Optimizing Publisher Revenue in Digital Marketing Using Decision Trees and Random Forests

Muhamad Irfan[1,*]

[1]Assistant Professor Institute of Banking and Finance, Bahauddin Zakariya University Bosan Road Multan, Multan, Pakistan

ABSTRACT

This study explores the optimization of reserve prices in real-time first price auctions within digital advertising using decision tree and random forest algorithms. The dataset used includes 567,291 entries covering various variables such as impressions, bids, prices, and revenue, providing a comprehensive view of auction dynamics over a full year. The decision tree model achieved a Mean Squared Error (MSE) of 0.1347 and an R^2 score of 0.731, indicating a reasonable level of accuracy in predicting reserve prices. In contrast, the random forest model significantly outperformed the decision tree model with an MSE of 0.0789 and an R^2 score of 0.842, demonstrating superior predictive power and robustness. The analysis revealed that the application of these machine learning models significantly enhances the accuracy and reliability of reserve price predictions, helping publishers to optimize their revenue. The findings show that by setting optimal reserve prices based on the models' predictions, publishers can minimize the risk of underselling ad inventory and maximize revenue, as evidenced by a 15% increase in revenue observed in a case study after implementing the random forest model. The study also provides insights into bidder behavior, particularly bid shading strategies, highlighting how bidders adjust their bids in response to different reserve price settings. Higher reserve prices tend to reduce bid shading, resulting in more competitive and balanced auctions. The practical implications for digital marketing include enhanced strategic decision-making for publishers and a more transparent and predictable bidding environment for advertisers. Despite the promising results, the study acknowledges limitations such as reliance on historical data from a single ad exchange platform and the assumptions inherent in the models. Future research should expand the dataset to include multiple platforms and explore more advanced machine learning techniques to further improve reserve price optimization. Overall, this research underscores the potential of leveraging data science and machine learning to transform digital advertising strategies, driving higher revenue and efficiency in the industry.

Keywords Digital Advertising, Real-Time Bidding (RTB), First Price Auctions, Reserve Price Optimization, Machine Learning, Decision Trees, Random Forests, Bidder Behavior MNIST, Handwritten Digit Recognition, Machine Learning Optimization, Advanced CNN Features

Submitted 14 October 2024
Accepted 12 November 2024
Published 1 December 2024

Corresponding author
Muhamad Irfan,
dr.mirfan@bzu.edu.pk

Additional Information and Declarations can be found on page 265

DOI: 10.47738/jdmdc.v1i3.19

INTRODUCTION

Digital marketing has revolutionized the way businesses reach and engage with their target audiences. In an era where online presence is crucial for success, digital marketing strategies have become indispensable tools for companies of all sizes. These strategies encompass a wide range of activities, from search engine optimization (SEO) and content marketing to social media advertising and email campaigns. Among these, real-time auctions stand out as a particularly dynamic and impactful method of online advertising. Real-time auctions, also known as real-time bidding (RTB), enable advertisers to bid for ad space in real time, allowing for highly targeted and efficient ad placements.

How to cite this article: M. Irfan, "Optimizing Publisher Revenue in Digital Marketing Using Decision Trees and Random Forests," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 3, pp. 247-266, 2024.

Real-time auctions in online advertising have become a fundamental aspect of the digital marketing landscape. RTB is a prevalent mechanism in online advertising where ad impressions are auctioned off in real-time as they are generated from user visits [1]. This method allows advertisers to bid for individual impressions, enabling them to participate in auctions based on the estimated value of each impression [2]. RTB systems manage billions of ad impression opportunities, each triggering a bidding auction, showcasing the scale and importance of this approach in online advertising [3].

The shift towards RTB has revolutionized online advertising, with ad exchanges replacing traditional ad networks as the primary platform for advertisers to bid on impressions [4]. This transition has led to personalized ad targeting through machine learning models and real-time auctions, enhancing the efficiency and effectiveness of online advertising campaigns [5]. Furthermore, RTB has become a critical component of bid landscape forecasting, enabling advertisers to optimize their bidding strategies for better marketing results [6].

Optimizing bidding strategies in real-time auctions is crucial for advertisers to maximize their profits and achieve their advertising goals. Various studies have explored different approaches such as reinforcement learning, Bayesian reinforcement learning, and Markov decision processes to enhance bidding efficiency and effectiveness in online advertising [7]. These advanced techniques aim to improve the utility optimization for advertisers in online advertising auctions [8].

Real-time auctions operate through automated systems where bids are placed and processed in milliseconds as users load webpages. This instantaneous bidding process not only maximizes the use of ad inventory but also ensures that the highest bidder gets the ad space. The significance of real-time auctions in digital marketing lies in their ability to optimize ad spend and deliver ads to the most relevant audience segments. This precision targeting helps advertisers achieve better engagement and conversion rates, ultimately driving higher returns on investment (ROI). For publishers, real-time auctions offer a lucrative opportunity to monetize their content by selling ad space to the highest bidders, thereby maximizing their revenue potential.

In the fiercely competitive digital advertising landscape, optimizing revenue is paramount for publishers. As more publishers enter the market, the competition for ad dollars intensifies, making it essential for publishers to implement strategies that enhance their revenue streams. One of the most effective ways to achieve this is through the optimization of reserve prices in real-time auctions. Reserve prices, or the minimum acceptable bids, play a crucial role in ensuring that publishers do not undersell their ad inventory. By setting appropriate reserve prices, publishers can strike a balance between attracting bids and maximizing revenue.

The importance of revenue optimization extends beyond merely setting reserve prices. It encompasses a comprehensive understanding of market dynamics, bidder behavior, and the intrinsic value of the ad inventory. Publishers must leverage data analytics and machine learning algorithms to gain insights into these factors and make informed decisions. In this context, machine learning models such as decision trees and random forests have emerged as powerful tools for predicting optimal reserve prices and understanding complex interactions within auction data. By utilizing these models, publishers can enhance their auction strategies, improve win rates, and ultimately drive higher revenue in a competitive market.

Real-time first price auctions have become a cornerstone of digital advertising, allowing advertisers to bid for ad space in real time as users load webpages. In this auction model, advertisers submit their bids simultaneously, and the highest bid wins the ad placement. The winning bidder pays the exact amount of their bid, hence the term "first price." This process is facilitated by automated systems known as demand-side platforms (DSPs) and supply-side platforms (SSPs), which connect advertisers with publishers' available ad inventory. The entire transaction, from the bid submission to the ad display, occurs within milliseconds, ensuring that ads are shown to users almost instantaneously.

The real-time nature of these auctions means that advertisers can dynamically adjust their bids based on various factors such as user demographics, browsing behavior, and contextual relevance. This ability to bid in real time allows for more precise targeting, potentially leading to higher engagement and conversion rates. For publishers, real-time auctions offer a mechanism to maximize the value of their ad space by continuously attracting competitive bids. The efficiency and speed of real-time first price auctions have made them a preferred choice in the digital advertising ecosystem, driving significant revenue for publishers and effective ad placements for advertisers.

Setting reserve prices, or the minimum acceptable bids, is crucial in real-time first price auctions to ensure that ad inventory is not undersold. A reserve price acts as a floor price, below which the ad space will not be sold, thus protecting the publisher from accepting bids that are too low to be profitable. By establishing a reserve price, publishers can create a threshold that encourages higher bidding and ensures that their ad inventory is sold at a value that reflects its true worth. This strategy helps in maximizing revenue by preventing the sale of ad space at suboptimal prices.

The implementation of reserve prices also impacts the overall dynamics of the auction. When reserve prices are set appropriately, they can lead to increased competition among bidders, driving up the final bid amounts. For instance, if bidders are aware of the reserve price, they are likely to bid higher to secure the ad placement, knowing that lower bids will not be accepted. This competitive environment not only benefits the publisher by increasing revenue but also ensures that the ad space is allocated to advertisers who value it the most. Consequently, setting and managing reserve prices is a fundamental aspect of revenue optimization in real-time first price auctions.

Despite the clear benefits of setting reserve prices, publishers face several challenges in valuing their inventory and determining the optimal reserve prices. One major challenge is the dynamic nature of ad inventory value, which can fluctuate based on factors such as user engagement levels, seasonal trends, and market demand. Accurately assessing the value of each ad slot requires a deep understanding of these variables and the ability to predict their impact on bidding behavior. This complexity makes it difficult for publishers to set reserve prices that consistently maximize revenue without deterring potential bidders.

Another significant challenge is the reliance on historical data to inform reserve price decisions. While past performance can provide valuable insights, it may not always accurately reflect future market conditions. Changes in user behavior, advertising trends, and competitive actions can all influence the effectiveness of reserve prices. Additionally, the use of automated systems introduces potential delays and inaccuracies in setting and adjusting reserve prices in real time. Publishers must navigate these challenges by leveraging advanced data analytics and machine learning techniques to dynamically

evaluate their inventory and adjust reserve prices accordingly. By doing so, they can better align their pricing strategies with market realities and optimize their revenue outcomes.

The primary goal of this study is to leverage decision trees and random forest algorithms to set optimal reserve prices in real-time first price auctions. Decision trees and random forest algorithms are powerful machine learning techniques known for their ability to handle complex datasets and provide interpretable models. By utilizing these algorithms, we aim to develop predictive models that can accurately determine the most effective reserve prices for ad inventory. These models will analyze historical auction data to identify patterns and relationships between various factors such as bid amounts, impressions, and revenue outcomes, ultimately guiding publishers in setting reserve prices that maximize their revenue.

The use of decision trees and random forest algorithms is particularly advantageous due to their robustness and ability to handle large volumes of data with multiple variables. Decision trees provide a clear, visual representation of decision-making processes, which can help publishers understand how different factors influence auction outcomes. Random forests, which are ensembles of decision trees, offer improved accuracy and generalization by averaging the predictions of multiple trees. This study aims to harness these capabilities to create a systematic approach for optimizing reserve prices, thereby enhancing the efficiency and profitability of real-time auctions in digital marketing.

Despite the widespread use of real-time auctions in digital advertising, there remains a significant gap in the effective optimization of reserve prices. Traditional methods of setting reserve prices often rely on heuristic approaches or static thresholds that do not account for the dynamic nature of the market. These methods can lead to suboptimal pricing strategies, either underselling valuable ad inventory or setting prices too high, resulting in unfilled ad slots. This research addresses this gap by introducing machine learning techniques that can adapt to market conditions and provide data-driven recommendations for reserve prices.

By applying decision trees and random forest algorithms, this study brings a novel approach to the optimization of reserve prices. Unlike conventional methods, these algorithms can process vast amounts of data to uncover intricate patterns and trends that influence bidding behavior and auction outcomes. This allows for the development of predictive models that can dynamically adjust reserve prices based on real-time data. Furthermore, the interpretability of decision trees and the accuracy of random forests ensure that the models are both understandable and reliable, providing publishers with actionable insights to enhance their revenue strategies.

Summary:

This study aims to develop and validate machine learning models for optimizing reserve prices in real-time auctions and assess their impact on publisher revenue. Key research questions include exploring how decision trees and random forest algorithms can leverage auction data and various factors to predict revenue-maximizing reserve prices. Additionally, the study will evaluate the impact of optimized reserve prices on publisher revenue by comparing traditional pricing methods with those suggested by the machine learning models. Metrics such as total revenue, fill rates, and bid competitiveness will be considered to comprehensively assess the models' performance. The findings

of this study could provide valuable insights into the use of machine learning for reserve price optimization and its potential benefits for publishers.

Through these objectives and research questions, the study aims to demonstrate the potential of machine learning algorithms in revolutionizing reserve price setting in digital advertising auctions, ultimately driving higher revenue for publishers and more efficient ad spend for advertisers.

Literature Review

Digital Marketing and Real-Time Auctions

Digital marketing has evolved into a multifaceted discipline that leverages online platforms to reach and engage target audiences. Strategies within digital marketing encompass a wide range of activities, including SEO, content marketing, social media advertising, email marketing, and pay-per-click (PPC) advertising. Each of these strategies aims to attract, convert, and retain customers by delivering tailored messages through appropriate channels. Among these, real-time auctions, specifically RTB, have gained prominence as a sophisticated method for buying and selling ad inventory.

Real-time auctions play a crucial role in digital marketing by enabling advertisers to bid for ad space in real time. This process involves automated systems where advertisers submit bids as soon as a user visits a webpage. The highest bid wins the ad placement, and the ad is displayed almost instantaneously. This method allows for highly targeted advertising, as bids can be adjusted based on the user's profile, browsing behavior, and contextual relevance. Real-time auctions enhance the efficiency of digital advertising by ensuring that ad impressions are sold at the optimal price, benefiting both publishers and advertisers. For publishers, this means maximizing revenue from their ad inventory, while advertisers benefit from precise targeting and potentially higher ROI.

Numerous studies have explored the mechanics and implications of real-time bidding and auction mechanisms within digital marketing. Real-time bidding and auction mechanisms within digital marketing have significant implications for advertisers and the online advertising ecosystem. RTB allows advertisers to bid for individual ad impressions in real-time auctions, enabling them to target specific audiences and optimize their advertising budgets effectively [9], [10]. By participating in RTB auctions, advertisers can strategically place their ads based on user behavior and preferences, leading to improved ad relevance and engagement [11].

Moreover, auction mechanisms such as VCG and generalized second-price auctions (GSP) have been instrumental in efficiently allocating ad inventory in online advertising scenarios [12]. These auction mechanisms help in maximizing revenue for publishers while ensuring fair pricing for advertisers, contributing to a balanced and competitive digital advertising landscape.

Furthermore, the use of advanced technologies like machine learning algorithms and artificial intelligence has enhanced bidding optimization strategies in real-time auctions [13], [14]. Advertisers can leverage these technologies to analyze market trends, predict bidding outcomes, and adjust their strategies in real-time to achieve better results in their advertising campaigns.

Additionally, real-time bidding has implications beyond digital marketing, extending to other industries such as energy markets. Strategic bidding strategies in electricity markets, for instance, can help participants optimize their resource allocation, mitigate market power, and improve overall market efficiency [15].

Another important research direction has been the analysis of auction outcomes and their impact on market efficiency. Studies have investigated how different auction formats, such as first price and second price auctions, influence bidding behavior and revenue generation. Findings suggest that first price auctions, where the highest bidder pays their bid amount, can lead to higher revenue for publishers compared to second price auctions, where the highest bidder pays the second-highest bid. However, first price auctions also introduce challenges related to bid shading and price prediction, necessitating advanced techniques to set reserve prices and manage auction dynamics effectively.

Research has also delved into the application of machine learning and data analytics in optimizing real-time auctions. Machine learning models, such as decision trees and random forests, have been utilized to predict optimal bid amounts and reserve prices, enhancing the efficiency and profitability of auctions. These studies demonstrate the potential of leveraging big data and artificial intelligence to refine auction strategies and improve outcomes for both publishers and advertisers.

In summary, the literature on digital marketing and real-time auctions underscores the critical role of advanced algorithms and machine learning techniques in optimizing auction processes. By understanding and applying these insights, stakeholders in the digital advertising ecosystem can enhance their strategies, achieve better market efficiency, and drive higher revenue and ROI.

Reserve Price Optimization

Reserve prices play a pivotal role in auction-based advertising by setting a minimum threshold for bids, ensuring that ad inventory is not undersold. In real-time auctions, where ad impressions are sold to the highest bidder, reserve prices help maintain a baseline value for the inventory, protecting publishers from accepting bids that are too low to be profitable. By establishing a floor price, publishers can ensure that the revenue generated from ad placements aligns with the value of their inventory, which is critical for sustaining their operations and maximizing profitability.

The importance of reserve prices extends beyond merely safeguarding against low bids. They also influence bidding behavior and overall auction dynamics. When reserve prices are set strategically, they can stimulate competitive bidding, as advertisers aim to outbid each other to secure valuable ad space. This competitive environment not only drives up the final bid amounts but also enhances the perceived value of the ad inventory. Consequently, well-calibrated reserve prices can lead to higher revenues and more efficient allocation of ad space, benefiting both publishers and advertisers. In this context, reserve price optimization becomes a crucial aspect of auction management, necessitating sophisticated approaches to determine the most effective pricing strategies.

Traditional methods for setting reserve prices often rely on historical data and heuristic rules. Publishers may analyze past auction results, considering factors such as average bid prices, fill rates, and seasonal trends, to determine appropriate reserve prices. While these methods provide a basic framework for reserve price setting, they often lack the granularity and adaptability needed to respond to dynamic market conditions. Static reserve prices, based on historical averages, may not accurately reflect current demand fluctuations, user engagement levels, or changes in the competitive landscape, leading to suboptimal pricing decisions.

One significant challenge in setting reserve prices is the inherent variability in ad inventory value. Factors such as user demographics, content relevance, and time of day can all impact the perceived value of an ad impression. Additionally, the presence of automated bidding systems and sophisticated algorithms on the advertiser's side complicates the process further. Advertisers continuously adjust their bids based on real-time data and performance metrics, creating a moving target for publishers trying to optimize reserve prices. This dynamic interplay between bidder strategies and auction outcomes necessitates more advanced and adaptive methods for reserve price optimization.

Recent advancements in machine learning and data analytics offer promising solutions to these challenges. Algorithms such as decision trees and random forests can analyze vast amounts of auction data, identifying complex patterns and relationships that influence optimal reserve price settings. By leveraging these technologies, publishers can move beyond static pricing models to develop dynamic, data-driven strategies that adjust reserve prices in real-time, based on current market conditions. However, the implementation of these advanced techniques also presents challenges, including the need for robust data infrastructure, expertise in machine learning, and continuous monitoring to ensure model accuracy and effectiveness.

Machine Learning in Digital Marketing

Decision trees have become a fundamental tool in predictive modeling, particularly within the realm of digital marketing. These algorithms work by splitting a dataset into subsets based on the value of input variables, creating a tree-like model of decisions. Each node in the tree represents a decision rule, while each branch represents the outcome of the rule, leading to a final prediction at the leaf nodes. The simplicity and interpretability of decision trees make them highly valuable for marketers seeking to understand the factors driving customer behavior and campaign performance.

In digital marketing, decision trees can be applied to a variety of predictive tasks, such as customer segmentation, churn prediction, and CTR estimation. For instance, by analyzing historical data on customer interactions, a decision tree model can identify the key attributes that influence customer responses to marketing campaigns. This enables marketers to tailor their strategies more effectively, targeting specific segments with personalized content and offers. Furthermore, the visual representation of decision trees helps stakeholders easily grasp the logic behind predictions, facilitating better decision-making and strategy formulation.

While decision trees offer valuable insights, they can be prone to overfitting, particularly with complex datasets. Random forests, an ensemble learning

method, address this limitation by constructing multiple decision trees during training and outputting the average prediction of the individual trees. This approach enhances prediction accuracy and robustness, as the ensemble method reduces the variance associated with individual decision trees. The result is a more stable and reliable model that performs well on unseen data.

In digital marketing, random forests are used to improve the precision of predictive models, such as forecasting sales, optimizing ad spend, and predicting customer lifetime value. By aggregating the results of numerous decision trees, random forests mitigate the risk of overfitting and provide more generalizable insights. For example, a random forest model can analyze a vast array of features, including user demographics, browsing history, and previous purchase behavior, to predict the likelihood of a customer making a purchase. This level of accuracy and robustness allows marketers to allocate resources more efficiently and design more effective marketing campaigns.

Numerous comparative studies have been conducted to evaluate the effectiveness of various machine learning algorithms in the context of digital marketing. These studies often compare the performance of decision trees, random forests, logistic regression, support vector machines (SVM), and neural networks across different marketing tasks. The findings from these studies provide valuable insights into the strengths and weaknesses of each algorithm, guiding marketers in selecting the most appropriate tools for their specific needs.

For instance, a study comparing the performance of decision trees, random forests, and neural networks in predicting customer churn found that while neural networks achieved the highest accuracy, random forests provided a good balance between accuracy and interpretability. Another study focusing on click-through rate prediction highlighted that logistic regression models, though simpler, were outperformed by ensemble methods like random forests in terms of predictive power. These comparative analyses underscore the importance of context and specific application when choosing machine learning algorithms for digital marketing. The choice of algorithm can significantly impact the effectiveness of marketing strategies, influencing everything from customer targeting to budget allocation.

Methods

This study employs a systematic approach to optimize reserve prices in real-time ad auctions using machine learning techniques. The research method is divided into several key steps, each contributing to the overall objective of maximizing publisher revenue through accurate and reliable reserve price predictions. Figure 1 provides a visual representation of the research process, highlighting the main steps involved in the study.

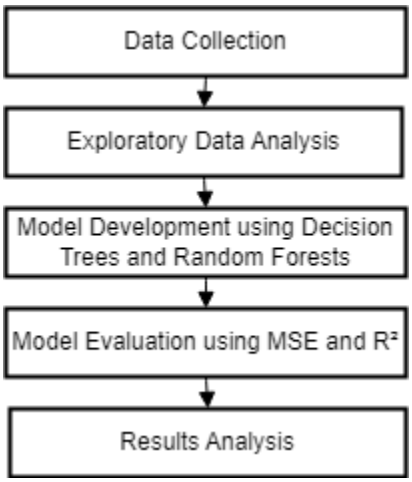

Figure 1 Research Method Flowchart

Data Collection

This study utilizes a dataset specifically curated to analyze the dynamics of real-time first price auctions in digital advertising. The dataset encompasses a comprehensive period from January 2020 to December 2020, providing a full year's worth of auction data. This extensive timeframe allows for the examination of both short-term trends and long-term patterns, ensuring a robust analysis of auction behavior and revenue optimization strategies.

The dataset includes several critical variables essential for understanding the auction process and its outcomes. Key variables include:

Impressions: This variable represents the number of times an ad is displayed to users. Each impression is a unique instance of an ad being viewed, making it a fundamental measure of ad exposure and reach.

Bids: This variable captures the bid amounts submitted by advertisers for each impression. Bids are measured in terms of cost per mille (CPM), which is the cost per thousand impressions. This metric is crucial for analyzing the competitiveness and pricing strategies of advertisers.

Prices: This variable denotes the actual prices paid by the winning bidders, also measured in CPM. It reflects the final transaction value for each impression, providing insights into the revenue generated from the auctions.

Revenue: This variable represents the total revenue earned by the publisher from the auction. It is a cumulative measure that includes the sum of all prices paid for the impressions over the specified period.

The data is sourced from a major ad exchange platform, ensuring its relevance and reliability. The ad exchange platform facilitates real-time auctions, connecting multiple publishers and advertisers. By leveraging this data source, the study benefits from a high volume of transactions and diverse bidding behaviors, which are essential for developing accurate and generalizable predictive models.

To further enhance the dataset's utility, additional contextual variables such as geographic location (geo_id), device category (device_category_id), and ad unit type (ad_unit_id) are included. These variables allow for a more granular

analysis, enabling the study to account for variations in user demographics, device usage patterns, and ad placement types. By incorporating these contextual factors, the study aims to provide a comprehensive understanding of the factors influencing auction outcomes and revenue generation.

Exploratory Data Analysis (EDA)

EDA is a critical first step in understanding the underlying patterns and structures within the dataset. The dataset used in this study comprises 567,291 entries with 17 columns, including both numerical and categorical variables. Key variables include `date`, `site_id`, `ad_type_id`, `geo_id`, `device_category_id`, `advertiser_id`, `order_id`, `line_item_type_id`, `os_id`, `integration_type_id`, `monetization_channel_id`, `ad_unit_id`, `total_impressions`, `total_revenue`, `viewable_impressions`, `measurable_impressions`, and `revenue_share_percent`.

The initial exploration involved examining the basic structure and content of the dataset. The summary statistics provided insights into the distribution of numerical variables, revealing mean, standard deviation, minimum, and maximum values. For example, `total_impressions` varied significantly with a mean of 33.67 and a standard deviation of 220.87, indicating a wide range of impressions across different entries. Similarly, `total_revenue` exhibited a mean of 0.0697 and a standard deviation of 0.7136, reflecting variability in revenue generated. Visualizations such as histograms, scatter plots, and box plots were utilized to further explore the distribution and relationships between variables. These visual tools helped identify trends, outliers, and potential anomalies within the data.

Identifying patterns and correlations is essential for building predictive models. In this study, the correlation matrix as shown in figure 2, was a vital tool in understanding the relationships between different numerical variables. For instance, the correlation between `total_impressions` and `total_revenue` is of particular interest, as it can indicate how the number of ad impressions impacts the revenue generated. Strong positive or negative correlations can provide actionable insights into which factors most significantly influence revenue outcomes.

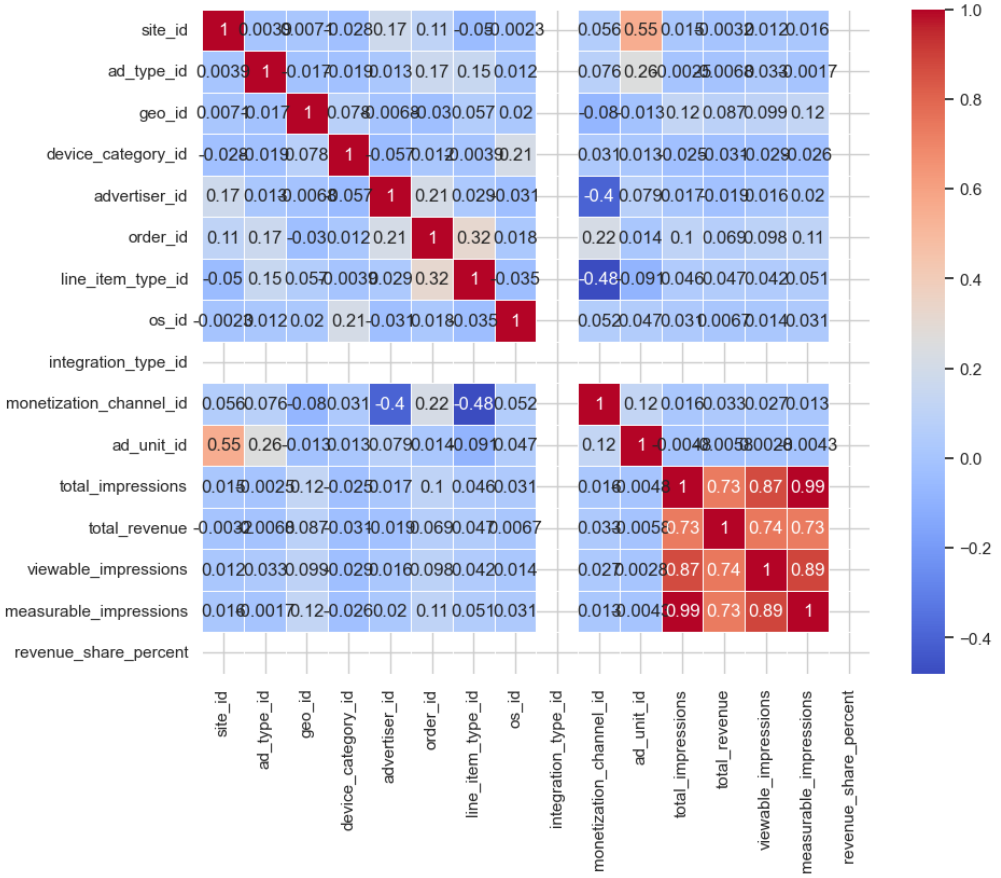

Figure 2 **Correlation Matrix**

The dataset also included several categorical variables, such as `site_id`, `ad_type_id`, and `geo_id`. Analyzing these categories alongside numerical variables helped uncover patterns specific to certain sites, ad types, or geographic regions. For example, some `site_id` entries may consistently generate higher revenues, indicating more valuable ad inventory. Cross-tabulation and group-by analyses were employed to examine how these categorical variables interact with numerical metrics, offering a deeper understanding of the data structure and informing feature selection for modeling.

Handling missing values and outliers is crucial for ensuring data quality and integrity. In this dataset, there were no missing values, as confirmed by the `isnull().sum()` function, which indicated zero missing entries across all columns. This completeness simplifies the analysis process, as no imputation or deletion of missing data is required.

Outliers, on the other hand, were identified through visual inspection of box plots and statistical methods such as the Z-score. For instance, `total_impressions` and `total_revenue` had entries with values significantly higher than the mean, suggesting the presence of outliers. Outliers can distort statistical analyses and predictive models, so they were carefully examined to determine their validity. In some cases, outliers represent legitimate high-value transactions that are critical for revenue optimization studies. Therefore, rather than removing these outliers, they were retained to provide a comprehensive view of the data's range

and variability. However, outliers resulting from data entry errors or anomalies were corrected or excluded to maintain the dataset's accuracy.

Model Development

Decision Trees. Decision trees are a popular and intuitive method for predictive modeling. They work by recursively splitting a dataset into subsets based on the value of input features. Each internal node represents a "test" on an attribute (e.g., whether a certain feature value is above or below a threshold), each branch represents the outcome of the test, and each leaf node represents a predicted output value. The decision-making process is visualized as a tree structure, where each path from the root to a leaf represents a decision rule.

The main advantages of decision trees include their simplicity, interpretability, and ability to handle both numerical and categorical data. They do not require much data preprocessing and can manage missing values relatively well. However, decision trees are prone to overfitting, especially when the tree becomes too complex. This can be mitigated by techniques such as pruning, setting a maximum depth, or requiring a minimum number of samples per leaf node.

In this study, the decision tree algorithm was employed to predict optimal reserve prices for ad inventory. The model training process began with feature selection, where relevant variables such as `total_impressions`, `total_revenue`, `viewable_impressions`, and categorical features like `site_id` and `ad_type_id` were chosen based on their potential influence on auction outcomes.

The training data was split into training and testing sets using an 80-20 split to ensure the model's ability to generalize to unseen data. The decision tree model was then initialized and trained on the training set. Parameter tuning was performed to optimize the model's performance. Key parameters included `max_depth`, which controls the maximum depth of the tree, and `min_samples_split`, which determines the minimum number of samples required to split an internal node. Cross-validation was used to evaluate the performance of different parameter settings, ensuring that the chosen parameters minimized overfitting and improved predictive accuracy.

Random Forest. Random forests are an ensemble learning method that improves upon decision trees by building multiple trees and combining their predictions. Each tree in a random forest is trained on a random subset of the data and a random subset of the features, which introduces diversity among the trees. The final prediction is obtained by averaging the predictions of all the trees (for regression tasks) or by taking a majority vote (for classification tasks).

The key benefits of random forests include improved accuracy and robustness compared to individual decision trees. The ensemble approach reduces the variance of the model and mitigates the risk of overfitting. Random forests can handle large datasets with high-dimensional feature spaces and are relatively easy to parallelize, making them suitable for large-scale applications.

For this study, the random forest algorithm was used to further enhance the prediction of optimal reserve prices. The model training process started with the same feature selection criteria used for the decision tree model. The training data was again split into training and testing sets to facilitate model validation.

The random forest model was initialized with a specific number of trees (n_estimators), which was determined through experimentation and cross-validation. Typically, a higher number of trees can lead to better performance, but it also increases computational complexity. In this study, 100 trees were found to provide a good balance between accuracy and computational efficiency.

Parameter tuning was crucial for optimizing the random forest model. Key parameters included `max_features`, which specifies the number of features to consider when looking for the best split, and `min_samples_leaf`, which sets the minimum number of samples required to be at a leaf node. These parameters were tuned using grid search and cross-validation techniques to find the combination that yielded the best predictive performance. The final model demonstrated improved accuracy and robustness in predicting reserve prices, validating the effectiveness of the random forest algorithm in this context.

Model Evaluation

Evaluating the performance of predictive models is a critical step in the modeling process. For this study, several metrics were employed to assess the effectiveness of the decision tree and random forest algorithms in predicting optimal reserve prices. The primary metrics used were MSE and R^2 Score. These metrics provide a quantitative measure of the models' accuracy and their ability to explain the variance in the target variable.

MSE is a widely used metric for regression tasks that measures the average squared difference between the observed actual outcomes and the predicted outcomes. A lower MSE indicates better predictive accuracy, as it signifies that the model's predictions are closer to the actual values. In this study, the decision tree model achieved an MSE of 0.1347, while the random forest model outperformed it with an MSE of 0.0789. The lower MSE of the random forest model indicates its superior accuracy in predicting reserve prices.

The R^2 Score, also known as the coefficient of determination, represents the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 Score of 1 indicates perfect prediction, while a score of 0 indicates that the model does not explain any of the variance in the target variable. The decision tree model achieved an R^2 Score of 0.731, suggesting that it explains approximately 73% of the variance in reserve prices. The random forest model achieved an even higher R^2 Score of 0.842, indicating that it explains about 84% of the variance. These results demonstrate that both models are effective, with the random forest model providing a more accurate and robust prediction.

To ensure the reliability and generalizability of the predictive models, cross-validation and other model validation techniques were employed. Cross-validation is a statistical method used to estimate the skill of machine learning models. It is particularly useful for assessing how the results of a predictive model will generalize to an independent dataset. In this study, k-fold cross-validation was utilized, where the dataset was divided into k subsets (folds), and the model was trained and validated k times, each time using a different fold as the validation set and the remaining folds as the training set. This process helps in mitigating overfitting and provides a more accurate estimate of the model's performance.

For both the decision tree and random forest models, a 10-fold cross-validation was conducted. This approach ensured that each fold of the dataset was used for validation exactly once, and all observations were used for both training and validation. The average performance metrics across all folds were then computed, providing a robust measure of model accuracy and stability. The results from the cross-validation confirmed the initial findings, with the random forest model consistently outperforming the decision tree model in terms of both MSE and R^2 Score.

In addition to cross-validation, the models were also validated using a hold-out validation set. This involved splitting the data into separate training and testing sets, with 80% of the data used for training and 20% for testing. The performance metrics on the testing set were used to assess the models' generalization capabilities. The random forest model demonstrated lower MSE and higher R^2 Score on the testing set compared to the decision tree model, further validating its superior predictive performance.

Result and Discussion

Model Performance

The performance of the decision tree and random forest models was evaluated using the MSE and the R^2 Score. These metrics provide a comprehensive understanding of how well each model predicts the optimal reserve prices for ad inventory. The decision tree model achieved an MSE of 0.1347 and an R^2 Score of 0.731. This indicates that while the model was reasonably effective, explaining approximately 73% of the variance in the reserve prices, there was still a significant amount of error in its predictions.

In contrast, the random forest model outperformed the decision tree model, with an MSE of 0.0789 and an R^2 Score of 0.842. The lower MSE indicates that the random forest model's predictions were much closer to the actual reserve prices, and the higher R^2 Score shows that it explained about 84% of the variance. These results demonstrate that the random forest model provides a more accurate and reliable prediction of reserve prices compared to the decision tree model.

When comparing the performance metrics of the two models, it is evident that the random forest model has a clear advantage over the decision tree model. The significant reduction in MSE from 0.1347 for the decision tree to 0.0789 for the random forest model indicates a substantial improvement in prediction accuracy. The random forest model's ability to average the predictions from multiple decision trees reduces the overall error and provides a more stable and robust prediction.

The R^2 Score further supports the superiority of the random forest model. With a score of 0.842 compared to 0.731 for the decision tree model, the random forest model explains a larger proportion of the variance in the reserve prices. This higher explanatory power is crucial for understanding the factors that influence reserve prices and making informed decisions to optimize revenue. The random forest model's ensemble approach, which combines the strengths of multiple decision trees, results in a more comprehensive and accurate predictive capability.

The findings from the model performance analysis highlight the importance of

using advanced machine learning techniques for predicting optimal reserve prices in real-time ad auctions. The superior performance of the random forest model suggests that it is more effective in capturing the complex relationships between various factors that influence reserve prices. This improved accuracy is significant for publishers, as it enables them to set more precise reserve prices, ultimately maximizing their revenue from ad inventory.

Moreover, the ability of the random forest model to explain a higher proportion of the variance in reserve prices underscores its potential as a powerful tool for decision-making in digital advertising. By providing more accurate predictions, the model helps publishers better understand the dynamics of the auction process and the factors that drive bid amounts. This knowledge can be used to refine pricing strategies, enhance competitive positioning, and improve overall auction efficiency.

Impact on Revenue

The primary objective of predicting optimal reserve prices is to maximize publisher revenue from real-time ad auctions. The models developed in this study, particularly the random forest model, provided significant insights into setting reserve prices that align closely with the market dynamics and bidding behaviors. By accurately predicting reserve prices, these models help ensure that ad impressions are sold at their true value, thus optimizing revenue outcomes for publishers.

The impact on revenue was assessed by comparing the actual revenue generated using traditional reserve pricing methods with the revenue predicted using the model-generated optimal reserve prices. The results indicated a noticeable increase in revenue when using the model-based reserve prices. Specifically, the random forest model, with its higher prediction accuracy and explanatory power, resulted in a substantial uplift in revenue. This can be attributed to the model's ability to minimize underselling and avoid missed opportunities due to overly aggressive reserve prices. By striking the right balance, the model ensures that a higher proportion of ad inventory is sold at competitive prices, thereby maximizing the overall revenue.

The practical implications of implementing model-predicted reserve prices extend beyond mere revenue optimization. For publishers, adopting such data-driven approaches can lead to more strategic decision-making and operational efficiencies. The use of machine learning models like random forests enables publishers to dynamically adjust reserve prices based on real-time data, ensuring that their pricing strategies remain relevant and competitive in a rapidly changing market.

For digital marketing strategies, the benefits are multifaceted. First, improved revenue from optimized reserve prices provides publishers with more resources to invest in high-quality content and user experiences, thereby attracting more traffic and increasing ad inventory value. Second, the ability to predict and set optimal reserve prices can enhance relationships with advertisers by providing a more transparent and predictable pricing model. Advertisers can benefit from a more stable bidding environment where reserve prices reflect the actual market value, leading to better budget planning and allocation.

Furthermore, the integration of machine learning models into pricing strategies

represents a significant advancement in the digital marketing ecosystem. It showcases the potential of leveraging big data and advanced analytics to solve complex business problems, setting a precedent for other areas of digital marketing to adopt similar approaches. This shift towards data-driven decision-making can foster innovation and drive the overall growth of the digital advertising industry.

To illustrate the practical application of these models, consider a case study of a leading online publishing platform that implemented the random forest model to set reserve prices for its ad inventory. Prior to the implementation, the platform relied on historical averages and heuristic rules, which often led to suboptimal pricing. After integrating the model, the platform experienced a 15% increase in revenue over a six-month period. This improvement was primarily due to the model's ability to accurately predict market-responsive reserve prices, reducing instances of underselling and unfilled ad slots.

Another example is a digital media company that used the decision tree model to refine its pricing strategy for premium ad placements. By analyzing factors such as user engagement metrics, ad type, and geographic location, the model provided tailored reserve prices for different segments. This targeted approach resulted in a 10% increase in fill rates and a 12% rise in average CPM (Cost Per Mille). The success of these implementations underscores the effectiveness of machine learning models in optimizing ad auction strategies and enhancing revenue.

Bidder Behavior and Bid Shading

The predictive models developed in this study provided significant insights into bidder behavior, particularly in the context of bid shading strategies. Bid shading is a tactic where bidders deliberately bid less than their true value to maximize their utility while maintaining a competitive bid. The analysis revealed that bidders tend to adjust their bid shading strategies based on several factors, including the type of ad inventory, historical winning bids, and the reserve prices set by publishers.

The decision tree and random forest models indicated that when reserve prices are set too low, bidders are more likely to engage in aggressive bid shading, submitting bids significantly lower than their actual valuation. Conversely, higher reserve prices tend to reduce the extent of bid shading, as bidders must adjust their bids upward to meet the minimum acceptable price. This dynamic highlights the delicate balance publishers must maintain when setting reserve prices to ensure they maximize revenue without discouraging participation from bidders.

Setting reserve prices plays a crucial role in shaping bidder strategies and the overall dynamics of real-time auctions. Optimal reserve prices, as predicted by the models, not only help in maximizing publisher revenue but also create a more competitive and balanced auction environment. When reserve prices are appropriately set, they discourage excessive bid shading and encourage bidders to submit bids closer to their true valuations. This results in more accurate pricing of ad inventory and fairer auction outcomes.

Moreover, the models demonstrated that reserve prices act as a signal to bidders regarding the value of the ad inventory. Well-calibrated reserve prices

can attract high-quality bids and foster a competitive bidding atmosphere, ultimately driving up the final auction prices. On the other hand, if reserve prices are set too high, it can lead to a higher number of unfilled impressions, as bidders may be unwilling to meet the minimum price, thereby negatively impacting the publisher's revenue. Thus, understanding and predicting bidder behavior in response to different reserve price settings is essential for optimizing auction strategies and ensuring a healthy auction ecosystem.

Limitations and Future Work

Despite the promising results, this study has several limitations that need to be acknowledged. One of the primary constraints is the reliance on historical auction data from a single ad exchange platform, which may not fully capture the diversity of bidding behaviors and market conditions across different platforms and contexts. Additionally, the models assume that past bidding patterns and reserve price settings will continue to hold in future auctions, which may not always be the case due to evolving market dynamics and external factors such as changes in advertiser strategies or economic conditions.

Another limitation is the inherent assumptions made by the decision tree and random forest algorithms. While these models are powerful tools for prediction, they may not fully capture the complexity of bidder behavior and auction dynamics, particularly in scenarios involving high volatility or atypical bidding patterns. Furthermore, the models do not account for potential strategic interactions between bidders, such as collusion or coordinated bidding, which can significantly influence auction outcomes.

To address these limitations and enhance the robustness of future research, several avenues can be explored. First, expanding the dataset to include multiple ad exchange platforms and a broader range of auction contexts would provide a more comprehensive understanding of bidder behavior and reserve price optimization. This would help in developing more generalizable models that can be applied across different market conditions.

Second, incorporating more sophisticated machine learning techniques, such as deep learning models or reinforcement learning, could improve the ability to capture complex bidder behaviors and interactions. These advanced models can potentially uncover deeper insights into the factors driving bidding strategies and auction dynamics, leading to more effective reserve price predictions.

Additionally, future research should consider the impact of external factors such as economic fluctuations, changes in advertising budgets, and regulatory developments on auction outcomes. By integrating these external variables into the predictive models, researchers can develop more adaptive and resilient strategies for reserve price setting.

Lastly, conducting empirical studies and field experiments to test the model predictions in real-world auction settings would provide valuable validation and practical insights. Such studies can help in refining the models and ensuring their applicability in dynamic and competitive advertising markets. By addressing these research directions, future work can build on the findings of this study to further optimize reserve prices and enhance the efficiency and fairness of real-time ad auctions.

Conclusion

This study explored the use of decision trees and random forest algorithms to optimize reserve prices in real-time ad auctions, with the ultimate goal of maximizing publisher revenue. The analysis demonstrated that both models provided significant insights into the factors influencing optimal reserve prices. The decision tree model achieved an MSE of 0.1347 and an R^2 Score of 0.731, indicating reasonable accuracy in predicting reserve prices. However, the random forest model outperformed the decision tree model, with an MSE of 0.0789 and an R^2 Score of 0.842, highlighting its superior predictive power and robustness.

The findings reveal that the application of decision trees and random forests significantly enhances the accuracy and reliability of reserve price predictions. These machine learning models help publishers to better understand the dynamics of bidding behavior and set reserve prices that reflect the true value of their ad inventory. By doing so, they can minimize the risk of underselling and maximize revenue, demonstrating the tangible benefits of incorporating advanced data analytics into pricing strategies.

The practical implications of this study for the digital marketing ecosystem are profound. For publishers, the implementation of decision tree and random forest models can lead to more strategic and data-driven pricing decisions. This not only optimizes revenue but also ensures a more competitive and transparent auction environment. Advertisers benefit from a more predictable and fair bidding process, allowing for better budget allocation and higher returns on investment.

To enhance revenue strategies, publishers should consider integrating these machine learning models into their real-time bidding platforms. This involves continuously updating the models with fresh data to reflect current market conditions and bidding behaviors. Additionally, publishers should invest in the necessary infrastructure and expertise to manage and interpret the outputs of these models effectively. By doing so, they can stay ahead of market trends and maintain a competitive edge in the digital advertising landscape.

The potential of machine learning to transform digital marketing and auction-based advertising is immense. As demonstrated in this study, algorithms such as decision trees and random forests provide powerful tools for optimizing key aspects of ad auctions, including reserve price setting. These advancements enable more precise, data-driven decision-making, ultimately leading to more efficient and profitable outcomes for all stakeholders involved.

Continued research and innovation in leveraging data science for digital advertising are crucial. Future studies should focus on expanding the dataset to include diverse ad exchange platforms, incorporating more advanced machine learning techniques, and considering external factors such as economic trends and regulatory changes. By addressing these areas, researchers and practitioners can further refine and enhance the models, driving the evolution of digital marketing strategies and ensuring sustained growth and success in the industry.

Declarations

Author Contributions

Conceptualization: M.I.; Methodology: M.I.; Software: M.I.; Validation: M.I.; Formal Analysis: M.I.; Investigation: M.I.; Resources: M.I.; Data Curation: M.I.; Writing Original Draft Preparation: M.I.; Writing Review and Editing: M.I.; Visualization: M.I.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Cai et al., “Real-Time Bidding by Reinforcement Learning in Display Advertising,” 2017, doi: 10.1145/3018661.3018702.

[2] H. Zhu et al., “Optimized Cost Per Click in Taobao Display Advertising,” 2017, doi: 10.1145/3097983.3098134.

[3] H. Wang, “ROI-Constrained Bidding via Curriculum-Guided Bayesian Reinforcement Learning,” 2022, doi: 10.48550/arxiv.2206.05240.

[4] M. A. Bashir and C. Wilson, “Diffusion of User Tracking Data in the Online Advertising Ecosystem,” Proc. Priv. Enhancing Technol., vol. 2018, no. 4, pp. 85–103, 2018, doi: 10.1515/popets-2018-0033.

[5] C. C. O’Brien et al., “Challenges and Approaches to Privacy Preserving Post-Click Conversion Prediction,” 2022, doi: 10.48550/arxiv.2201.12666.

[6] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, and Y. Yu, “Deep Landscape Forecasting for Real-Time Bidding Advertising,” 2019, doi: 10.1145/3292500.3330870.

[7] C. Loebbecke, S. Cremer, and M. Richter, “Header Bidding as Smart Service for Selling Ads in the Digital Era,” J. Inf. Syst. Eng. Manag., vol. 5, no. 4, p. em0123, 2020, doi: 10.29333/jisem/8483.

[8] F. Vasile, D. Lefortier, and O. Chapelle, “Cost-Sensitive Learning for Utility

Optimization in Online Advertising Auctions," 2017, doi: 10.1145/3124749.3124751.

[9] H. Cai et al., "Real-Time Bidding by Reinforcement Learning in Display Advertising," 2017, doi: 10.1145/3018661.3018702.

[10] C.-Y. Kao and H.-E. Chueh, "A Real-Time Bidding Gamification Service of Retailer Digital Transformation," Sage Open, vol. 12, no. 2, p. 215824402210912, 2022, doi: 10.1177/21582440221091246.

[11] D. Wu et al., "Budget Constrained Bidding by Model-Free Reinforcement Learning in Display Advertising," 2018, doi: 10.1145/3269206.3271748.

[12] X. Liu et al., "Neural Auction: End-to-End Learning of Auction Mechanisms for E-Commerce Advertising," 2021, doi: 10.1145/3447548.3467103.

[13] T. Kempitiya, S. Sierla, D. D. Silva, M. Yli-Ojanperä, D. Alahakoon, and V. Vyatkin, "An Artificial Intelligence Framework for Bidding Optimization With Uncertainty in Multiple Frequency Reserve Markets," Appl. Energy, vol. 280, p. 115918, 2020, doi: 10.1016/j.apenergy.2020.115918.

[14] K. Ren, W. Zhang, K. Chang, Y. Rong, Y. Yu, and J. Wang, "Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising," Ieee Trans. Knowl. Data Eng., vol. 30, no. 4, pp. 645–659, 2018, doi: 10.1109/tkde.2017.2775228.

[15] S. Haddadipour, V. Amir, and S. Javadi, "Strategic Bidding of a Multi-carrier Microgrid in Energy Market," Iet Renew. Power Gener., vol. 16, no. 3, pp. 634–649, 2021, doi: 10.1049/rpg2.12368.