov 2016.

[18] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995.

[19] D. Bergemann and S. Morris. Robust predictions in games with incomplete information. *Econometrica*, 81(4):1251–1308, 2013.

[20] R. Berk. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.

[21] M. Bernard, F. Hett, and M. Mechtel. Social identity and social free-riding. *European Economic Review*, 90:4 – 17, 2016. Social identity and discrimination.

[22] M. Bertrand and S. Mullainathan. Are Emily and Greg more employable than lakisha and jamal? A field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[23] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[24] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[25] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[26] G. E. Box, W. H. Hunter, S. Hunter, et al. *Statistics for experimenters*, volume 664. John Wiley and sons New York, 1978.

[27] M. B. Brewer. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, 55(3):429–444, 1999.

[28] J. P. Campbell. Modeling the performance prediction problem in industrial and organizational psychology. 1990.

[29] S. Carton, J. Helsby, K. Joseph, A. Mahmud, Y. Park, J. Walsh, C. Cody, C. Patterson, L. Haynes, and R. Ghani. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 67–76. ACM, 2016.

[30] U. S. Census. 2019 US population estimates continue to show the nation's growth is slowing. *United States Census Bureau*, 2019.

[31] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: Causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16, 2010.

[32] B. Chandar, U. Gneezy, J. A. List, and I. Muir. The drivers of social preferences: Evidence from a nationwide tipping field experiment. Technical report, National Bureau of Economic Research, 2019.

[33] B. K. Chandar, A. Hortaçsu, J. A. List, I. Muir, and J. M. Wooldridge. Design and analysis of cluster-randomized field experiments in panel data settings. Technical report, National Bureau of Economic Research, 2019.

[34] G. Charness and Y. Chen. Social identity, group behavior and teams. *Annual Review of Economics*, 2020.

[35] G. Charness, L. Rigotti, and A. Rustichini. Individual behavior and group membership. *The American Economic Review*, 97:1340 – 1352, September 2007.

[36] M. K. Chen, P. E. Rossi, J. A. Chevalier, and E. Oehlsen. The value of flexible work: Evidence from uber drivers. *Journal of Political Economy*, 127(6):2735–2794, 2019.

[37] M. K. Chen and M. Sheldon. Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. In *Ec*, page 455, 2016.

[38] R. Chen and Y. Chen. The potential of social identity for equilibrium selection. *The American Economic Review*, 101(6):2562–2589, October 2011.

[39] R. Chen, Y. Chen, Y. Liu, and Q. Mei. Does team competition increase pro-social lending? Evidence from online microfinance. *Games and Economic Behavior*, 101:311–333, 2017.

[40] Y. Chen, F. M. Harper, J. Konstan, and S. X. Li. Social comparisons and contributions to online communities: A field experiment on movielens. *Amer. Econ. Rev.*, 100(4):1358–98, 2010.

[41] Y. Chen, S. X. Li, T. X. Liu, and M. Shih. Which hat to wear? Impact of natural identities on coordination and cooperation. *Games and Economic Behavior*, 84(0):58 – 86, 2014.

[42] Y. Chen, F. Lu, and J. Zhang. Social comparisons, status and driving behavior. *Journal of Public Economics*, 155:11–20, 2017.

[43] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

[44] Z. Cheng, Y. Yang, C. Tan, D. Cheng, A. Cheng, and Y. Zhuang. What makes a good team? A large-scale study on the effect of team composition in honor of kings. In *The World Wide Web Conference*, pages 2666–2672. ACM, 2019.

[45] L. Chidambaram and L. L. Tung. Is out of sight, out of mind? An empirical study of social loafing in technology-supported groups. *Information systems research*, 16(2):149–168, 2005.

[46] S. Chowdhury, A. Mukherjee, and R. Sheremeta. In-group versus out-group preferences in intergroup conflict: An experiment. 2021.

[47] S. M. Chowdhury. The economics of identity and conflict. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press (OUP), 2021.

[48] R. B. Cialdini and M. R. Trost. Social influence: Social norms, conformity and compliance. *The Handbook of Social Psychology*, pages 151–192, 1998.

[49] P. Cohen, R. Hahn, J. Hall, S. Levitt, and R. Metcalfe. Using big data to estimate consumer surplus: The case of uber. Working Paper 22627, National Bureau of Economic Research, September 2016.

[50] C. Cook, R. Diamond, J. Hall, J. A. List, and P. Oyer. The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. Review of Economic Studies (forthcoming).

[51] S. DellaVigna and E. Linos. Rcts to scale: Comprehensive evidence from two nudge units. Working Paper 27594, National Bureau of Economic Research, July 2020.

[52] N. H. P. . Development and U. S. C. Bureau. New York City housing and vacancy survey. *NYC Housing Preservation & Development*, May 2018.

[53] N. H. P. . Development and U. S. C. Bureau. New York City housing and vacancy survey. *United States Census Bureau*, May 2018.

[54] T. R. Dillahunt, V. Kameswaran, L. Li, and T. Rosenblat. Uncovering the values and constraints of real-time ridesharing for low-resource populations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2757–2769. ACM, 2017.

[55] M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[56] C. C. Eckel and P. J. Grossman. Managing diversity by creating team identity. *Journal of Economic Behavior & Organization*, 58(3):371–392, November 2005.

[57] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Stanford University Online Manuscript, 2010.

[58] I. Erev, G. Bornstein, and R. Galili. Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology*, 29(6):463–478, November 1993.

[59] B. Fang, Q. Ye, R. Law, et al. Effect of sharing economy on tourism industry employment. *Annals of Tourism Research*, 57:264–267, 2016.

[60] G. Fang, I. E. Annis, J. Elston-Lafata, and S. Cykert. Applying machine learning to predict real-world individual treatment effects: Insights from a virtual patient cohort. *Journal of the American Medical Informatics Association*, 26(10):977–988, 2019.

[61] Z. Fang, L. Huang, and A. Wierman. Prices and subsidies in the sharing economy. *Performance Evaluation*, page 102037, 2019.

[62] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

[63] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

[64] D. Fudenberg and A. Liang. Predicting and understanding initial play. 2018.

[65] R. B. Gallupe, G. DeSanctis, and G. W. Dickson. Computer-based support for group problem-finding: An experimental investigation. *MIS quarterly*, pages 277–296, 1988.

[66] E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review*, 106(5):114–18, 2016.

[67] E. L. Glaeser, S. D. Kominers, M. Luca, and N. Naik. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137, 2018.

[68] L. Goette, D. Huffman, S. Meier, and M. Sutter. Competition between organizational groups: Its impact on altruistic and antisocial motivations. *Management Science*, 58(5):948–960, 2012.

[69] A. Goldszmidt, J. A. List, R. D. Metcalfe, I. Muir, V. K. Smith, and J. Wang. The value of time in the united states: Estimates from nationwide natural field experiments. Technical report, National Bureau of Economic Research, 2020.

[70] J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

[71] R. Guo, J. Li, and H. Liu. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, page 232–240, 2020.

[72] J. Gyourko and P. Linneman. Rent controls and rental housing quality: A note on the effects of New York City's old controls. *Journal of Urban Economics*, 27(3):398–409, 1990.

[73] B. Halperin, B. Ho, J. A. List, and I. Muir. Toward an understanding of the economics of apologies: evidence from a large-scale natural field experiment. Technical report, National Bureau of Economic Research, 2019.

[74] J. Hamari, M. Sjöklint, and A. Ukkonen. The sharing economy: Why people participate in collaborative consumption. *Journal of the association for information science and technology*, 67(9):2047–2059, 2016.

[75] N. Heller. Is the gig economy working? *The New Yorker*, 2017.

[76] B. E. Hermalin. Toward an economic theory of leadership: Leading by example. *The American Economic Review*, 88(5):1188–1206, 1997.

[77] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[78] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[79] ILOSTAT. Labor force, total. *Retrieved from The world bank group at https://data.worldbank.org/indicator/SL.TLF.TOTL.IN?end=2019*, 2019.

[80] ILOSTAT. Population, total. *Retrieved from The world bank group at https://data.worldbank.org/indicator/SP.POP.TOTL*, 2019.

[81] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[82] S. Jiang, L. Chen, A. Mislove, and C. Wilson. On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi. In *Proceedings of the 2018 World Wide Web Conference*, pages 863–872. International World Wide Web Conferences Steering Committee, 2018.

[83] T. Joachims and A. Swaminathan. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1199–1201, 2016.

[84] J. S. Kang, P. Kuznetsova, M. Luca, and Y. Choi. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1443–1448, 2013.

[85] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD*, pages 1168–1176. ACM, 2013.

[86] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[87] S. B. Kotsiantis. Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.

[88] M. Kraus and S. Feuerriegel. Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104:38–48, 2017.

[89] R. E. Kraut, S. R. Fussell, S. E. Brennan, and J. Siegel. Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. *Distributed work*, pages 137–162, 2002.

[90] N. Kumar, N. Jafarinaimi, and M. Bin Morshed. Uber in bangladesh: The tangled web of mobility and justice. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):98, 2018.

[91] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international*

conference on knowledge discovery and data mining, pages 1909–1918. ACM, 2015.

[92] D. Lambert and D. Pregibon. More bang for their bucks: Assessing new features for online advertisers. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pages 7–15, 2007.

[93] M. Lecuyer, M. Tucker, and A. Chaintreau. Improving the transparency of the sharing economy. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1043–1051. International World Wide Web Conferences Steering Committee, 2017.

[94] L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934, 2015.

[95] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

[96] T. X. Liu, Z. Wan, and C. Yang. The efficiency of a dynamic decentralized two-sided matching market. University of Rochester Working Paper, 2018.

[97] M. Makar, A. Swaminathan, and E. Kıcıman. A distillation approach to data efficient individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4544–4551, 2019.

[98] A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. Schwamb, C. Lintott, and A. Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, 2013.

[99] T. McCue. 57 million U.S. workers are part of the gig economy. *Forbes*, 2018.

[100] A. Meyers, D. Cutts, D. A. Frank, S. Levenson, A. Skalicky, T. Heeren, J. Cook, C. Berkowitz, M. Black, P. Casey, et al. Subsidized housing and children's nutritional status: Data from a multisite surveillance study. *Archives of Pediatrics & Adolescent Medicine*, 159(6):551–556, 2005.

[101] B. Moldovanu, A. Sela, and X. Shi. Contests for status. *Journal of Political Economy*, 115(2):338–363, 2007.

[102] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[103] J. S. Neyman. On the application of probability theory to agricultural experiments. Essay on Principles. Section 9.(translated and edited by D.M. Dabrowska and T.P. Speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

[104] M. O. of Data Analytics. Tenant harassment project. 2018.

[105] C. H. Park, K. Son, J. H. Lee, and S.-H. Bae. Crowd vs. crowd: Large-scale cooperative design through open team competition. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1275–1284. ACM, 2013.

[106] J. Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[107] A. Pinsonneault, H. Barki, R. B. Gallupe, and N. Hoppen. Electronic brainstorming: The illusion of productivity. *Information Systems Research*, 10(2):110–133, 1999.

[108] A. K. Ponnuswami, K. Pattabiraman, D. Brand, and T. Kanungo. Model characterization curves for federated search using click-logs: Predicting user engagement metrics for the span of feasible operating points. In *Proceedings of the 20th international conference on World wide web*, pages 67–76, 2011.

[109] G. Quattrone, D. Proserpio, D. Quercia, L. Capra, and M. Musolesi. Who benefits from the sharing economy of airbnb? In *Proceedings of the 25th international conference on world wide web*, pages 1385–1394. International World Wide Web Conferences Steering Committee, 2016.

[110] T. S. Rai and A. P. Fiske. Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 118(1):57–75, 2011.

[111] A. Ramachandran, A. Kumar, H. Koenig, A. De Unanue, C. Sung, J. Walsh, J. Schneider, R. Ghani, and J. P. Ridgway. Predictive analytics for retention in care in an urban HIV clinic. *Scientific reports*, 10(1):1–10, 2020.

[112] A. J. Ravenelle. *Hustle and Gig: Struggling and Surviving in the Sharing Economy*. University of California Press, Oakland, California, first edition, 2019.

[113] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[114] M. Rokicki, S. Zerr, and S. Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. In *Proceedings of the 24th international conference on world wide web*, pages 906–915, 2015.

[115] J. Roman. In pursuit of smart. *National Fire Protection Association Journal*, 2014.

[116] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[117] D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[118] E. J. Sambolec, N. L. Kerr, and L. A. Messé. The role of competitiveness at social tasks: Can indirect cues enhance performance? *Journal of Applied Sport Psychology*, 19(2):160–172, 2007.

[119] C. D. Scales Jr, T. Moin, A. Fink, S. H. Berry, N. Afsar-Manesh, C. M. Mangione, and B. P. Kerfoot. A randomized, controlled trial of team-based competition to increase learner participation in quality-improvement education. *International Journal for Quality in Health Care*, 28(2):227–232, 2016.

[120] E. T. Schneiderman. NYS attorney general tenant's right guide. *NYC Government*, 2018.

[121] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085, 2017.

[122] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 275–284, 2011.

[123] M. Shayo. A model of social identity with an application to political economy: Nation, class, and redistribution. *American Political Science Review*, 103:147–174, 5 2009.

[124] R. M. Sheremeta. Behavior in group contests: A review of experimental research. *Journal of Economic Surveys*, 32(3):683–704, 2018.

[125] V. Subrahmanian and S. Kumar. Predicting human behavior: The next frontiers. *Science*, 355(6324):489–489, 2017.

[126] A. Sundararajan. *The sharing economy: The end of employment and the rise of crowd-based capitalism*. MIT Press, 2017.

[127] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, pages 3231–3239, 2015.

[128] H. Tajfel and J. Turner. The social identity theory of intergroup behavior. In S. Worchel and W. Austin, editors, *The Social Psychology of Intergroup Relations*. Nelson- Hall, Chicago, 1986.

[129] L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1587–1594, 2013.

[130] D. Tasse, A. Sciuto, and J. I. Hong. Our house, in the middle of our tweets. In *ICWSM*, pages 691–694, 2016.

[131] D. L. Teele. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. Yale University Press, 2014.

[132] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[133] G. G. Van Ryzin and T. Kamber. Subtenures and housing outcomes for low income renters in New York City. *Journal of Urban Affairs*, 24(2):197–218, 2002.

[134] H. R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016.

[135] M. Vojnović. *Contest theory: Incentive mechanisms and ranking methods*. Cambridge University Press, 2015.

[136] A. Walker. In New York, rents are increasing twice as fast as wages. *Curbed New York*, 2017.

[137] Z. Wang, K. Fu, and J. Ye. Learning to estimate the travel time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 858–866, 2018.

[138] J. Wegge, C. Roth, B. Neubach, K.-H. Schmidt, and R. Kanfer. Age and gender diversity as determinants of performance and health in a public organization: The role of task complexity and group size. *Journal of Applied Psychology*, 93(6):1301, 2008.

[139] M. Wood, J. Turnham, and G. Mills. Housing affordability and family well-being: Results from the housing voucher evaluation. *Housing Policy Debate*, 19(2):367–412, 2008.

[140] E. Wyly, K. Newman, A. Schafran, and E. Lee. Displacing new york. *Environment and Planning A*, 42(11):2602–2623, 2010.

[141] Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236, 2015.

[142] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[143] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

[144] T. Ye, W. Ai, Y. Chen, M. Qiaozhu, and J. Zhang. Inter-team status competition for ride sharing: A large-scale field experiment at didi. *AEA RCT Registry*, November 2018.

[145] T. Ye, W. Ai, L. Zhang, N. Luo, L. Zhang, J. Ye, and Q. Mei. Predicting individual treatment effects of large-scale team competitions in a ride-sharing economy. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2368–2377, 2020.

[146] T. Ye, K. Reinecke, and L. P. Robert Jr. Personalized feedback versus money: the effect on reliability of subjective data in online experimental platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 343–346, 2017.

[147] T. Ye and L. P. Robert Jr. Does collectivism inhibit individual creativity? The effects of collectivism and perceived diversity on individual creativity and satisfaction in virtual ideation teams. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2344–2358, 2017.

[148] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE international conference on data mining*, pages 435–442. IEEE, 2003.

[149] L. Zhang, T. Song, Y. Tong, Z. Zhou, D. Li, W. Ai, L. Zhang, G. Wu, Y. Liu, and J. Ye. Recommendation-based team formation for on-demand taxi-calling platform. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 59–68, 2019.

[150] Z. Zhang and B. Li. A quasi-experimental estimate of the impact of p2p transportation platforms on urban consumer patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1683–1692. ACM, 2017.

[151] Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405. International World Wide Web Conferences Steering Committee, 2015.

[152] D. Zheng, T. Hu, Q. You, H. A. Kautz, and J. Luo. Towards lifestyle understanding: Predicting home and vacation locations from user's online photo collections. In *ICWSM*, pages 553–561, 2015.