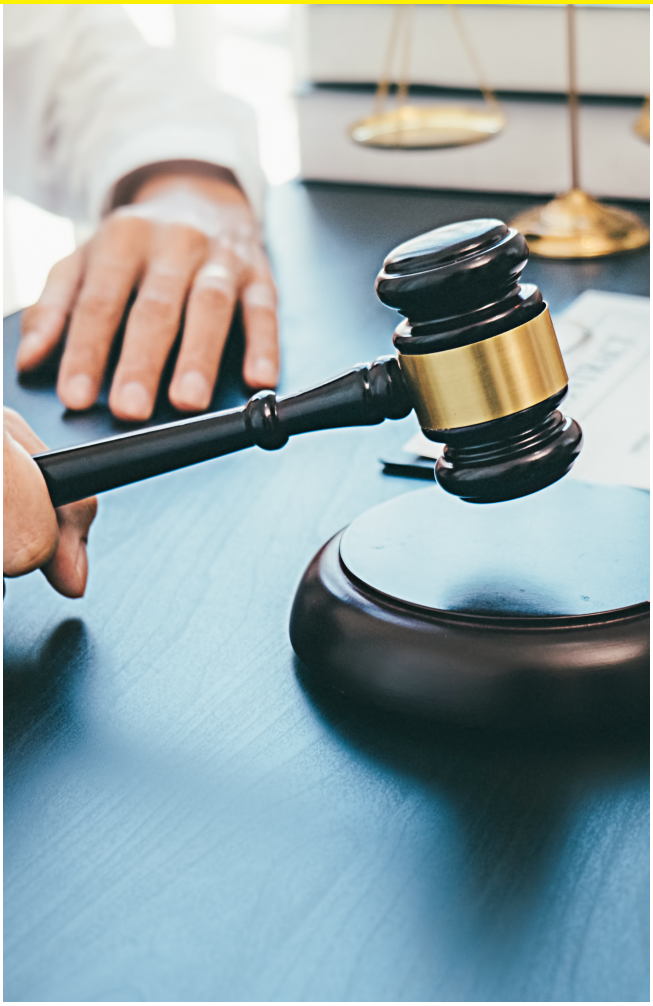
PROJECT REPORT

AUCTION FRAUD DETECTION

SPRINGBOARD
SCHOOL OF DATA

BY GABRIELLE WALD

INTRODUCTION



CONTEXT

On an auction website, human bidders are becoming increasingly frustrated with their inability to win auctions vs. their software-controlled counterparts. As a result, usage from the site's core customer base is plummeting. To rebuild customer happiness, the site owners need to eliminate computer generated bidding from their auctions.

GOAL

HUMAN OR ROBOT?

The goal is to identify online auction bids that are placed by "robots", helping the site owners easily flag these users for removal from their site to prevent unfair auction activity.

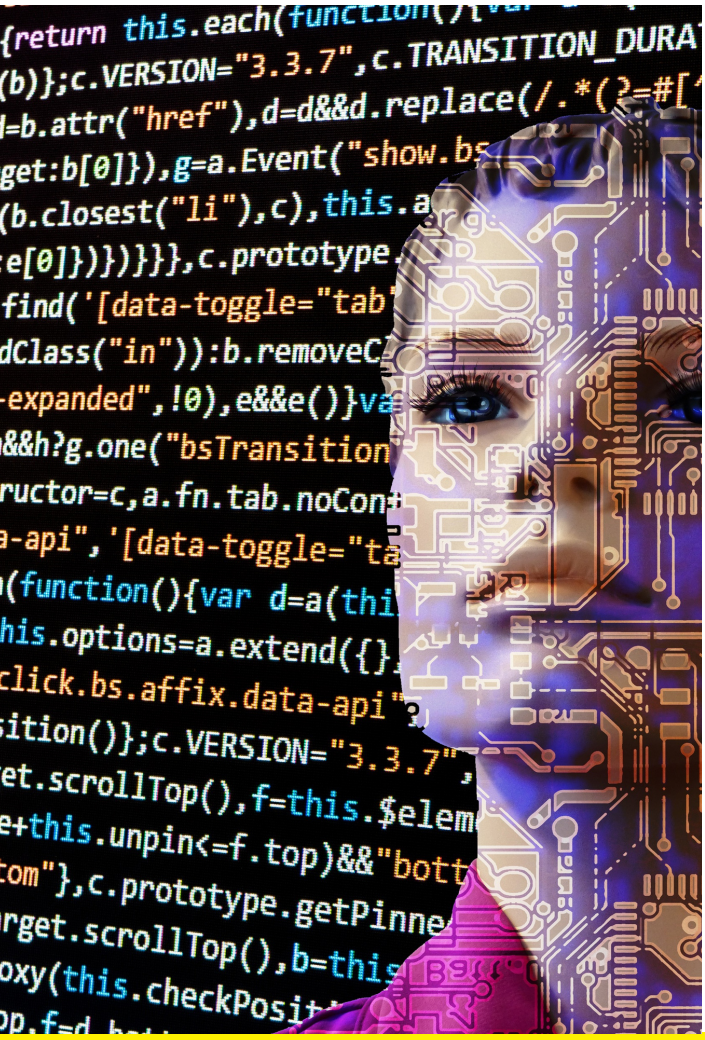
DATA SOURCE

This project idea is part of an Engineering competition created by Facebook and Kaggle in 2015. The provided dataset is said to be one of the richest data of its kind and a world class machine learning problem with a great potential for feature engineering. The data was retrieved from the Kaggle website in csv format.

DATA

ABOUT THE DATA

There are two datasets. The bidder dataset includes a list of bidder information, including their unique id, payment account, address, and outcome. The bid dataset includes 7.6 million bids on different auctions. Bids are all made from mobile devices. The online auction platform has a fixed increment of dollar amount for each bid, so it does not include an amount for each bid.



This was a challenging dataset because the raw data contains mostly obfuscated fields to protect privacy, unique identifiers, and categorical variables. Data wrangling and EDA were performed to understand and prepare the data for machine learning modeling.

TOTAL OF 7.6 MILLION BIDS

DATA FIELDS

For bidder dataset:

- bidder_id
- payment_account
- address
- outcome

For the bid dataset:

- bid_id
- bidder_id
- auction
- merchandise
- device
- time
- country
- ip
- url

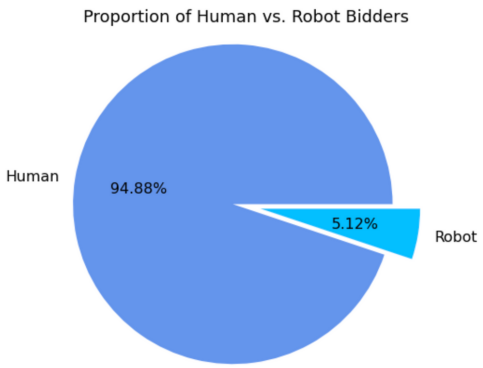
DATA WRANGLING

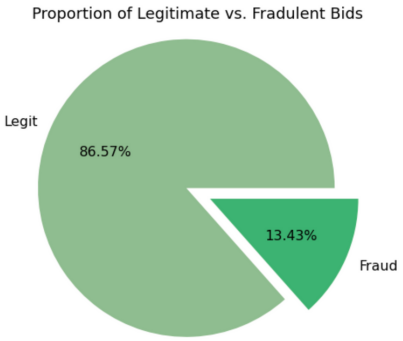
The datasets were merged on bidder_id to make it easier to access information. Although the dataset was fairly clean, there were 29 bidders with no bids (1.5%). Lacking any data on bids, and the fact that all 29 bidders are humans (and our interest is in identifying robots), the rows corresponding to these missing values were dropped. All columns were kept. Regarding the obfuscated fields, they are an issue for interpretability, but will still be useful for modeling.

EXPLORATORY DATA ANALYSIS

INITIAL HYPOTHESIS

1. Total number of bids: Robots might have significantly higher numbers of bids compared to humans.
2. Number of bids per auction: Robots might have a higher number of bids each auction.
3. Distinct IPs: Robots might bid from more diverse IP addresses.
4. Merchandise: Robots might bid more often on certain merchandises.



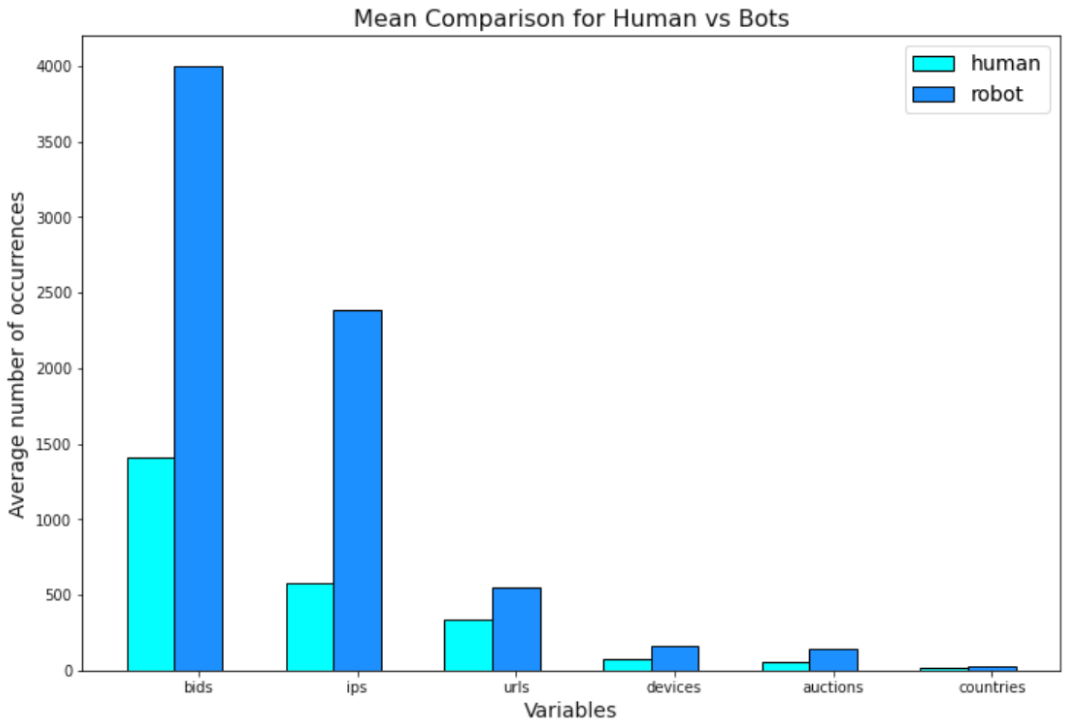


FIRST IMPRESSIONS

The data is highly unbalanced from the bidders and bids perspective. Humans represent 94.9% of the data, while only 5.1% are robots. Legitimate bids represent 86.5% of the data, while 13.4% are fraudulent. There are 12,740 unique auctions represented, and over 3 million bids in the train dataset. Bids come from 5,729 device models in 199 countries. There are 663,873 URLs, 1,030,950 IP addresses, and 10 distinct merchandise categories.

HUMAN VS. ROBOT DESCRIPTIVE STATISTICS

- Robots have higher mean and median numbers than humans for all the variables.
- The difference in the median is greater than the mean. That is because of some extreme outliers in the human class.



FINDINGS

The initial assumptions held true. There is a significant difference in the mean and median number of occurrences for various features between humans and robots. The mean number of bids per robot is 4004 and for humans is 1443, the median number of bids per robot is 716, while for humans it is only 14. The mean number of IP addresses per robot is 2388, for humans is 581, the median for robots is 290, and 11 for humans. A similar pattern occurs in the other variables (number of auctions, countries, devices, and urls) with robots having more bidding activity on average than humans.

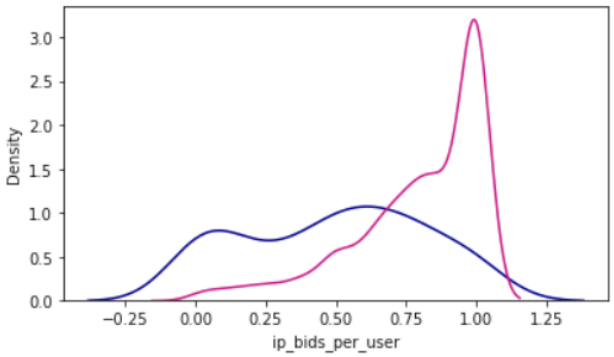
FEATURE ENGINEERING

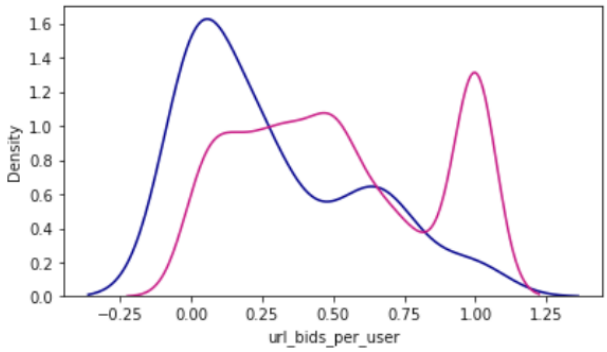
Informed by the EDA, the major differences between robots and humans were considered. Features were created around the number of occurrences, mean and median which showed to be significantly apart in the analysis. These features will be fed into the classification model so the model can learn to classify the bidders into robots or humans.

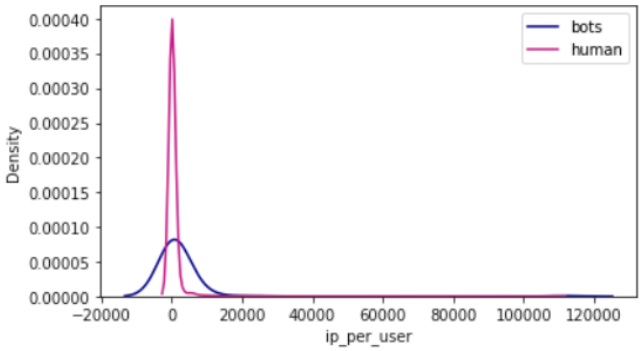
Some of the features created:

- Mean/median number of bids per auction per user
- Number of unique auctions per user
- Proportion of unique ip addresses to bids per user
- Mean/median number of IP addresses per auction per user
- Total unique ip and url per user
- Mean/median number of url per device per user
- Mean number of auctions for each country per user

FEATURE DISTRIBUTION







MACHINE LEARNING
CLASSIFICATION MODELS

PREPROCESSING

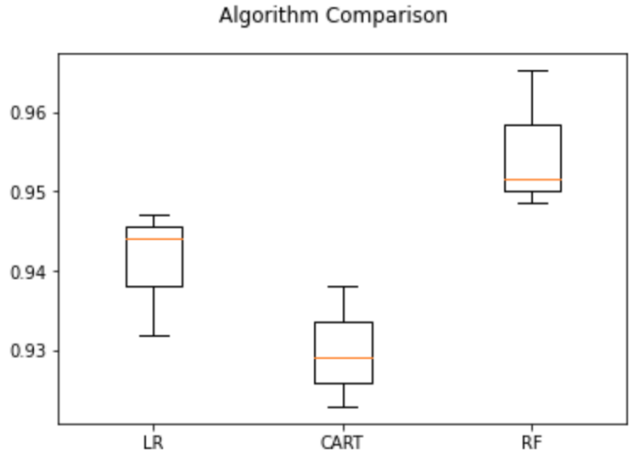
- 16 features were created summarizing information at the bidder_id level.
- The merchandise column was remapped to have one product per column with the number of bids for each bidder.
- Dataset was filtered at bidder_id level, 1984 rows.
- The original columns were dropped, except for the outcome column.
- Split into train (80%) and test data (20%)

MODELS

1. Logistic Regression
2. Random Forest
3. Decision Trees

Based on an initial modeling generated to identify the most promising algorithm accuracy, Random Forest Classifier comes first, followed by Logistic Regression, and lastly the Decision Tree Classifier.

```
LR: 0.941027 (0.006549)
CART: 0.929940 (0.006212)
RF: 0.955136 (0.007262)
```



SUMMARY RESULTS

LOGISTIC REGRESSION

Accuracy: 0.94

- True Negatives: 372
- False Positives: 2
- False Negatives: 22
- True Positives: 2

RANDOM FOREST

Accuracy: 0.95

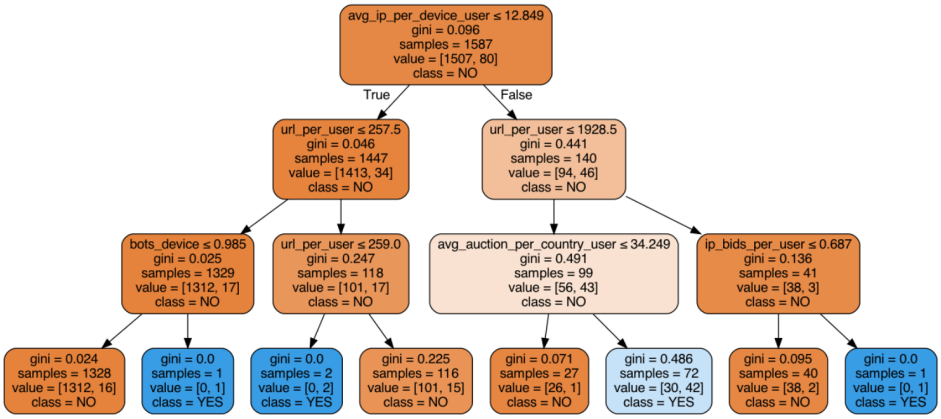
- True Negatives: 374
- False Positives: 0
- False Negatives: 19
- True Positives: 4

DECISION TREE

Accuracy: 0.94

- True Negatives: 366
- False Positives: 8
- False Negatives: 15
- True Positives: 8

The decision tree below returns the highest number of true positives. Although not great, it correctly classifies 8 robots out of 23. The splits are happening at **avg_url_per_device_user**, **url_per_user**, **computers**, **avg_auction_per_country_user**, and **ip_bids_per_user**.



In terms of feature importance, **avg_bids_per_user** shows up as the most important feature for the decision tree and random forest classifier. Other important features for both models are **url_per_device_user**, and **avg_ip_per_device_user**. Considering the mean and median differences in behavior for humans and robots for the variables bids, url, ip, and devices. This result is not surprising!

EVALUATION

Average number of bids, mean or median url per device and mean number of ip addresses per device are factors distinguishing humans and robots behavior.

Accuracy is not the best metric for this case. The target variable (robot or human) is not balanced. We want to consider metrics like Recall (aka Sensitivity) and the Area Under the Curve (AUC). First though, let's have a better understanding of the terms that form the basis for these.

- **True Positive:** label predicted 'robot' and it is in fact 'robot' (predicted 1.0 and it's 1.0).
- **True Negative:** label was predicted 'human' and it is in fact 'human' (predicted 0.0 and it's 0.0).
- **False Positive**: label was predicted 'robot' but it is in fact 'human' (predicted 1.0 but it's 0.0). Type 1 error or incorrect rejection of Null Hypothesis.
- **False Negative**: label was predicted 'human' but it is in fact 'robot' (predicted 0.0 but it's 1.0). Type 2 error or failure to reject of Null Hypothesis.\

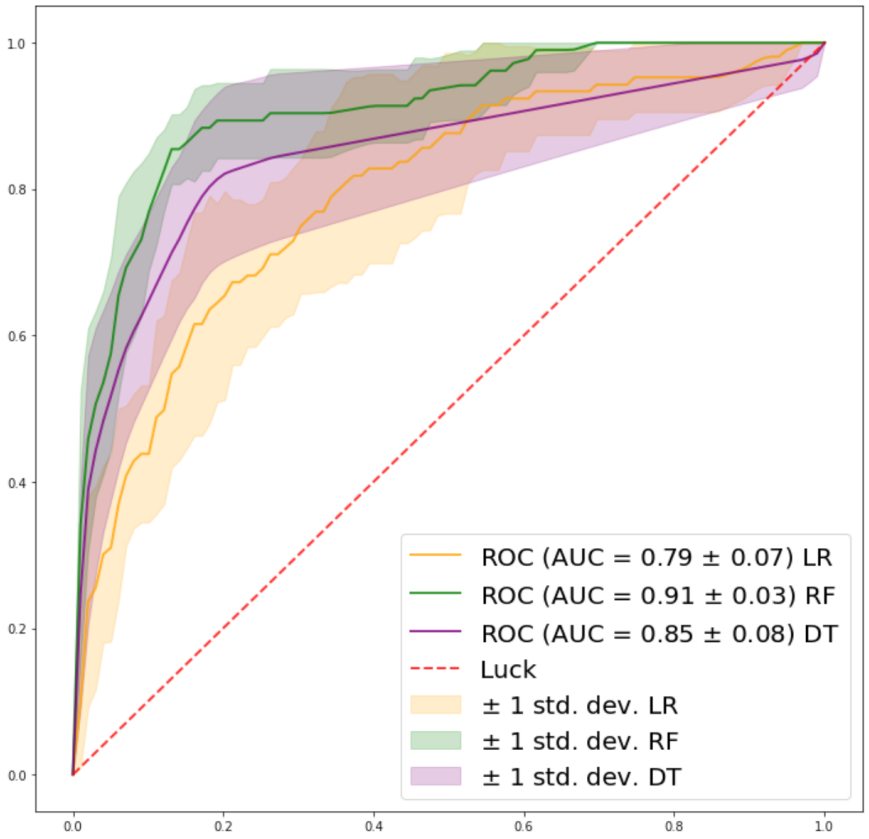
What are the best metrics to evaluate our models?

Recall is the ability of the model to identify all relevant instances, that is True Positive Rate, aka Sensitivity. It quantifies the number of correct positive predictions made out of all positive predictions that could have been made. We want the highest number of robots correctly classified as robots.

Recall = True Positives/ (True Positives + False Negatives)

AUC measures the ability of a classifier to distinguish between classes. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. It is plotted with Sensitivity against False Positive rate (1-Specificity). An AUC near 1 means the model has a good measure of separability.

AREA UNDER THE CURVE



The Random Forest classifier has the highest performance with an AUC between 0.88 and 0.94.

This means that RF is classifying robots correctly while keeping false positives low (humans are not being classified as robots). This makes sense, in fact the random forest classifier has zero false positives. On the other hand, it only classifies 4 out of 23 robots, that's a 19% success rate. It has 19 false negatives (robots classified as humans).

The second best AUC is for the Decision Tree classifier with a higher range that is between 0.77 and 0.93. The decision tree has a higher recall. It correctly identifies 8 robots, but it also misclassifies 8 humans as robots. Let's get into more details before selecting a model.

DISCUSSION

Recall and AUC are the metrics we want to consider. In fraud detection it is critical to correctly classify fraud or what is causing frauds, in this case we want to identify the “robot” users. Because Recall is the measure of relevant instances (True Positives) it is a good metric to evaluate the models. We want a model with a high Recall rate.

The AUC metric provides an understanding of how the model is performing with True Positives and False Positives (humans classified as robots). Ideally the best model classifies all 'robots' as 'robots' (True Positive), and it does not make type 1 error (False Positive), classifying 'humans' as 'robots'. A model that correctly identifies all robots as robots and makes no type I error would return an AUC of 1. That's what we want to see, an AUC score that is near 1.

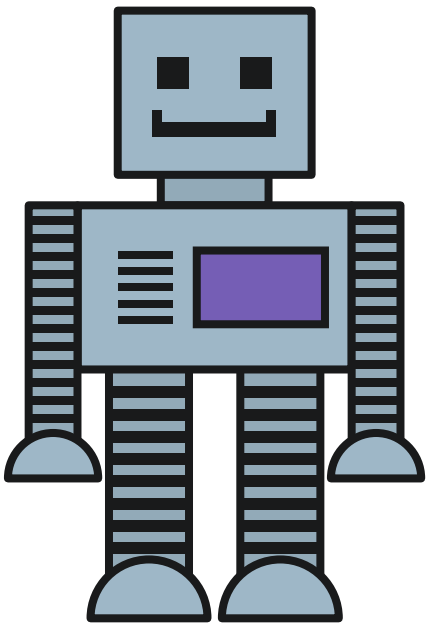
In case of misclassification, the preference is for type I error (False Positive: humans classified as robots) vs type II error (False Negative: robots classified as humans). The cost of keeping robots on the website is more costly than having humans misclassified as robots banned *temporarily* from the website. Keeping robots misclassified as humans on the website will not improve clients satisfaction.

In addition, in case of type I error, the use of CAPTCHA can be a good approach, allowing humans misclassified as robots to pass through the initial ban and continue to use the website.

CONCLUSION & NEXT STEPS

Considering all, the Decision Tree classifier has the highest True Positive rate (Recall) and a good AUC score, and the Random Forest classifier has the highest AUC score. However, given the unbalanced nature of this problem the best performing model in identifying True Positives (Decision Tree classifier) still only identifies 8 out of 23 robots. That's a 35% success rate, which is not good enough to move to production. If requiring more data was possible, it would be useful to have a larger dataset. The sample size for this problem is relatively small (1984 users with 103 labeled robots), with 80% reserved for training, the test data is left with 397 users with 23 labeled as robots. Nevertheless, there are still some things that can be done with the current data to further improve the models like:

- Create new features from the time column which due to time constraint I was not able to add. But time features showed to be very useful for other Kaggle competitors working on this same problem. Features like the following are very likely to improve classification:
  - Maximum number of bids made within a 20 minute span;
  - Median time between a user's bid and that user's previous bid;
  - Proportion of bidder's bid by day (~9 days of bid);
  - Number of simultaneous bids;
- Try other machine learning models like XGBoost, Support Vector Machine and Naive Bayes.



Thank you Nadav for your support and mentoring.