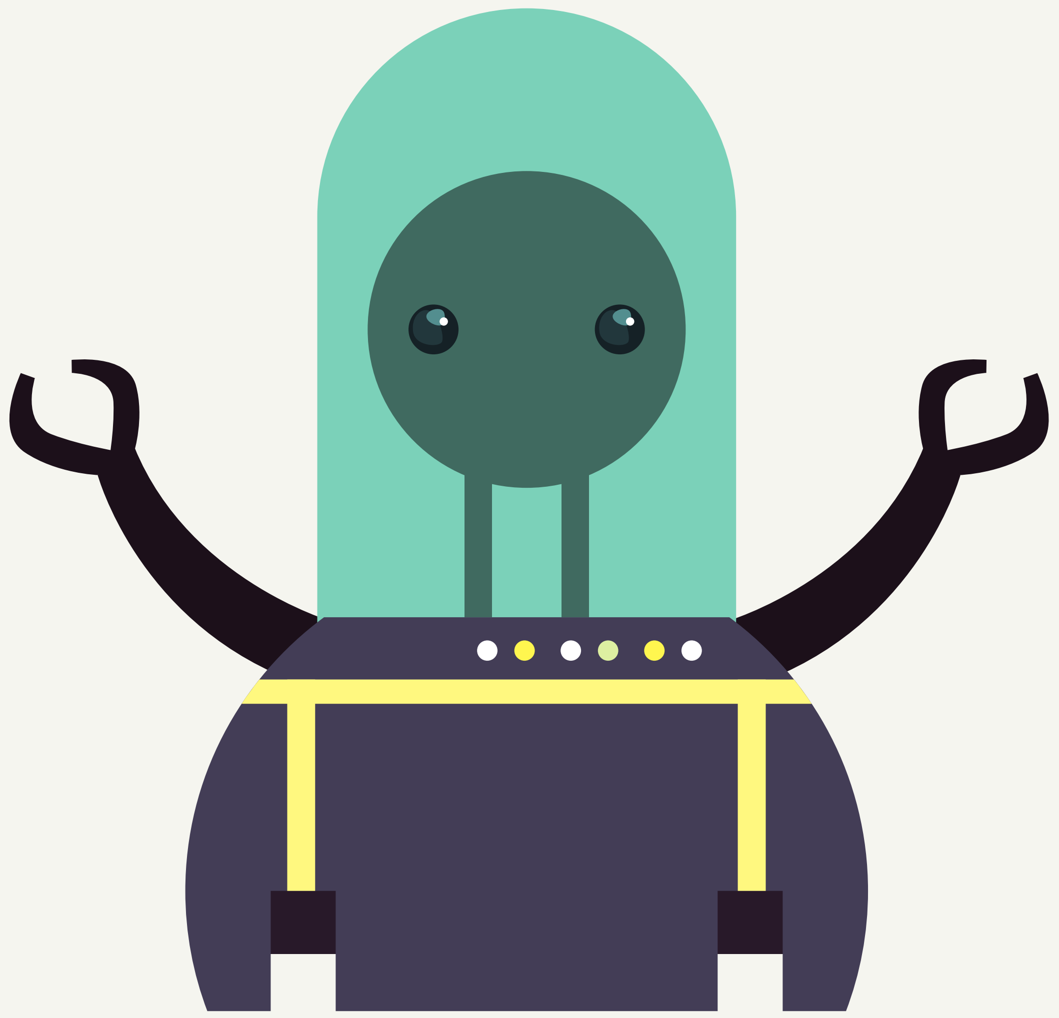
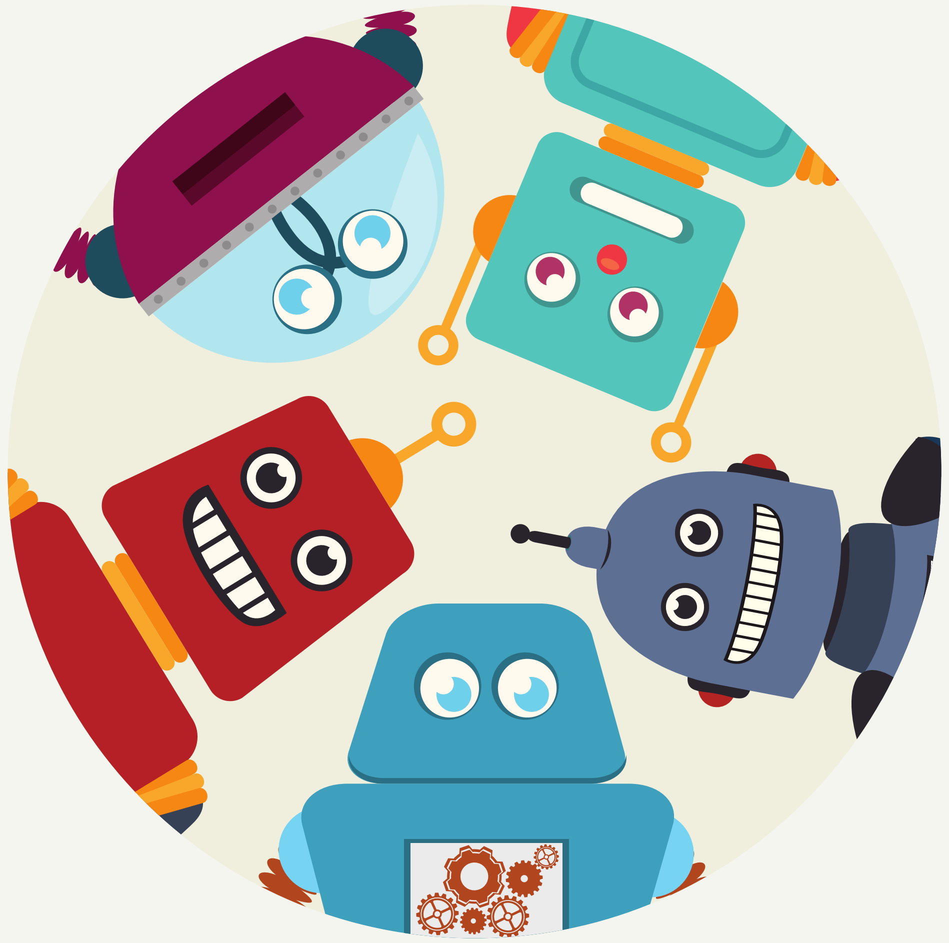
Auction Fraud Detection

"Human or Robot"

Springboard School of Data



by Gabrielle Wald



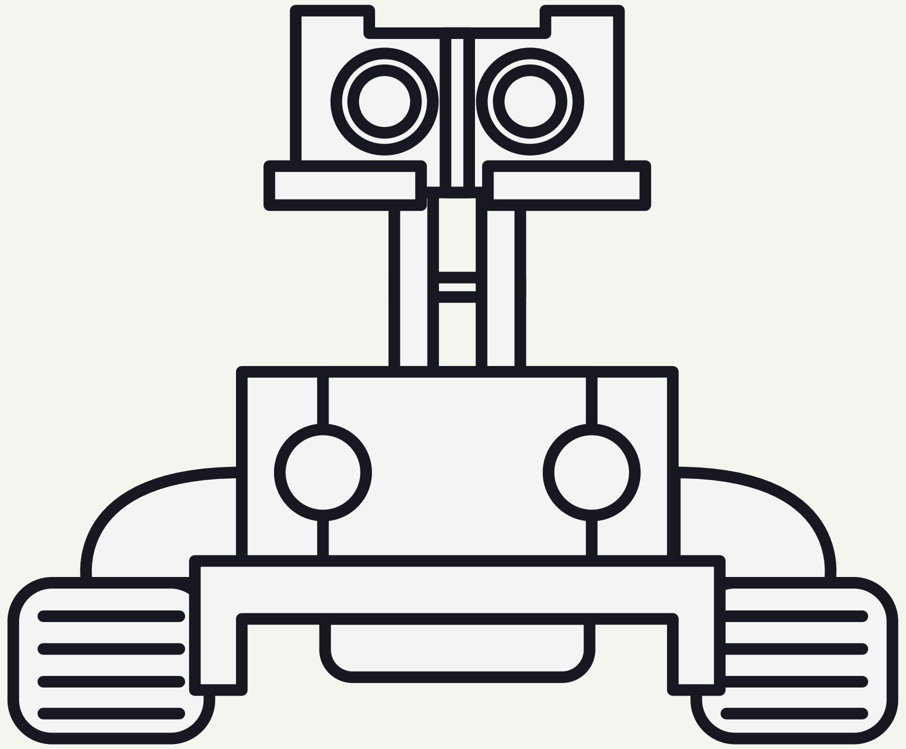
Introduction

Background

Overview of the project

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. In order to rebuild customer happiness, the site owners need to eliminate computer generated bidding from their auctions.



The Problem

What we want to solve

- Customer churn
- Customer satisfaction

Assumption

- Problem can be improved by eliminating robots from the website.

Goal

Identify users that are "robots", so they can be removed from the auction site.

Significance of the Project

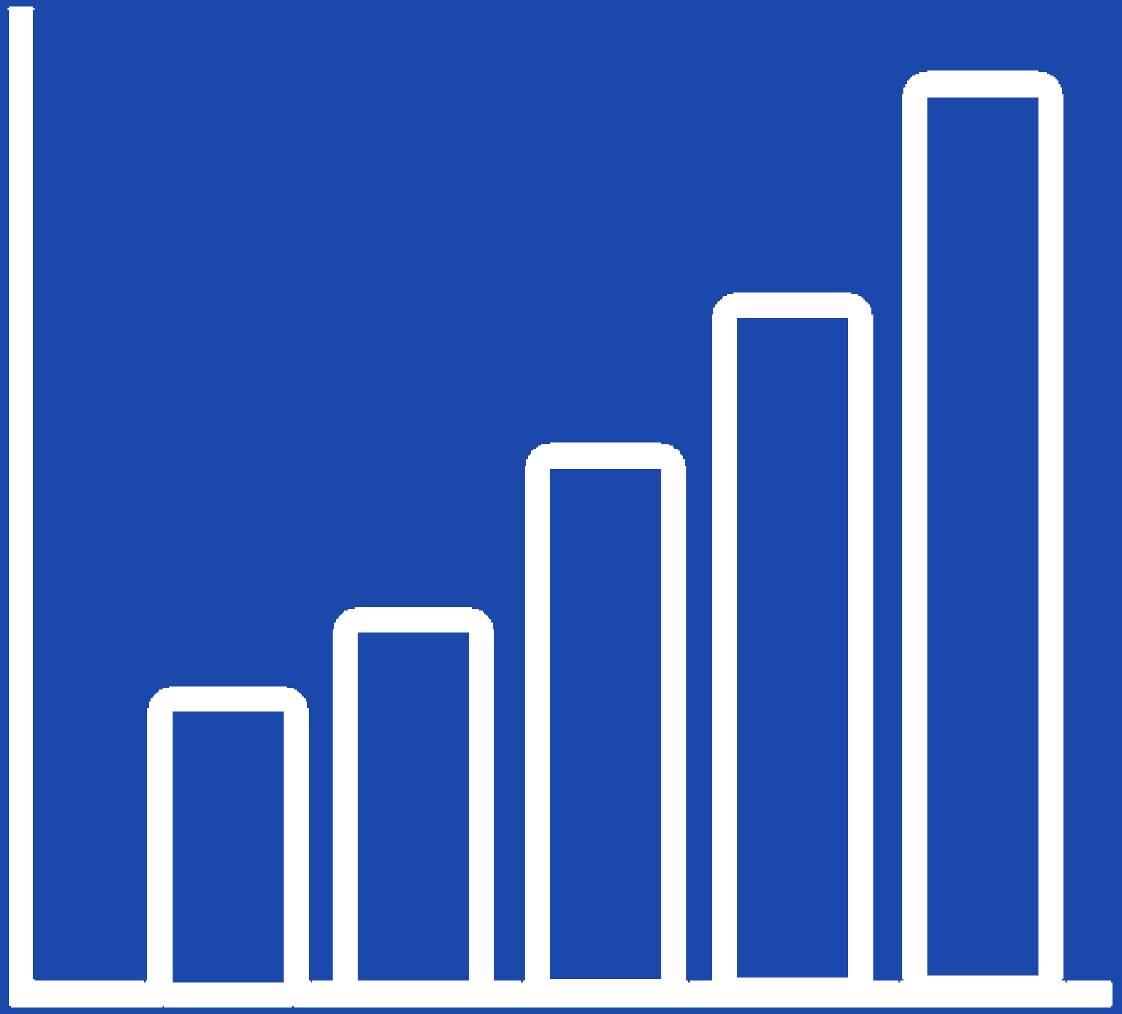
For customers

- Customer satisfaction and experience improves.

For website owners

- Customer churn decreases.
- Clients stay on website, improves profitability.

Data Understanding



Data Source

This project is part of an Engineering competition created by Facebook and Kaggle in 2015.

1. The data was retrieved from the Kaggle website in csv format.
2. One of the richest data of its kind, a world class machine learning problem.
3. Great potential for feature engineering.

About the Data

There are two datasets:
- bidder dataset (train and test)
- bid dataset
- Over 3 million bids (data points)

Challenge

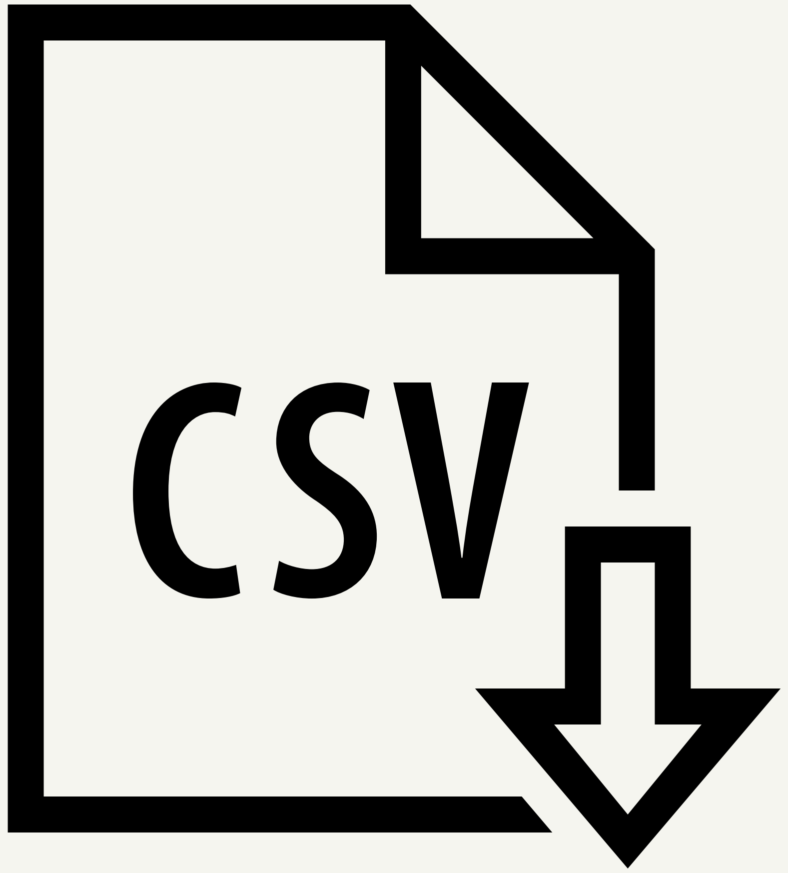
- Obfuscated fields for privacy
- Unique identifiers

Data Wrangling

- Merge on bidder_id
- EDA was performed to identify patterns and inform feature engineering.
- Data was later on rearranged at bidders level (1984 rows).

Data Cleaning

- The data came relatively clean.
- 29 missing data points were dropped (mapped to human data).



DATA FIELDS

For bidder dataset:

- bidder_id
- payment_account
- address
- outcome

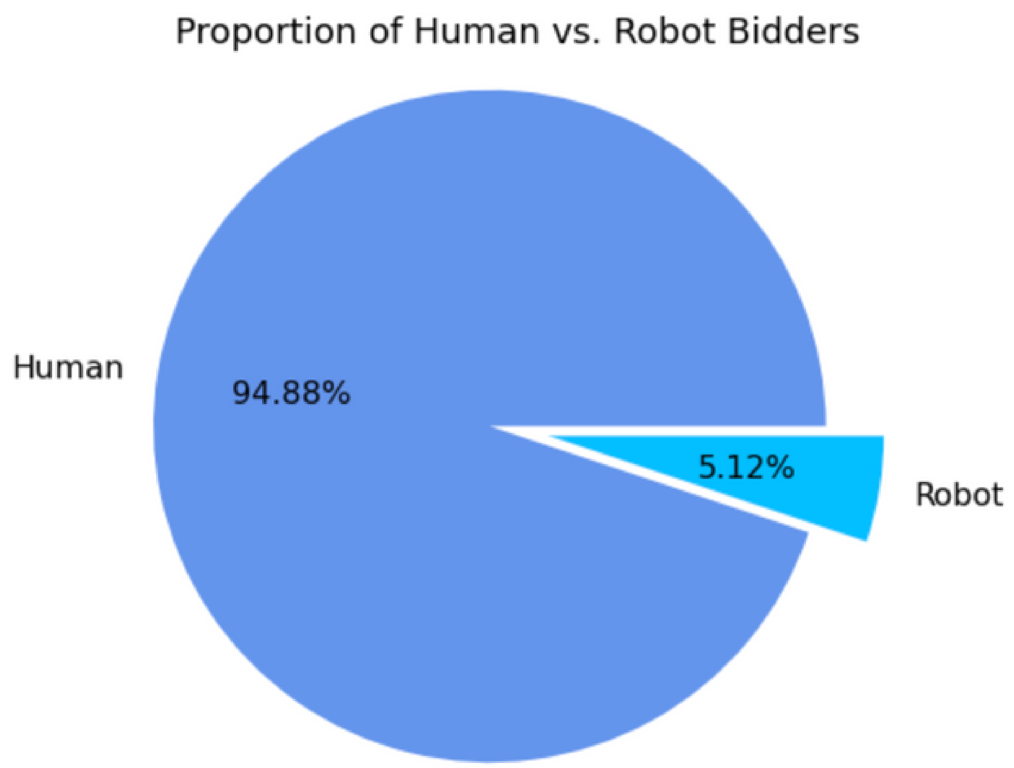
For the bid dataset:

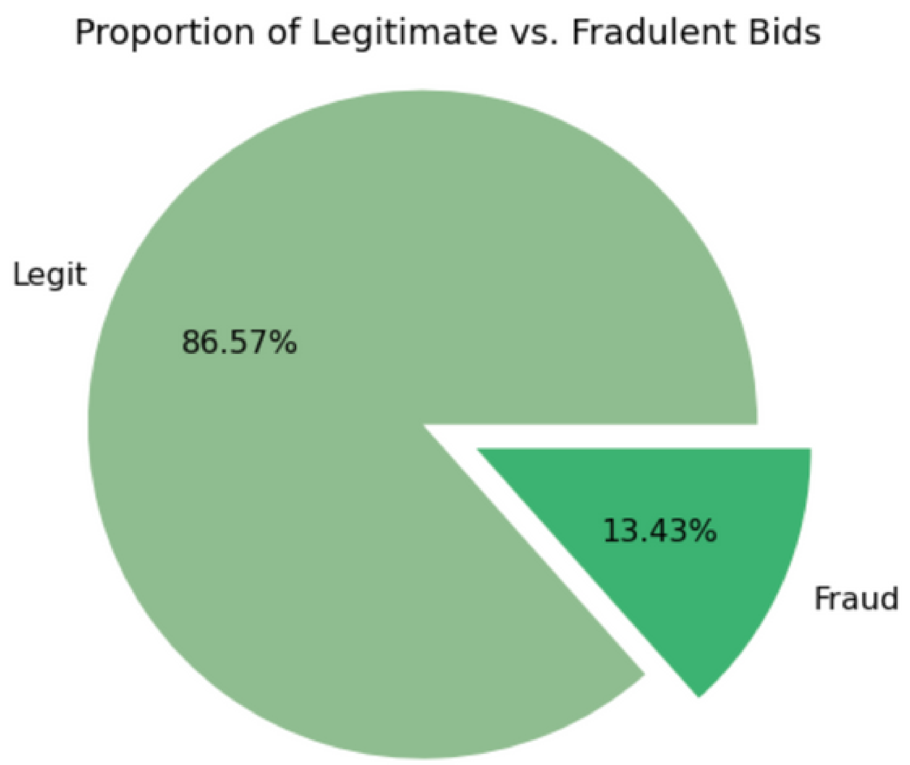
- bid_id
- bidder_id
- auction
- merchandise
- device
- time
- country
- ip
- url

Exploratory Data Analysis

Initial Hypotheses

1. Total number of bids
2. Number of bids per auction
3. Number of distinct IP addresses
4. Favorite Merchandise
5. Number of devices

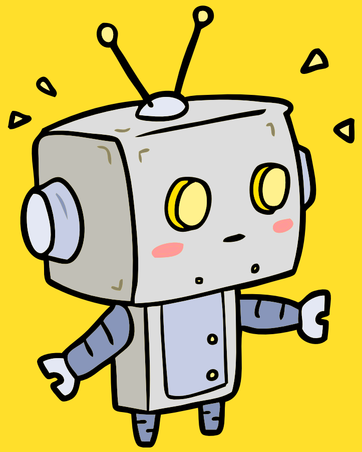




First Impressions

- The data is highly unbalanced.
- 12,740 auctions
- 5,729 devices
- 199 countries
- 663,873 unique URLs
- 1,030,950 unique IP addresses

EDA FINDINGS



Initial hypotheses held true.

The mean and median number of occurrences between human and robots differ significantly for several variables.

Highlight 1

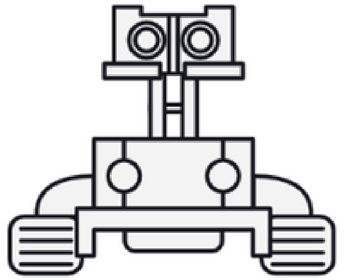
- Average number of bids per robot = 4,004, median = 716.
- Average number of bids per human = 1443, median = 14

Highlight 2

- Average number of IP addresses per robot = 2,388, median = 290.
- Average number of IP addresses per human = 581, median = 11

HUMAN VS. ROBOT DESCRIPTIVE STATISTICS

```
Mean of bids per robot: 4004.04
Median of bids per robot: 716.0
Mode of bids per robot: 1
Robot user with more bids: 161935
Robot user with less bids: 1
-----------------------------
Mean of auctions per robot: 145.04
Median of auctions per robot: 74.0
Mode of auctions per robot: 1
Robot user with more auctions: 1018
Robot user with less auctions: 1
-----------------------------
Mean of countries per robot: 26.48
Median of countries per robot: 13.0
Mode of countries per robot: 1
Robot user with more countries: 179
Robot user with less countries: 1
-----------------------------
Mean of IPs per robot: 2387.80
Median of IPs per robot: 290.0
Mode of IPs per robot: 1
Robot user with more IPs: 111918
Robot user with less IPs: 1
-----------------------------
Mean of devices per robot: 163.61
Median of devices per robot: 78.0
Mode of devices per robot: 1
Robot user with more devices: 1144
Robot user with less devices: 1
-----------------------------
Mean of urls per robot: 544.58
Median of urls per robot: 88.0
Mode of urls per robot: 1
Robot user with more urls: 8551
Robot user with less urls: 1
-----------------------------
```



```
Mean of bids per human: 1413.51
Median of bids per human: 14.0
Mode of bids per human: 1
Human user with more bids: 515033
Human user with less bids: 1
-----------------------------
Mean of auctions per human: 58.07
Median of auctions per human: 9.0
Mode of auctions per human: 1
Human user with more auctions: 1623
Human user with less auctions: 1
-----------------------------
Mean of countries per human: 12.68
Median of countries per human: 3.0
Mode of countries per human: 1
Human user with more countries: 164
Human user with less countries: 1
-----------------------------
Mean of IPs per human: 581.26
Median of IPs per human: 11.0
Mode of IPs per human: 1
Human user with more IPs: 109159
Human user with less IPs: 1
-----------------------------
Mean of devices per human: 73.95
Median of devices per human: 8.0
Mode of devices per human: 1
Human user with more devices: 2618
Human user with less devices: 1
-----------------------------
Mean of urls per human: 335.19
Median of urls per human: 4.0
Mode of urls per human: 1
Human user with more urls: 81376
Human user with less urls: 1
-----------------------------
```

Feature Engineering

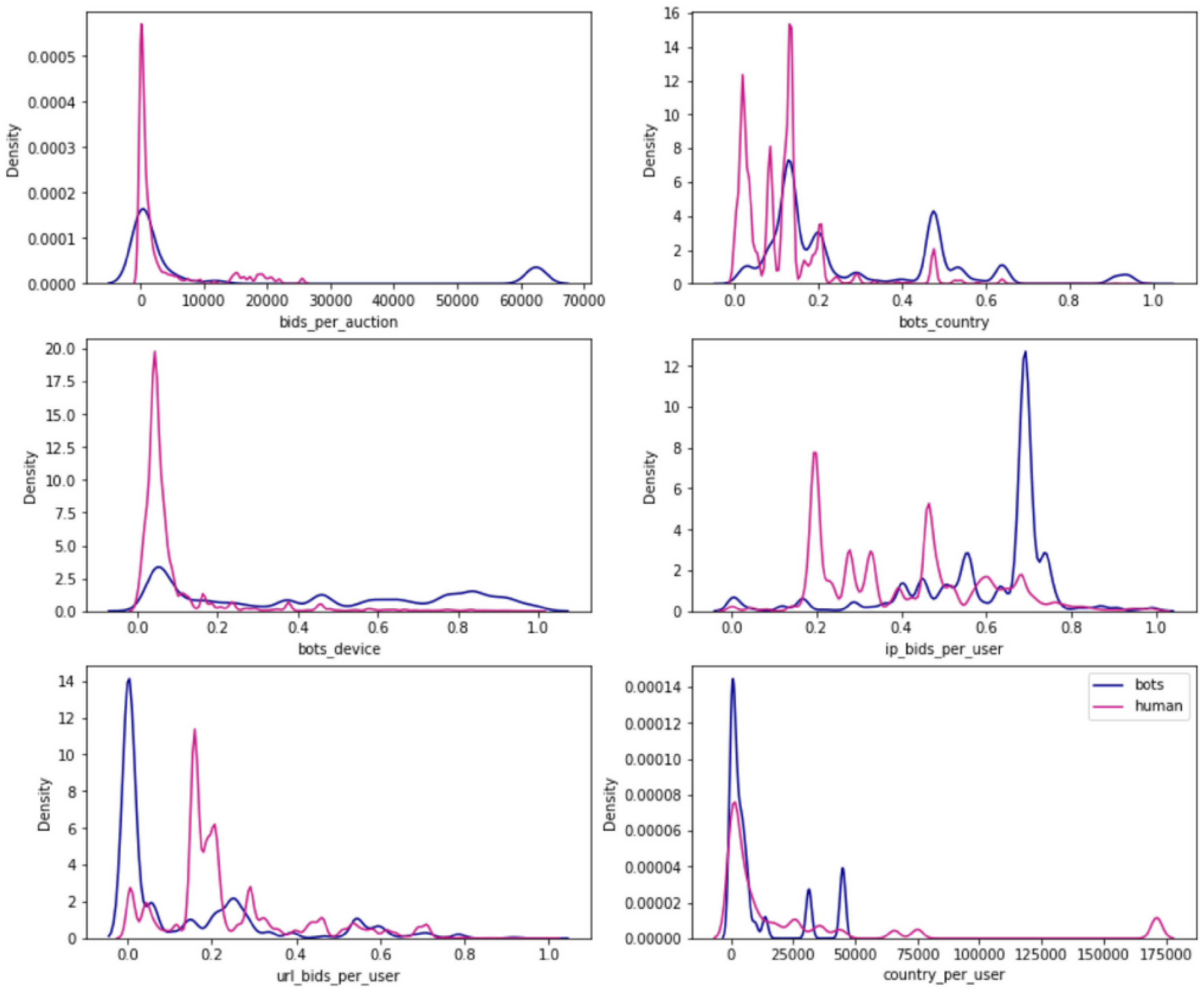
23 Features were created summarizing information at the bidder_id level

The major differences between robots and humans are in the number of occurrences.

Features were created to highlight the differences in behaviors.

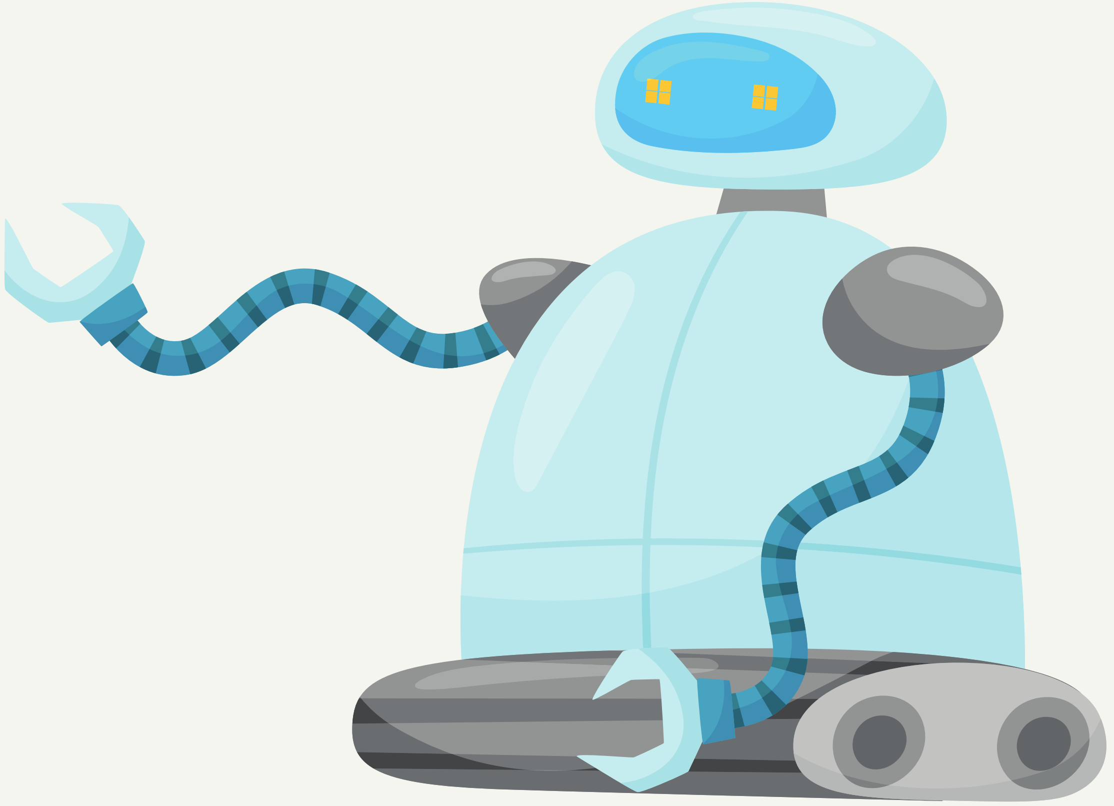
Some of the Features:

- Number of bids per auction by bidder_id
- Number of countries per bidder_id
- Number of IP addresses per bidder_id
- Number of URLs per bidder_id
- Number of same IP addresses per auction for bidder_id
- Mean number of bids per bidder_id
- Median number of bids per bidder_id
- Mean number of countries per bidder_id
- Median number of countries per bidder_id
- Mean number of IP per device per bidder_id



Advanced Analytics & Insights

Machine Learning



Binary Classification

Predicting a class: "human" or "robot"

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1 score

Models applied

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier

Models Summary

Accuracy is 0.937;

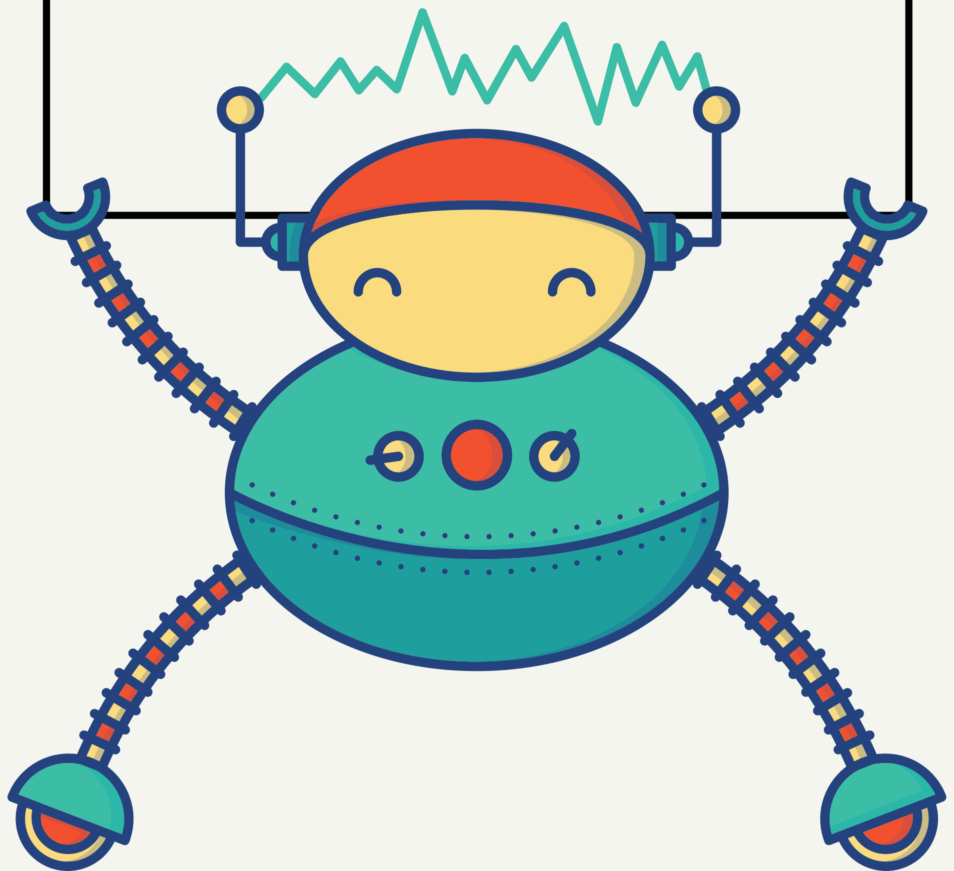
- True Negatives: 371
- False Positives: 3
- False Negatives: 22
- True Positives: 1

LOGISTIC REGRESSION

Accuracy is 0.953;

- True Negatives: 374
- False Positives: 0
- False Negatives: 20
- True Positives: 3

RANDOM FOREST CLASSIFIER



Accuracy is 0.945;

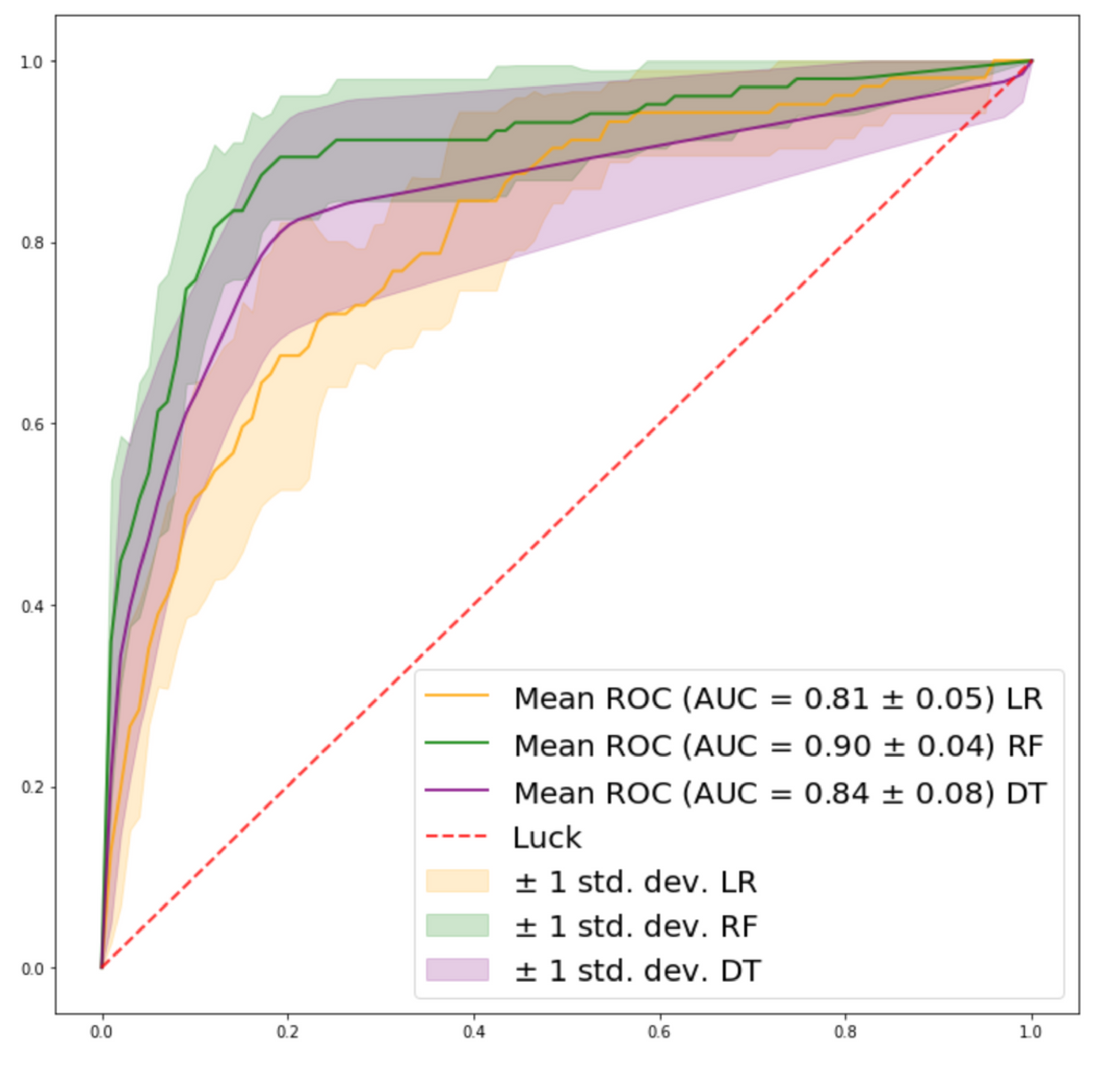
- True Negatives: 367
- False Positives: 7
- False Negatives: 15
- True Positives: 8

DECISION TREE CLASSIFIER

Evaluation

ROC - AUC Curve

- Random Forest has the highest AUC value. The best performance at distinguishing between the classes.



What are the best metrics to evaluate our model?

Precision: ability of the model to return only relevant instances. In this case, we want to minimize false negative, and don't want 'robots' to be classified as 'humans'.
Recall: ability of the model to identify all relevant instances, that is True Positive Rate, aka Sensitivity. We want the least false positive, minimize 'humans' classified as 'robots'.
F1 Score: returns a harmonic mean of precision and recall, indicating a balance between Precision & Recall. Therefore, a model that has a high F1 score can be a good model for us too.

Conclusion &

Suggestions for future improvement

Conclusion

1. Highest value of true positives and a low value for false positives.
2. High recall model: keep false positive low, but in case we misclassify a human for a robot, it isn't as damaging as keeping 'robot' bidders on the website.
3. Decision tree classifier has the highest true positives.
4. Add authentication steps to avoid banning misclassified human bidders.

Questions?

Thank you!