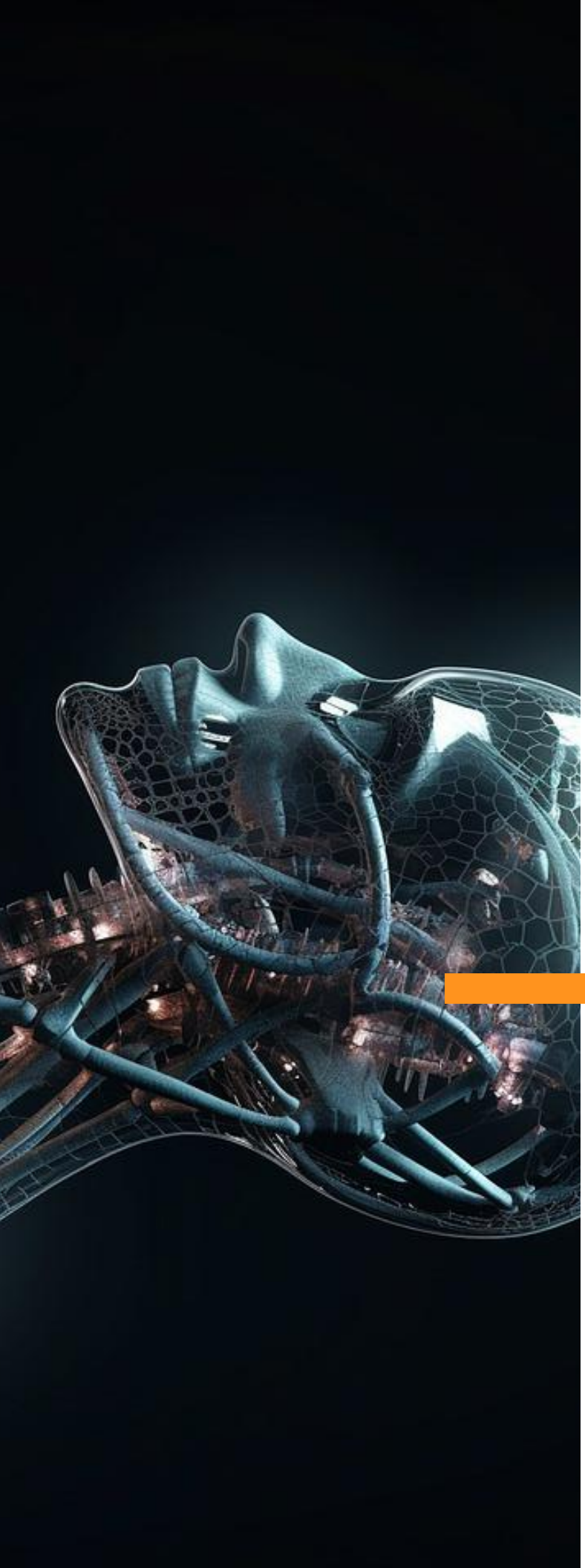


Final Report

Human or Robot?

Joyce Huang
huangj@brandeis.edu

Evan Goddard
evangoddard@brandeis.edu

Kaggle Project
Facebook Recruiting IV: Human or Robot?

Introduction

Online auctions are an increasingly popular way to purchase goods, because theoretically, they provide a fair and open marketplace for buyers and sellers to engage in business transactions. However, as online auctions have grown in popularity, so has the threat of fraudulent activity. Identifying fraudulent activity is important not only to prevent financial losses but also to maintain trust in the integrity of the online auction system.

The “Facebook Recruiting IV: Human or Robot?” competition hosted by Kaggle presented a dataset of auction bidding information to detect whether a bidder is a human or a robot. The dataset contains information on the bidding activity of multiple bidders and is labeled as human or robot. The challenge was to develop a predictive model to identify whether a given bidder is a human or robot based on the bidding data. This is a challenging problem because fraudulent activity can be hard to detect, especially when the fraudsters are sophisticated enough to blend in with legitimate users. In addition, there are many factors that can influence the outcome of an auction, making it difficult to determine which features are most important in distinguishing between human and robot bidders. The architecture of our algorithm can be considered a next step because it goes beyond simply using raw features from the bids dataset. By creating new features that capture the bidding behavior of bidders, our approach has the potential to better differentiate between bots and humans. In addition, the use of RandomForestClassifier, a powerful ensemble method, and hyperparameter tuning, leads to potentially better performance compared to simpler classifiers.

Approach

Our approach to solving this problem involved several steps. First, we performed data cleaning and exploratory analysis to gain insights into the data and identify potential outliers. Then, we selected relevant features and used density plots to check the significance of the features. Next, we performed over-sampling using the RandomOverSampler tool and ensemble averaging of multiple Random Forest models to reduce the variance of our predictions. Finally, we performed hyperparameter tuning for each model separately using GridSearchCV to improve the performance of our models.

Diving into the details of our algorithm architecture, we started by loading the bids dataset and computing bid statistics (count, mean, and median) for each bidder. We then merged these statistics with the train_set and test_set, filling any missing values with the median value of the respective columns. We also created several new features that might help identify bots: the number of simultaneous bids (bids with a time difference of 0); the number of first and last bids (counting the number of times a bidder placed the first or last bid in an auction); the auction duration (time difference between the first and last bids in an auction); and the bid ratios (calculating the number of bids placed in the first and second half of auctions). Furthermore, we performed several data analysis and visualization tasks, such as identifying outliers, computing

mean merchandise values for bots and humans, and plotting histograms and density plots for various features. When doing model building, we used a RandomForestClassifier to predict if a bidder is a bot or a human. We first applied RandomOverSampler to balance the training data, and then used GridSearchCV to perform hyperparameter tuning for the RandomForestClassifier. The best hyperparameters and corresponding accuracy score are printed.

Analysis

This section of the report details the data sources we used, the motivations behind our selected methods, the evaluation metrics, how we ran our experiments, and the results we obtained.

A. Data

This project utilizes two datasets. The first dataset is a bidder dataset containing bidder information, such as their ID, payment account, and address. The second dataset, the bid dataset, consists of 7.6 million bids made on various auctions through mobile devices. The auction platform enforces a fixed dollar increment for each bid, and the data comes from an online platform not affiliated with Facebook. The bidder dataset includes fields for bidder_id, payment_account, address, and outcome, with the latter indicating whether a bidder is human or a bot. The bid dataset comprises fields for bid_id, bidder_id, auction, merchandise, device, time, country, IP, and URL. The aim of the project is to analyze these datasets and determine whether a bidder is human or a bot based on their bidding behavior, auction participation, and device usage.

B. Algorithms

We used a combination of machine learning algorithms and feature engineering to address the problem of identifying bots and humans in online auctions. The primary model used was Random Forest. This method was chosen due to its excellent performance in handling large-scale data, ability to capture complex relationships, and robustness to noise and outliers. For feature engineering, we extracted valuable information from the raw datasets, such as counts, time-related statistics, and categorical variables. This process aimed to provide the models with meaningful input features that better represent the bidding behavior of bidders and capture the differences between bots and humans.

In terms of evaluation metrics, we used the area under the ROC curve (AUC-ROC) to assess the models' performance. This metric was chosen because it provides a good balance between the true positive rate (sensitivity) and the false positive rate (1-specificity), making it suitable for classification problems with imbalanced classes, such as this one, where the number of human bidders likely exceeds the number of bots.

The choice of using ensemble methods like Random Forest stems from their capability to combine multiple weak learners to form a strong learner, enhancing the predictive performance. Random Forest provides a good balance between variance and bias through bagging, while

LightGBM is a gradient boosting algorithm that can optimize for a given loss function, resulting in improved performance. Additionally, LightGBM is particularly efficient in terms of computational resources and can handle large datasets with a high number of features effectively.

In summary, the chosen methods and evaluation metric were tailored to address the specific challenges of the task, which were handling large-scale data, capturing complex relationships, dealing with imbalanced classes, and optimizing computational efficiency.

C. Experiment

Hypothesis*: The choice of features and models was driven by the hypothesis that bid volume and timing would be strongly related to bot activity.*

As mentioned prior, our experiment consisted of several key stages: data preprocessing, feature engineering, model selection, hyperparameter tuning, and evaluation metrics. The goal of our experiment was to accurately predict whether a bidder was a bot or a human.

During the feature engineering process, we generated new features for each bidder, such as count, mean, and median, allowing us to train our model with a more comprehensive understanding of the bidders' average historical activity. Furthermore, we took into account the timing of bids by considering simultaneous bids and the number of first and last bids placed by each bidder. We hypothesized that there would be a strong correlation between bid volume and the likelihood of bot activity, as well as a relationship between the timing of bids and the presence of bots.

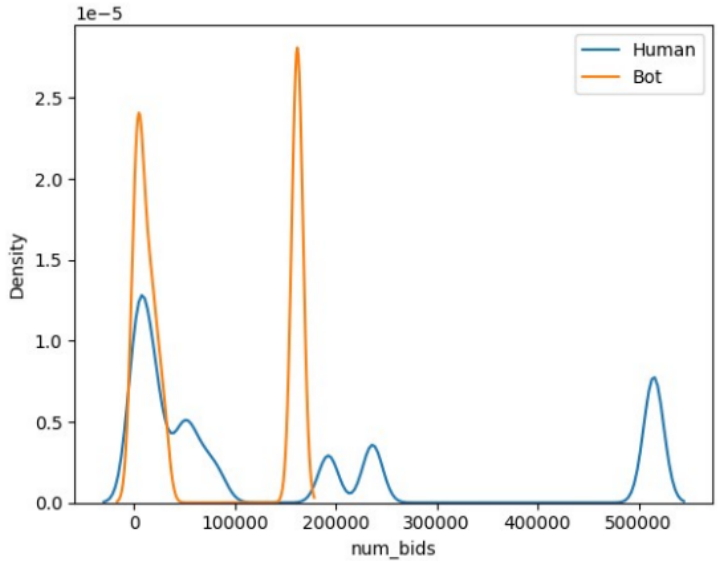
By incorporating these features and considering their potential impact on the classification task, our experiment aimed to build a highly accurate model capable of distinguishing between human and bot bidders in online auctions.

Next, we divided the dataset into training and testing sets. We opted to use RandomForestClassifier as our primary model due to the benefits mentioned in our Approach and used RandomOverSampler to balance the training data. Hyperparameter tuning was performed using GridSearchCV, and the best combination of hyperparameters was determined.

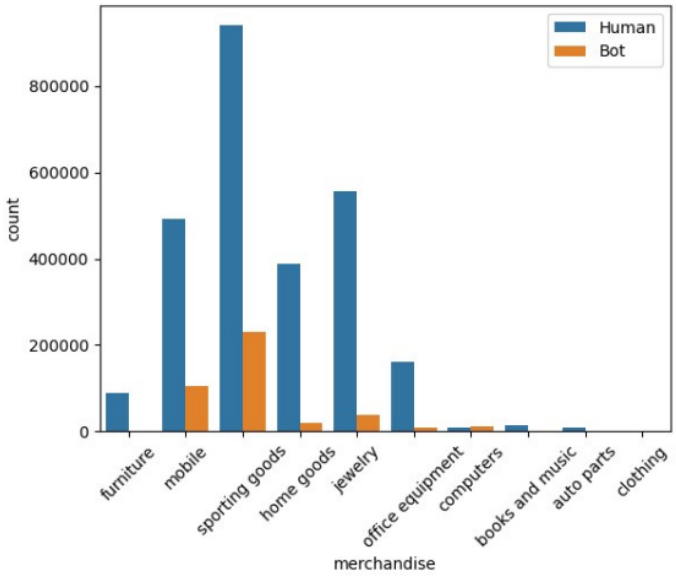
We used the ensemble averaging and hyperparameter tuning techniques with the hope that they would contribute to more robust and accurate predictions with our test sample with the incorporation of our modified features. After running our models, we selected the best models and used them to predict auction bot probabilities on the test set.

D. Results

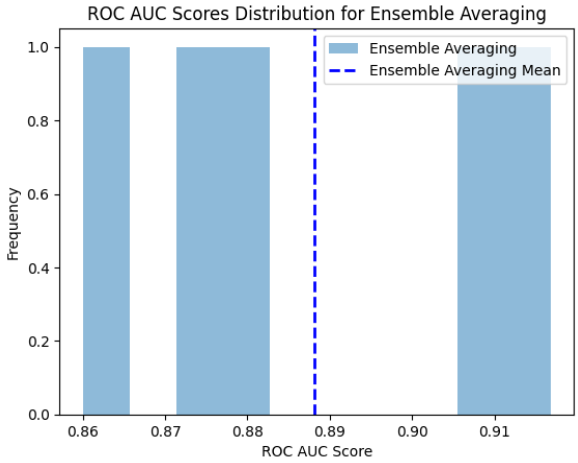
Based on our strong AUC scores, we have concluded that our hypothesis is partially correct that bid volume and timing are strong predictors of bot activity. Our choice to implement features such as num_simultaneous_bids and bid_mean helped drive the success of our model. As seen in the chart below, bots would disproportionately contribute bids in certain ranges.

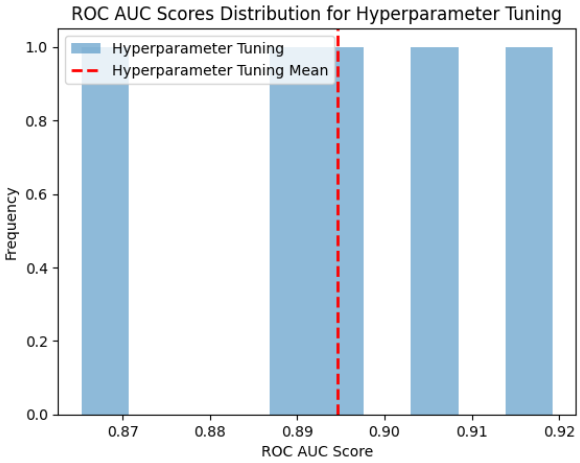


Other features, such as merchandise, were notable but only proved relevant in a small percentage of cases. In the case of merchandise, a nearly equal amount of computer bidders were slightly more likely to be bots. However, there were too few samples of computer merchandise to make a significant impact, as seen below.



Plotting of ROC AUC Scores Distribution for Ensemble Averaging and Hyperparameter Tuning:





As seen in the plots above, in our trials, hyperparameter tuning tended to outperform ensemble averaging. This is likely due to the fact that hyperparameter tuning allows for more fine-tuned control over the model's performance by optimizing its parameters for the specific data set and problem at hand. Compared to similar models on Kaggle, we had a slightly lower AUC average.

While our final project model has shown promising results in discerning bots from humans in online auctions, we acknowledge its potential for further enhancement. We could expand our approach to include other feature engineering and selection techniques. By deploying more advanced techniques, we could cultivate additional features that could potentially amplify the model's performance. Refining our dataset more meticulously by incorporating robust outlier detection and removal strategies could also improve model precision. Exploring a broader spectrum of algorithms, such as gradient boosting machines or support vector machines could improve the model's predictive capabilities. Finally, given the imbalanced nature of our dataset, our model's performance could be more accurately gauged using alternative metrics, such as the F1-score or precision-recall curves. Our steadfast goal is to enhance our model's performance and attain a higher AUC average, potentially by combining multiple models. This commitment underscores our ongoing efforts to improve bot detection in online auctions.

Team Contributions

Team member	Contribution
Joyce Huang	Data Cleaning, Feature Engineering, Model Validation, Final Model, Introduction, Approach, Analysis (Data, Algorithms)
Evan Goddard	Data Cleaning, Feature Engineering, Model Validation, Final Model, Analysis (Algorithms, Experiment, Results)

References

https://www.kaggle.com/competitions/facebook-recruiting-iv-human-or-bot/overview

https://www.kaggle.com/code/chongzhenjie/human-or-robot-random-forest/notebook