Model Trees for Personalization

Ali Aouad
London Business School, London, UK, aaouad@london.edu

Adam N. Elmachtoub
Department of Industrial Engineering and Operations Research and Data Science Institute, Columbia University, New York, NY, adam@ieor.columbia.edu

Kris J. Ferreira
Harvard Business School, Harvard University, Boston, MA, kferreira@hbs.edu

Ryan McNellis
Department of Industrial Engineering and Operations Research and Data Science Institute, Columbia University, New York, NY, rtm2130@columbia.edu

As more commerce and media consumption are being conducted online, a wealth of new opportunities are emerging for personalized advertising. We propose a general methodology, *Model Trees for Personalization* (MTP), for tackling a broad class of personalized decision-making problems including personalized advertising. The MTPs learn an interpretable market segmentation driven by differences in user behavior. Using this methodology, we design two new algorithms for fundamental problems in personalized advertising – Choice Model Trees (CMTs) for the user ad-response prediction problem, and Isotonic Regression Model Trees (IRMTs) for the bid landscape forecasting problem. We provide a customizable, computationally-efficient, and open-source code base for training MTPs in Python. We train our IRMT algorithm on historical bidding data from three different ad exchanges and show that the IRMT achieves 5-29% improvement in bid landscape forecasting accuracy over the model that a leading demand-side platform (DSP) provider currently uses in production.

1. Introduction

Recent growth of online commerce and media consumption have resulted in an expansion of opportunities for personalized advertising. Online retailers such as Amazon and eBay display "interest-based" ads on their homepage, which are personalized using the visiting user's purchase history and demographic information. Streaming services such as Hulu, YouTube, and Spotify can also personalize ads based on the media content being consumed and other aspects of the user's activity history. Additionally, in online advertising exchanges, bids for ad spots can be customized on the basis of various features encoding the ad spot and the site visitor.

Personalized decision-making often lies at the intersection of two fundamental technical challenges: *market segmentation* (clustering users into segments based on user characteristics) and *response modeling* (the probabilistic modeling of a user's response to a personalized decision). For example, if a retailer wishes to personalize product ads for its users, it could (1) segment users into interpretable and homogeneous groups, and (2) model the ad-response behavior of users in each group. One common approach is to perform the tasks of market segmentation and response modeling separately, using a clustering algorithm (e.g., k-means) for market segmentation and then fitting a response model (e.g., logistic regression) within each cluster. However, such a market segmentation is driven only by feature dissimilarity rather than differences in user response behavior.

We propose a general methodology, Model Trees for Personalization (MTP), that builds interpretable model trees for *joint* market segmentation and response modeling, which can be used for a variety of personalized advertising applications. Decision tree splits are applied by the MTP to segment the market according to available *contextual* attributes for personalization (e.g., features encoding the user). A response model is fit in each segment to probabilistically model user response (e.g., to ad exposure) as a function of the decision variables (e.g., ads that were offered). We propose a training procedure for MTPs which yields a market segmentation driven by predicting user response behavior – the decision tree splits of the MTP are decided through optimizing the predictive accuracy of the resulting collection of response models.

We provide an open-source implementation of our training procedure in Python (Aouad et al. [n. d.]). The code base is modular and easily customized to fit different personalized decision-making applications. Several features have been included for improved scalability, including the option of using parallel processing and warm starts for training the MTP models.

To demonstrate the versatility of our methodology, we design two new, specialized MTP algorithms for applications in personalized advertising. First, we propose a new algorithm, *Choice Model Trees* (CMTs), for ad-response prediction. An ad response could be defined as an ad click, conversion, or any other form of user engagement with an ad. Our model uses decision tree splits to segment users on the basis of their features (e.g., prior purchase history), and within each segment a Multinomial Logit (MNL) choice model is fit to model the users' ad-response behavior. Due to space limitations, we mention this application only in passing and focus our attention on the next application below.

Second, we propose a new algorithm, *Isotonic Regression Model Trees* (IRMTs), for the bid landscape forecasting problem. A "bid landscape" refers to the probability distribution of the highest (outside) bid that an ad spot will receive when being auctioned at an advertising exchange. The bid landscape forecasting problem is important to Demand Side Platforms (DSPs) – ad campaign management platforms – in estimating the minimum bid necessary to win different types of ad spots. A significant challenge is presented when ad spot transactions occur through first-price auctions – in such cases the highest outside bid is never revealed, and the DSP only sees whether their submitted bid resulted in an auction win or loss outcome. Current trends suggest that most major ad exchanges will switch to first-price auctions by the end of 2019 (Sluis 2019). We propose a new model, IRMTs, for the bid landscape forecasting problem under first-price auction dynamics. Our model uses a decision tree to segment auctions according to features about the visiting user (e.g., user's location) and the ad-spot being auctioned (e.g., width/height in pixels). An isotonic regression model is used to model the bid landscapes of the auctions within each segment. IRMTs are fully non-parametric, operating without assumptions about the distribution of the bid landscapes or of their relationship with the auction features. We apply our IRMT to an ad-spot transaction data set collected by a large DSP provider, and we demonstrate that our model consistently achieves a 5-29% improvement in bid landscape forecasting accuracy over the DSP's current approach across multiple ad exchanges.[1]

2. Literature Review

In this work, we propose a general framework for building model trees for personalized decision-making problems. Model trees refer to a generalization of decision trees which allow for non-constant leaf prediction models. Arguably the most common model tree algorithms explored in the literature are linear model trees (Quinlan et al. 1992) and logistic model trees (Chan and Loh 2004, Landwehr et al. 2005), which propose using linear and logistic regression leaf models with decision trees. Zeileis et al. (2008) develop a general framework, model-based recursive partitioning (MOB), for training model trees with parametric leaf models such as linear and logistic regression. While Zeileis et al. (2008) conduct decision tree splits through parameter instability tests, our method chooses splits which directly minimize the predictive error of the resulting collection of leaf models.

We are among the first to propose using model trees for market segmentation and for personalized decision-making problems. Similar to our CMT algorithm, Mišić (2016) proposes using model trees with choice model leaves for personalizing assortment decisions. In contrast, MTPs offer a more general framework for building model trees for personalization problems outside of assortment optimization, and the code for training MTPs has been empirically validated on real data and made open source. Kallus (2017) and Bertsimas et al. (2019) propose new decision tree algorithms for personalized decision-making, but the decisions are limited to a small, finite set of treatments. The framework proposed in our work allows for continuous high-dimensional decision spaces (e.g., vectors of prices). There have been several non-tree-based approaches to data-driven market segmentation proposed in the literature, including using mixture models (Kallus and Udell 2016, Noor et al. 2014, Bernstein et al. 2018) and model-based embeddings (Jagabathula et al. 2017) for personalization.

A significant contribution of our work is in developing a new model, IRMTs, for bid landscape forecasting with respect to first-price auctions. To the best of our knowledge, building model trees with isotonic regression leaf models has not been proposed in the prior literature, and the idea of using isotonic regression to model first-price auction dynamics is also novel. Wang et al. (2016) also propose a model tree algorithm for bid landscape forecasting. In contrast to IRMTs, Wang et al.

[1] For confidentiality reasons, the name of the DSP provider is not reported in this paper.

(2016) use KL-Divergence as the criterion for split evaluation, and their leaf models are designed to model second-price auctions rather than first price. Cui et al. (2011) also propose using decision trees for this application; however, their segmentation procedure only takes into account the *average* outside bid in each segment rather than the full bid landscape distributions.

Finally, by proposing the CMT model, our work contributes to the literature on ad-response modeling. Since an exhaustive literature review is beyond scope here, we refer the readers to the practice-oriented papers McMahan et al. (2013), Chapelle et al. (2015), Lu et al. (2017), and to the references therein.

3. Methodology

3.1. Problem Formulation

We now provide a general formulation of a personalized decision-making problem, which we break down into three components. First, the agent observes variables $x \in \mathbb{R}^m$ which serve as the *context* for the decision. The agent then makes a decision encoded by features p, and finally a user's response y is observed as a result of the decision. We emphasize that our approach can handle categorical, ordinal, and continuous data with respect to x, p, and y. In the application of ad-response prediction, the contextual variables x consist of features about the user (e.g., prior purchase history), the decision p is a collection of feature vectors encoding the displayed assortment of ads, and the response y indicates whether the user responded to an ad and if so which ad was chosen. For the bid landscape forecasting problem, the contextual variables x encodes the features describing the current user and auctioned ad-spot (e.g., the ad spot's width/height, encompassing website, and location on page), the decision $p \geq 0$ is the submitted bid price, and the response $y \in \{0, 1\}$ indicates the outcome of the auction (win/loss).

Our objective is to build an interpretable model for personalized decision-making problems that accomplishes two goals:

1. **Market Segmentation**. Our model should yield an interpretable market segmentation of the contextual variables x. Here, we define a market segmentation as a partition of the context space $\mathbb{R}^m$ into a finite number of disjoint segments. Given this partition, a response model is computed over each segment. Beyond the benefit of interpretability, market segmentation allows us to fit simple response models for each market since the user features have already been accounted for in the segmentation. In contrast, one can avoid market segmentation and fit a single, high-dimensional model (with many interaction terms) for personalization, although this approach can be computationally challenging, less interpretable, and have empirically weaker performance.

2. **Response Modeling**. Our model should accurately estimate the probability of each response y for all contexts x and decisions p, $P(y|x, p)$. Note that for the bid landscape forecasting problem,

$P(y|x,p)$ yields the distribution (c.d.f.) of the highest outside bid price p_o, as $P(y = \text{win}|x,p) = P(p_o \leq p|x)$. If the expected reward from a realized response y is given by the function $r(x,p,y)$, then the optimal personalized decision for each context is given by $p^*(x) = \arg\max_p E[r(x,p,Y)|x,p]$. Clearly, accurately estimating $P(y|x,p)$ is a critical step toward approximating this expectation and computing near-optimal decisions.

Section 3.2 discusses our MTP approach which tackles these tasks *jointly*, with the market segmentation being informed by the resultant response models. This arguably yields a more informative market segmentation – users in the same segment of the CMT can be interpreted as having similar ad-response behaviors, and auctions in the same segment of our IRMT model can be interpreted as having similar bid landscapes. Section 3.3 presents an algorithm for training MTPs from historical data.

3.2. Model Trees For Personalization (MTPs)

We tackle the personalized decision-making problems previously described using an approach we call *Model Trees For Personalization* (MTPs). MTPs perform market segmentation according to successive decision tree splits on the contextual variables x. Each split partitions the space of contexts with respect to a single contextual variable; continuous and ordinal contexts are split using inequalities (e.g., "Age $\leq$ 40?"), while categorical contexts are split using equalities (e.g., "Gender = Male?"). Each resulting market segment l – referred to as a *leaf* of the MTP – contains a response model $f_l(y|p)$ estimating the distribution of the response y given the decision p for users in segment l. Since different market segments may exhibit different distributions of the response y, the response models $f_l(y|p)$ may vary significantly across segments.

To use the MTP for prediction, i.e. to estimate $P(y|x,p)$ for a given context x and decision p, one simply needs to follow the decision tree splits to the leaf l to which the context x belongs and output $f_l(y|p)$. For example, with respect to the MTP in Figure 1, a user with context $x = \{\text{Age} = 30, \text{Location} = \text{USA}, \text{Gender} = \text{Male}\}$ would belong to segment $l = 2$, so response model $f_2(y|p)$ would be used to make predictions with respect to that user.

As Figure 1 demonstrates, the market segmentation produced by MTPs is interpretable and easily visualized. MTPs also have a number of desirable properties as estimators. Their decision tree splitting procedure is non-parametric, allowing MTPs to model potentially highly non-linear relationships in the mapping from contexts to segments. MTPs also naturally model interactions among the contextual variables – for example, in the MTP in Figure 1, the variable *age* interacts with both *location* and *gender*.

MTPs provide a general framework from which new algorithms can be designed for personalized decision-making problems. To do so, the practitioner simply needs to specify a family of response

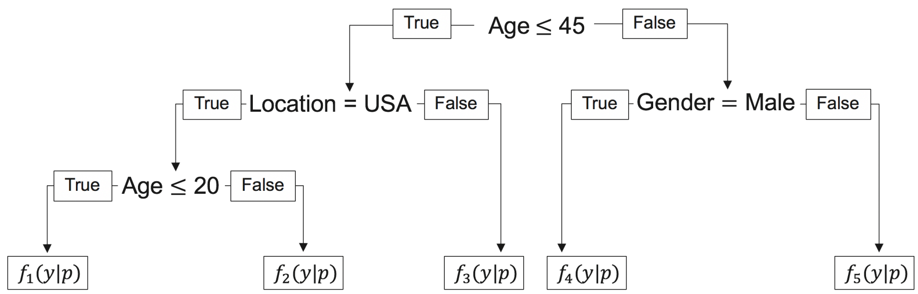


Figure 1: An example of an MTP with five market segments. Decision tree splits are performed with respect to the contextual variables *age* (numeric), *location* (categorical), and *gender* (categorical). Each of the resulting market segments contains a unique model $f_l(y|p)$ of the distribution of the response given the decision variables.

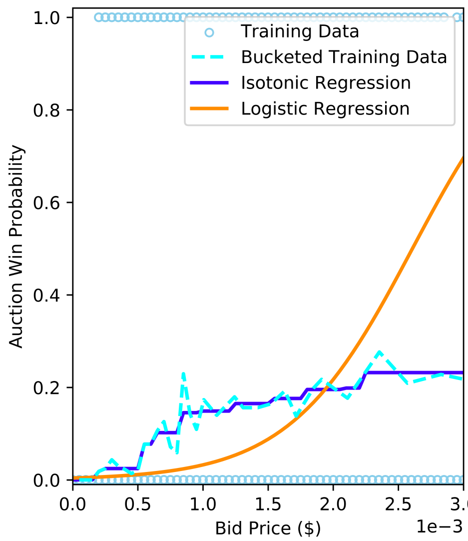
models for the given problem at hand (as well as a loss function for training the response models – this is covered in Section 3.3). As a proof of concept, we design from our MTP framework two new algorithms for fundamental problems in personalized advertising, outlined in the two subsections below.

3.2.1 Isotonic Regression Model Trees (IRMTs) for Bid Landscape Forecasting

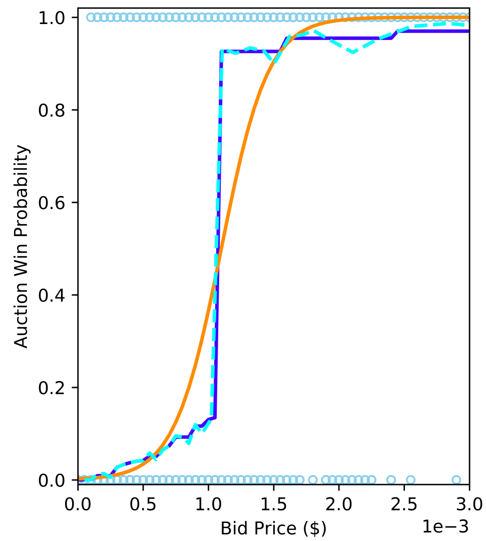
We propose a new algorithm, *Isotonic Regression Model Trees* (IRMTs), for bid landscape forecasting. The model tree segments ad-spot auctions according to contexts such as the auctioned ad-spot's dimensions in pixels and the visiting user's location. Within each leaf of the model tree, an isotonic regression model estimates the bid landscape of the auctions belonging to that leaf. Let $p \geq 0$ denote an auction bid, and let y be a binary variable which equals 1 if and only if the bid won the auction. The isotonic regression model in each leaf l, denoted by $f_l(y|p)$, estimates the probability that a given bid of p will result in an outcome of y for auctions in that leaf.

An isotonic regression model is a free-form curve fitted to historical data in the following way: the curve is the best *monotonically-increasing* curve that minimizes the training set prediction error (as defined by mean-squared error). The constraint of monotonicity is natural for this application, as the probability of an auction win should increase when the submitted bid p increases. Isotonic regression models are non-parametric and uniformly consistent estimators, feasibly capturing any stochastic, monotone function given sufficient data (Brunk 1970, Hanson et al. 1973). Also, the decision tree segmentation procedure of MTPs is non-parametric, imposing no distributional assumptions about the data. Thus, IRMTs offer a fully non-parametric, interpretable algorithm for bid landscape estimation.

Figure 2 plots the estimated isotonic regression models in two different leaves of an IRMT trained on historical bidding data collected by an anonymous DSP. As the figure demonstrates, different types of auctions can have differently-shaped bid landscapes, and the isotonic regression models are



(a) IR model (first leaf)



(b) IR model (second leaf)

Figure 2: Estimated bid landscapes in two leaves of an IRMT fit on bid data collected by a large DSP. The isotonic regression models are fit on training sets of auction outcomes (blue circles) within each leaf. Also included in the figures are logistic regression models trained on the same data. The models are compared against a curve (blue dashed line) constructed by bucketing the training set bids and computing the fraction of auction wins in each bucket.

flexible enough to capture these differences in curve shape. The figure also suggests that parametric models can fail to exhibit this level of robustness – a logistic regression model trained on the same data fails to adequately capture the (approximately) concave bid landscape shown in Figure 2a.

We mention in passing that IRMTs also offer a powerful new tool for customer demand modeling and personalized pricing. In these settings, the contextual variables x are features encoding the visiting customer, the decision p is the price of the offered product, and the response y is a binary indicator of whether the customer purchased the product at that price. IRMTs offer a non-parametric alternative for demand modeling which (1) naturally captures the monotonic (decreasing) relationship between product price and customer purchase probability through isotonic regression, and (2) finds an interpretable market segmentation driven by differences in customers' demand models.

3.2.2 Choice Model Trees (CMTs) for Ad-Response Prediction

Although not the primary focus of this work, we offer another application of MTPs in ad-response prediction settings, referred to as *Choice Model Trees* (CMTs). The CMT segments users on the basis of available demographic information (e.g., age and location) as well as any activity history on the site (e.g., prior purchases or search queries). Within each leaf, a Multinomial Logit (MNL) choice model estimates the ad-response behavior of the users contained in that market segment. Let $p = \{p_k\}_{k\in[K]}$ denote the collection of feature vectors encoding an offered assortment of K ads, with $p_k \in \mathbb{R}^d$ representing the feature vector encoding ad $k \in [K] := \{1,...,K\}$ in the assortment. If the ads correspond to different products, for example, then the elements of p_k might include the displayed products' price, color and brand. Let $y \in \{0,1,\dots,k\}$ denote the user's ad response when being presented with the assortment p, where 0 corresponds to the "no-response" alternative. Each leaf l of the CMT contains an MNL instance, $f_l(y|p)$, estimating the probability of each ad-response outcome y given the features p describing the assortment of ads. Given the parameters $\beta_l \in \mathbb{R}^d$ of the MNL model in leaf l, the probability of observing each ad-response outcome is:

$$\begin{cases} f_l\,(y=k \mid p) = \dfrac{e^{\beta_l^T p_k}}{1+\sum_{h\in[K]} e^{\beta_l^T p_h}}, \forall\, k \in [K] \\ f_l\,(y=0 \mid p) = \dfrac{1}{1+\sum_{h\in[K]} e^{\beta_l^T p_h}} \end{cases} \tag{1}$$

We refer the reader to (Train 2009, Chap. 3) for a derivation of these probabilities from random-utility maximization principles.

3.3. Training Procedure

We present an algorithm for training the MTPs outlined in Section 3.2. Assume there are n training set observations, and denote the collection of all such observations by $[n] = \{1,\dots,n\}$. Let $i \in [n]$ denote an individual observation which consists of a context x_i, decision p_i, and response y_i. The training algorithm is fed the data $\{(x_i,p_i,y_i)\}_{i\in[n]}$ and learns (1) a segmentation of the contextual features x_i, and (2) the response models $f_l(y|p)$ within each segment. In Section 3.3.1, we first tackle problem (2) in isolation, showing how the final response models are optimized to accurately estimate the distributions of responses given decisions in each leaf. We then propose in Section 3.3.2 a training procedure for learning the market segmentation, which is driven by optimizing the accuracy of the resulting collection of response models.

3.3.1 Learning the Response Models

In what follows, we denote by $R_l \subseteq [n]$ the subset of training set observations which belong to leaf l of the MTP, and we designate by $f_l(y|p)$ the corresponding response model. Given a class $\mathcal{F}$ of response models, the goal is to find the best response model $f_l \in \mathcal{F}$ which most accurately models the data $\{(p_i,y_i)\}_{i\in R_l}$. Define a loss function $\ell(p_i,y_i;f_l)$ which penalizes discrepancies between the observed response y_i and the predicted response distribution $f_l(y|p_i)$. We assume that this loss function is *additive*, i.e. the loss incurred on the entire training data should be interpreted as the sum of the prediction losses for each individual observation. Then, the response model is trained through solving the following optimization problem:

$$\mathcal{L}(R_l) := \min_{f_l\in\mathcal{F}} \sum_{i\in R_l} \ell(p_i,y_i;f_l) \tag{2}$$

To tailor our MTP training algorithm to specific applications, the practitioner simply needs to specify a class of response models $\mathcal{F}$ and a loss function $\ell(p_i, y_i; f_l)$ for evaluating models $f_l \in \mathcal{F}$. Below are examples for how these would be defined for the CMT and IRMT models:

- **IRMT**: Since the response y_i is binary, then without loss of generality we may identify $\mathcal{F}$ as a class of functions $f_l(p)$ estimating the probability of $y=1$ given the user belongs to segment (leaf) l. Isotonic regression fits a monotonically increasing function to the training data which minimizes mean-squared error. Consequently, we define $\mathcal{F}$ as the set of all monotonically-increasing functions $f_l : \mathbb{R} \to [0,1]$, and the loss function is defined as $\ell(p_i, y_i; f_l) := \big(y_i - f_l(p_i)\big)^2$.
- **CMT**: The class of response models $\mathcal{F}$ are the set of MNL choice models characterized by coefficients $\beta \in \mathbb{R}^d$ that satisfy Equation (1). MNL models are typically trained using the loss function of negative log-likelihood, defined as $\ell(p_i, y_i; f_l) := -\log(f_l(y = y_i | p_i))$.

3.3.2 Learning the Segmentation

We now describe our market segmentation algorithm. From Equation (2), $\mathcal{L}(R_l)$ represents the total loss when training a response model on the collection of observations R_l. The goal of our market segmentation algorithm is to find the MTP which segments the data into L leaves, $R_1, ..., R_L$, whose response models collectively minimize training set loss:

$$\min_{(R_1,\ldots,R_L) \in \mathcal{P}(n)} \sum_{l=1}^{L} \mathcal{L}(R_l) \ , \tag{3}$$

where $\mathcal{P}(n)$ is the collection of partitions satisfying $\bigsqcup_l R_l = [n]$.

It is well-known that this optimization problem is NP-Complete, since training optimal classification trees is a special case which is known to be NP-Complete (Laurent and Rivest 1976).[2] Thus, we rely on a technique known as *recursive partitioning* to approximate an optimal market segmentation. The procedure is directly analogous to the CART algorithm for greedily training classification trees, recursively finding the best decision-tree split with the smallest loss across the resulting leaves.

Denote the j-th attribute of the i-th context by $x_{i,j}$. Starting with all of the data, consider a decision tree split (j, s) encoded by a splitting variable j and split point s which partitions the data into two leaves:

$$R_1(j,s) = \{i \in [n] \mid x_{i,j} \le s\} \ \text{and} \ R_2(j,s) = \{i \in [n] \mid x_{i,j} > s\} \, ,$$

if variable j is numeric, or

$$R_1(j,s) = \{i \in [n] \mid x_{i,j} = s\} \ \text{and} \ R_2(j,s) = \{i \in [n] \mid x_{i,j} \ne s\} \, ,$$

[2] To formulate a classification tree as an MTP, let each response model map to a constant $K \in \{0,1\}$ and define the loss function as classification loss.

if variable j is categorical. We wish to find the decision tree split (j, s) which minimizes the following optimization problem:

$$\min_{j,s} \mathcal{L}(R_1(j,s)) + \mathcal{L}(R_2(j,s))$$

This can be solved via an exhaustive search over all potential splitting variables and split points, choosing the split (j, s) which achieves the best value of the objective function. After a split is selected in this manner, the procedure is then recursively applied in the resulting leaves until a stopping criteria is met. Examples of stopping criteria include a maximum tree depth limit or a minimum number of training set observations per leaf. To prevent overfitting, the CART pruning technique detailed in Breiman et al. (1984) can be applied to the MTP using a held-out validation set of data. To avoid lengthy technical details, we refer the reader to (Breiman et al. 1984) for an in-depth description of the pruning method.

3.3.3 Code Base for Training MTPs

We provide an open-source implementation of this training procedure in Python (Aouad et al. [n. d.]). The implementation is general, allowing practitioners to specify the class of response models $\mathcal{F}$, loss function $\ell(p_i, y_i; f_l)$, and response model training procedure which is best suited for their particular applications. Our code offers several features for improved scalability on high-dimensional data sets. First, we develop a parallelization strategy to leverage distributed computational resources. The idea is that the generation of distinct subtrees of the MTPs are independent sub-problems that can be computed in parallel, in a fully asynchronous manner. Hence, as our learning algorithm proceeds with nodes located at a larger depth of the tree, we are able to effectively distribute the computational load across the available resources. Second, we take advantage of warm-starts to reduce the number of gradient descent iterations needed to calibrate the response models in each node. Specifically, the parameter estimates of the parent's response model are provided as initial conditions for the gradient descent algorithm when fitting the response models of each of its children. Among all response models computed in the tree, parent nodes are arguably the most similar and informative estimates available. We find that the warm-starts significantly reduce the overall computational cost associated with learning the response models. Finally, our code has a mini-batch option, allowing to sample a relatively small portion of the population within each node in order to fit the corresponding response model. The batch size is tuned as a function of the characteristics of the response model. For example, in the context of IRMTs for the bid-landscape prediction problem, the response models take the form of univariate relationships (auction win probability as a function of bid price), which can be estimated with sufficient accuracy from relatively few observations (we observe empirically that observations in the order of 10,000, sampled uniformly at random, are sufficient to obtain accurate estimates). Similarly, in the context of CMTs, the choice models are calibrated using a stochastic gradient descent method, where the batch size and the number of iterations can be determined in each application on the basis of the number of choice alternatives and alternative-specific features.

4. Case Study

We evaluate the empirical performance of our MTP methodology on several data sets. Specifically, we apply the IRMT algorithm for bid landscape forecasting on historical bidding data from three separate ad exchanges, and we find that the IRMT algorithm achieves a 5-29% decrease in out-of-sample prediction error over a leading DSP provider's approach across all 21 days of testing data (details below).

For each ad exchange (referred to as exchanges 1, 2, and 3), an IRMT is trained on a data set of historical bids submitted by the DSP between 1/13/2019 and 1/24/2019, which amount to a training set of 60-370 million bids per exchange. The IRMT is pruned using a validation set consisting of 15% of the training data. Finally, the IRMT is evaluated on test sets of bids submitted between 1/25/2019 and 1/31/2019 amounting to 40-160 million bids per exchange.

Each observation in the data is encoded by (1) the ad-spot auction features available to the bidder, (2) the submitted bid price, and (3) the auction outcome (win/loss). There are ten auction features used as contexts for segmentation which can be categorized as follows:

- **Information regarding the ad spot**: Area and aspect ratio of the ad spot, ad spot fold position, and ID of the encompassing site. Due to the high dimensionality of the site IDs (with thousands of unique values per exchange), we first pre-cluster the site IDs before applying the IRMT and the benchmark algorithms to the training data.
- **Information regarding the site visit**: Time-of-day and day-of-week of site visit, country of visiting user, and ad channel (e.g., video, mobile, search).
- **Information regarding private marketplace deals**: ID encoding a private deal between an advertiser and a publisher which might affect the dynamics of the auction.

The IRMT algorithm's predictive performance is compared with several benchmarks trained and tested on the same data sets:

- **Const**: A model which predicts a constant win probability for all bid prices equal to the average training set win rate.
- **IR**: An isotonic regression model fit on the entire training set to estimate the auction win rate given the submitted bid price. This is a "context-free" model and does not incorporate the auction features; thus, the discrepancy between this model's performance and the IRMT's illustrates the value of segmentation in bid landscape forecasting. The assumption that auction features have a negligible effect on bid landscape forecasting accuracy has some precedent in the prior literature (Zhang et al. 2014).

- **IRKM**: Performs k-means clustering on the auction features and then fits an isotonic regression model within each cluster. The number of clusters k is chosen by optimizing performance on the same validation set used for IRMT pruning. K-means clustering is a common approach for market segmentation; this benchmark segments auctions based on feature dissimilarity rather than differences in their estimated bid landscapes.
- **DSP**: The bid landscape forecasting model which the DSP used in production during the testing period (1/25/2019-1/31/2019), which was also trained using the same data as our training set.

We also include additional algorithmic benchmarks testing the impact of using logistic regression models for bid landscape forecasting as opposed to isotonic regression models. Logistic regression is one of the most common parametric approaches for probabilistically modeling binary response data. The benchmark LR fits a single, "context-free" logistic regression model to the entire data; the benchmark LRKM performs k-means clustering on the auction features and fits a logistic regression model in each cluster; and the benchmark LRMT runs our MTP algorithm with logistic regression leaf models. These benchmarks are directly analogous to the IR, IRKM, and IRMT benchmarks, respectively.

The IRMT was trained on each exchange separately using our open-source Python implementation, specifying a minimum leaf size of 10000 observations and no depth limit. The training procedure terminated after 12-35 hours of computational time across the three exchanges. The tree was then pruned on a validation set, taking 6-35 minutes to complete per exchange. The final IRMTs were of depths 52-78 and contained 800-4100 leaves depending on the exchange. The reasonable computation times of our training and pruning procedures illustrate the scalability of our implementation when presented with large-scale high-dimensional data.

The IRMT and trained benchmarks were evaluated on the test data set using mean-squared-error (MSE), which measures the average squared difference between the algorithms' win probability estimates and the realized auction outcomes.[3] The test set MSEs obtained by the algorithms are given in Table 1, in which we report (1) overall MSE measured across the entire test data, and (2) the MSEs for each individual day of test data (1/25/19-1/31/19). The algorithms were also compared on the basis of their test-set ROC curves using the AUC (area under curve) metric. The ROCs and AUCs obtained by the algorithms are given in Figure 3.

The IRMT achieves a lower prediction error than any of the benchmarks in all settings tested. In particular, the IRMT attains a lower MSE than any benchmark for each of the 21 individual days of test data. The IRMT achieves a 5-29% improvement in overall MSE and 2-14% improvement in AUC over the DSP's approach across the three exchanges. The IRMT also achieves a 7-13%/7-15%

[3] This metric is also referred to as *Brier score* and is a proper scoring rule for evaluating probabilistic predictions.

Table 1 Test set mean squared errors (MSEs) of our algorithm (IRMT) and the benchmarks on three ad exchanges. The column "Avg." measures the average MSE across the entire test set, and the column "% Imp." measures the percentage improvement (decrease) in MSE from the IRMT relative to each benchmark.

(a) Test set MSEs: Exchange 1

Model	1/25	1/26	1/27	1/28	1/29	1/30	1/31	Avg.	% Imp.
IRMT	**0.0465**	**0.0476**	**0.0432**	**0.0474**	**0.0482**	**0.0539**	**0.0482**	**0.0480**	
LRMT	0.0508	0.0508	0.0458	0.0504	0.0523	0.0588	0.0521	0.0518	7.3%
Const	0.0613	0.0613	0.0552	0.0599	0.0626	0.0718	0.0631	0.0625	23%
IR	0.0538	0.0545	0.0492	0.0529	0.0540	0.0619	0.0550	0.0546	12%
LR	0.0586	0.0584	0.0526	0.0571	0.0590	0.0680	0.0597	0.0593	19%
IRKM	0.0489	0.0497	0.0446	0.0488	0.0494	0.0556	0.0497	0.0497	3.4%
LRKM	0.0535	0.0540	0.0478	0.0522	0.0536	0.0603	0.0536	0.0537	11%
DSP	0.0564	0.0558	0.0508	0.0560	0.0569	0.0640	0.0592	0.0572	16%

(b) Test set MSEs: Exchange 2

Model	1/25	1/26	1/27	1/28	1/29	1/30	1/31	Avg.	% Imp.
IRMT	**0.0276**	**0.0253**	**0.0341**	**0.0318**	**0.0366**	**0.0419**	**0.0405**	**0.0339**	
LRMT	0.0301	0.0273	0.0368	0.0344	0.0393	0.0450	0.0437	0.0366	7.3%
Const	0.0316	0.0285	0.0391	0.0364	0.0414	0.0471	0.0451	0.0384	12%
IR	0.0305	0.0275	0.0371	0.0349	0.0397	0.0449	0.0432	0.0368	7.9%
LR	0.0320	0.0287	0.0394	0.0366	0.0417	0.0473	0.0455	0.0387	12%
IRKM	0.0281	0.0258	0.0345	0.0321	0.0369	0.0423	0.0408	0.0343	1.2%
LRKM	0.0306	0.0278	0.0372	0.0347	0.0396	0.0453	0.0440	0.0370	8.4%
DSP	0.0296	0.0285	0.0377	0.0341	0.0379	0.0428	0.0416	0.0359	5.6%

(c) Test set MSEs: Exchange 3

Model	1/25	1/26	1/27	1/28	1/29	1/30	1/31	Avg.	% Imp.
IRMT	**0.1200**	**0.1090**	**0.1098**	**0.1184**	**0.1230**	**0.1311**	**0.1268**	**0.1199**	
LRMT	0.1375	0.1198	0.1203	0.1303	0.1347	0.1386	0.1347	0.1310	8.5%
Const	0.1591	0.1361	0.1422	0.1510	0.1521	0.1631	0.1587	0.1520	21%
IR	0.1396	0.1232	0.1291	0.1348	0.1396	0.1500	0.1425	0.1372	13%
LR	0.1478	0.1262	0.1318	0.1418	0.1459	0.1567	0.1501	0.1431	16%
IRKM	0.1307	0.1155	0.1182	0.1267	0.1318	0.1408	0.1346	0.1285	6.7%
LRKM	0.1419	0.1208	0.1275	0.1371	0.1386	0.1498	0.1443	0.1373	13%
DSP	0.1661	0.1662	0.1759	0.1605	0.1646	0.1724	0.1763	0.1689	29%

improvement in MSE/AUC relative to the IR benchmark and a 1-7%/0.6-5% improvement relative to IRKM. The strong performance of IRMT over IR demonstrates the value of segmentation in bid landscape forecasting. Moreover, the superior performance of IRMT over IRKM illustrates the utility of applying more supervised segmentation procedures driven by accurately capturing differences in the underlying segments' bid landscapes. Notably, each benchmark using isotonic regression achieves better empirical performance than its logistic regression counterpart; in particular, our IRMT algorithm achieves a 7.3-8.5% improvement in mean-squared-error over the LRMT benchmark. This

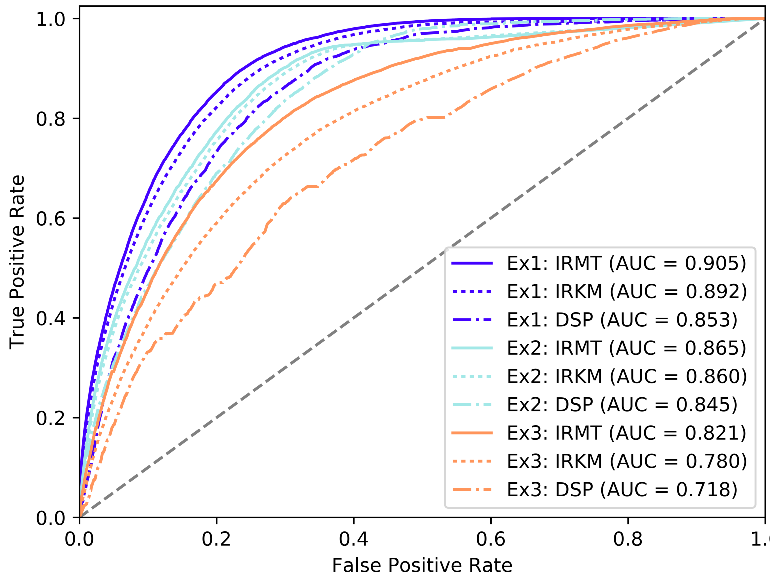


Figure 3: Test set ROC curves and AUCs of our algorithm (IRMT) and the benchmarks on three ad exchanges. The benchmark IR, not shown in the figure due to space constraints, achieved AUCs of 0.844, 0.776, and 0.716 on exchanges 1,2, and 3, respectively.

finding illustrates that isotonic regression models can offer substantial improvements in predictive accuracy over other parametric approaches for bid landscape forecasting.

Acknowledgments

Elmachtoub and McNellis were partially supported by NSF grant CMMI-1763000.

References

Ali Aouad, Adam N Elmachtoub, Kris Ferreira, and Ryan McNellis. [n. d.]. GitHub repository. ([n. d.]). `https://github.com/rtm2130/MTP`

Fernando Bernstein, Sajad Modaresi, and Denis Sauré. 2018. A dynamic clustering approach to data-driven assortment personalization. *Management Science* (2018).

Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. 2019. Optimal prescriptive trees. *INFORMS Journal on Optimization* (2019), ijoo–2018.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press, Chapter 10, 279–294.

HD Brunk. 1970. Estimation of isotonic regression. nonparametric Techniques in Statistical Inference. *Cambridge Univ. Press* 177 (1970), 195.

Kin-Yee Chan and Wei-Yin Loh. 2004. LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics* 13, 4 (2004), 826–852.

Olivier Chapelle, Eren Manavoglu, and Romer Rosales. 2015. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 61.

Ying Cui, Ruofei Zhang, Wei Li, and Jianchang Mao. 2011. Bid landscape forecasting in online ad exchange marketplace. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 265–273.

David Lee Hanson, Gordon Pledger, FT Wright, et al. 1973. On consistency in monotonic regression. *The Annals of Statistics* 1, 3 (1973), 401–421.

Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. 2017. A Model-based Projection Technique for Segmenting Customers. *arXiv preprint arXiv:1701.07483* (2017).

Nathan Kallus. 2017. Recursive partitioning for personalization using observational data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1789–1798.

Nathan Kallus and Madeleine Udell. 2016. Revealed preference at scale: Learning personalized preferences from assortment choices. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 821–837.

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine learning* 59, 1-2 (2005), 161–205.

Hyafil Laurent and Ronald L Rivest. 1976. Constructing optimal binary decision trees is NP-complete. *Information processing letters* 5, 1 (1976), 15–17.

Quan Lu, Shengjun Pan, Liang Wang, Junwei Pan, Fengdan Wan, and Hongxia Yang. 2017. A Practical Framework of Conversion Rate Prediction for Online Display Advertising. In *Proceedings of the ADKDD'17*. ACM, 9.

H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1222–1230.

Velibor V Mišić. 2016. *Data, models and decisions for large-scale stochastic optimization problems*. Ph.D. Dissertation. Massachusetts Institute of Technology.

Waheed Noor, Matthew N Dailey, and Peter Haddawy. 2014. Learning Predictive Choice Models for Decision Optimization. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1932–1945.

John R Quinlan et al. 1992. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, Vol. 92. World Scientific, 343–348.

Sarah Sluis. 2019. Google Switches To First-Price Auction. AdExchanger. (2019). `https://adexchanger.com/online-advertising/google-switches-to-first-price-auction/`

Kenneth E Train. 2009. *Discrete choice methods with simulation*. Cambridge university press, Chapter 2, 23–25.

Yuchen Wang, Kan Ren, Weinan Zhang, Jun Wang, and Yong Yu. 2016. Functional bid landscape forecasting for display advertising. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 115–131.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17, 2 (2008), 492–514.

Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1077–1086.