Alley, Michael; Biggs, Max; Hariss, Rim et al.

Working Paper

Pricing for Heterogeneous Products : Analytics for Ticket Reselling

Pricing for heterogeneous products: analytics for ticket reselling

Michael Alley
StubHub

Max Biggs
University of Virginia, Darden School of Business

Rim Hariss
McGill University, Desautels Faculty of Management

Charles Herrmann
BCG Gamma

Michael Li
Operations Research Center, MIT

Georgia Perakis
Sloan School of Management, MIT

Problem definition: We present a data-driven study of the secondary ticket market. In particular, we are primarily concerned with accurately estimating price sensitivity for listed tickets. In this setting, there are many issues including endogeneity, heterogeneity in price sensitivity for different tickets, binary outcomes and non-linear interactions between ticket features. Our secondary goal is to highlight how this estimation can be integrated into a prescriptive trading strategy for buying and selling tickets in an active marketplace. **Academic/practical relevance:** We present a novel method for demand estimation with heterogeneous treatment effect in the presence of confounding. In practice, we embed this method within an optimization framework for ticket reselling, providing StubHub with a new framework for pricing tickets on its platform. **Methodology:** We introduce a general double/orthogonalized machine learning method for classification problems. This method allows us to isolate the causal effects of price on the outcome by removing the conditional effects of the ticket and market features. Furthermore, we introduce a novel loss function which can be easily incorporated into powerful, off-the-shelf machine learning algorithms, including gradient boosted trees. We show how in the presence of hidden confounding variables instrumental variables can be incorporated. **Results:** Using a wide range of synthetic data sets, we show this approach beats state-of-the-art machine learning and causal inference approaches for estimating treatment effects in the classification setting. Furthermore, using NBA ticket listings from the 2014-2015 season, we show that probit models with instrumental variables, previously used for price estimation of tickets in the resale market, are significantly less accurate and potentially misspecified relative to our proposed approach. Through pricing simulations we show our proposed method is able to achieve an 11% ROI by buying and selling tickets, while existing techniques are not profitable. **Managerial implications:** The knowledge of how to price tickets on its platform offers a range of potential opportunities for our collaborator, both in terms of understanding sellers on their platform and in developing new products to offer them.

Keywords: Pricing, Revenue Management, Heterogeneous Treatment Effects

1. Introduction

Ticket marketplaces are a rapidly growing industry with listings for over 10 million events per month (Erskine 2015). They host an exchange for sellers who wish to sell tickets to interested buyers. Most of these tickets are previously purchased and not sold from the primary seller, but rather they are being resold. Although these marketplaces have an abundance of observational data on how sellers price tickets, understanding market dynamics remains a challenge. Many firms have difficulty in estimating demand and its sensitivity to price. Additionally, researchers have highlighted the high elasticity of demand in these markets, in particular how it varies heterogeneously across different tickets (Diehl et al. 2015, Jiaqi Xu et al. 2019). Through our collaboration with StubHub[1], we notice that estimating price sensitivity in demand is particularly challenging for the following reasons:

-*Heterogeneity in treatment effect* - Different seats have different sensitivity to price. In the data, we observe that tickets that are likely to sell tend to be less price sensitive. However, even for tickets which have similar estimated probabilities of selling, there are still large differences in price sensitivity, suggesting that there is further interaction between the ticket features and the price which our model aims to capture. Heterogeneous treatments are more difficult to estimate than homogeneous treatments due to the relative sparsity of individuals who react in the same way.

-*Hidden confounding factors* - A common issue firms face when trying to make inferences on data is confounding by hidden variables, which influence both the treatment (price) and the outcome (demand). This often occurs because firms are not able to exhaustively record or access all variables which are relevant to the pricing decision being made. This is true even with StubHub, which has a rich dataset with thousands of explanatory variables; econometric tests still suggest the presence of endogeneity (see section §6.2). These hidden confounding factors can lead to biased estimation of the parameters of interest.

-*Large, high dimensional datasets* - The dataset from StubHub contains millions of tickets, which potentially allows for the estimation of very rich models with complex interaction effects. Recent advances in machine learning and non-parametric estimation have enabled these effects to be captured with high predictive accuracy. However, high predictive accuracy does not imply the underlying price sensitivity can be captured consistently. It is well

[1] StubHub, the world's largest ticket marketplace, operates in 48 countries and offers live events across the globe. The events offered span music, sports, theater and other live entertainment shows

known that advanced machine learning methods could exhibit insufficient sensitivity toward key parameters (Niculescu-Mizil and Caruana 2005). The difficulty is further exacerbated by the high dimensionality of the data, as StubHub operates in tens of thousands of separate events, and each with tens of different sections of tickets.

- *Continuous treatments and binary outcomes* - The majority of the literature on causal treatment effect estimation focuses on binary or discrete treatments and continuous outcomes (Imbens and Rubin 2015). In our setting however, the treatment is price which is continuous, and the outcome (whether the ticket is sold or not) is binary. Binary outcomes require a non-linear response function to transform the response variables, while a continuous treatment means it is significantly harder to group tickets with similar treatments. Furthermore, unlike a primary ticket seller, and in contrast to many traditional revenue management settings, at any given time there is only a small percentage of the stadium $(5 - 10\%)$ available on the market. This further amplifies the effects of heterogeneity and makes it hard to pool similar tickets together into an aggregate demand model.

In this work, our primary goal is to accurately isolate and capture the treatment effect of price within a prediction framework that estimates the probability of an individual ticket selling. Our secondary goal is to integrate this model into a prescriptive trading strategy for buying and selling tickets in an active marketplace. Accurately estimating the causal price effects is very important when selling tickets, as incorrect estimates of the price sensitivity may lead to sub-optimal pricing decisions (Bertsimas and Kallus 2016). The knowledge of how to price tickets on its platform offers a range of potential opportunities for our collaborator, both in terms of understanding sellers on their platform and in developing new products to offer them. One potential application would be to offer sellers a bid at the time of listing wherein StubHub would advance the seller a guaranteed payment and then take on the risk of selling the inventory. This offers security to the seller as they no longer bear the risk that their ticket will not sell and the burden of managing the pricing.

Contributions: In this paper, we make several contributions.

1. *Novel classification model accounting for heterogeneous treatment effects.* To the best of our knowledge, we are the first to propose a semi-parametric model for measuring heterogeneous treatment effects using the concept of orthogonalization in the classification setting. To accomplish this goal, we develop a novel loss function that allows us to estimate the treatment effect directly. This loss function can be incorporated into a wide range

of off-the-shelf machine learning methods and allows us to leverage the power of cutting edge algorithms.

(a) We develop the *Two-stage estimation for classification algorithm* which accounts for any general functional form of the treatment effect. The model isolates the causal effects of the treatment on the outcome by removing the conditional effects of the control variables. In the presence of hidden confounders, instrumental variables can also be incorporated. Section §3.1 formulates the classification model and show how to estimate each of its parameters.

(b) *Consistency results and finite-sample guarantee on the estimation error.* We show that the estimates of the classification model are consistent, under mild assumptions on the functional form of the true treatment effect. We also provide a theoretical rate of convergence. The results are portrayed in section §3.2.

(c) *Numerical performance of the estimation procedure.* We integrate the resulting loss function into `lightgbm`, a widely used and highly accurate gradient boosting method. Over a wide range of synthetic data experiments, we show how our approach outperforms state-of-the-art machine learning methods for estimating treatment effects in classification tasks. We dedicate section §5 to the numerical experiments results.

2. *Practical contribution: StubHub case study.*

(a) We show how to *embed this classification procedure within an optimization framework for ticket reselling*. We apply our estimation procedure to predict whether an individual ticket is sold under a particular price (treatment). We then develop two pricing strategies, based on our underlying probability model with the heterogeneous price sensitivity, that can be embedded into a master problem with the goal to decide which tickets to buy. Section §4 describes the details of the pricing and the revenue management models.

(b) *Improved return on investment.* Using NBA ticket listings from the 2014-2015 calendar year, we observe strong predictive accuracy of our classification method. Furthermore, we show in pricing simulations that the proposed trading strategy (using our underlying predictive classification method) is able to achieve 11% ROI, while relying on existing techniques for estimating price sensitivity are not profitable. This shows that in an industry with slim margins, even slight misspecification of the demand model can result in significant losses. We observe a small but significant improvement from

7% to 11% ROI when using a time-varying pricing policy relative to a constant pricing policy, suggesting improvements can be made by considering the perishable inventory effects of tickets. All details of the implementation using StubHub data can be found in section §6.

We would also like to note that our approach for heterogeneous treatment estimation in classification settings can be generalized beyond the scope of ticket reselling. For example, in medicine there are many outcomes which are binary in nature (occurrence of a health event such as a heart attack or cancer) and treatments that need to be personalized according to a patient's symptoms and their corresponding health profile. The approach introduced in this paper could also be used in other revenue management settings where there is a high degree of differentiation and/or personalization of a product, leading to a classification model that is more accurate than an aggregate demand model.

Managerial insights: Our results present valuable insights on the state of the marketplace and show there is potential to improve overall marketplace efficiency for StubHub. We estimate that 12% of tickets are more than 10% under priced relative to optimal pricing. This is a significant arbitrage opportunity resulting in a surplus that is currently captured by 3rd party market makers, a practice StubHub could reduce without hindering liquidity. Similarly, we estimate that 14% of tickets are currently more than 10% overpriced relative to optimal pricing. This suggests there is a significant opportunity to help sellers through "better" pricing practices. Using our suggested algorithms, it is possible to correctly estimate whether a ticket will sell at a given price with 91% accuracy, an insight which is valuable in helping StubHub develop products which enhance the user experience. We show that previously used methods for price estimation of tickets in the resale market are significantly less accurate and potentially misspecified relative to our proposed approach. More broadly, we hope this study highlights the potential difficulties involved in causal price sensitivity estimation, as well as its implications on pricing practice.

2. Literature review

Our work lies at the intersection of three streams of literature; causal inference, pricing for events and data-driven pricing.

2.1. Causal Inference

In the operations management community, there is growing awareness of the need to identify the causal relationships of treatments or interventions that are used in making business

decisions. In the presence of endogeneity, instrumental variables are required to identify the treatment effect. Starting with continuous outcomes (Wright 1928, Reiersøl 1945), there is a body of work which extended instrumental variable models to binary outcomes using linear probit models (Rivers and Vuong 1988, Lee 1981, Amemiya 1978). These models are often referred to in the literature as control function approaches. Similar to our setting, these models have a continuous treatment but treatment effects are homogeneous. Due to recent advances in machine learning and advantages in utilizing high dimensional data, there have been extensions to non-parametric models with instrumental variables and binary outcomes (Blundell and Powell 2003, 2004). Our main contribution over this literature is showing the effectiveness of an orthogonalization step in the estimation, which is beneficial regardless of whether confounding is present, and the heterogeneity in effect estimation. We outline the estimation procedure of Rivers and Vuong (1988) and Blundell and Powell (2003) in more detail in section §5, where we use these approaches as benchmarks in simulation experiments. For a comprehensive review of instrumental variable models with binary outcomes, we refer the reader to Imbens and Wooldridge (2007).

A relatively recent stream of literature for estimating treatment effects is orthogonalized or double machine learning methods. These methods aim to isolate the effects of the treatment on the outcome by removing the conditional effects of the control variables. The methodology was first introduced by Robinson (1988) for estimating parametric components in partially linear models. Chernozhukov et al. (2018) provides examples of estimating (homogeneous) treatment effects in the presence of high dimensional controls. Application of these ideas in a pricing setting with a continuous demand model can be found in Chernozhukov et al. (2017). Oprescu et al. (2018) provides some extensions of this approach to a heterogeneous treatment effect estimation using random forests. Nie and Wager (2017) develops a specific loss function for the regression setting of heterogeneous treatment effect estimation with binary treatments, using the orthogonalized outcome and treatment. In contrast, our approach provides novel loss functions and algorithms for the classification setting for continuous rather than binary treatments.

2.2. Ticket Pricing for sporting events

There have been several papers studying pricing for event sales for sports events (Bouchet et al. 2016, Barlow 2000, Duran et al. 2012, Sainam et al. 2010, Kemper and Breuer 2016, Jiaqi Xu et al. 2019). These papers focus on the problem faced by primary sellers rather than

resellers. They typically assume demand can be aggregated by section in the venue of the event. While this is reasonable in the primary sellers case due to the large number of tickets they are selling, this is not as reasonable in the reselling case where the tickets being sold at any time are extremely sparse and dependent on market features.

In terms of recent work on secondary ticket sales markets, Zhu (2014) studies the case of strategic consumers who potentially defer purchasing decisions. Diehl et al. (2015) studies the NFL resale market using an aggregate demand model and IV regression. These models use traditional econometric techniques and transformations of linear specifications, which aren't necessarily able to capture the complex interactions which exist in large scale data, relative to the machine learning based techniques we develop.

2.2.1. Comparison with Sweeting (2012) Similar to our work, Sweeting (2012) develops a dynamic pricing model for selling a single ticket on the MLB resale market. Nevertheless, the paper's focus is on capturing the seller's behavior rather than prescribing the optimal prices. Sweeting (2012) uses a probit linear model with instrumental variables (`probit_iv`) to capture the effect of price on the probability of selling an individual ticket. The price sensitivity (treatment effect) is homogeneous (not dependent on ticket or market features) and the effects of covariates on price are characterized linearly. In contrast, our specification is able to capture heterogeneous price sensitivity with high flexibility (using general non-parametric models such as boosted trees and deep neural networks) of how the heterogeneous treatment effect and nuisance parameters depend on covariates. The collaboration with StubHub as well as recent related research have made us aware of the high elasticity of demand in secondary markets, and in particular, how it varies heterogeneously across different tickets (Diehl et al. 2015, Jiaqi Xu et al. 2019).

To estimate the parameters in the `probit_iv` model, Sweeting (2012) uses Full Maximum Likelihood Estimation (FMLE). Heckman (1977) observed however that estimating structural parameters using FMLE methods is computationally difficult. In particular, our model's flexible specification using non-parametric functions for the nuisance parameters amplifies this computational difficulty. To address the estimation's challenges in this flexible specification, we use orthogonalization to develop a novel loss function for classification which is easily embedded within the state-of-the-art machine learning algorithms, and is robust to nuisance parameters. In particular, consistency results from Sweeting (2012) only hold if the true model

has a linear form. In comparison, the results of this paper hold under more general assumptions allowing non-linearity and heterogeneity.

There are also many papers on data-driven pricing problems which incorporate high dimensional features of the product or customer, for example (Cohen et al. 2016, Javanmard and Nazerzadeh 2016, Ban and Keskin 2017).

3. Heterogeneous treatment effect estimation for classification

In this section, we consider a latent variable model that is a semi-parametric variant of the model proposed in Rivers and Vuong (1988), which explicitly models the confounding effect of ticket features on price. We assume access to n identically and independent samples $W_i = (Y_i, X_i, T_i, Z_i), i = 1, ..., n$, where $Y_i \in \{0, 1\}$ is a binary outcome, $T_i \in \mathbb{R}$ is the treatment, $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is the control, and $Z_i \in \mathbb{R}^k$ is a set of instrumental variables. We propose the following model where Y^* is an unobserved latent random variable such that:

$$Y^* = g(X) + \tau(X)T + \epsilon \tag{1}$$

$$T = m(X) + \beta^T Z + \upsilon \tag{2}$$

$$Z = h(X) + u \tag{3}$$

$$Y = \begin{cases} 1, & \text{if } Y^* > 0 \\ 0, & \text{if } Y^* \leq 0 \end{cases} \tag{4}$$

$$\begin{pmatrix} \epsilon \\ \upsilon \end{pmatrix} \sim N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{pmatrix} 1 & \rho\sigma_\upsilon \\ \rho\sigma_\upsilon & \sigma_\upsilon^2 \end{pmatrix}\right), \; u \sim N(0,1), \; \mathbb{E}[u\upsilon|X] = \; \mathbb{E}[u\epsilon|X] = 0 \tag{5}$$

In the ticket selling context, Y is the observation of whether the ticket is sold or not. X corresponds to a set of features with respect to the ticket and the market state (for a more detailed description, see section §6.3). T is the price chosen for that ticket. The heterogeneous treatment effect is captured through $\tau(X)$, which can be interpreted as the price sensitivity dependent on the features of the ticket. The term $m(X)$ can be interpreted as the fair market value of the ticket (the portion of the ticket's price that can be explained by the features of the ticket), while υ can be interpreted as the idiosyncratic choice by the seller of how to price the ticket relative to the fair market value. The term $g(X)$ corresponds to the contribution to the likelihood of selling of the ticket features.

Finally, the model can suffer from omitted variable bias, where there are unobserved variables which affect both Y and T. These endogenous factors cause correlation between residual errors ϵ and υ, leading to bias in the estimate of the treatment effect $\tau(X)$. To overcome

this, suitable instruments Z are required, which only affect Y through the interaction with T. Through variation of Z, we can observe the impact of changing T on Y. These relationships are captured through the condition $\mathbb{E}[uv|X] = \mathbb{E}[u\epsilon|X] = 0$. The term $h(X)$ captures possible correlation between the instruments and features of the ticket. We go over the specific set of instruments we use in the StubHub case study in section §6.2.

The model in (1-5) is difficult to estimate due to the semi-parametric form of equations (1-3). The non-linear terms $g(X), \tau(X), m(X), h(X)$ are challenging to estimate jointly using maximum likelihood estimation. Simpler models where these terms are linear and the treatment is homogeneous, such as in Rivers and Vuong (1988), can perform poorly when the true models are non-linear as is often the case with real data. Fully non-parametric models, such as boosted trees and neural networks, are often able to achieve good predictive accuracy, but do not explicitly exploit the structure of the model to estimate $\tau(X)$, leading to poor approximate estimates of the heterogeneous treatment effect.

Our model provides a general approach that works for heterogeneous treatment effect estimation in the presence of endogeneity using instrumental variables. However, we note that this method can also be applied in the special case where ϵ and υ are independent (no endogeneity), in which case an instrumental variable Z is not required. We show in section §5 strong empirical evidence that this methodology is more effective in the uncorrelated case than other machine learning approaches which do not use orthogonalization.

3.1. Method

We follow the literature on double/orthogonalized machine learning to isolate the causal effects of the treatment on the outcome by removing the conditional effects of the control variables. Machine learning algorithms are known to be good predictors, but this literature also shows they can be adapted to be good treatment effect estimators that are robust to confounding. Using orthogonalization, we derive a novel loss function for treatment effects estimation for classification. We first partial out the effect of X from T to obtain the orthogonalized regressor $T - \mathbb{E}[T|X]$. Furthermore, we partial out the effect of X on Y^* to obtain an orthogonalized outcome $Y^* - \mathbb{E}[Y^*|X]$. Since $\mathbb{E}[Y^*|X] = g(X) + \tau(X)m(X) + \tau(X)\beta^T h(X)$ and $\mathbb{E}[\epsilon|X] = 0$, we can rearrange (1) and (2) as follows:

$$Y^* - \mathbb{E}[Y^*|X] = \tau(X)(T - \mathbb{E}[T|X]) + \epsilon. \tag{6}$$

This new expression has the advantage of eliminating $g(X)$ and thus removing the direct effect of confounding. In the classification setting, since Y^* is a latent variable, $\mathbb{E}[Y^*|X]$ is

generally not known, but $\mathbb{E}[Y|X]$ can often be approximated using non-parametric machine learning techniques. Under some assumptions that we will present below, we next link $\mathbb{E}[Y^*|X]$ to $\mathbb{E}[Y|X]$. In what follows, we take $k=1$ (one instrumental variable) but one can easily generalize our results to the case of $k>1$.

PROPOSITION 1. *If* $\begin{pmatrix}\epsilon\\ \upsilon\end{pmatrix} \sim N\left(\begin{matrix}0\\ 0,\end{matrix}\begin{pmatrix}\sigma_\epsilon^2 & \rho\sigma_\epsilon\sigma_\upsilon\\ \rho\sigma_\epsilon\sigma_\upsilon & \sigma_\upsilon^2\end{pmatrix}\right)$ *, and* $u\sim\mathcal{N}(0,\sigma_u^2)$ *is uncorrelated with* ϵ,υ*, then:*

$$\mathbb{E}[Y^*|X] = \sqrt{\sigma_u^2\tau(X)^2\beta^2+\sigma_\epsilon^2+2\tau(X)\sigma_\epsilon\sigma_\upsilon+\tau(X)^2\sigma_\upsilon^2}\cdot\Phi^{-1}(\mathbb{E}[Y|X]).$$

The proof can be found in appendix A. We can also decompose $\epsilon=\frac{\rho\sigma_\epsilon}{\sigma_\upsilon}\upsilon+\tilde{\epsilon}$ into orthogonal components, where $E[\upsilon\tilde{\epsilon}]=0$, and $\tilde{\epsilon}|X\sim N(0,(1-\rho^2)\sigma_\epsilon^2)$ (see Bertsekas and Tsitsiklis (2002)). With this modification, the treatment T is orthogonal to the error $\tilde{\epsilon}$. The expression in Proposition 1 can be to substituted into (6) to obtain a new expression for the latent variable:

$$Y^* = \Phi^{-1}(\mathbb{E}[Y|X])\sqrt{\sigma_\epsilon^2+2\rho\sigma_\epsilon\sigma_\upsilon\tau(X)+({\sigma_\upsilon}^2+\sigma_u^2\beta^2)\tau(X)^2}+\tau(X)(T-\mathbb{E}[T|X])+\frac{\rho\sigma_\epsilon}{\sigma_\upsilon}\upsilon+\tilde{\epsilon}. \tag{7}$$

In the remainder of the paper, we assume for simplicity and without loss of generality that $\sigma_\epsilon=1$. This is a standard assumption in the literature of probit models with instrumental variables (Rivers and Vuong 1988)[2]. For brevity, denote $w{=}(\beta,\sigma_u,\sigma_\upsilon)$. From expression (7), we can derive an appropriate loss function resulting from the normally distributed random component $\tilde{\epsilon}$:

$$l(Y,X,T,\tau,\rho,w,\upsilon) = Y\log(\Phi(f(X,T,\tau,\rho,w,\upsilon))+(1-Y)\log(1-\Phi(f(X,T,\tau,\rho,w,\upsilon))) \tag{8}$$

$$f(X,T,\tau,\rho,w,\upsilon) = \frac{\Phi^{-1}(\mathbb{E}[Y|X])k(X,\tau,\rho,w)+\tau(X)(T-\mathbb{E}[T|X])+\frac{\rho}{\sigma_\upsilon}\upsilon}{\sqrt{1-\rho^2}} \tag{9}$$

$$\textit{and}\quad k(X,\tau,\rho,w) = \sqrt{1+2\rho\sigma_\upsilon\tau(X)+({\sigma_\upsilon}^2+\sigma_u^2\beta^2)\tau(X)^2}. \tag{10}$$

We note that most of these nuisance parameters can be estimated prior to optimization of this loss function using machine learning algorithms. $\hat{r}(X)\approx\mathbb{E}[Y|X]$ can be estimated using non-parametric classification prediction methods of Y on X. An estimator $\hat{q}(X)\approx\mathbb{E}[T|X]$ can be found by non-parametric regression of T on X, $\hat{h}(X)\approx\mathbb{E}[Z|X]$ by Z on X while $\hat{u}$ can be calculated as the residuals.

[2] Although in general σ_ϵ is not identifiable, it is not necessary to estimate the treatment effect. In the case it is not equal to one, we estimate a scaled treatment effect $\hat{\tau}(X)=\frac{\tau(X)}{\sigma_\epsilon}$, which reflects the change in probability with treatment.

In the case where $m(X)$ is linear, linear regression of T on X and Z can be used for estimates $\hat{m}(X)$ and $\hat{\beta}$, while $\hat{v}$ are the residuals. In the general non-linear case, an additional orthogonalization step can be used on (2):

$$T - \mathbb{E}[T|X] = \beta(Z - h(X)) + \upsilon, \tag{11}$$

so $\hat{\beta}$ can be calculated by regressing $T - \hat{q}(X)$ on $Z - \hat{h}(X)$, while again $\hat{v}$ are the residuals. To ensure that the estimates of $\hat{\beta}, \hat{v}$ are indeed consistent, we randomly split the data into 3 equally-sized partitions and use different datasets to estimate $\hat{q}(X)$, $\hat{h}(X)$, and the nuisance parameters $\hat{\beta}, \hat{v}$. This ensures that $\hat{q}(X)$ is independent from $\hat{h}(X)$, and independent from T, Z which serve to estimate the nuisance parameters, enabling standard consistency results.

For the unknown variances σ_v and σ_u, the naive estimation procedure ($\hat{\sigma}_\upsilon = \sqrt{\frac{1}{n}\sum_i \hat{v}_i^2}$, $\hat{\sigma}_u = \sqrt{\frac{1}{n}\sum_i \hat{u}_i^2}$) does result in a consistent estimator if we follow the sample splitting procedure above. The analytical results are detailed in section §3.2. The finite-sample performance of this estimator can be further improved if a bootstrapping method is applied for a finite sample correction. For example, we can show that the estimator $\hat{\sigma}_u$ satisfies:

$$\mathbb{E}_{u,X}\left[\hat{\sigma}_u^2\right] = \sigma_u^2 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{u,X}\left[(\hat{h}(X_i) - h(X_i))^2\right].$$

Given our assumptions, this estimator is indeed consistent as $\hat{h}(X) \to h(X)$, but suffers from a finite sample error of $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{u,X}\left[(\hat{h}(X_i) - h(X_i))^2\right]$. One can use the bootstrap method (see details in appendix C) to estimate this finite sample error which further improves the accuracy of our estimator. In section §5, we run computations on synthetic datasets to illustrate the practical performance of our estimation procedure (including variance estimation). More formally, our estimation algorithm can be summarized as follows:

Algorithm 1 Two stage estimation for classification algorithm with instrumental variables

1: *Randomly divide the dataset into three equal-sized sets* S_1, S_2, S_3

2: *Fit* $\hat{r}(X) \approx E[Y|X]$, $\hat{q}(X) \approx \mathbb{E}[T|X]$ *via appropriate non-parametric methods using set* S_1.

3: *Fit* $\hat{h}(X) \approx \mathbb{E}[Z|X]$ *via appropriate non-parametric methods using set* S_2.

4: *Using set* S_3, *calculate* $\hat{w} = (\hat{\beta},\ \hat{\sigma_u},\ \hat{\sigma_v})$ *and* $\hat{v}$.

5: *Define the approximate log-likelihood function:*

$$\begin{aligned}\hat{l}(Y, X, T, Z, \tau, \rho, \hat{w}, \hat{v}, \hat{r}, \hat{q}) =& Y \log(\Phi(\hat{f}(X, T, Z, \tau, \rho, \hat{w}, \hat{v}, \hat{r}, \hat{q})) \\ &+ (1 - Y) \log(1 - \Phi(\hat{f}(X, T, Z, \tau, \rho, \hat{w}, \hat{v}, \hat{r}, \hat{q}))) \end{aligned} \tag{12}$$

where $\hat{f}(X, T, Z, \tau, \rho, \hat{w}, \hat{v}, \hat{r}, \hat{q}) = \frac{\Phi^{-1}(\hat{r}(X))k(X,\tau,\rho,\hat{w}) + \tau(X)(T - \hat{q}(X)) + \frac{\rho}{\hat{\sigma}_v}\hat{v}}{\sqrt{1-\rho^2}}$ and k is defined in Eq. (10). Then using all samples from S_1, S_2, S_3, find the approximate maximum likelihood estimator:

$$(\hat{\tau}(\cdot), \hat{\rho}) = \arg\min_{\tau,\rho} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \hat{l}(Y_i, X_i, T_i, Z_i, \tau, \rho, \hat{w}, \hat{v}_i, \hat{r}, \hat{q}) \right\} \tag{13}$$

In this procedure, steps 1-4 make use of sample splitting to learn independent estimates for relevant parameters, while step 5 optimizes the approximate empirical log-likelihood loss function. Optimizing this loss function requires finding the optimal function $\hat{\tau}(\cdot)$ and the correlation parameter $\hat{\rho}$. As ρ is one dimensional, we recommend using a line search algorithm to find $\hat{\rho}$. In particular, for any fixed ρ, we could find the optimal $\hat{\tau}(\cdot) = \arg\min_\tau -\frac{1}{n}\sum_{i=1}^{n} \hat{l}(Y_i, X_i, T_i, Z_i, \tau, \rho, \hat{w}, \hat{v}_i, \hat{r}, \hat{q})$, and evaluate the loss function. Then ρ can be optimized using Brent's method (for example).

Various machine learning optimization methods and base functions can be used for estimating the function $\tau(X)$. This step can efficiently be solved by adapting off the shelf methods, using a customized loss function. In the ticket reselling setting, we observe that restricting $\hat{\tau}(\cdot)$ to be a tree ensemble and optimizing the second stage using the `lightgbm` package (Ke et al. 2017) works well in practice as shown in section §6.3. The `lightgbm` package is an example of gradient boosted trees (Friedman 2001). The wide effectiveness of ensembles are displayed in Kaggle competitions with 60% of winning solutions using gradient boosting implementations (Rogozhnikov and Likhomanenko 2017). The specific details of how to incorporate the loss function minimization into the `lightgbm` package are given in appendix G. Alternatively, a

deep learning approach could be taken by restricting $\hat{\tau}(\cdot)$ to be a neural network and optimizing the second stage loss function using `tensorflow` (Abadi et al. 2016). For simpler settings, such as when the treatment effect is constant ($\hat{\tau}(\cdot) = \tau$), a one dimensional search of the loss function can be used to find the optimal value (see section §5.1 for experiments).

Note that this method presents a general way of estimating heterogeneous treatments for classification tasks, as it can be generalized to other variants of the problem depending on distributions of ϵ and v. For Gaussian errors, we are able to derive a closed form of the loss function (as Gaussian variables have a closed form convolution). It is possible to derive analogous equations for the case where ϵ and v are not Gaussian if we assume full knowledge of the joint distribution of ϵ and v. Nevertheless, this CDF might be difficult to calculate in practice. We show how to derive these equations with general error distributions in appendix H. Next, we prove consistency results of the estimators in (13).

3.2. Consistency and Stability of Estimator $\hat{\tau}(\cdot)$

In what follows, we show that under mild regularity conditions, the estimator $\hat{\tau}(\cdot)$ that results from Algorithm 1 would approximate well, with enough data, the true treatment $\tau(\cdot)$. This proof relies on standard assumptions in the econometrics literature (Engle 1994) about consistency, compactness and identifiability of the treatment effect. This in turn ensures that the approximate likelihood function defined in (12) is close to the true likelihood function. Another feature we show is that $\hat{\tau}(\cdot)$ converge as long as $\hat{r}(X)$ and $\hat{q}(X)$ converge (along with the other estimated nuisance parameters). Our proof holds for a broad class of treatment effects, in which the treatment can be characterized by finitely many parameters $\tau(X) = t(\theta_1, \cdots, \theta_p, X)$. We first summarize these standard assumptions regarding the underlying structure of the treatment effect and the nuisance parameters, which facilitate the subsequent analysis.

Assumption 1 *For the latent variable model in (1)-(5), we assume the following:*

(a) (Finite Parameterization) $\tau(X)$ *is a finitely parameterized function* $\tau(X) = t(\theta_1, \cdots, \theta_p, X) = t(\boldsymbol{\theta}, X)$ *where* $p \in \mathbb{Z}^+$, *and* t *is known.*

(b) (Compactness) $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \rho) \in \mathcal{D}_\Theta$ *and* $X \in \mathcal{D}_X$ *where* $\mathcal{D}_\Theta$ *and* $\mathcal{D}_X$ *are compact.*

(c) (Continuity) $r(X), q(X), h(X)$ *are continuous,* $0 < r(X) < 1$ *and* $t(\boldsymbol{\theta}, X)$ *is continuous in* $\boldsymbol{\theta}$ *for every* X, *and is continuous in* X *for every* $\boldsymbol{\theta}$ *with probability 1.*

(d) (Identifiability) $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \Leftrightarrow \quad t(\boldsymbol{\theta}, \cdot) = t(\boldsymbol{\theta}_0, \cdot)$.

(e) (Consistency of Estimation) $\sup_{X \in \mathcal{D}_x} |\hat{r}(X) - r(X)| \xrightarrow{p} 0$, $\sup_{X \in \mathcal{D}_x} |\hat{q}(X) - q(X)| \xrightarrow{p} 0$, *and* $\sup_{X \in \mathcal{D}_x} |\hat{h}(X) - h(X)| \xrightarrow{p} 0$.

To derive rates of convergence, we further require our non-parametric methods to converge at least at the $\sqrt{n}$ rate. This is achieved by many popular machine learning methods under suitable conditions, including random forests (Biau 2012) and gradient boosted trees under early stopping (Zhang et al. 2005).

Assumption 2 (Rate of Convergence) *Assume that the non-parametric estimators satisfy, for some* $\delta > 0$, $\sup_{X \in \mathcal{D}_x} n^{1/(2+\delta)} |\hat{r}(X) - r(X)| \xrightarrow{p} 0$, $\sup_{X \in \mathcal{D}_x} n^{1/(2+\delta)} |\hat{q}(X) - q(X)| \xrightarrow{p} 0$, *and* $\sup_{X \in \mathcal{D}_x} n^{1/(2+\delta)} |\hat{h}(X) - h(X)| \xrightarrow{p} 0$.

The following lemma then shows that under the regularity assumptions, the nuisance parameters can be consistently estimated. Furthermore, their rate of convergence can also be properly derived if we assume the non-parametric estimators have a sufficient rate of convergence:

LEMMA 1. *Consider the latent variable model in (1)-(5). If Assumptions 1(b), 1(c) and 1(e) hold, then* $\hat{\beta}$, $\hat{\sigma}_u$, $\hat{\sigma}_v$ *and* $\hat{v}$ *from Algorithm 1 are consistent estimators. Furthermore, if Assumption 2 holds, then the estimators converge with a rate at least as fast as* $n^{1/(2+\delta)}$.

The proof of the lemma is in appendix B. Having set up these assumptions, the following theorem proves the finite sample consistency of the treatment effect estimator:

THEOREM 1. *For the latent variable model in (1)-(5), we can parameterize the log-likelihood as* $l(Y, X, T, \tau, \rho, w, v) = l(Y, X, T, \boldsymbol{\Theta}, w, v)$ *where* $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \rho)$ *and similarly for* $\hat{l}, \hat{f}, k$. *For iid samples* $(X_i, Y_i, T_i, Z_i)_{i=1\cdots n} \in \mathcal{D}_X \times \mathcal{D}_Y \times \mathcal{D}_T \times \mathcal{D}_Z$, *define the estimator* $\hat{\tau}$ *as:*

$$\hat{\tau}(X) = t(\hat{\theta}_1, \cdots, \hat{\theta}_p, X) = t(\hat{\boldsymbol{\theta}}, X) \tag{14}$$

$$s.t.\ (\hat{\boldsymbol{\theta}}, \hat{\rho}) = \hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathcal{D}_\Theta}{\arg\max} \left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{l}(Y_i, X_i, T_i, Z_i, \boldsymbol{\Theta}, \hat{w}, \hat{v}_i, \hat{r}, \hat{q}) \right\} \tag{15}$$

Under Assumption 1:

a. *The log likelihood* $(\boldsymbol{\theta}, \rho) = \boldsymbol{\Theta} \mapsto \mathbb{E}[l(Y, X, T, \boldsymbol{\Theta}, w, v)]$ *has a unique minima* $(\boldsymbol{\theta}^*, \rho^*)$.
b. $\hat{\tau}(X) = t(\hat{\boldsymbol{\theta}}, X)$ *is a consistent estimator of* $\tau^*(X) = t(\boldsymbol{\theta}^*, X)$, *and* $\hat{\rho}$ *is a consistent estimator of* ρ^*, *as in:*

$$\sup_{X \in \mathcal{D}_X} \|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}^*, X)\| \xrightarrow{p} 0 \quad and \quad \|\hat{\rho} - \rho^*\| \xrightarrow{p} 0$$

Furthermore, if Assumption 2 holds, then we have:

$$\sup_{X \in \mathcal{D}_X} \sqrt{n} \|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}^*, X)\| \xrightarrow{p} 0 \quad and \quad \sqrt{n} \|\hat{\rho} - \rho^*\| \xrightarrow{p} 0$$

The proof can be found in appendix D. Through extensive numerical simulations on synthetic data (see section §5), we show that gradient boosted trees are able to find the unique optimal solution fast. Next, we leverage our classification method to develop an optimization framework that identifies arbitrage opportunities for StubHub.

4. Revenue management for ticket reselling

In this section, we introduce an optimization framework for revenue management for ticket reselling. The framework relies on two building blocks: (i) Estimating potential optimal revenue for each ticket, (ii) Developing a trading algorithm for tickets given their potential revenues. To achieve part (i), we first examine two pricing models for choosing the optimal price to sell an individual ticket: a static pricing model, where prices are held constant over the time horizon and a time-variant pricing model, where prices change over time. Both pricing models use our underlying classification method for whether an individual ticket is sold or not under a particular price. These pricing models can then be embedded into a master problem with the goal to decide which tickets to buy given the estimated potential revenues.

4.1. Assumptions

We now discuss a few key assumptions in our pricing approach:

First, since our collaborators do not aim to transact large volumes of tickets, we assume that the trading algorithm will buy/sell small volumes relative to the secondary ticket market size. This is necessary to prevent possible endogeneity issues; if a market maker were to buy too many tickets, the underlying market mechanisms would change. Thus, it would be difficult to justify using historical data to estimate price sensitivity once a new entrant becomes a sizeable market participant.

Second, we develop both pricing models for pricing a single ticket. If the number of tickets being sold by the seller is small and types of tickets are diverse, the cross effects of tickets in the portfolio will also be small and the relative effects on the overall market will be minor. We evaluate the strength of these cross-effects on the data in appendix J, and show that based on our data single-ticket optimization is justified.

Finally, we assume that demand is independent over time periods, but allow it to change over time according to a forecast. While simple, this model is sufficiently rich to capture the time dynamics of perishable inventory, and how prices fall closer to the event time to improve the chances of selling. Furthermore, the simplicity resulting from these assumptions is necessary to solve the model tractably on the scale of the millions of tickets that could possibly

be purchased and sold on the market. It is desirable to be able to make these calculations instantaneously to be able to offer the seller a price before they list their ticket.

In what follows, we introduce first the static, and then the dynamic pricing algorithms. Subsequently, we analyze the optimization problem for deciding the ticket portfolio.

4.2. Static pricing

To find the optimal revenue for a single ticket, the retailer optimizes the expected revenue of a ticket by choosing the optimal price $T^* = \arg\max_T \mathbb{E}[R|T, X, Z]$ where X are features of the ticket, Z is the related instrument and R is the resulting revenue. Denoting $Y = 1$ the event that the ticket sells, then $\mathbb{E}[R|T, X, Z] = T \cdot P(Y = 1|T, X, Z)$ where $P(Y = 1|T, X, Z)$ is the probability that the ticket sells at price T. Under the two-stage classification model, we can rearrange (9) and substitute for the maximum likelihood estimator of $P(Y = 1|T, X, Z)$. We then find an estimate of the maximum expected revenue $\hat{R}^*(X, Z)$ by solving the following problem:

$$\hat{R}^*(X, Z) = \max_T \; T \cdot \Phi\Big(\tilde{f}(X, T, \hat{\tau}, \hat{\rho}, \hat{\sigma_v}, \hat{v}, \hat{q}, \hat{r}, \hat{h})\Big) \tag{16}$$

where $\tilde{f}(X, T, \hat{\tau}, \hat{\rho}, \hat{\sigma_v}, \hat{v}, \hat{q}, \hat{r}, \hat{h}) \equiv \frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})(T-\hat{q}(X))-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}$ and k defined in Eq. (10). The following theorem describes the reseller's optimal pricing strategy.

THEOREM 2 **(Optimal Price for the reseller)**. *For a ticket defined by covariates X:*

1. *If $\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}} \geq 0$, then the revenue is increasing with price and the reseller should choose the highest price possible for the ticket.*
2. *If $\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}} < 0$, then the revenue is unimodal with optimum T^* satisfying:*

$$\bar{T} = \frac{\Phi^{-1}(\hat{r}(X))k(X, \hat{\tau}, \hat{\rho}, \hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{-2(\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})}$$
$$+ \frac{\sqrt{\left(\Phi^{-1}(\hat{r}(X))k(X, \hat{\tau}, \hat{\rho}, \hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))\right)^2 + 8(1 - \hat{\rho}^2)}}{-2(\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})}$$

The proof can be found in appendix E. In practice, in the pricing setting, the demand or likelihood of selling (a ticket in this setting) is negatively correlated with price, which means that $\hat{\tau}(X) < 0$, and furthermore our experiments suggest $\hat{\rho}$ is negative for the ticket reselling setting so the condition $\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}} < 0$ is always satisfied in practice. There are advantages to having the revenue be a unimodal function, it can be optimized efficiently using a line search algorithm such as Brent's method (Brent 1971). In addition, having an upper bound helps limit the search space.

4.3. Time-varying pricing

We formulate the dynamic pricing problem for a single ticket as a Markov decision problem over a finite horizon H. We denote the binary variable W_t whether the ticket will sell in a given period t. The ticket features are explicitly split into M_t market features which vary with time (discussed further in appendix F) and X the ticket features which do not change over time. Note, for notational simplicity we suppress the dependence on the ticket i, but there is a unique X and M_t for each ticket. The probability of selling the ticket in a single period at a price T_t is $P(W_t = 1|T_t, X, M_t)$ and the system's state is described by M_t. Denote $J_t(M_t)$ the expected optimal revenue to go from period $t+1$ to H given state M_t. Then, the dynamic programming formulation (using Bellman's equation) is:

$$\begin{aligned} J_t(M_t) = &\max_{T_t} \; T_t \cdot P(W_t = 1|T_t, X, M_t) \\ &+ (1 - P(W_t = 1|T_t, X, M_t))\mathbb{E}[J_{t+1}(M_{t+1})], \; 1 \le t \le H, \text{ and } J_{H+1}(M_{H+1}) = 0; \end{aligned} \tag{17}$$

where the first term in (17) is the revenue achieved if the ticket sells while the second term is the expected revenue from the future. It is possible to estimate the probability of selling in a single period $P(W_t = 1|T_t, X, M_t)$ from the probability of selling in the remaining horizon $P(Y_t = 1|T_t, X, M_t)$:

$$P(W_t = 1|T_t, X, M_t) = \frac{P(Y_t = 1|T_t, X, M_t) - P(Y_{t-1} = 1|T_{t-1}, X, M_{t-1})}{1 - P(Y_{t-1} = 1|T_{t-1}, X, M_{t-1})} \tag{18}$$

While theoretically solvable, the dynamic programming solution is not tractable at the scale of our problem. A common approach in the literature is to solve this model using the *deterministic equivalent LP*, which we denote as $\mathcal{DLP}$. It is used by Ma et al. (2018) in the pricing setting, and by several others (Gallego and Van Ryzin 1994, Wang et al. 2018) in the revenue management and scheduling literature.

$$\bar{R}_H(X_i) = \max \sum_{j \in \mathcal{L}} \sum_{t=1}^{H} q_{tj} T_{tj} x_{tj} \tag{19}$$

$$\text{s.t} \sum_{j \in \mathcal{L}} \sum_{t=1}^{H} q_{tj} x_{tj} \le 1 \tag{20}$$

$$\sum_{j \in \mathcal{L}} x_{tj} \le 1 \qquad \forall t \in \{1, ..., H\} \tag{21}$$

$$x_{tj} \ge 0 \qquad \forall t \in \{1, ..., H\}, \; j \in \mathcal{L} \tag{22}$$

In this formulation, the variable x_{tj} selects a price T_{tj} and quantity $q_{tj} = P(W_t = 1|T_{tj}, X, \mathbb{E}M_t)$ from a discrete ladder $j \in \mathcal{L}$ for each time period, while the objective is the expected revenue over the selling time horizon. Constraint (20) is an inventory constraint which requires the ticket can be sold at most once in probability, while constraint (21) requires no more than one price selected per period.

It is well known that $OPT_{\mathcal{LP}}$, the optimal objective value for $\mathcal{DLP}$ is an upper bound on the optimal dynamic policy OPT_{DP}; i.e. $OPT_{\mathcal{LP}} \geq OPT_{DP}$ (Gallego and Van Ryzin 1994, Gallego et al. 2004). Furthermore, we can use the $\mathcal{DLP}$ solution to devise a dynamic pricing policy with a revenue guarantee of $\frac{1}{2}$ with respect to the optimal dynamic policy. This follows from the Proposition below established by (Ma et al. 2018):

PROPOSITION 2. *(Ma et al. 2018) Let r^* be the optimal objective value of $\mathcal{DLP}$. For each time period t, we define the price $\hat{T}_t$ such that:*

$$\hat{T}_t = T_{t,j_t} \text{ where } j_t \in \arg\max_{j \in \mathcal{L}} (T_{tj} - \frac{r^*}{2}) q_{tj}$$

Then the expected revenue of the dynamic pricing policy $(\hat{T}_t)_{t=1,\cdots,H}$ earns an expected revenue of at least $OPT_{LP}/2$.

Remark. For our setting, we approximate the true (uknown) q_{tj} with the estimated $\hat{q}_{tj}$ from our classification model. Theorem 1 ensures that $\hat{q}_{tj}$ converge to $\hat{q}_{tj}$ with the rate $\frac{1}{\sqrt{n}}$, n being the size of the data. Assuming that StubHub obtains sales data at a rate proportional to the time passed ($n \propto t$), then the estimated $\hat{q}_{tj}$ converges at the rate $\frac{1}{\sqrt{t}}$ to the true value. Thus, by linearity, the maximum error of using $\hat{q}_{tj}$ rather than q_{tj} in $\mathcal{DLP}$ is proportional to $\frac{1}{\sqrt{t}}$. This allows StubHub to utilize $\mathcal{DLP}$ efficiently with sufficient time (data) to construct the dynamic pricing strategy in Proposition 2.

One can then calculate the expected revenue of the ticket that this pricing strategy achieves:

$$\hat{R}^*_H(X) = \mathbb{E}\left[\sum_{t=1}^{H} \hat{T}_t \cdot W_t(\hat{T}_t, X, M_t)\right], \tag{23}$$

where the expectation is taken over the sample paths of the (ticket dependent) market dependent features M_t (for example, the quantity of tickets sold in the game). Since the expectation needs to be estimated on all sample paths which are not observed through data, we can use Monte Carlo simulation to generate these sample paths using the following evolution equation:

$$M_t = M_0 + \mu t + \sigma t \epsilon.$$

μ and σ denote the parameters fitted through historical data for the ticket, M_0 indicates the market feature at time of listing and ϵ is a standard $(0, 1)$ normally distributed variable. Therefore, each market feature is assumed to follow a Gaussian random walk with drift, as is commonly the case in reproducing dynamic data (see e.g. Bell et al. 2003).

4.4. Global optimization for purchasing a ticket portfolio

We introduce the following optimization problem to decide which tickets to purchase and add to our portfolio. In practice, this optimization problem is resolved frequently over the time horizon to respond to market changes.

$$\max_{z_i} \sum_{s \in S} \sum_{i \in A_s} (\hat{R}^*_H(X_i) - T^0_i - b) z_i \tag{24}$$

$$\text{s.t.} \sum_{i \in A_s} z_i + |I_s| \leq C_s \qquad \forall s \in S \tag{25}$$

$$z_i \in \{0, 1\}, \tag{26}$$

In this formulation, z_i is a binary variable that decides if each ticket i should be purchased, A_s is the set of available tickets on the market for a particular section of an event $s \in S$, and I_s is the number of tickets purchased in previous stages but were unsold by the algorithm for that section. The objective represents the optimal net expected revenue relative to the listed price T^0_i of the ticket. If the expected revenue sufficiently exceeds the listing price plus a buffer b, the ticket is considered for purchase. By tuning the buffer parameter b, we limit the rate at which tickets are purchased. Constraint (25) limits the number of tickets held in inventory for a given section. This is a knapsack problem that can be decomposed by section and solved efficiently in practice, once the price optimization and the expected revenue have been calculated for each ticket using either (16) or (23). As for the tickets previously purchased through the algorithm but that are unsold, we re-optimize their listing prices based on the updated market features and the time until the event.

5. Testing the method: experiments on synthetic data

In this section, we evaluate the performance of the estimation procedure from section §3.1 on a variety of synthetic datasets. In fact, it is important to use synthetic datasets first to establish the validity of our method since in real-world datasets, the parameters of interest (for example price sensitivity) are seldom known. Furthermore, real-world datasets rarely have counterfactuals, which are the outcomes that would occur if a different treatment is taken (in our example, whether a ticket sell if a different price is prescribed). Without this, the

evaluation of the estimation of treatment effects is difficult. However, in a synthetic data environment, these can be generated due to knowledge of the underlying probability model used to create the data. We begin with evaluating our model's ability to estimate a constant treatment effect, then we progress to the case of heterogeneous treatment effect using gradient boosted trees.

5.1. Homogeneous treatment effect simulations

We study a case of the model explained in equations (1) to (5) where the functional form of the nuisance parameters is linear:

$$Y^* = \beta_g' X + \tau T + \epsilon, \quad T = \beta_m' X + \beta Z + \upsilon, \quad Z = \beta_h + u, \begin{pmatrix} \epsilon \\ \upsilon \end{pmatrix} \sim N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{pmatrix} 1 & \rho\sigma_\upsilon \\ \rho\sigma_\upsilon & \sigma_\upsilon^2 \end{pmatrix}\right)$$

The endogeneity is captured by the correlation between ϵ and υ, where $\tau = 2, \rho = 0.5,\ \sigma_\upsilon = 1,\ u \sim N(0,1), X \sim N(0, I_d)$. To prove our methodology works on a wide range of simulated datasets, we generate 100 datasets with $n = 1000$ data-points each. Furthermore, the coefficients for each dataset are drawn from the following distributions: $\beta_h \sim N(0,1),\ \beta_g, \beta_m \sim N(0, I_d)$ but only have 5 non-zero components. This aims to replicate a realistic sparse scenario, where the signal depends only on a few of the observed variables, and the others are noise.

5.1.1. Benchmark methods: We test against the seminal work by Rivers and Vuong (1988) from the literature on probit models with instrumental variables. It provides an estimation procedure based on the decomposition of $\epsilon = \frac{\rho}{\sigma_\upsilon}\upsilon + \tilde{\epsilon}$ into independent components, where $E[\upsilon\tilde{\epsilon}] = 0$ and:

$$Y^* = \beta_g' X + \tau T + \frac{\rho}{\sigma_\upsilon}\upsilon + \tilde{\epsilon}$$

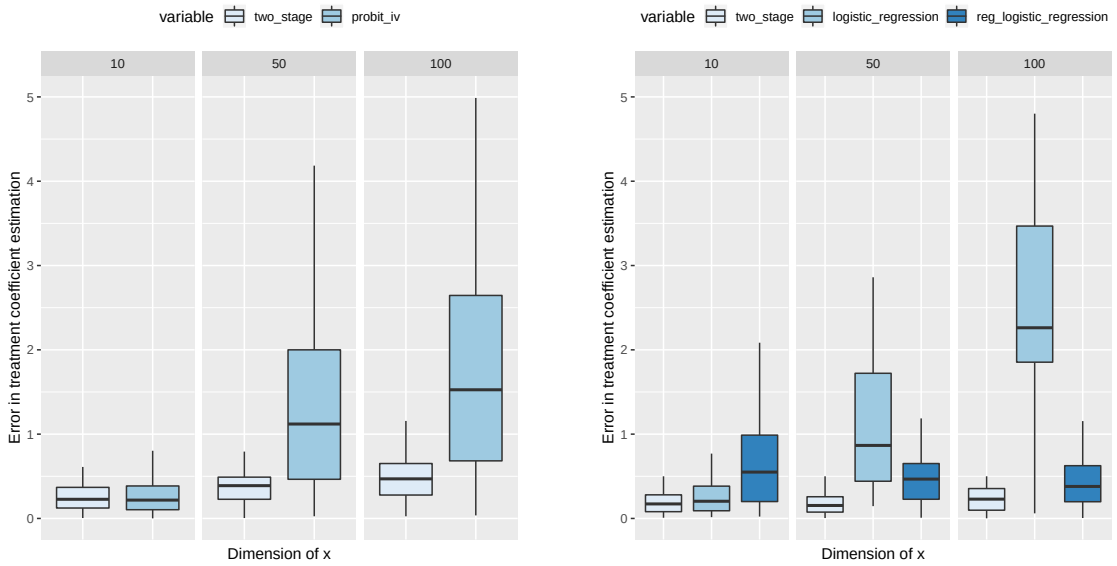
Rivers and Vuong (1988) proposed the following two stage procedure (`probit_iv`):

1. Regress T on X and Z to find $\hat{\upsilon} = T - (X'\hat{\beta_m} + \hat{\beta}Z)$, and $\hat{\sigma}_\upsilon = \frac{||\hat{\upsilon}||}{\sqrt{n}}$.
2. Probit regression of Y on $X, Z, \hat{\upsilon}$ to obtain estimates of $\hat{\beta}_g, \hat{\tau}, \hat{\lambda} = \frac{\hat{\rho}}{\sigma_\upsilon}$.

This method is different from our proposed approach as it does not have the orthogonalization step to reduce sensitivity to the nuisance parameters.

5.1.2. Results: Figure 1a shows how the error in the estimated treatment coefficient changes with the dimension of the data. We used `probit_iv` and the proposed two stage method (`two_stage`) to estimate the treatment coefficient for each dataset. To optimize the

second stage loss function (13), we only require a simple line search to find the correct treatment coefficient. For the first stage, $\hat{r}(x) \approx E[Y|X]$ is estimated using probit regression while $\hat{q}(x) \approx E[T|X]$ and $\hat{h}(x) \approx E[Z|X]$ are estimated using OLS regression. $\hat{\sigma}_u$ and $\hat{\sigma}_v$ are unknown and estimated to be on average 1.014 (standard deviation 0.036) and 0.974 (standard deviation 0.043) respectively, using the bootstrap procedure outlined in appendix C. Figure 1a shows the `probit_iv` estimator is unable to accurately estimate the treatment effect when the dimension is high. In this setting, it is beneficial to do the orthogonalization step, which is why `two_stage` is able to more accurately estimate the treatment effect in a larger dimension.



(a) With endogeneity (b) Without endogeneity

Figure 1: Homogeneous treatment effect estimate accuracy ($n = 1000$)

5.1.3. Experiments with no endogeneity: We also study a case with no endogeneity or instrumental variables ($\rho = 0,\ Z = 0$) and show that we are still able to achieve an improvement relative to traditional single stage logistic regression (`logistic_regression`) and logistic regression with elastic-net penalty (`reg_logistic_regression`, Zou and Hastie 2005), implemented using the `Glmnet` package (Friedman et al. 2009), with 10 fold cross-validation). In Figure 1b, the logistic regression is able to achieve relatively high accuracy when the number of samples n is large relative to d (the dimension of X), but performs poorly as d grows. As expected, the elastic-net logistic regression is less sensitive to d, but has a bias in the treatment coefficient's estimation across all datasets, which results in poor performance

relative to the two stage estimator. This highlights that the benefits of the two stage approach extend beyond the scenario with endogeneity, and that an orthogonalization step is useful here as well.

5.2. Heterogeneous treatment effect: gradient boosted trees

We now move to the more general case where the treatment effect depends on the covariates X. We explore this using gradient boosted tree methods. To do so, we implement our customized loss function within the `lightgbm` package for boosted trees. For the details of the implementation, we refer the reader to appendix G.

5.2.1. Numerical experiments with heterogeneous treatment effect: We use the model in (1) - (5), with $u \sim N(0,1)$, $X_i \sim N(0, I_d)$, $h(x) = x_1$, $\beta = 1$, $\rho = 0.5$, $\sigma_v = 1$.

We explore a number of different synthetic datasets:

- Dataset 1: simple treatment effect $(d = 2)$, $g(x) = x_1 + x_2$, $m(x) = x_1$, $\tau(x) = x_1$.
- Dataset 2: linear relationships $(d = 10)$, $g(x) = x_1 - x_2 + x_3$, $m(x) = x_1 + x_2 + x_3$, $\tau(x) = x_1 - x_2 + x_3$.
- Dataset 3: non-linear $g(x)$, linear $m(x), \tau(x)$ $(d = 2)$, $g(x) = \sin(10(x_1 + x_2)^2))$, $m(x) = x_1$, $\tau(x) = x_1$.
- Dataset 4: non-linear relationships $(d = 2)$, $g(x) = \sin(10(x_1 + x_2)^2))$, $m(x) = (x_1 - x_2)^2$, $\tau(x) = \cos(x_1 + x_2)$.

We test our different methods on the ability to correctly predict the change in probability associated with a change in the treatment. Specifically, for a particular change in treatment δ_k and a probability estimator $\Phi(\hat{f}(X, Z, T))$, we define the *average error in estimating the treatment change* as:

$$\frac{1}{n}\sum_{i=1}^{n} |(\Phi(\hat{f}(X_i, Z_i, T_i + \delta_k)) - \Phi(\hat{f}(X_i, Z_i, T_i)) - (E[Y|X = X_i, Z = Z_i, T = T_i + \delta_k] - E[Y|X = X_i, Z = Z_i, T = T_i])|$$

In other terms, we first look at the change in the predicted probability when we change each treatment by a specific amount δ_k, while holding the control features and instrument (X_i, Z_i) constant. We then take the mean absolute difference with the true change in probability associated with the known probability distribution. We display this metric in the Y axis in Figure 3. We repeat this over a range of treatment changes, δ_k, from -2 to 2 in 0.2 increments, using a randomly generated sample of 1,000,000 data points.

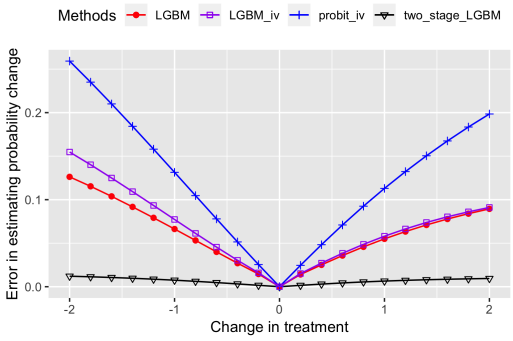
5.2.2. Benchmark methods: In addition to comparing against `probit_iv` which is limited by its linear form, we also compare against Blundell and Powell (2003). This extends Rivers and Vuong (1988) so that $g(X)$ and $h(X)$ are non-parametric functions, by replacing the steps in the procedure by the non-parametric equivalent (`LGBM_iv`):

1. Estimate $\hat{q}(X,Z) = \hat{m}(x) + \hat{\beta}Z$ non-parametrically. Find $\hat{v} = T - \hat{q}(X,Z)$, and $\hat{\sigma}_v = \frac{||\hat{v}||}{\sqrt{n}}$.
2. Estimate $\hat{f}(X,Z,T,\hat{v}) = g(X) + \tau(X)T + \frac{\rho}{\hat{\sigma}_v}\hat{v}$ non-parametrically.

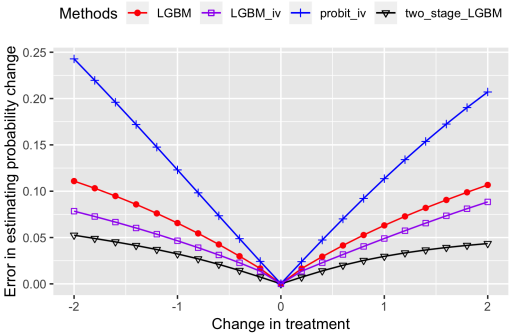
We note that the main difference with (`two_stage_LGBM`) is that the `LGBM_iv` approach above does not have an orthogonalization step, and does not identify the treatment effect $\tau(X)$, although, as in section §5.2, this can be approximated by changing T in the function $\hat{f}(X,Z,T,\hat{v})$. Finally, we compare against a `lightgbm` classifier with X,T as explanatory variables, a method denoted `LGBM`. 50 rounds of boosting were used for all procedures.

5.2.3. Results: Figure 1 shows that `two_stage_LGBM` performs well over a range of synthetic datasets with endogeneity as long as a suitable instrument is available. The `probit_iv` method performs poorly on all models where the nuisance parameters are non-linear or the treatment effect is heterogeneous. The orthogonalization step appears to give `two_stage_LGBM` an advantage over `LGBM_iv` in almost all cases. `LGBM` is generally inferior to `LGBM_iv` due to its inability to leverage the instrumental variable to remove the effects of endogeneity. The noise parameters are estimated as part of the estimation procedure. Using the bootstrap procedure for dataset (1)-(4), the estimates for $\hat{\sigma_u}$ are $1.00110, 1.00008, 1.00017, 1.00026$ respectively, and for $\hat{\sigma_v}$ they are $1.00017, 1.00441, 0.99960, 1.03516$. Further comparisons with the naive estimator are available in appendix N.

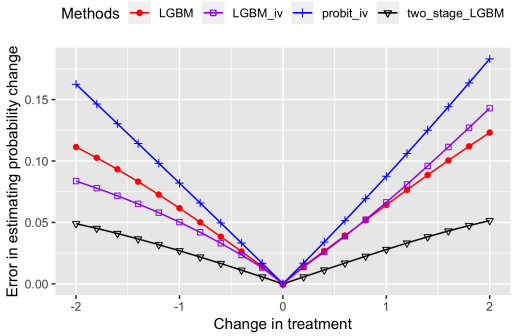
5.2.4. Experiments with no endogeneity: We also test our approach when there is no endogeneity ($\rho = 0$) and compare to `LGBM_logit` and `LGBM_probit` using boosted trees with a logistic or probit loss function respectively. Due to the lack of existing methods on estimating heterogeneous treatment effects in the classification setting with continuous treatments, we also tried some state of the art methods from the literature meant for the regression setting, such as generalized random forests (Athey et al. 2016) and `rlearner` (Nie and Wager 2017). These methods were not insightful in the classification setting so are omitted from the following results. The results are shown in Figure 3. We observe that our two stage approach is able to consistently better estimate the treatment effect than single stage gradient boosted trees across all data sets in this setting.



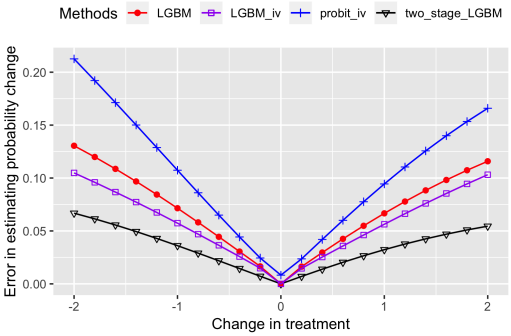
(a) Dataset 1: simple variable effect treatment



(b) Dataset 2: linear relationships



(c) Dataset 3: non-linear $g(x)$, linear $m(x), \tau(x)$



(d) Dataset 4: non-linear relationships

Figure 2: Estimating treatment effect for changing sample size

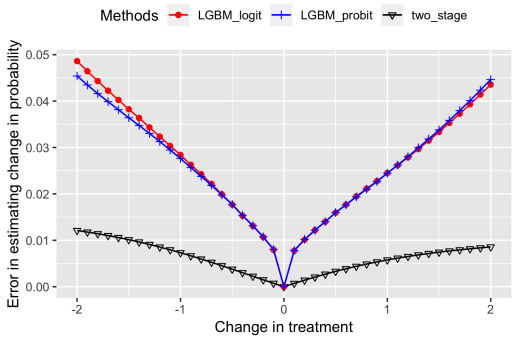
We have additional numerical experiments in appendix L.1 and appendix L.2. We also explore the effect of sample size on the ability to predict the change in probability in appendix L.1. In appendix L.2, we explore how the treatment effect estimation impacts the (sub)optimality of the solution in synthetic pricing simulations.

6. StubHub Case Study

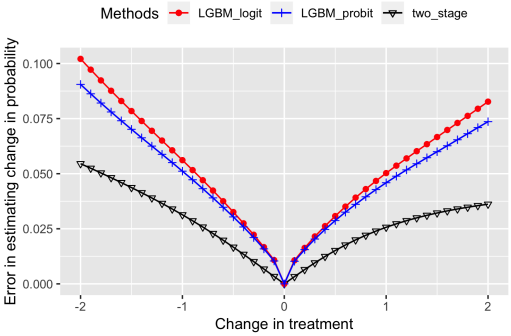
We investigate how our method performs in practice using StubHub data from the 2014-2015 NBA season. We first study the performance of the treatment effect estimation procedure portrayed in section §3, then the optimization problem for ticket reselling from section §4. We finally discuss the possible managerial implications for StubHub.

6.1. Introduction

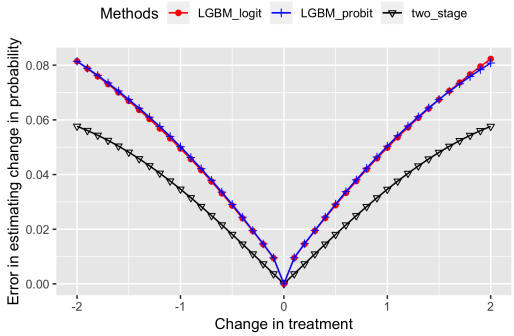
To test the performance of our algorithm, we construct a back-testing environment on the StubHub market for the 2014-2015 NBA season. In total, there are 1230 games and StubHub have 8.5 million tickets. We use 34 variables from the database as predictive features X (the



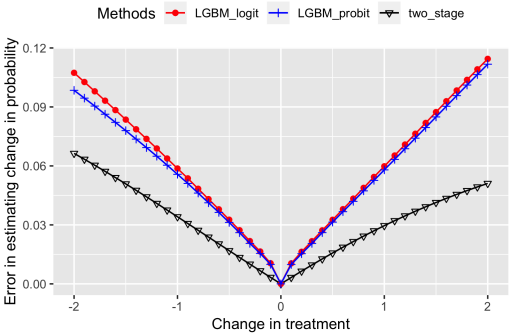
(a) Dataset 1: simple variable effect treatment



(b) Dataset 2: linear relationships



(c) Dataset 3: non-linear $g(x)$, linear $m(x), \tau(x)$



(d) Dataset 4: non-linear relationships

Figure 3: Estimating change in probability with a change in treatment

full table of variables is in appendix F). The outcome variable Y is the binary indicator if the ticket was eventually sold, and T is the listed price for that ticket.

We construct the instrumental variable Z from how the seller has historically priced tickets relative to the market. Specifically, Z is defined as the average of the difference between the seller listed price and the median listed price (in the section of the ticket) across all historical tickets sold by the seller in the NBA market prior to the 2014-2015 season. If the seller has not listed a ticket before, we let the value be 0. This is a suitable instrument because buyers cannot view the seller's history when buying a ticket, so the only mechanism where Z can affect the probability of a ticket being sold is through the price T. For benchmarks, we use the same methods listed in section §5.1.

6.2. Verifying the Instrumental Variable

We conduct experiments to show that Z is a valid and significant instrumental variable. First, we show that Z is significant in predicting the price variable T. Table 2a displays the coefficient

of the instrument in the regression of ticket and market features and the instrument on price. A Wald test of significance, with and without the IV, gives a F-value of 2481.8 and p-value of $< 2.2 \times 10^{-16}$. Typically the F-value for the joint significance of the instruments should be greater than 10 (Stock and Watson 2015). We also conducted a likelihood ratio test, obtaining an F-value of 2477.802 and p-value of $< 2.2 \times 10^{-16}$.

(a) Regression of ticket and market features and the instrument (the average historical price difference on price)

Variable	Estimate
Average historical price difference	0.009499*** (0.0001907)
Observations	698286
R^2	0.6508
Adjusted R^2	0.6508
Note:	*p<0.1; **p<0.05; ***p<0.01

(b) Probit regression of price, ticket and market features and price residual (from IV model) on selling outcomes

Variable	Estimate
Price residual	0.002902*** (0.0004329)
Observations	698286
AUC	0.8005
Misclassifcation rate	0.1942
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Analysis of Instrumental Variable

To further justify the use of an instrumental variable approach, we demonstrate that endogeneity exists in the data. We use the test from Guilkey et al. (1992), which incorporates the residual from the price prediction into the probit estimation for ticket selling likelihood. Table 2b shows strong evidence against the coefficient of the residual being zero (p-val$= 2.04 \times 10^{-11}$), suggesting there is endogeneity in the model.

6.3. Predictive accuracy

We split the historical ticket data into 70% training and 30% testing to evaluate the predictive performance of the various methods. The testing set is selected so that all tickets in such set are listed later than the training set to ensure no temporal dependence.

In Table 3, we observe the predictive power of the various models on the NBA dataset for 2014-2015[3]. We observe comparable performance in both misclassification error and AUC across all methods that use gradient boosted trees, with the `two_stage_LGBM` method achieving the highest accuracy. The corresponding ROC curves are shown in Figure 4. We also show in appendix O.1 that it is well calibrated.

[3] There are a number of parameters which needs to be estaimed, including: $\hat{\sigma_u} = 58.6$, $\hat{\rho} = -0.32$ and $\hat{\sigma_v} = 0.38$. The small size of $\hat{\sigma_v}$, is explained by that better results are obtained if $log(price)$ is used.

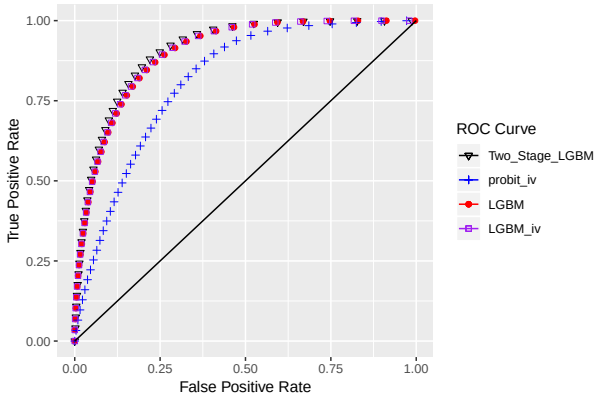


Figure 4: Performance on NBA data with iv: ROC curve

Algorithm	Misclassification error	AUC
two_stage_LGBM	0.135	0.908
LGBM	0.144	0.898
probit_iv	0.190	0.795
LGBM_iv	0.142	0.901

Table 3: Predictive accuracy on NBA data

We note that parametric models which do not capture the interaction terms or heterogeneous treatment effects, such as `probit_iv`, are considerably less accurate and possibly misspecified. This is particularly relevant as previous econometric studies for ticket selling have used such models (Sweeting 2012), which may lead to erroneous treatment effect estimation. Overall, this high predictive accuracy is promising for real-world applications. We also show the performance of a simpler model with fewer features in appendix I.

6.4. Back-testing optimization

To understand the performance of our trading algorithm in buying/reselling tickets, we leverage data from 2014-2015 NBA season to conduct backtesting experiments. The backtesting framework consists of two models:

- A purchasing model that uses one of the pricing models above to optimally price tickets and then decide whether it is favorable to purchase and resell at a higher price.
- An independent baseline model simulating the ground truth of if tickets will sell.

Additional technical details about this backtesting procedure are given in appendix M. We further construct a RShiny application to live monitor the backtesting procedure with customizable features. A screenshot of the application is shown in appendix K.

For the experiment, we select the `LGBM_iv` model (Blundell and Powell 2003) as the baseline model for its high accuracy in section §6.3. We divide the dataset into two equal-sized sets, with the training set chronologically before the testing set. We train the baseline model on the training dataset while we evaluate our models on the testing dataset, assuming the baseline to be the ground truth. We restrict the trading algorithm to only purchase 50 tickets per day, due to the computational burden of both the dynamic pricing optimization and the

simulation of selling probabilities. However, we foresee that this could be scaled significantly with parallelization. The results are shown in Table 4:

Optimization Model Method	Pricing Strategy	Total Profit	% of Bought Tickets Sold	ROI
two_stage_LGBM	Constant	13.7K	58%	7%
	Time varying	19.5K	64%	11%
probit_iv	Constant	-247K	22%	-70%
	Time varying	-116K	38%	-29%

Table 4: Comparison of backtesting profit between methods with LGBM_iv baseline (limited to purchasing 50 tickets a day)

Table 4 shows that two_stage_LGBM performed significantly better than probit_iv, a specification previously used in the ticket resale market. A dynamic pricing strategy based on probit_iv's probabilities loses 29% of capital invested, while the two_stage_LGBM approach achieves a positive 11% return on investment, as evaluated by an independent baseline model. probit_iv underestimates the price sensitivity and as a result just sells 38% of the tickets it purchases, while two_stage_LGBM sells 64% of the tickets. Furthermore, we observe that this misestimation is worse for high value tickets, which have a significant impact on ROI. This highlights that in an industry with slim margins, even slight misspecification of the demand model can result in significant losses. We observe a small but significant improvement from 7% to 11% ROI when using a time-varying pricing policy relative to a constant policy, suggesting there is improvement to be made by considering the perishable inventory effects of tickets.

7. Conclusions and Discussions

Despite the availability of large datasets on the behaviour of buyers and sellers, estimating causal relationship between price and probability of selling for tickets on the resale market is challenging. We introduce a novel loss function for heterogeneous treatment effect estimation with binary outcomes which can be easily incorporated into off-the-shelf machine learning algorithms, including neural networks and gradient boosted trees. We prove this loss function is consistent under mild assumptions, and establish the rate of convergence. In our numerical experiments on a wide range of synthetic data sets, the two stage approach is able to consistently outperform the single stage counterpart and methods from causal inference in terms of estimation of treatment effect. On historical data, we show that our approach is price sensitive with potential to earn up to an 11% return by trading on tickets on the platform. This is

significant as probit models with instrumental variables previously used for estimating price sensitivity in this setting are not profitable. We hope that these results increase awareness of the importance of accurately estimating the casual relationship between price and sales in the revenue management and pricing community.

The results for heterogeneous treatment effect estimation for classification are widely applicable beyond the scope of ticket reselling. We believe they could be useful in areas such as medicine, revenue management and other econometric applications.

Finally, we discuss the effect of StubHub's strategy on the efficiency of the market. The primary objective of StubHub is to utilize the two stage method to serve as a market maker on its platform by buying under-priced tickets, and recommending sellers to sell around equilibrium. There is evidence in the literature (e.g. Logue 1975) that market makers provide a vital information channel and could increase market efficiency. The same literature also points out that if the market maker takes a monopolistic position, then it could also decrease the efficiency in the market by affecting the market equilibrium.

Currently, there are already many market makers on StubHub's ticket reselling platform (by buying lower priced tickets and repricing them). In addition, to the best of our knowledge, StubHub has no plans to be the sole market maker. Therefore, one expect StubHub's strategy would increase market efficiency, though the final impact may be limited due to StubHub's current self-imposed limit of trading less than 1% of the market.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. (2016) Tensorflow: a system for large-scale machine learning. *OSDI*, volume 16, 265–283.

Amemiya T (1978) The estimation of a simultaneous equation generalized probit model. *Econometrica: Journal of the Econometric Society* 1193–1205.

Athey S, Tibshirani J, Wager S (2016) Generalized random forests. *arXiv preprint arXiv:1610.01271* .

Ban GY, Keskin NB (2017) Personalized dynamic pricing with machine learning .

Barlow GL (2000) Capacity management in the football industry. *Yield management: Strategies for the service industries* 303–314.

Bell WH, Cameron DG, Millar AP, Capozza L, Stockinger K, Zini F (2003) Optorsim: A grid simulator for studying dynamic data replication strategies. *The International Journal of High Performance Computing Applications* 17(4):403–416.

Bertsekas DP, Tsitsiklis JN (2002) *Introduction to probability*, volume 1 (Athena Scientific Belmont, MA).

Bertsimas D, Kallus N (2016) The power and limits of predictive approaches to observational-data-driven optimization. *arXiv preprint arXiv:1605.02347* .

Biau G (2012) Analysis of a random forests model. *Journal of Machine Learning Research* 13(Apr):1063–1095.

Blundell R, Powell JL (2003) Endogeneity in nonparametric and semiparametric regression models. *Econometric society monographs* 36:312–357.

Blundell RW, Powell JL (2004) Endogeneity in semiparametric binary response models. *The Review of Economic Studies* 71(3):655–679.

Bouchet A, Troilo M, Walkup BR (2016) Dynamic pricing usage in sports for revenue management. *Managerial Finance* 42(9):913–921, URL http://dx.doi.org/10.1108/MF-01-2016-0017.

Brent RP (1971) An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal* 14(4):422–425.

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*, 161–168 (ACM).

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1):C1–C68.

Chernozhukov V, Goldman M, Semenova V, Taddy M (2017) Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988* .

Cohen M, Lobel I, Paes Leme R (2016) Feature-based dynamic pricing .

Diehl MA, Maxcy JG, Drayer J (2015) Price elasticity of demand in the secondary market: Evidence from the national football league. *Journal of Sports Economics* 16(6):557–575.

Duran S, Swann JL, Yakıcı E (2012) Dynamic switching times from season to single tickets in sports and entertainment. *Optimization Letters* 6(6):1185–1206.

Engle RF (1994) *Handbook of Econometrics, Volume 4* (North Holland), ISBN 0444887660.

Erskine J (2015) Stubhub announces hottest summer concerts with taylor swift at #1. https://www.businesswire.com/news/home/20150528005481/en/StubHub-Announces-Hottest-Summer-Concerts-Taylor-Swift#.VdRNgpfEkTA, accessed: 2018-08-28.

Friedman J, Hastie T, Tibshirani R (2009) glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* 1(4).

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Gallego G, Iyengar G, Phillips R, Dubey A (2004) Managing flexible products on a network URL http://dx.doi.org/10.13140/2.1.2833.0560.

Gallego G, Van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science* 40(8):999–1020.

Guilkey DK, Mroz TA, Taylor L (1992) Estimation and testing in simultaneous equations models with discrete outcomes using cross section data. *Unpublished manuscript* .

Heckman JJ (1977) Dummy endogenous variables in a simultaneous equation system.

Imbens G, Wooldridge J (2007) Control function and related methods. *What's new in Econometrics* .

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press), URL `http://dx.doi.org/10.1017/CBO9781139025751`.

Javanmard A, Nazerzadeh H (2016) Dynamic pricing in high-dimensions. *arXiv preprint arXiv:1609.07574* .

Jiaqi Xu J, Fader PS, Veeraraghavan S (2019) Designing and evaluating dynamic pricing policies for major league baseball tickets. *Manufacturing & Service Operations Management* .

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. *NIPS*.

Kemper C, Breuer C (2016) How efficient is dynamic pricing for sport events? designing a dynamic pricing model for bayern munich. *International Journal of Sport Finance* 11(1).

Lee LF (1981) Simultaneous equations models with discrete and censored dependent variables. *Structural analysis of discrete data with econometric applications* 346.

Logue DE (1975) Market-making and the assessment of market efficiency. *The Journal of Finance* 30(1):115–123.

Ma W, Simchi-Levi D, Zhao J (2018) Dynamic pricing under a static calendar. *Available SSRN 3251015* .

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Nie X, Wager S (2017) Learning objectives for treatment effect estimation. *preprint arXiv:1712.04912* .

Oprescu M, Syrgkanis V, Wu ZS (2018) Orthogonal random forest for heterogeneous treatment effect estimation. *arXiv preprint arXiv:1806.03467* .

Reiersøl O (1945) *Confluence analysis by means of instrumental sets of variables*. Ph.D. thesis, Almqvist & Wiksell.

Rivers D, Vuong QH (1988) Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics* 39(3):347–366.

Robinson PM (1988) Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 931–954.

Rogozhnikov A, Likhomanenko T (2017) Infiniteboost: building infinite ensembles with gradient descent. *arXiv preprint arXiv:1706.01109* .

Sainam P, Balasubramanian S, Bayus BL (2010) Consumer options: Theory and an empirical application to a sports market. *Journal of Marketing Research* 47(3):401–414.

Stewart GW (1990) Matrix perturbation theory .

Stock JH, Watson MW (2015) *Introduction to econometrics*.

Sweeting A (2012) Dynamic pricing behavior in perishable goods markets: Evidence from secondary markets for major league baseball tickets. *Journal of Political Economy* 120(6):1133–1172.

Wang X, Truong VA, Bank D (2018) Online advance admission scheduling for services with customer preferences. *arXiv preprint arXiv:1805.10412* .

Wright PG (1928) *Tariff on animal and vegetable oils* (Macmillan Company, New York).

Zhang T, Yu B, et al. (2005) Boosting with early stopping: Convergence and consistency. *The Annals of Statistics* 33(4):1538–1579.

Zhu JD (2014) Effect of resale on optimal ticket pricing: Evidence from major league baseball tickets. Technical report, Working paper, Texas A&M University.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.

Appendix A: Derivation of loss function with endogeneity

We begin by noting that $\mathbb{E}[Z|X] = h(X),\ \mathbb{E}[T|X] = m(X) + \beta h(X)$ and

$$\mathbb{E}[Y^*|X] = g(X) + \tau(X)m(X) + \beta\tau(X)h(X).$$

Then we have:

$$\begin{aligned}
\mathbb{E}[Y|X] &= P\Big(Y^* > 0|X\Big) && (27)\\
&= P(g(X) + \tau(X)T + \epsilon > 0|X) && (28)\\
&= P(g(X) + \tau(X)(m(X) + \beta(h(X) + u) + \upsilon) + \epsilon > 0|X) && (29)\\
&= P(\beta\tau(X)u + \tau(X)\upsilon + \epsilon > -(g(X) + \tau(X)m(X) + \beta\tau(X)h(X))|X) && (30)\\
&= P(\beta\tau(X)u + \tau(X)\upsilon + \epsilon > -\mathbb{E}[Y^*|X]|X) && (31)
\end{aligned}$$

We know that

$$\mathbb{E}[\beta\tau(X)u + \tau(X)\upsilon + \epsilon|X] = \beta\tau(X)\mathbb{E}[u|X] + \tau(X)\mathbb{E}[\upsilon|X] + \mathbb{E}[\epsilon|X] = 0$$

and

$$\begin{aligned}
Var[\beta\tau(X)u + \tau(X)\upsilon + \epsilon|X] &= \beta^2\tau(X)^2 Var[u|X] + Var[\tau(X)\upsilon + \epsilon|X] + 2Cov(u, \tau(X)\upsilon + \epsilon|X) && (32)\\
&= \beta^2\tau(X)^2 Var[u|X] + \tau(X)^2 Var[\upsilon|X] + Var[\epsilon|X] + 2Cov(\upsilon, \epsilon|X) && (33)\\
&= \sigma_u^2\tau(X)^2\beta^2 + \tau(X)^2\sigma_\upsilon^2 + \sigma_\epsilon^2 + 2\rho\tau(X)\sigma_\epsilon\sigma_\upsilon && (34)
\end{aligned}$$

Due to normality, $\beta\tau(X)u + \tau(X)\upsilon + \epsilon|X \sim \mathcal{N}(0, \sigma_u^2\tau(X)^2\beta^2 + \tau(X)^2\sigma_\upsilon^2 + \sigma_\epsilon^2 + 2\rho\tau(X)\sigma_\epsilon\sigma_\upsilon)$. It follows:

$$\mathbb{E}[Y|X] = \Phi\left(\frac{\mathbb{E}[Y^*|X]}{\sqrt{\sigma_u^2\tau(X)^2\beta^2 + \sigma_\epsilon^2 + 2\tau(X)\sigma_\epsilon\sigma_\upsilon + \tau(X)^2\sigma_\upsilon^2}}\right) \tag{35}$$

where Φ is the cdf of standard normal variable. We denote Φ^{-1} the inverse of Φ, then we conclude that

$$\mathbb{E}[Y^*|X] = \sqrt{\sigma_u^2\tau(X)^2\beta^2 + \sigma_\epsilon^2 + 2\tau(X)\sigma_\epsilon\sigma_\upsilon + \tau(X)^2\sigma_\upsilon^2} \cdot \Phi^{-1}(\mathbb{E}[Y|X]). \tag{36}$$

Appendix B: Proof of Lemma 1

In what follows, we detail the proofs for the second part of the lemma: by Assumption 2, the estimators converge to the true values with a rate at least as fast as $n^{1/2+\delta}$. The first part of the lemma (the consistency of the estimators when Assumption 2 does not hold) follows immediately.

Note that in Algorithm 1, $\hat{\beta}$ is calculated by regressing $T - \hat{q}(X)$ against $Z - \hat{h}(X)$, while the true β can be calculated by regressing $T - q(X)$ against $Z - h(X)$. Therefore, we have formulas:

$$\begin{aligned}
\beta &= [(Z - h(X))^T(Z - h(X))]^{-1}(Z - h(X))^T(T - q(X))\\
\hat{\beta} &= [(Z - \hat{h}(X))^T(Z - \hat{h}(X))]^{-1}(Z - \hat{h}(X))^T(T - \hat{q}(X))
\end{aligned}$$

By Assumption 2, we have that $\sup_{X\in\mathcal{D}_x} n^{1/(2+\delta)}|\hat{q}(X) - q(X)| \xrightarrow{p} 0$, and $\sup_{X\in\mathcal{D}_x} n^{1/(2+\delta)}|\hat{h}(X) - h(X)| \xrightarrow{p} 0$. Therefore, using the continuous mapping theorem, we have:

$$\sup_{X\in\mathcal{D}_x} n^{1/(2+\delta)}|(Z - \hat{h}(X))^T(T - \hat{q}(X)) - (Z - h(X))^T(T - q(X))| \xrightarrow{p} 0 \tag{37}$$

Matrix Perturbation Theory (see e.g. Stewart 1990) gives rise to:

$$\begin{aligned}&\|[(Z-\hat{h}(X))^T(Z-\hat{h}(X))]^{-1}-[(Z-h(X))^T(Z-h(X))]^{-1}\|\\&\leq\underbrace{\|(Z-\hat{h}(X))^T(Z-\hat{h}(X))-(Z-h(X))^T(Z-h(X))\|}_{O(\frac{1}{n^{1/(2+\delta)}})\text{ by Assumption 2}}\times\\&\qquad\|[(Z-\hat{h}(X))^T(Z-\hat{h}(X))]^{-1}\|\|[(Z-h(X))^T(Z-h(X))]^{-1}\|\end{aligned}$$

Therefore, as long as we assume that $\|[(Z-\hat{h}(X))^T(Z-\hat{h}(X))]^{-1}\|$ and $\|[(Z-h(X))^T(Z-h(X))]^{-1}\|$ are both absolutely bounded above by some constant (which is true as long as $\hat{\beta}$ and β exist), then it follows:

$$\sup_{X\in\mathcal{D}_x} n^{1/(2+\delta)}\|[(Z-\hat{h}(X))^T(Z-\hat{h}(X))]^{-1}-[(Z-h(X))^T(Z-h(X))]^{-1}\|\xrightarrow{p}0 \tag{38}$$

Combining the consistency results in Equation (37) and (38) gives rise to:

$$n^{1/(2+\delta)}\|\hat{\beta}-\beta\|\xrightarrow{p}0, \tag{39}$$

The residual is defined as $v=T-q(X)-(Z-h(X))\beta$, while our estimator $\hat{v}$ has the formula $\hat{v}=T-\hat{q}(X)-(Z-\hat{h}(X))\hat{\beta}$. Thus, we have:

$$\begin{aligned}\|v-\hat{v}\|&=\|T-q(X)-(Z-h(X))\beta-T-\hat{q}(X)-(Z-\hat{h}(X))\hat{\beta}\|\\&\leq\underbrace{\|q(X)-\hat{q}(X)\|+\|\beta\|\|h(X)-\hat{h}(X)\|}_{O(\frac{1}{n^{1/(2+\delta)}})\text{ by Assumption 2}}+\|Z-\hat{h}(X)\|\underbrace{\|\beta-\hat{\beta}\|}_{O(\frac{1}{n^{1/(2+\delta)}})\text{ by (39)}}\end{aligned}$$

This in turn implies $n^{1/(2+\delta)}\|\hat{v}-v\|\xrightarrow{p}0$.

Next, we focus on proving consistency results for the variance estimates $\hat{\sigma}_u$ and $\hat{\sigma}_v$. The estimator for σ_u is defined as: $\hat{\sigma}_u^2=\frac{1}{n}\sum_{i=1}^n(Z_i-\hat{h}(X_i))^2$ where n is the number of samples in the dataset. Then:

$$\mathbb{E}_{u,X}\left[\hat{\sigma}_u^2\right]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{u,X}\left[(Z_i-\hat{h}(X_i))^2\right] \tag{40}$$

$$=\frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}_{u,X}\left[(Z_i-h(X_i))^2\right]+\mathbb{E}_{u,X}\left[(\hat{h}(X_i)-h(X_i))^2\right]+2\mathbb{E}_{u,X}\left[u_i(\hat{h}(X_i)-h(X_i))\right]\right) \tag{41}$$

Note that by construction of Algorithm 1, $\hat{h}(X_i)$ is independent of u_i due to $\hat{h}(X)$ being estimated using a separate dataset, and thus we have:

$$\mathbb{E}_{u,X}\left[\hat{\sigma}_u^2\right]=\sigma_u^2+\frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}_{u,X}\left[(\hat{h}(X_i)-h(X_i))^2\right]+2\mathbb{E}_u[u]\mathbb{E}[(\hat{h}(X_i)-h(X_i))]\right) \tag{42}$$

$$=\sigma_u^2+\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{u,X}\left[(\hat{h}(X_i)-h(X_i))^2\right]=\sigma_u^2+O\left(\frac{1}{n^{2/(2+\delta)}}\right) \tag{43}$$

Where the final equality is by Assumption 2. Similarly, the estimator for σ_v is defined as:

$$\hat{\sigma}_v^2=\frac{1}{n}\sum_{i=1}^n(T_i-\hat{q}(X_i)-\hat{\beta}_i^T(Z_i-\hat{h}(X_i)))^2.$$

We have:

$$\mathbb{E}_{u,v,X}[\hat{\sigma}_v^2]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{u,v,X}\left[(T_i-\hat{q}(X_i)-\hat{\beta}_i^T(Z_i-\hat{h}(X_i)))^2\right] \tag{44}$$

$$=\frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}_{u,X}\left[(T_i-m(X_i)-\beta_i^TZ_i)^2\right]+\mathbb{E}_{u,v,X}\left[(\hat{q}(X_i)-q(X_i))^2\right]\right. \tag{45}$$

$$+\mathbb{E}_{u,v,X}\left[(\hat{\beta}_i^T(Z_i-\hat{h}(X_i))-\beta_i^T(Z_i-h(X_i)))^2\right]+2\mathbb{E}_{u,X}[\mathrm{Cov}_v(v_i,\hat{q}(X_i))] \tag{46}$$

$$+2\mathbb{E}_{u,X}[\mathrm{Cov}_v(v_i,\hat{\beta}_i^T(Z_i-\hat{h}(X_i)))] \tag{47}$$

$$\left.+2\mathbb{E}_{u,v,X}[(\hat{q}(X_i)-q(X_i))(\hat{\beta}_i^T(Z_i-\hat{h}(X_i))-\beta_i^T(Z_i-h(X_i)))]\right) \tag{48}$$

By construction of Algorithm 1, $\hat{q}(X), \hat{\beta}, \hat{h}(X)$ are independent of v due to these estimators being estimated on a separate dataset. Furthermore, $\hat{q}(X)$ is independent from $\hat{\beta}$. Thus, we have:

$$= \sigma_v^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}_{u,v,X}\left[(\hat{q}(X_i) - q(X_i))^2\right] \tag{49}$$

$$+ \mathbb{E}_{u,v,X}\left[(\hat{\beta}_i^T(Z_i - \hat{h}(X_i)) - \beta_i^T(Z_i - h(X_i)))^2\right] \tag{50}$$

$$+ 2\mathbb{E}_{u,v,X}[(\hat{q}(X_i) - q(X_i))(\hat{\beta}_i^T(Z_i - \hat{h}(X_i)) - \beta_i^T(Z_i - h(X_i)))]\Big) \tag{51}$$

$$\leq \sigma_v^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}_{u,v,X}\left[(\hat{q}(X_i) - q(X_i))^2\right] \tag{52}$$

$$+ \mathbb{E}_{u,v,X}\left[(\hat{\beta}_i^T(Z_i - \hat{h}(X_i)) - \beta_i^T(Z_i - h(X_i)))^2\right] \tag{53}$$

$$+ \mathbb{E}_{u,v,X}[(\hat{q}(X_i) - q(X_i))^2] + \mathbb{E}_{u,v,X}[(\hat{\beta}_i^T(Z_i - \hat{h}(X_i)) - \beta_i^T(Z_i - h(X_i))^2]\Big) \tag{54}$$

By Assumption 2, it follows that:

$$= \sigma_v^2 + O\left(\frac{1}{n^{2/(2+\delta)}}\right) \quad Q.E.D. \tag{55}$$

Appendix C: Variance Estimates Using the Bootstrap Method

1. Split up the data to sets S_1, S_2, and S_3.
2. Sample S_1 to create b bootstrap samples $\{S_1^{(k)}\}_{k=1}^b$. Use these samples to estimate
$$\hat{h}^{(k)}(X) = \arg\min_h \frac{1}{|S_1^{(k)}|}\sum_{i\in S_1^{(k)}} L(h(x_i), z_i)$$
$$\hat{q}^{(k)}(X) = \arg\min_q \frac{1}{|S_1^{(k)}|}\sum_{i\in S_1^{(k)}} L(q(x_i), t_i)$$
where L are binary cross entropy loss functions.
3. Sample S_2 to create b bootstrap samples $\{S_2^{(k)}\}_{k=1}^b$.Use these samples to estimate
$$\hat{\beta}^{(k)} = \arg\min_\beta \frac{1}{|S_2^{(k)}|}\sum_{i\in S_2^{(k)}} (t_i - \hat{q}^{(k)}(x_i) - \beta(z_i - \hat{h}^{(k)}(x_i)))^2.$$
4. Using S_3, estimate:
$$\hat{\mathbb{E}}[\hat{\mathrm{Var}}(\hat{h}(X))] = \frac{1}{|S_3|b}\sum_{i\in S_3}\sum_{k=1}^{b}(\hat{h}^{(k)}(x_i) - \frac{1}{b}\sum_{k=1}^{b}\hat{h}^{(k)}(x_i))^2$$
$$\hat{\mathbb{E}}[(Z - \hat{h}(X))^2] = \frac{1}{|S_3|b}\sum_{i\in S_3}\sum_{k=1}^{b}(z_i - \hat{h}^{(k)}(x_i))^2$$
$$\hat{\sigma}_u^2 = \hat{\mathbb{E}}[(Z - \hat{h}(X))^2] - \hat{\mathbb{E}}[\hat{\mathrm{Var}}(\hat{h}(X))]$$
$$\hat{\mathbb{E}}[\hat{\mathrm{Var}}(\hat{q}(X))] = \frac{1}{|S_3|b}\sum_{i\in S_3}\sum_{k=1}^{b}(\hat{q}^{(k)}(x_i) - \frac{1}{b}\sum_{k=1}^{b}\hat{q}^{(k)}(x_i))^2$$
$$\hat{\mathbb{E}}[\hat{\mathrm{Var}}(\hat{\beta}^T(Z - \hat{h}(X))] = \frac{1}{|S_3|b}\sum_{i\in S_3}\sum_{k=1}^{b}(\hat{\beta}^{(k)}(z_i - \hat{h}^{(k)}(x_i)) - \frac{1}{b}\sum_{k=1}^{b}\hat{\beta}^{(k)}(z_i - \hat{h}^{(k)}(x_i)))^2$$
$$\hat{\mathbb{E}}[(T - \hat{q}(X) - \hat{\beta}^T(Z - \hat{h}(X)))^2] = \frac{1}{|S_3|b}\sum_{i\in S_3}\sum_{k=1}^{b}(t_i - \hat{q}^{(k)}(x_i) - \hat{\beta}^{(k)}(z_i - \hat{h}^{(k)}(x_i)))^2$$
$$\hat{\sigma}_v = \hat{\mathbb{E}}[(T - \hat{q}(X) - \hat{\beta}^T(Z - \hat{h}(X)))^2] - \hat{\mathbb{E}}[\hat{\mathrm{Var}}(\hat{q}(X))] - \hat{\mathbb{E}}[\hat{\mathrm{Var}}(\hat{\beta}^T(Z - \hat{h}(X))]$$

Appendix D: Proof of Theorem 1

The proof that $(\boldsymbol{\theta}, \rho) = \boldsymbol{\Theta} \mapsto \mathbb{E}[l(Y, X, T, \boldsymbol{\Theta}, w, v)]$ has a unique global minimum follows from the properties of the log likelihood function (see Lemma 2.2 in Engle 1994). Define:

$$\tilde{\Theta} = (\tilde{\boldsymbol{\theta}}, \tilde{\rho}) = \underset{\boldsymbol{\Theta} \in \mathcal{D}_\Theta}{\arg\max} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(Y_i, X_i, T_i, \boldsymbol{\Theta}, w, v_i) \right\} \tag{56}$$

Note that $\tilde{\boldsymbol{\theta}}$ is the maximum likelihood estimator of θ. The proof is divided into three steps:

1. We verify the sufficient assumptions to leverage the asymptotic theory of maximum likelihood estimators, and thus prove that $\tilde{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\Theta}$.
2. We use the uniform convergence assumptions of $\hat{h}$, $\hat{q}$, and $\hat{r}$ along with the maximum likelihood estimator consistency above to prove that $\hat{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\Theta}$
3. Using the continuity and compactness assumptions, we show that $\hat{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\Theta}$ implies:

$$\sup_{X \in M} \|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}, X)\| \xrightarrow{p} 0 \quad \text{and} \quad \|\hat{\rho} - \rho\| \xrightarrow{p} 0$$

Proof of 1 There are 4 assumptions that once satisfied, are sufficient to use the asymptotic theory of maximum likelihood estimators (Engle 1994):

(a) The domain of $\boldsymbol{\Theta}$ is compact.

Proof: This follows directly from the assumption that $\boldsymbol{\theta} \in D_\Theta$, and D_Θ is compact.

(b) The function $l(Y, X, T, \boldsymbol{\Theta}, w, v)$ is continuous in $\boldsymbol{\Theta}$ for all tuples (X, T, Y, v).

Proof: As $\tau = t(\boldsymbol{\theta}, X)$ is continuous everywhere in $\boldsymbol{\theta}$ by assumption, the addition, composition of it with continuous functions stay continuous. Thus $l(Y, X, T, \boldsymbol{\Theta}, w, v)$ is continuous.

(c) There exists an integrable function $D(X, T, v)$ with respect to (X, T, v) such that $\forall \tau$:

$$|l(Y, X, T, \boldsymbol{\Theta}, w, v)| < D(X, T, v) \qquad \forall\, (X, T, Y, Z, \Theta) \in \mathcal{D}_X \times \mathcal{D}_T \times \mathcal{D}_Y \times \mathcal{D}_Z \times \mathcal{D}_\Theta$$

Proof: The domain of $\tau = t(\boldsymbol{\theta}, X)$ is compact (both $\mathcal{D}_\Theta$ and $\mathcal{D}_X$ are compact, so the product space is compact). Thus, there exists $(\boldsymbol{\theta}^*, X^*)$ such that:

$$|t(\boldsymbol{\theta}^*, X^*)| \geq |t(\boldsymbol{\theta}, X)| \qquad \forall\, (\boldsymbol{\theta}, X) \in D_\Theta \times \mathcal{D}_X$$

Then we have:

$$\begin{aligned} \tilde{f}(T, X, v) &= \frac{|\Phi^{-1}(r(X))| \sqrt{1 + 2|\rho\sigma_v| t(\boldsymbol{\theta}^*, X^*) + (\sigma_v{}^2 + \sigma_u^2 \beta^2) t(\boldsymbol{\theta}^*, X^*)^2} + |t(\boldsymbol{\theta}^*, X^*)||T - q(X)| + \frac{|\rho|}{\sigma_v}|v|}{\sqrt{1 - \rho^2}} \\ &\geq |f(T, X, v, \Theta)| \end{aligned}$$

for all $\boldsymbol{\Theta} \in D_\Theta$. As log is strictly increasing, we then have, for all $\boldsymbol{\theta} \in D_\Theta$:

$$D(X, T, v) = |\log(\Phi(-\tilde{f}(T, X, v)))| > |l(Y, X, T, \boldsymbol{\Theta}, w, v)|$$

(d) $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}_0 \quad \Leftrightarrow \quad l(\cdot, \boldsymbol{\Theta}) \neq l(\cdot, \boldsymbol{\Theta}_0)$.

Proof: The reverse direction is trivial. For the forward direction, we would show that there exists (Y,T,X,w,v) such that $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}_0 \quad \Rightarrow \quad l(Y,X,T,\boldsymbol{\theta},w,v) \neq l(Y,X,T,\boldsymbol{\Theta}_0,w,v)$. We first take $Y=1$, and thus the log-likelihood function becomes:

$$l(1,X,T,\boldsymbol{\Theta},w,v) = \log(\Phi(f(X,T,\boldsymbol{\theta},\rho,w,v)))$$

As log and Φ are injective functions, we only need to show that for all $(\boldsymbol{\theta},\rho) \neq (\boldsymbol{\theta}_0,\rho_0)$ there exists (T,X,w,v) such that $f(X,T,\boldsymbol{\theta},\rho,w,v) \neq f(X,T,\boldsymbol{\theta}_0,\rho_0,w,v)$.

First, we would consider the case where $\rho \neq \rho_0$. Then let $X \in \mathcal{D}_X$, $T \in \mathcal{D}_T$, $w \in \mathbb{R}$ and let:

$$v > \frac{\sigma_v}{\left|\frac{\rho}{\sqrt{1-\rho^2}} - \frac{\rho_0}{\sqrt{1-\rho_0^2}}\right|} \cdot \Big(\left| \frac{\Phi^{-1}(\mathbb{E}[Y|X])k(X,\boldsymbol{\theta},\rho,w) + t(\boldsymbol{\theta},X)(T-\mathbb{E}[T|X])}{\sqrt{1-\rho^2}} \right| \cdot \\ + \left| \frac{\Phi^{-1}(\mathbb{E}[Y|X])k(X,\boldsymbol{\theta}_0,\rho_0,w) + t(\boldsymbol{\theta}_0,X)(T-\mathbb{E}[T|X])}{\sqrt{1-\rho_0^2}} \right| \Big).$$

Then we have:

$$\begin{aligned} &|f(X,T,\boldsymbol{\theta},\rho,w,v) - f(X,T,\boldsymbol{\theta}_0,\rho_0,w,v)| \\ &\geq \frac{\left|\frac{\rho}{\sqrt{1-\rho^2}} - \frac{\rho_0}{\sqrt{1-\rho_0^2}}\right|}{\sigma_v} v - \left| \frac{\Phi^{-1}(\mathbb{E}[Y|X])k(X,\boldsymbol{\theta},\rho,w) + t(\boldsymbol{\theta},X)(T-\mathbb{E}[T|X])}{\sqrt{1-\rho^2}} \right| \\ &- \left| \frac{\Phi^{-1}(\mathbb{E}[Y|X])k(X,\boldsymbol{\theta}_0,\rho_0,w) + t(\boldsymbol{\theta}_0,X)(T-\mathbb{E}[T|X])}{\sqrt{1-\rho_0^2}} \right| > 0 \end{aligned}$$

as required. Therefore, the only remaining case is when $\rho = \rho_0$ but $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. By identifiability of τ, there exists $X \in \mathcal{D}_X$ such that $t(\boldsymbol{\theta},X) \neq t(\boldsymbol{\theta}_0,X)$. Then let:

$$T > \left| \frac{\Phi^{-1}(\mathbb{E}[Y|X])(k(X,\boldsymbol{\theta},\rho,w) - k(X,\boldsymbol{\theta}_0,\rho,w)) - \mathbb{E}[T|X](t(\boldsymbol{\theta},X) - t(\boldsymbol{\theta}_0,X))}{t(\boldsymbol{\theta},X) - t(\boldsymbol{\theta}_0,X)} \right|.$$

Thus we have:

$$\begin{aligned} &|f(X,T,\boldsymbol{\theta},\rho,w,v) - f(X,T,\boldsymbol{\theta}_0,\rho_0,w,v)| \\ &= \left| \frac{\Phi^{-1}(\mathbb{E}[Y|X])(k(X,\boldsymbol{\theta},\rho,w) - k(X,\boldsymbol{\theta}_0,\rho,w)) + (t(\boldsymbol{\theta},X) - t(\boldsymbol{\theta}_0,X))(T-\mathbb{E}[T|X])}{\sqrt{1-\rho^2}} \right| \\ &> \frac{1}{\sqrt{1-\rho^2}} \left(|(t(\boldsymbol{\theta},X) - t(\boldsymbol{\theta}_0,X))T| \right. \\ &\left. - |\Phi^{-1}(\mathbb{E}[Y|X])(k(X,\boldsymbol{\theta},\rho,w) - k(X,\boldsymbol{\theta}_0,\rho,w)) - \mathbb{E}[T|X](t(\boldsymbol{\theta},X) - t(\boldsymbol{\theta}_0,X))| \right) > 0 \end{aligned}$$

Thus we have $\tilde{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\Theta}$. Furthermore, according to asymptotic normality of MLE estimators, we have for some covariance matrix $\boldsymbol{\Sigma}$:

$$\sqrt{n}(\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \xrightarrow{p} N(\mathbf{0},\boldsymbol{\Sigma})$$

Proof of 2 As $\mathcal{D}_X, \mathcal{D}_\Theta$ are compact, we have $|\tau| \leq c < \infty$ for some c. Moreover, since $\mathcal{D}_X$ is compact, and $0 < r(X) < 1$, let the maximum and minimum value of $r(X)$ be $b < 1$ and $a > 0$ respectively. Then we note that $\Phi^{-1}(r(X))$ is bounded.

Recall $\hat{\Theta} = (\hat{\boldsymbol{\theta}},\hat{\rho}) = \arg\max_{\boldsymbol{\Theta} \in \mathcal{D}_\Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \hat{l}(Y_i,X_i,T_i,Z_i,\boldsymbol{\Theta},\hat{w},\hat{v}_i,\hat{r},\hat{q}) \right\}$. By the uniform convergence assumptions of $\hat{q}(X)$, $\hat{r}(X)$, $\hat{\rho}$, $\hat{\sigma_u}$, $\hat{\sigma_v}$, $\hat{\beta}$, $\hat{v}$, and compactness of $\boldsymbol{\Theta}$, it is not hard to see that we have uniform convergence of the loss function over $\boldsymbol{\Theta}$ with probability 1:

$$\sup_{\boldsymbol{\Theta}\in D_\Theta} \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{l}(Y_i,X_i,T_i,Z_i,\boldsymbol{\Theta},\hat{w},\hat{v}_i,\hat{r},\hat{q}) - l(Y_i,X_i,T_i,Z_i,\boldsymbol{\Theta},w,v)\Big) \xrightarrow{p} 0$$

As $\boldsymbol{\Theta}$ uniquely maximizes the likelihood, we thus have $|\hat{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}| \xrightarrow{p} 0$. Therefore, we have $\hat{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\Theta}$. If in addition we have for some $\delta > 0$ $\sup_{X\in\mathcal{D}_x} n^{1/2+\delta}|\hat{q}(X) - q(X)| \xrightarrow{p} 0$, $\sup_{X\in\mathcal{D}_x} n^{1/2+\delta}|\hat{r}(X) - r(X)| \xrightarrow{p} 0$, and similarly for all nuisance parameters, then $n^{1/2+\delta}|\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}| \xrightarrow{p} 0$. Thus, we have $\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \xrightarrow{p} 0$ using the asymptotic normality of MLE above.

Proof of 3 Then, as $\tau : \mathcal{D}_\Theta \times \mathcal{D}_X$ is continuous in $\boldsymbol{\theta}$, we have that $\|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}, X)\| \xrightarrow{p} 0 \quad \forall X \in M$. Now as τ has a compact domain (products of compact spaces are compact), pointwise convergence implies uniform convergence: $\sup_{X\in\mathcal{D}_X} \|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}, X)\| \xrightarrow{p} 0$, as τ is continuous in X (with probability 1). By consistency of our estimator Θ as above, we also have: $\hat{\rho} \xrightarrow{p} \rho$.

With the additional assumptions, we have in addition that:

$$\sqrt{n}\|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}, X)\| \xrightarrow{p} 0 \quad \forall X \in \mathcal{D}_X \quad \Rightarrow \quad \sup_{X\in\mathcal{D}_X} \sqrt{n}\|t(\hat{\boldsymbol{\theta}}, X) - t(\boldsymbol{\theta}, X)\| \xrightarrow{p} 0$$

Similarly for $\hat{\rho}$.

Appendix E: Optimal Price Strategy

Recall $\hat{w} = (\hat{\beta}, \hat{\sigma}_u, \hat{\sigma}_v)$. We have for a given ticket, the optimal price T^* is the following:

$$T^* = \text{argmax}_T \ \hat{R}^*(X,T) = \max_T \ T \cdot \Phi\Big(\tilde{f}(X,T,\hat{\tau},\hat{\rho},\hat{w},\hat{v},\hat{q},\hat{r},\hat{h})\Big) \tag{57}$$

where $\tilde{f}(X,T,\hat{\tau},\hat{\rho},\hat{w},\hat{v},\hat{q},\hat{r},\hat{h}) \equiv \frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})(T - \hat{q}(X)) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}$ and k defined in Eq. (10).

1. If $\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}} \geq 0$, then since $\Phi(\cdot)$, the standard normal cdf, is increasing , by composition, the objective $\hat{R}^*(X,\cdot)$ is increasing in price T, and the reseller should choose the highest price possible for the ticket.

2. If $\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}} < 0$, we calculate the first and second derivatives of $\hat{R}^*(X,\cdot)$ with respect to price T. We have:

$$\begin{aligned}\frac{\partial \hat{R}^*}{\partial T}(X,T) = \Phi\Big(&\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})(T - \hat{q}(X)) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\Big) \\ &+ T \cdot \frac{\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}} \cdot \Phi'\Big(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})(T - \hat{q}(X)) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\Big)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 \hat{R}^*}{\partial^2 T}(X,T) = &2\frac{\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}} \cdot \Phi'\Big(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})(T - \hat{q}(X)) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\Big) \\ &+ T \cdot \Big(\frac{\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\Big)^2 \cdot \Phi''\Big(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})(T - \hat{q}(X)) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\Big)\end{aligned}$$

Since for all $u \in \mathbb{R}$ $: \Phi''(u) = -u \cdot \Phi'(u)$, we can rewrite the last equation as :

$$\begin{aligned}\frac{\partial^2 \hat{R}^*}{\partial^2 T}(X,T) = g(T,X) \cdot \Big(&- T \cdot \Big(\frac{\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\Big)^2 \cdot \Big(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w}) + (\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}})(T - \hat{q}(X)) - \frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z - \hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\Big) \\ &+ 2 \cdot \frac{\hat{\tau}(X) + \frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\Big)\end{aligned}$$

where $g(T,X) = \Phi'\left(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})(T-\hat{q}(X))-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\right) > 0 \quad \forall T, X.$

The sign of $\frac{\partial^2 \hat{R}^*}{\partial^2 T}$ depends on the sign of :

$$-T\cdot\left(\frac{\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\right)^2\cdot\left(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})(T-\hat{q}(X))-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\right)+2\cdot\frac{\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}$$

which can be rewritten as:

$$\left(\frac{\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\right)\Bigg(2-\left(\frac{\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\right)^2\cdot T^2 - \left(\frac{\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}}}{\sqrt{1-\hat{\rho}^2}}\right)\cdot T\cdot\left(\frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X)-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))}{\sqrt{1-\hat{\rho}^2}}\right)\Bigg)$$

which is a quadratic function in T. We can then show that there exist $\underline{T} < \bar{T}$ such that $\frac{\partial^2 \hat{R}^*}{\partial^2 T}(X,T) > 0$ for $T < \underline{T}$, $\frac{\partial^2 \hat{R}^*}{\partial^2 T}(X,T) \leq 0$ for $T \in [\underline{T}, \bar{T})$ and $\frac{\partial^2 \hat{R}^*}{\partial^2 T}(X,T) > 0$ for $T > \bar{T}$, and we characterize $\underline{T}$ and $\bar{T}$ as the following:

$$\underline{T} = \frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X)-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))}{-2(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})} - \frac{\sqrt{\left(\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X)-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))\right)^2+8(1-\hat{\rho}^2)}}{-2(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})}$$

and

$$\bar{T} = \frac{\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X)-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))}{-2(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})} + \frac{\sqrt{\left(\Phi^{-1}(\hat{r}(X))k(X,\hat{\tau},\hat{\rho},\hat{w})+(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})\hat{q}(X)-\frac{\hat{\rho}\cdot\hat{\beta}}{\hat{\sigma_v}}(Z-\hat{h}(X))\right)^2+8(1-\hat{\rho}^2)}}{-2(\hat{\tau}(X)+\frac{\hat{\rho}}{\hat{\sigma_v}})}$$

One can show easily that $\underline{T} < 0$ and $\bar{T} > 0$. We can also show that $\frac{\partial \hat{R}^*}{\partial T}(X,0) > 0$ and $\frac{\partial \hat{R}^*}{\partial T}(X,\bar{T}) < 0$. We conclude that $\frac{\partial \hat{R}^*}{\partial T}(X,T) = 0$ has a unique solution T^* and we have $\frac{\partial \hat{R}^*}{\partial T}(X,T) > 0$ for $T < T^*$ and $\frac{\partial \hat{R}^*}{\partial T}(X,T) \geq 0$ for $T \geq T^*$, which makes the revenue function unimodal. In addition, we have $T^* \leq \bar{T}$.

Appendix F: Features Utilized in Real-World Experiment

Environment State Features (X)	**Game Features (X)**
Day of Week	Home Team
Week of Year	Away Team
Month of Year	Average 3pts Attempted (Home)
Days Listed Before Game	Average 3pts Made (Home)
	Win Percentage (Home)
Market and Pricing Features (M_t)	
Price of Ticket	section of Ticket
Row of Ticket	# of Tickets in Listing
25th/50th/75th percentile price listed in section/game	25th/50th/75th percentile price sold in section/game
Quantity sold in section/game	Quantity listed in section/game
Historical Game Median Price	Total Value sold in section/game
Normalized Quality Score by section/Game	

Appendix G: Boosted tree ensemble implementation:

A boosted tree with J rounds of learning has the functional form of a tree ensemble, where $\kappa_j(\cdot)$ are individual decision trees. In our case, we approximate the heterogeneous treatment effect $\tau_J(\cdot)$, using a tree ensemble.

$$\tau_J(X) = \sum_{j=1}^{J} \kappa_j(X)$$

The accuracy of the model is captured through the loss function defined in (13) (here with some terms omitted for brevity). Gradient boosted trees train the model in an additive manner, where in the J^{th} iteration, another decision tree $\kappa_J(\cdot)$ is added to the existing (and fixed) tree ensemble $\tau_{J-1}(\cdot)$, in a greedy manner to minimize the loss function, with an additional regularization term $\Omega(\kappa_J)$ included:

$$\mathcal{L}^J = \sum_{i=1}^{n} l(Y_i, \tau_{J-1}(X_i) + \kappa_J(X_i)) + \Omega(\kappa_J)$$

A second order approximation is used to approximate the loss function:

$$\mathcal{L}^J \approx \sum_{i=1}^{n} \left[l(Y, \tau_{J-1}(X_i)) + g_i \kappa_J(X_i) + \frac{1}{2} h_i \kappa_J(x)^2 \right] + \Omega(\kappa_J)$$

where

$$g_i = \frac{\partial l(Y_i, \tau_{J-1}(X_i))}{\partial \tau_{J-1}(X_i)}, \qquad h_i = \frac{\partial^2 l(Y_i, \tau_{J-1}(X_i))}{\partial \tau_{J-1}(X_i)^2}$$

are the first and second order partial derivatives of the loss function with respect to $\tau_{J-1}(X_i)$. Ignoring constant terms, we obtain the simplified objective at iteration J:

$$\sum_{i=1}^{n} \left[g_i \kappa_J(X_i) + \frac{1}{2} h_i \kappa_J(X_i)^2 \right] + \Omega(\kappa_J)$$

Therefore, if g_i and h_i are calculated from data, the tree in the next round $\kappa_J(X_i)$ can be trained to minimize this regularized loss function. The `lightgbm` package has an customizable function, `objective`, that takes current predictions $\tau_{J-1}(X_i)$ and labels Y_i as inputs and provides outputs g_i and h_i. Given g_i, h_i, there are other functions in the `lightgbm` package that fit the tree for this round $\kappa_J(X_i)$, and continue with the algorithm.

We calculate g_i and h_i using the function `fderiv` from the `pracma` package, which numerically differentiates the function (13) with respect to $\tau_{J-1}(X_i)$ using the central difference formula, with automatically chosen step size.

In our experiments, we implement Algorithm 1 using gradient boosted trees, which we denote `two_stage_LGBM`. With this approach, the nuisance parameters $(\hat{r}(X), \hat{q}(X), \hat{h}(X))$ are estimated using a standard implementation of the `lightgbm` package, while the loss function (13) is optimized and $\hat{\tau}(X)$ is estimated using a customized loss function defined within `lightgbm`, as outlined above.

Supplementary Appendices

Appendix H: Derivation of loss function with general error distributions

Let's consider the following model:

$$Y^* = g(X) + \tau(X)T + \epsilon \tag{58}$$

$$T = m(X) + \beta^T Z + \upsilon \tag{59}$$

$$Z = h(X) + u \tag{60}$$

$$Y = \begin{cases} 1, & \text{if } Y^* > 0 \\ 0, & \text{if } Y^* \leq 0 \end{cases} \tag{61}$$

$$(\epsilon, \upsilon) \text{ have have a mean 0 and their joint Probability Density Function (PDF) is } f_{(\epsilon,\upsilon)}, \tag{62}$$

$$u \text{ has a mean 0 and } f_u \text{ is its PDF, } u \perp \upsilon | X \text{ and } u \perp \epsilon | X \tag{63}$$

In this case, we assume full knowledge of the joint distribution of (ϵ, υ) and the distribution of u. We also note that u is independent of υ and of ϵ. We can partial out the effect of X on Y^* to obtain an orthogonalized outcome $Y^* - \mathbb{E}[Y^*|X]$. Since $\mathbb{E}[Y^*|X] = g(X) + \tau(X)m(X) + \tau(X)\beta^T h(X)$ and $\mathbb{E}[\epsilon|X] = 0$, we can rearrange (58) and (59) as follows:

$$Y^* - \mathbb{E}[Y^*|X] = \tau(X)(T - \mathbb{E}[T|X]) + \epsilon. \tag{64}$$

Similar to Proposition 1, we can try to characterize $E[Y^*|X]$ with respect to $E[Y|X]$. We have:

$$\begin{aligned}
\mathbb{E}[Y|X] &= P\Big(Y^* > 0 | X\Big) && (65)\\
&= P(g(X) + \tau(X)T + \epsilon > 0 | X) && (66)\\
&= P(g(X) + \tau(X)(m(X) + \beta(h(X) + u) + \upsilon) + \epsilon > 0 | X) && (67)\\
&= P(\beta\tau(X)u + \tau(X)\upsilon + \epsilon > -(g(X) + \tau(X)m(X) + \beta\tau(X)h(X)) | X) && (68)\\
&= P(\beta\tau(X)u + \tau(X)\upsilon + \epsilon > -\mathbb{E}[Y^*|X] | X) && (69)
\end{aligned}$$

Since $\tau(X)\upsilon + \epsilon$ and $\beta\tau(X)u$ are independent, we can use convolution to characterize the PDF of $\beta\tau(X)u + \tau(X)\upsilon + \epsilon$. In particular, let $W = \tau(X)\upsilon + \epsilon$, and $R = W + \beta\tau(X)u$. It is straightforward to show that the PDF of W is $f_W(w) = \int_{-\infty}^{\infty} f_{(\epsilon,\upsilon)}(w - \tau(X)\upsilon, \upsilon)d\upsilon$. Then, the PDF of R is

$$\begin{aligned}
f_R(r) &= \frac{1}{\beta\tau(X)} \int_{-\infty}^{\infty} f_W(r-u) \cdot f_u\big(\frac{u}{\beta\tau(X)}\big) du && (70)\\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(\epsilon,\upsilon)}(r - \tau(X)(\beta u + \upsilon), \upsilon) \cdot f_u(u) d\upsilon du. && (71)
\end{aligned}$$

If we denote the CDF of R as $F_R(x) = \int_{-\infty}^{x} f_R(r)dr$, then we can write:

$$\mathbb{E}[Y|X] = 1 - F_R(-\mathbb{E}[Y^*|X]) \tag{72}$$

and thus:

$$\mathbb{E}[Y^*|X] = -F_R^{-1}(1 - \mathbb{E}[Y|X]) \tag{73}$$

If we substitute (73) in (64), we have:

$$Y^* = -F_R^{-1}(1-\mathbb{E}[Y|X]) + \tau(X)(T-\mathbb{E}[T|X]) + \epsilon. \tag{74}$$

In addition, we can also decompose $\epsilon = \frac{\rho\sigma_\epsilon}{\sigma_v}v + \tilde{\epsilon}$ into orthogonal components, where $E[v\tilde{\epsilon}] = 0$, and $\tilde{\epsilon}|X$ has a pdf $f_{\tilde{\epsilon}}(\tilde{\epsilon}) = \int_{-\infty}^{\infty} f_{(\epsilon,v)}(\tilde{\epsilon} + \frac{\rho\sigma_\epsilon}{\sigma_v}v, v)dv$. With this modification, the treatment T is orthogonal to the error $\tilde{\epsilon}$. We obtain a new expression for the latent variable:

$$Y^* = -F_R^{-1}(1-\mathbb{E}[Y|X]) + \tau(X)(T-\mathbb{E}[T|X]) + \frac{\rho\sigma_\epsilon}{\sigma_v}v + \tilde{\epsilon} \tag{75}$$

Then we can derive an appropriate loss function resulting from the distribution of $\tilde{\epsilon}$ using its CDF $F_{\tilde{\epsilon}}(\tilde{\epsilon}) = \int_{-\infty}^{\tilde{\epsilon}} f_{\tilde{\epsilon}}(t)dt$:

$$\begin{aligned} l(Y,X,T,\tau,\rho,\beta,\sigma_\epsilon,\sigma_v,v) =& Y\log(1-F_{\tilde{\epsilon}}(-f(X,T,\tau,\rho,\beta,\sigma_\epsilon,\sigma_v,v)) + \\ & (1-Y)\log(F_{\tilde{\epsilon}}(-f(X,T,\tau,\rho,\beta,\sigma_\epsilon,\sigma_v,v))) \\ with \quad f(X,T,\tau,\rho,\beta,\sigma_\epsilon,\sigma_v,v) =& -F_R^{-1}(1-\mathbb{E}[Y|X]) + \tau(X)(T-\mathbb{E}[T|X]) + \frac{\rho\sigma_\epsilon}{\sigma_v}v \end{aligned}$$

We note that most of these nuisance parameters can be estimated prior to optimization of this loss function using machine learning algorithms. Similar to the model with normal errors, we can estimate $\hat{r}(X) \approx \mathbb{E}[Y|X]$, $\hat{q}(X) \approx \mathbb{E}[T|X]$ as well as β. However, we would need full knowledge of the joint distributions to fully characterize the loss function.

Appendix I: Feature Importance and Few Feature Formulation

I.1. Feature importance

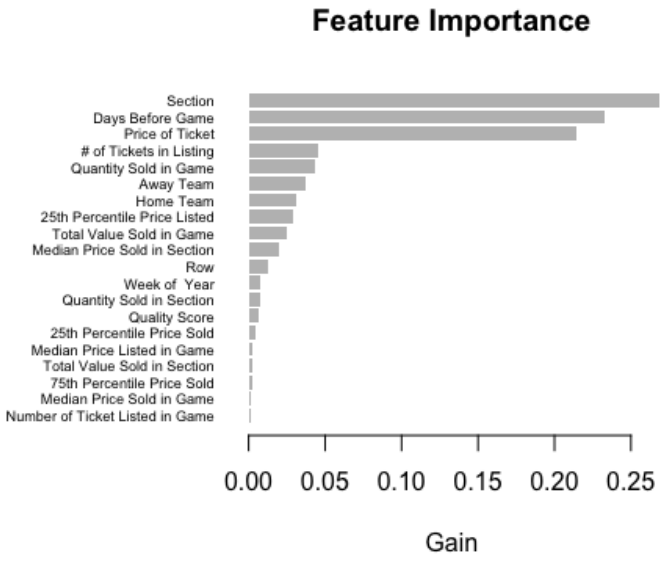


Figure 5: Feature importance as evaluated by boosted trees

Figure 5 shows the relative feature importance as evaluated by a `lightgbm` gradient boosted trees predictor. As expected, price is a very important factor for determining the probability of a ticket selling. The section of the ticket, corresponding to how close the seat is to the court, is also very important, as is the number of days until the game. When there is only a few days until the game, the ticket has a lower probability of selling. Previous sales and prices, and the teams playing are also important predictors.

I.2. Few Feature Formulation

Given the high feature importance of some features compared to others, we would test a model with few features as a baseline for comparison. The specific features chosen are:

- Section
- Days Before Game
- Price of Ticket

This corresponds to the three features with the highest variable importance as illustrated in 5. We then present the predictive accuracy results under this case:

Algorithm	Misclassification error	AUC
`two_stage_LGBM`	0.181	0.832
`LGBM`	0.187	0.830
`probit_iv`	0.217	0.745
`LGBM_iv`	0.197	0.820

Table 5: Predictive accuracy on NBA data

We see that the top 3 features, although overwhelmingly important by variable importance metric, is insufficient for a well-performing model, as even the best-performing model under this restriction is less accurate than the worst model under the full set of features. This suggests that the additional list of features are essential for an accurate model.

Appendix J: Correlation between Tickets

In this section, we investigate the validity of our assumption that the selling behavior between different tickets listed at the same time are not highly correlated, and thus the decision to use a single-item optimization model is justified.

In a market with many players (thousands in the case of a ticket exchange), each seller is a price taker and will not have an significant effect on the market. Apart from capturing the (assumed) exogenous market effects, a seller may be concerned about cross effects between tickets they are selling, and whether they need to solve a joint pricing problem. To explore this, we look at pairs of tickets which are listed at the same time and try to see if the covariance is significant.

The sample covariance of Y_i, Y_j given prices T_i, T_j are as below:

$$\hat{\mathrm{Cov}}(Y_i, Y_j \mid T_i, T_j) = \hat{\mathbb{P}}[Y_i = 1, Y_j = 1 \mid T_i, T_j] - \hat{\mathbb{P}}[Y_i = 1 \mid T_i]\hat{\mathbb{P}}[Y_j = 1 \mid T_j] \tag{76}$$

If $\hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j) \approx 0$, then the joint optimization problem to maximize expected revenue over the price of these two tickets (T_i, T_j) degenerates to a single-ticket optimization problem:

$$\begin{aligned}
&\max_{T_i,T_j} \hat{\mathbb{P}}[Y_i = 1, Y_j = 1 \mid T_i, T_j](T_i + T_j) + \hat{\mathbb{P}}[Y_i = 1, Y_j = 0 \mid T_i, T_j]T_i + \hat{\mathbb{P}}[Y_i = 0, Y_j = 1 \mid T_i, T_j]T_j \\
&\approx \max_{T_i,T_j} \hat{\mathbb{P}}[Y_i = 1 \mid T_i]\hat{\mathbb{P}}[Y_j = 1 \mid T_j](T_i + T_j) + \hat{\mathbb{P}}[Y_i = 1 \mid T_i]\hat{\mathbb{P}}[Y_j = 0 \mid T_j]T_i + \hat{\mathbb{P}}[Y_i = 0 \mid T_i]\hat{\mathbb{P}}[Y_j = 1 \mid T_j]T_j \\
&= \max_{T_i} \hat{\mathbb{P}}[Y_i = 1 \mid T_i]T_i + \max_{T_j} \hat{\mathbb{P}}[Y_j = 1 \mid T_j]T_j
\end{aligned}$$

To test the condition $\hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j) \approx 0$ experimentally on our dataset, we focus on the mean of the sample covariance among tickets to gain sufficient power. That is, for a set of tickets M, we investigate:

$$\hat{\mu} = \frac{1}{|\{i,j\} \mid i,j \in M \;\&\; \text{Colist}(i,j) = 1|} \sum_{i,j \in M : \text{Colist}(i,j)=1} \hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j) \tag{77}$$

Here $\text{Colist}(i,j)$ is the binary function such that $\text{Colist}(i,j) = 1$ if ticket i and ticket j has a minimum of one day in overlap in listing time, and that ticket i and ticket j belong to the same game. We only consider ticket pairs in the same game, as inter-game correlation is expected to be low due to the nature of the events, while most of the correlation (if there is any) should be intra-game.

If the correlation between tickets are small or non-existent, we expect $\hat{\mu} \approx 0$. Thus we would utilize the Two One-Sided Equivalence Test (TOST):

$$H_0 : |\hat{\mu}| > \Delta \qquad H_1 : |\hat{\mu}| < \Delta$$

And we intend to find the minimum value of Δ such that H_0 can be rejected on the 5% level. We would utilize two separate probit regressions to estimate $\mathbb{P}[Y_i = 1, Y_j = 1 | T_i, T_j]$ and $\mathbb{P}[Y_i = 1 | T_i]$, and use the residuals to determine the variance of $\hat{\mu}$. under strong law of large numbers, $\hat{\mu}$ is approximately normal, and thus we would utilize a t-test to test this hypothesis. The details of the calculations on the variance of $\hat{\mu}$ is contained in J.1.

We trained the two probit regressions on historical data and evaluated it against 4.47 Million pairs of tickets, formed through a random set of tickets with size $|M| = 100,000$. The first probit model (n=4,470,000) is used to estimate $\hat{\mathbb{P}}[Y_i = 1, Y_j = 1 \mid T_i, T_j]$, while the second probit model (n=100,000) estimates $\hat{\mathbb{P}}[Y_i = 1 \mid T_i]$.

The minimum Δ is:

$$\Delta_m = 0.0110$$

Thus, we can reject the hypothesis that $|\hat{\mu}| > 0.01$ on the 5% level. Since $\Delta_m \approx 0$, this provides statistical evidence that the ticket market, on average, is uncorrelated, and thus we can utilize single-ticket optimization.

J.1. Technical details for the Two One-Sided Equivalence Test

In this section, we rigorously derive the distribution of $\hat{\mu}$ under the null distribution subject to standard assumptions. For the covariance term:

$$\hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j) = \hat{\mathbb{P}}[Y_i = 1, Y_j = 1 \mid T_i, T_j] - \hat{\mathbb{P}}[Y_i = 1 \mid T_i]\hat{\mathbb{P}}[Y_j = 1 \mid T_j] \tag{78}$$

We assume that our estimates approximately follows the normal distribution with iid samples:

$$\hat{\mathbb{P}}[Y_i = 1, Y_j = 1 \mid T_i, T_j] \dot{\sim} N(\mathbb{P}[Y_i = 1, Y_j = 1 \mid T_i, T_j], \sigma_1^2) \qquad \hat{\mathbb{P}}[Y_i = 1 \mid T_i] \dot{\sim} N(\mathbb{P}[Y_i = 1 \mid T_i], \sigma_2^2) \quad \forall\, i, j \tag{79}$$

Then we have that:

THEOREM 3. *Assume Equation 79 holds. The bias and the variance of the covariance is then:*

$$\mathbb{E}[\hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j)] = \text{Cov}(Y_i, Y_j \mid T_i, T_j)$$
$$\mathbb{V}[\hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j)] = \sigma_1^2 + \sigma_2^4$$

Proof: The proof of unbiasedness is trivial and would be omitted. For the variance, we note the following lemma:

LEMMA 2. *For independent normal variables $X, Y \sim N(0, \sigma^2)$, we have that:*

$$XY = c(Q - R), \tag{80}$$

where $c = \frac{1}{2}\sigma^2$, and $Q, R \sim \chi_1^2$ are independent.

Then using the lemma, we have:

$$\begin{aligned}\mathbb{V}[\hat{\text{Cov}}(Y_i, Y_j \mid T_i, T_j)] &= \mathbb{V}\left[\hat{\mathbb{P}}[Y_i = 1, Y_j = 1 \mid T_i, T_j]\right] + \mathbb{V}\left[\hat{\mathbb{P}}[Y_i = 1 \mid T_i]\hat{\mathbb{P}}[Y_j = 1 \mid T_j]\right] \\ &= \sigma_1^2 + c^2(\mathbb{V}[Q] + \mathbb{V}[R]) \\ &= \sigma_1^2 + \sigma_2^4\end{aligned}$$

Where $Q, R \sim \chi_1^2$ and $c = \frac{1}{2}\sigma_2^2$. Thus, under the assumption that there is no correlation ($\text{Cov}(Y_i, Y_j \mid T_i, T_j) = 0$), we derive the first and second moment of $\hat{\mu}$ as:

$$\mathbb{E}[\hat{\mu}] = 0, \qquad \mathbb{V}[\hat{\mu}] = \frac{\sigma_1^2 + \sigma_2^4}{|\{i, j\} \mid i, j \in M \;\&\; \text{Colist}(i, j) = 1|} \tag{81}$$

Then the variances σ_1^2 and σ_2^2 are replaced with their estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_1^2$ from the models for the single probabilities $\mathbb{P}[Y_i = 1 \mid T_i]$ and cross-probabilities $\mathbb{P}[Y_i = 1, Y_j = 1 \mid T_i, T_j]$ respectively. In our simulation, the estimated numbers were $\sigma_2^2 = 0.1098$ and $\sigma_1^2 = 0.2033$ respectively.

Appendix K: RShiny App

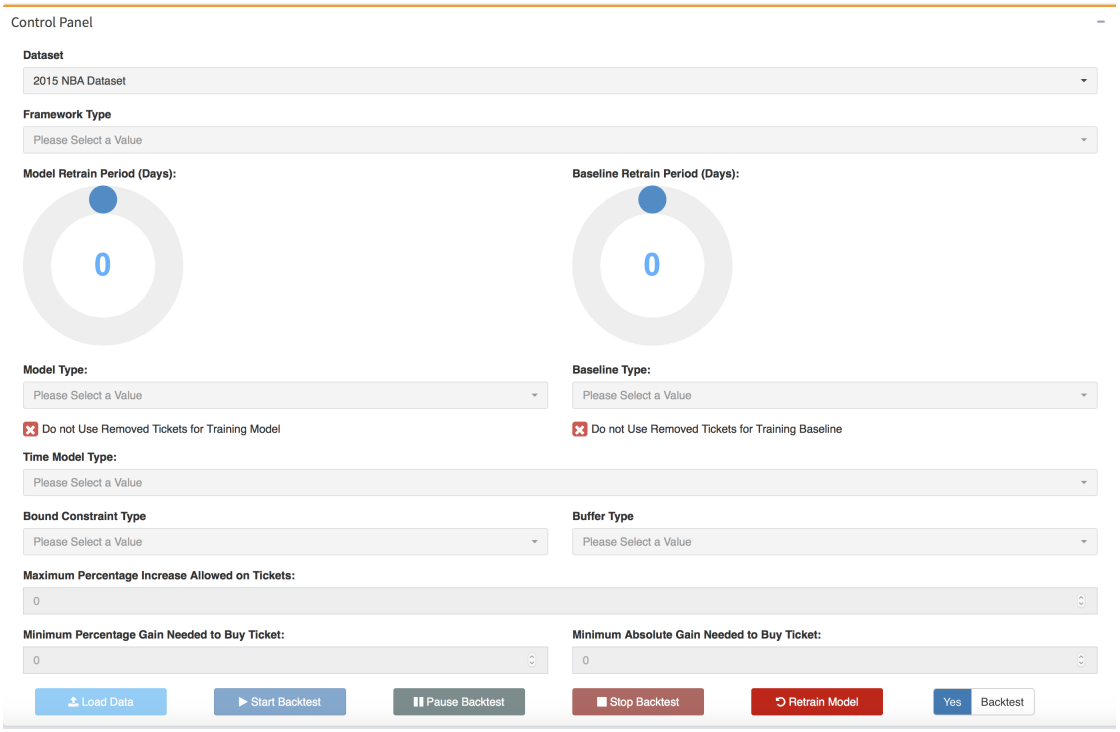


Figure 6: RShiny app

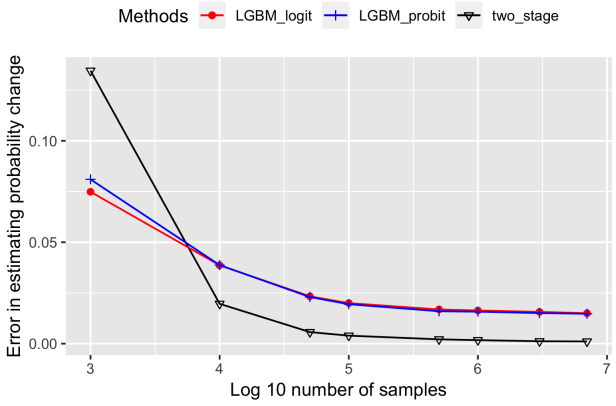
Appendix L: Additional numerical experiments

L.1. Treatment effect error with sample size

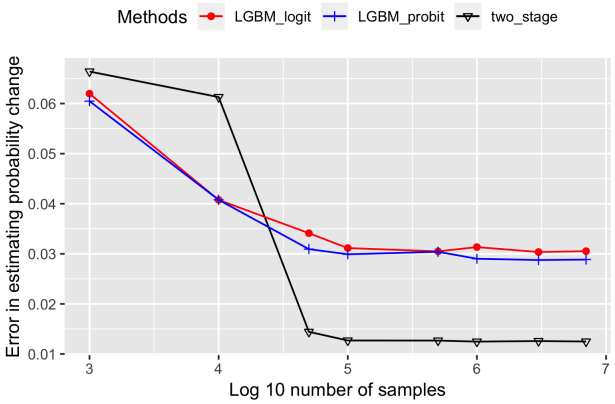
We also explore the effect of sample size on the ability to predict the change in probability associated with a fixed treatment change of 0.5 in Figure 7. We observe that the two stage approach is generally able to converge to high accuracy estimates. In dataset 2 and 4, we observe that the two stage approach may require more samples than a single stage, since it requires high accuracy estimates of $\mathbb{E}[T|X]$ and $\mathbb{E}[Y|X]$ which are also affected by a small sample size.

L.2. Optimal treatment prescription simulations

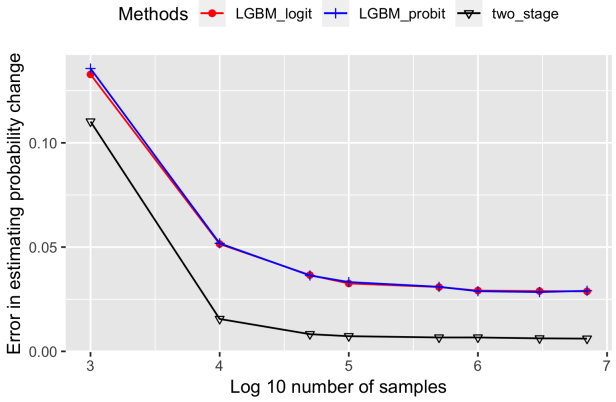
To explore the prescriptive element of our algorithm, we can calculate the optimal treatment for an adaptation of our synthetic data to a revenue maximization setting. Using notation from section §4.2, we find the optimal price $T^* = \arg\max_T T \cdot P(Y=1|X,T)$, where $P(Y=1|X,T)$ is calculated analytically for the respective synthetic datasets. We can compare against the estimated optimal treatment $\hat{T} = \arg\max_T T \cdot \hat{f}(X,T)$, where $\hat{f}(X,T)$ is the estimate of $P(Y=1|X,T)$ for each of our models. In the case of the two stage approach, this is equivalent to solving (16). We compare the optimality gap in revenue for each estimated optimal treatment. We use the same experimental set up as above and use an exhaustive line search to find optimal T in each optimization problem:



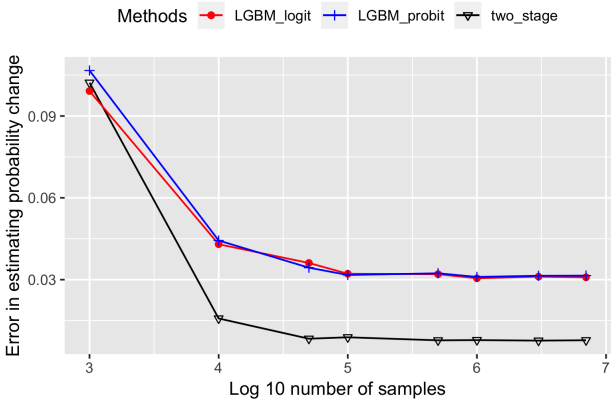
(a) Dataset 2: simple variable effect treatment



(b) Dataset 3: linear relationships



(c) Dataset 4: non-linear $g(x)$, linear $m(x), \tau(x)$



(d) Dataset 5: non-linear relationships

Figure 7: Estimating treatment effect for changing sample size

dataset	two stage	one stage probit	one stage logit
constant treatment	0	0	0
linear relationships	**0.0017**	0.0082	0.0082
non-linear relationships	**0.0028**	0.0104	0.0134

Table 6: Difference between revenue at prescriptive price $\hat{T}$ and true optimal price T^*

We can observe that in each of our synthetic datasets, the improvement in estimation of the treatment effect translates to an improvement in selecting the optimal treatment. In the next section we outline how we adapted this approach to StubHub's data.

Appendix M: Backtesting Framework

To understand the performance of our algorithm on real-life data without immediately deploying it to production, we need to conduct backtesting experiments to see how different models perform over periods in a trading environment. To do so, we generated some simplifying assumptions that allowed us to construct a backtesting framework for testing the performance of the models on past years' data efficiently. The backtesting framework uses two models; the testing model which is used to buy/price tickets and a baseline model which is used to evaluate whether the tickets will sell. The backtesting procedure is as follows:

1. We enter all the parameters needed to execute the backtesting framework, including:
   - **Model Retrain Period** - How frequent do we want to retrain the model, if we want to do so?
   - **Baseline Retrain Period** - How frequent do we want to retrain the baseline, if we want to do so?
   - **Test Model Type** - The algorithm used to train the model for predictive decisions.
   - **Baseline Model Type** - The algorithm used to train the baseline for counterfactuals.
   - **Time Model Type** - The algorithm used to estimate the time needed for a sold ticket to be actually sold.
   - **Buffer Type** - What type of buffer do we want to have for our optimization to be comfortable to buy a ticket? A percentage or an absolute gain, or both?
   - **Maximum Percentage Increase Allowed on Tickets** - What is the upper limit of our optimization?
   - **Minimum Percentage Gain Needed to buy Tickets** - A minimum percentage gain needed (of its optimal value over current price) for the optimization to decide to buy the ticket, if the appropriate buffer type is chosen.
   - **Minimum Absolute Gain Needed to buy Tickets** - A minimum absolute gain needed (of its optimal value over current price) for the optimization to decide to buy the ticket, if the appropriate buffer type is chosen.
2. The specified testing model is trained on 1,000,000 ticket samples in the immediate past of the starting date. The specified baseline model is trained on all ticket samples contained in the entire dataset so it serves as an oracle.
3. At the start of the day (00:00), we assume all the tickets that are listed on that day are available.
4. We execute the testing model to evaluate all the tickets listed on the market and buy tickets according to optimization procedure in section §4.4.
5. After the ticket is bought, the testing model immediately uses the calculated optimal price to relist the ticket either dynamically or in a static way. For the dynamic setting, we discretize the selling period decisions to the following cutoffs:

   $$(100, 90, 80, 70, 60, 50, 45, 40, 35, 30, 29, 28, \cdots, 1, 0)$$

   We consider selling prices discretized by 5 dollar intervals centered around the original price, with a minimum of 50% below the listing price and a maximum of 50% over the listing price. The selling probabilities for each period is decided through evaluating the baseline model at the endpoints of these periods, while accounting

for market state changes. While evaluating, the daysBefore feature is set to the number of days on the end points, and for each ticket i and its market feature M_i, we define it using the following evolution equation:

$$M_i(t) = M_i(0) + \mu_i t + \sigma_i t \epsilon_i,$$

where μ_i and σ_i are parameters fitted through historical data, and $M_i(0)$ indicates the market state feature at time of listing. ϵ_i is a standard $(0, 1)$ Normally distributed variable. Therefore, each market feature is assumed to be a Gaussian random walk. We run 10 random walks each for each ticket to calculate the selling probability matrix. The results are averaged.

6. The testing model tries to sell everything in its inventory at the calculated optimal price, and the baseline model evaluates the selling. We evaluate the selling decision based on 1 random walk for each ticket. The evaluation selling probabilities are based on the baseline model.
7. At the end of the day (23:59), we assume all the tickets that are removed on that day are removed, and all tickets that perish that day are also removed.
8. The day count moves forward by one day, and we return to Step 3. If a retraining period for the testing model or the baseline model is specified, and the period is due on that day, we return to Step 2.

Appendix N: Comparing Variance Estimates Under Naive Estimator and Bootstrap Correction

Dataset	Naive Estimation	Bootstrap Correction
Dataset 1	1.0015	1.00017
Dataset 2	1.0109	1.0044
Dataset 3	1.0009	0.9996
Dataset 4	1.0429	1.0352

Table 7: Estimation of σ_v Under Various Datasets (Average over 10 Runs)

Dataset	Naive Estimation	Bootstrap Correction
Dataset 1	1.0015	1.0011
Dataset 2	1.0006	1.0001
Dataset 3	1.0005	1.0002
Dataset 4	1.0007	1.0003

Table 8: Estimation of σ_u Under Various Datasets (Average over 10 Runs)

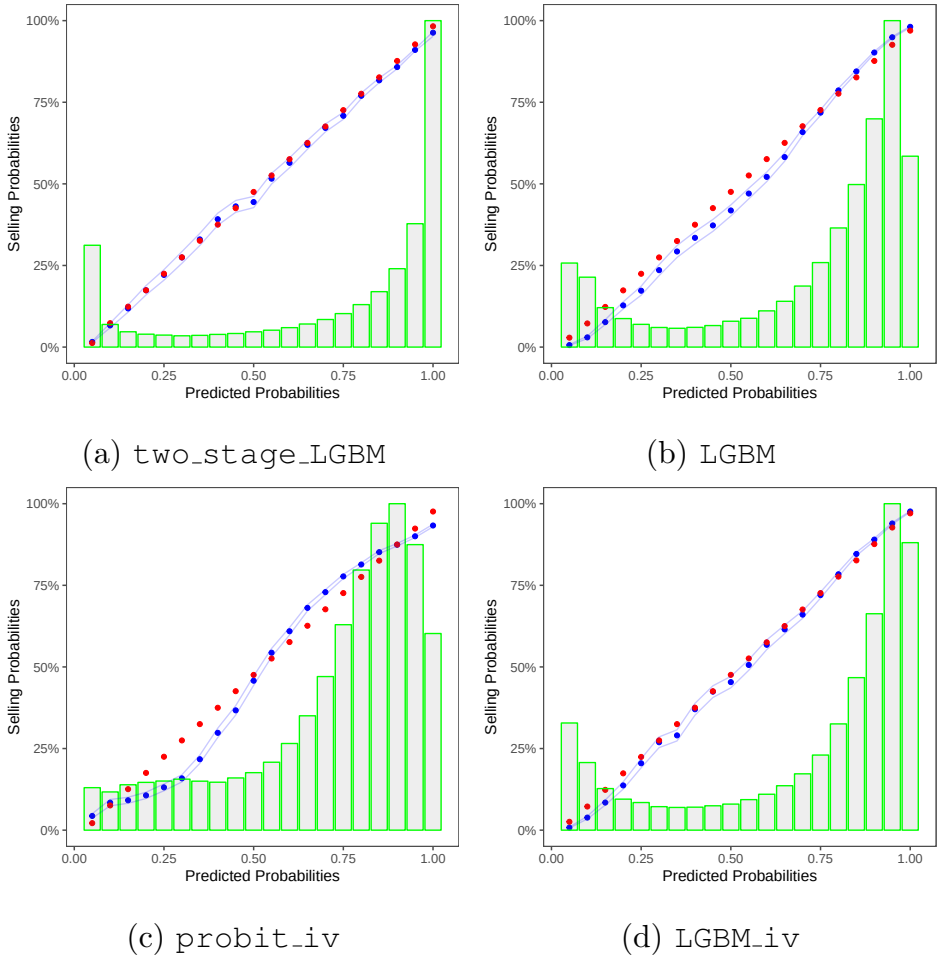
Appendix O: Probability Calibration for Estimation Methods

O.1. Probability calibration

Our algorithm uses the predicted probability of a ticket selling to set the optimal price for a ticket. Therefore, it is very important to be able to estimate these probabilities accurately, which is different from just having a low misclassification rate. For example, it is a well-documented fact that gradient boosted trees do not produce well calibrated probabilities (Caruana and Niculescu-Mizil 2006).

Figure 8 shows a calibration plot, which explores the difference between the predicted and empirical probabilities of the model. We separate the tickets into 21 bins based on the nearest 5%-increment of the predicted selling probability ($0-2.5\%$,$2.5-7.5\%$,$\cdots$,$97.5-100\%$). Then we calculate the empirical selling probability of the tickets in the bin and plot the result as the blue dots. For example, for the probit IV model, tickets with a predicted probability of selling between 22.5-27.5% actually were sold, on average, 15% of the time. We then include the red dots, which plots an oracle that has perfect calibration. A well-calibrated algorithm would have near complete overlap between the blue dots and the red dots. We qualitatively observe that the two-stage LGBM is significantly closer to an oracle classifier than the other models.

To quantify the effect of miscalibration, we devise the following miscalibration statistic. Consider a set of tickets $\mathcal{T} = \{t_i\}_{i=1\cdots n}$, and predicted probabilities $\hat{p}_i$ for ticket t_i. We create bins $M_j = [l_j, u_j)$ that separate the predicted probabilities into equally spaced intervals. Then the empirical probability of bin M_j is $p_j = \frac{1}{|i:\hat{p}_i \in M_j|} \sum_{i:\hat{p}_i \in M_j} Y_i$ where Y_i is the observed outcome for ticket t_i. We define the predicted probability of bin M_j, as $\bar{p}_j = \frac{1}{|i:\hat{p}_i \in M_j|} \sum_{i:\hat{p}_i \in M_j} \hat{p}_i$. Thus, miscalibration is defined as $\frac{1}{n} \sum_j |i : \hat{p}_i \in M_j| \cdot |\bar{p}_j - p_j|$. This can be seen as a weighted-average difference between the average model probability and the true (empirical) probabilities.



(a) `two_stage_LGBM` (b) `LGBM`

(c) `probit_iv` (d) `LGBM_iv`

Figure 8: Calibration Curves on NBA Data of different prediction models

In table 9, we compare the miscalibration of the various methods. We observe that the miscalibration of the two-stage method is significantly less than that of the other methods at just over 1%.

Algorithm	Miscalibration
`two_stage_LGBM`	0.0112
`LGBM`	0.0251
`probit_iv`	0.0426
`LGBM_iv`	0.0141

Table 9: Probability Calibration on NBA data