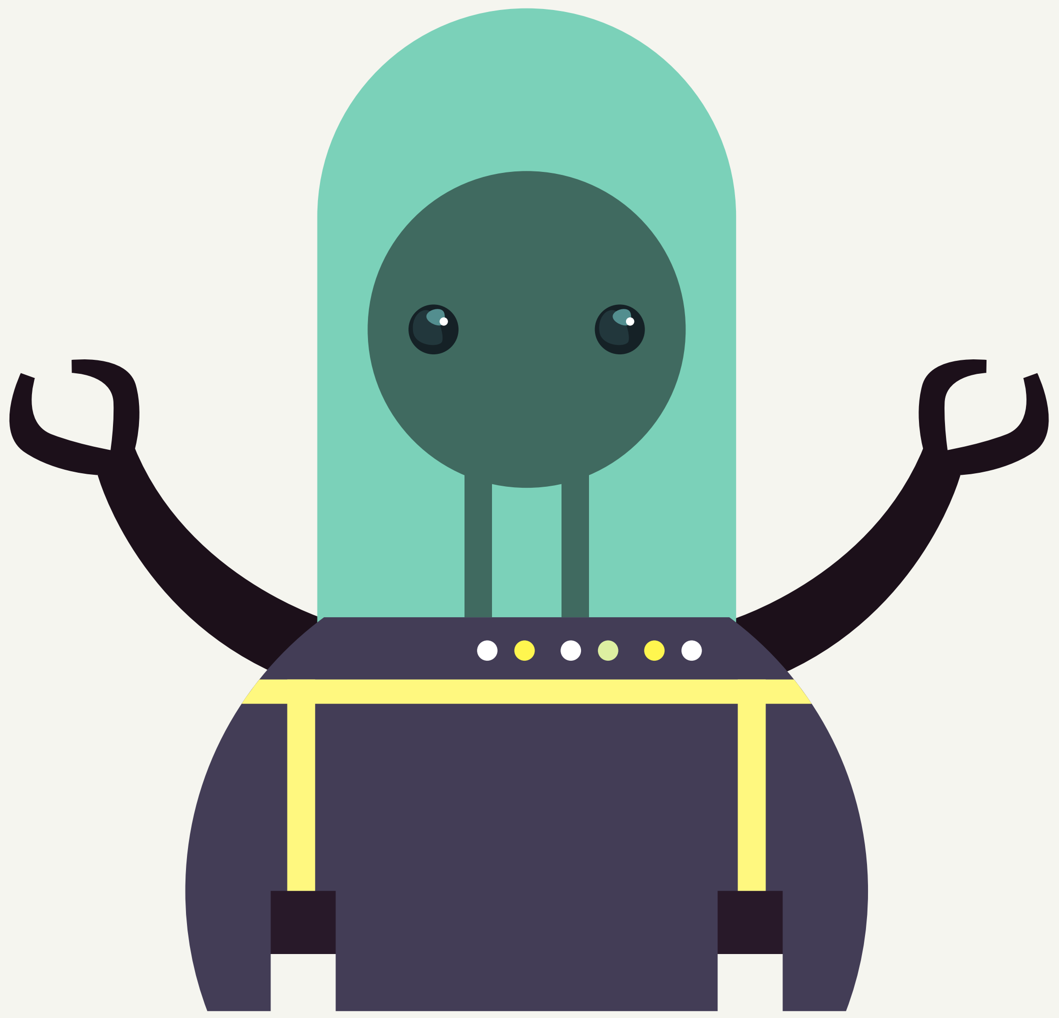
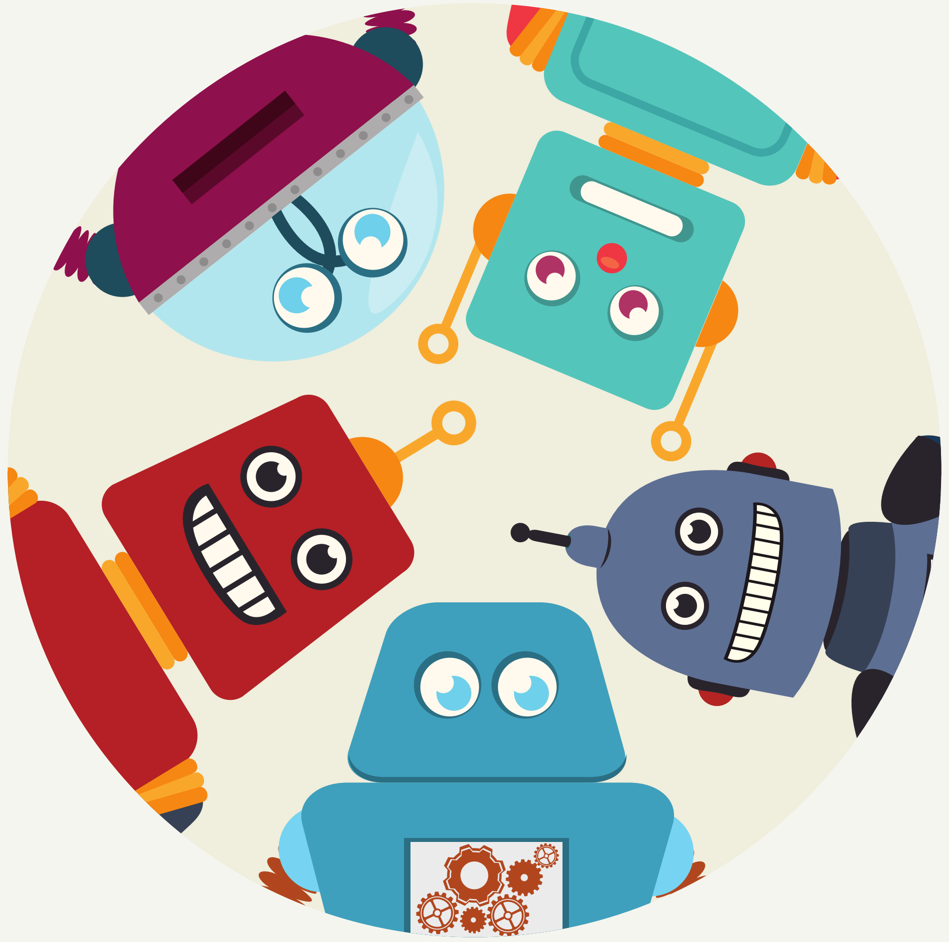
Auction Fraud Detection

"Human or Robot"

Springboard School of Data

by Gabrielle Wald



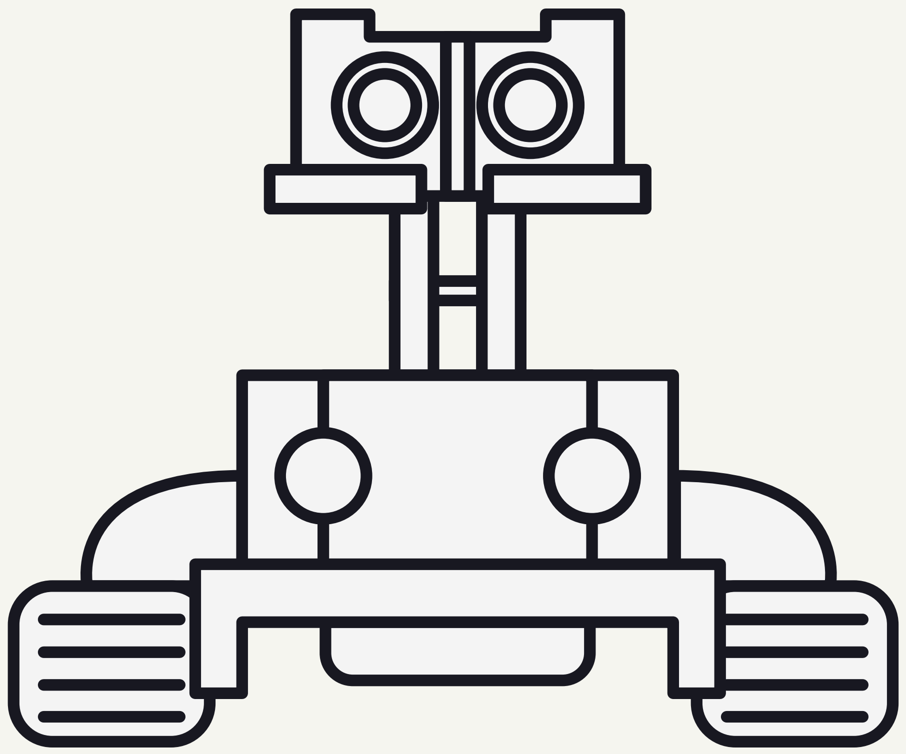


Introduction

Background

Overview of the project

- Auction website
- Frustrated human bidders
- Inability to win auctions vs. their software-controlled accounts
- Customer base is plummeting
- Rebuild customer happiness by eliminating computer generated bidding from their auctions.



The Problem

What we want to solve

- Prevent customer churn and improve 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

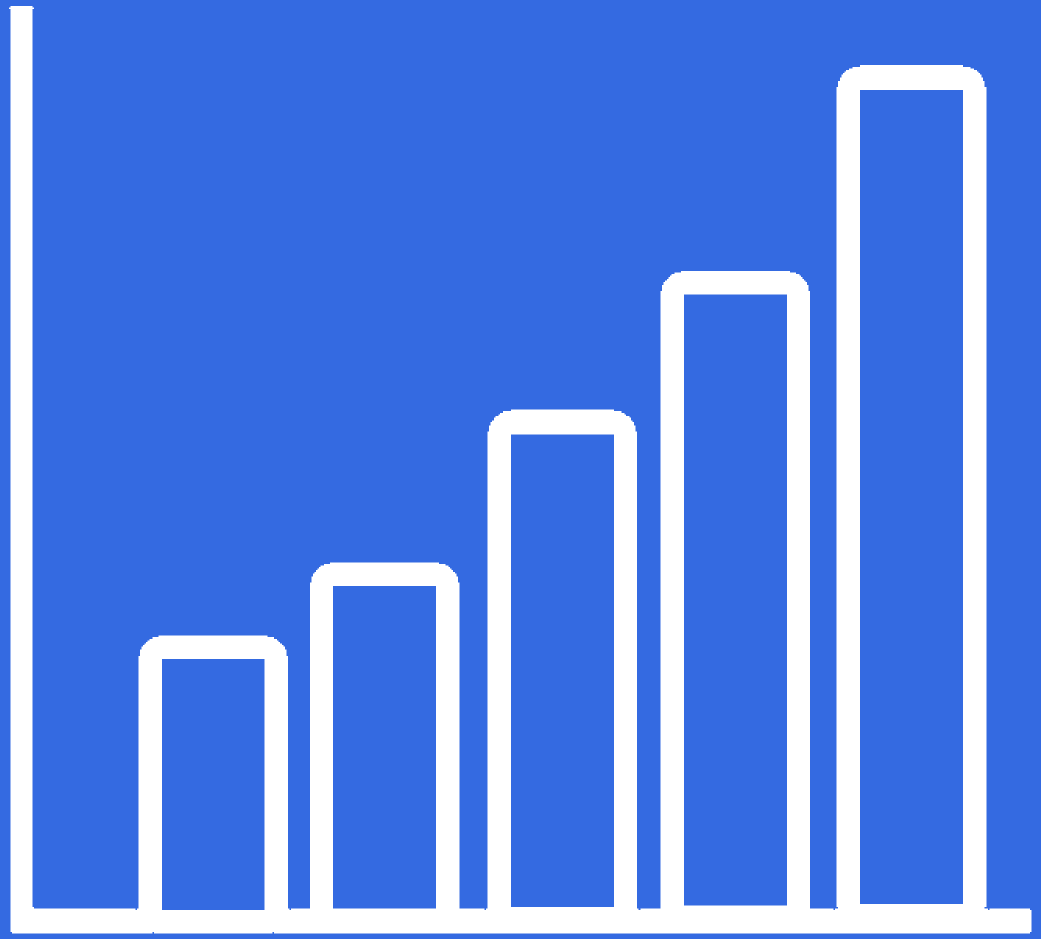
For customers

- 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.
3. Great potential for feature engineering.

About the Data

There are two datasets:
- bidder dataset (train and test)
- bid dataset
- Over 7 million bids (data points)

Challenge

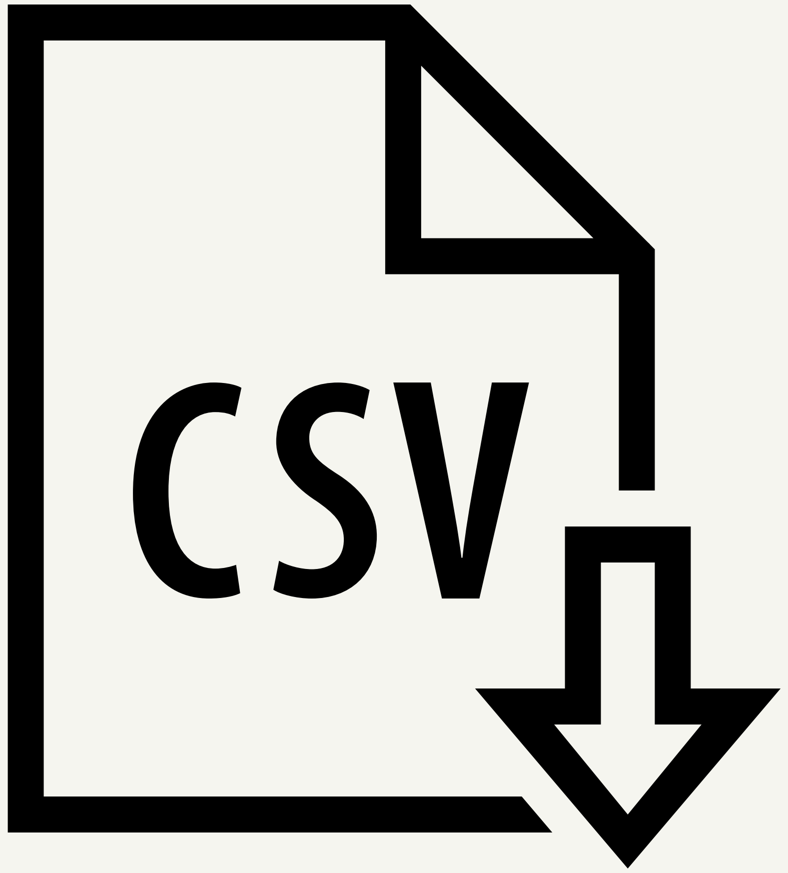
- Obfuscated fields for privacy
- Unique identifiers

Data Wrangling

- Merge on bidder_id
- Performed EDA to identify patterns and inform feature engineering.
- Dataset prepared for modeling at bidder level (1984 rows)

Data Cleaning

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



DATA FIELDS

For bidder dataset:

- bidder_id
- payment_account
- address
- outcome

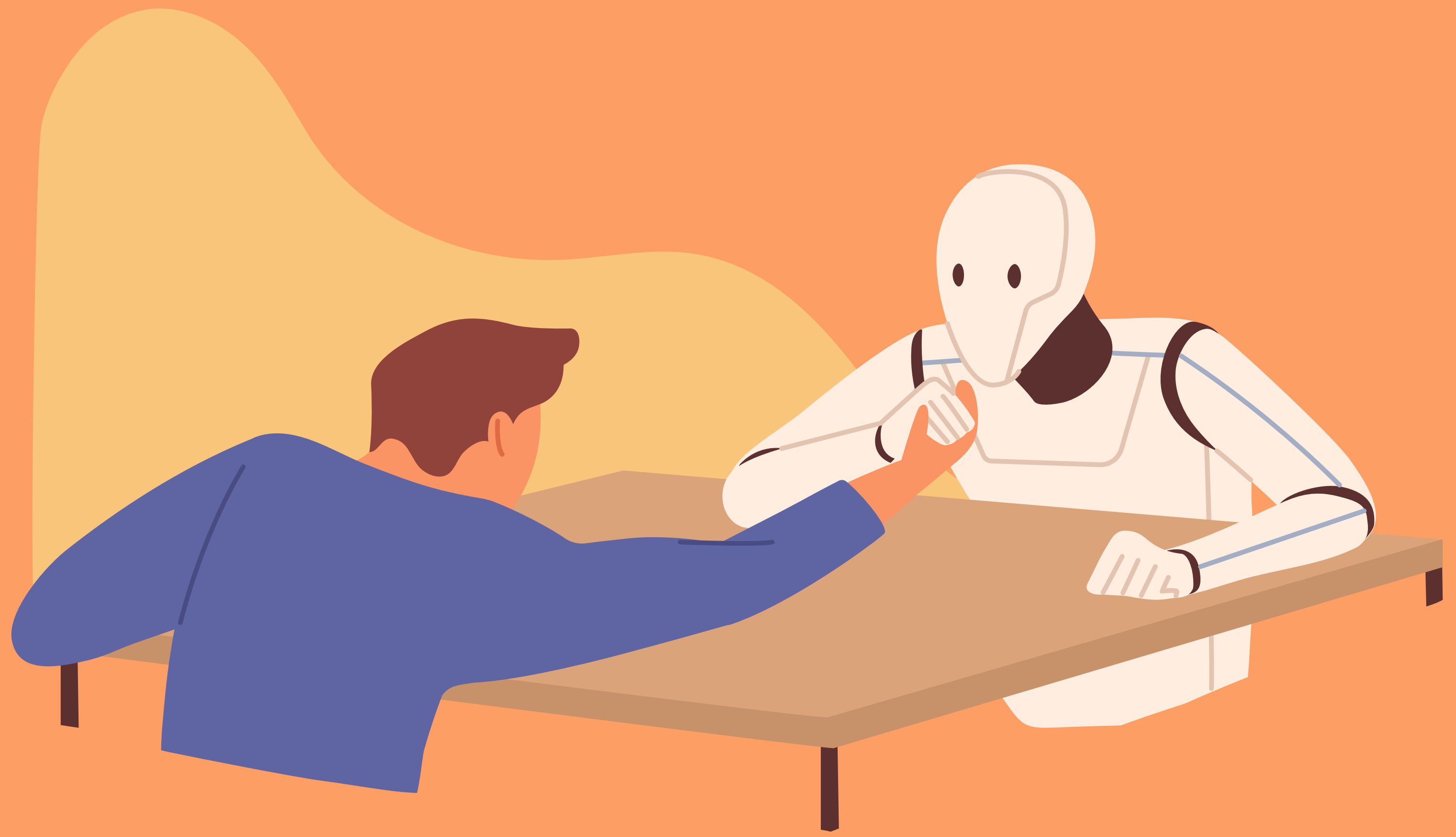
For the bid dataset:

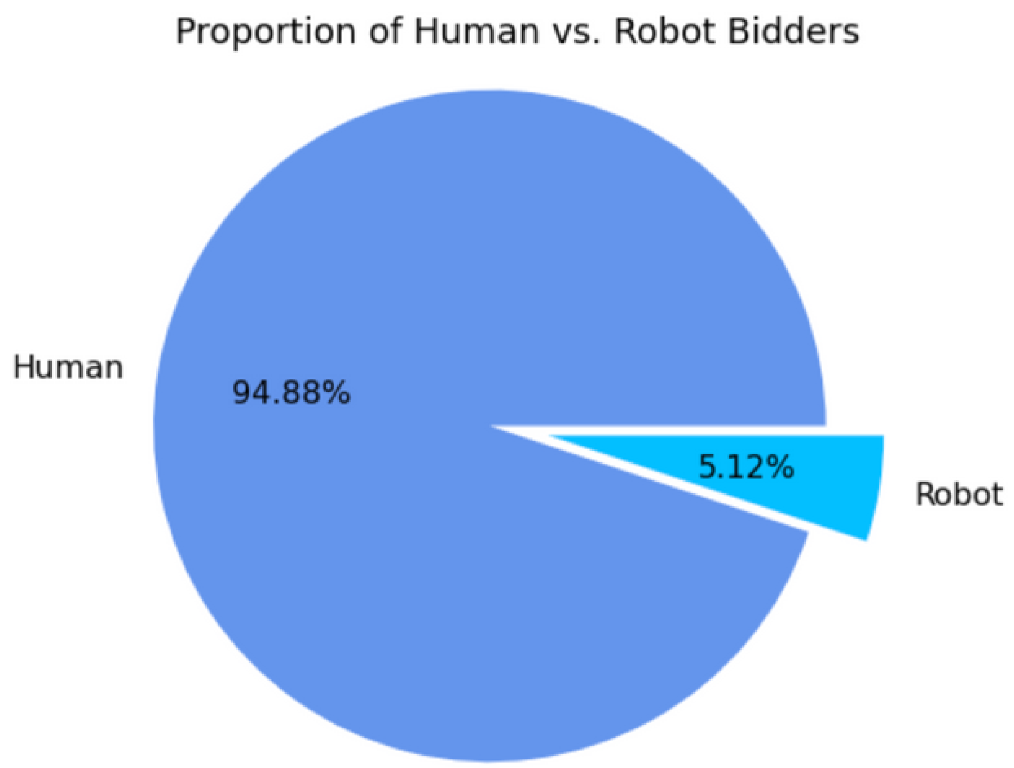
- bid_id
- bidder_id
- auction
- merchandise
- device
- time
- country
- ip
- url

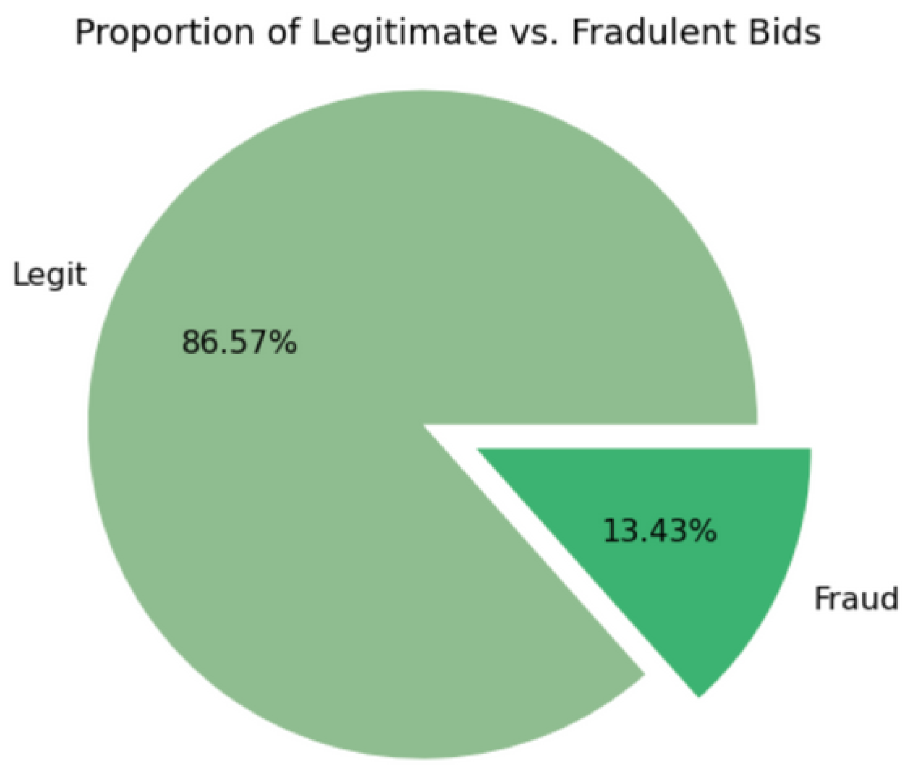
Exploratory Data Analysis

Initial Assumptions

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





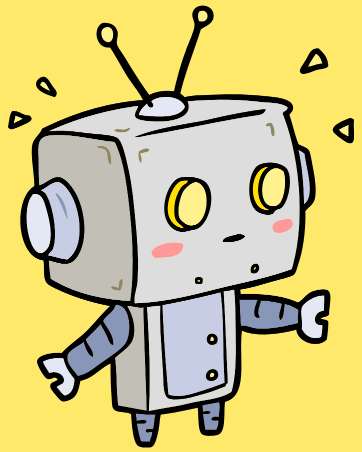


First Impressions

- 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

Robots vs. Humans



Initial assumptions held true.

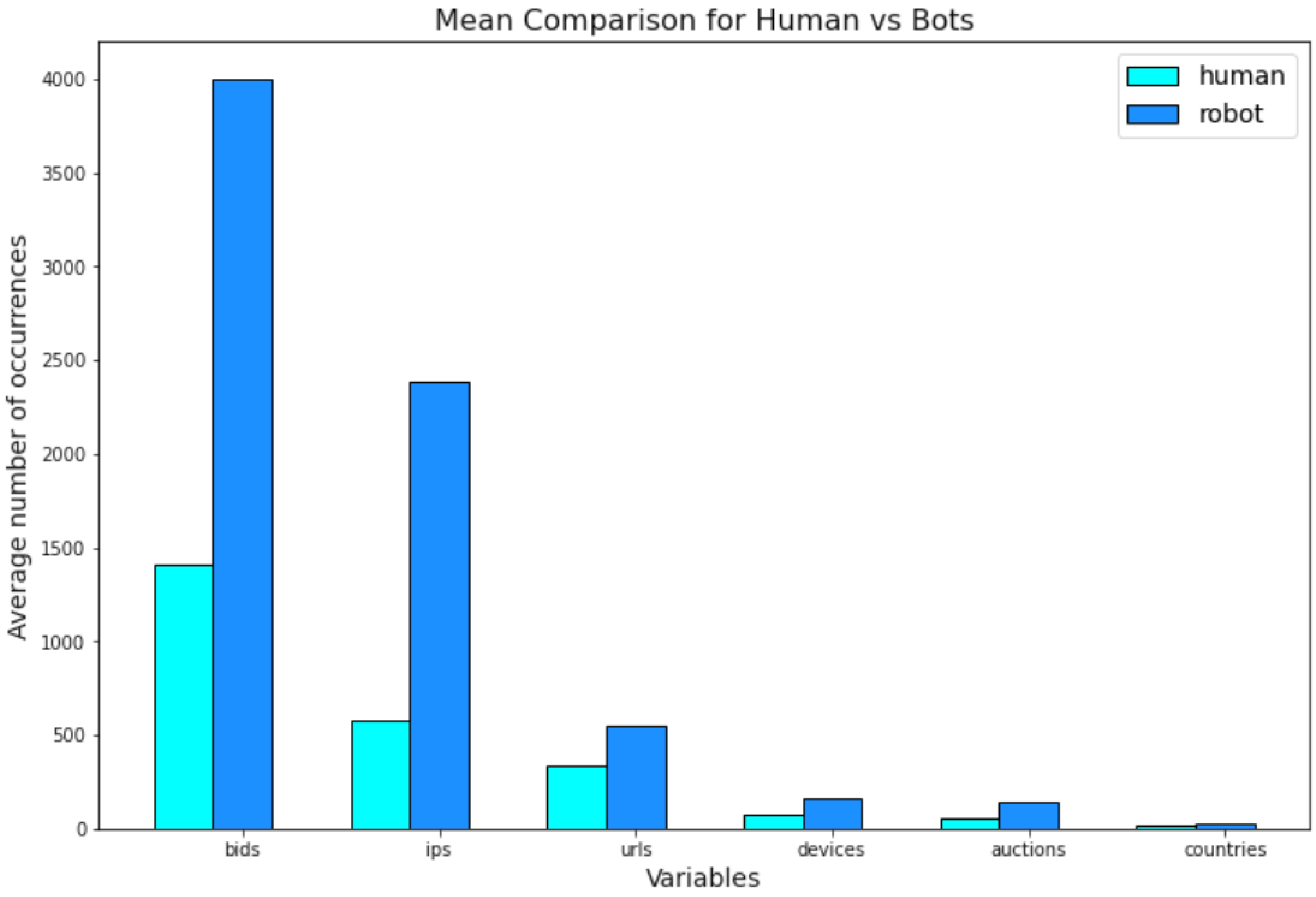
The mean and median number of occurrences between human and robots differ significantly for several variables.

Bids

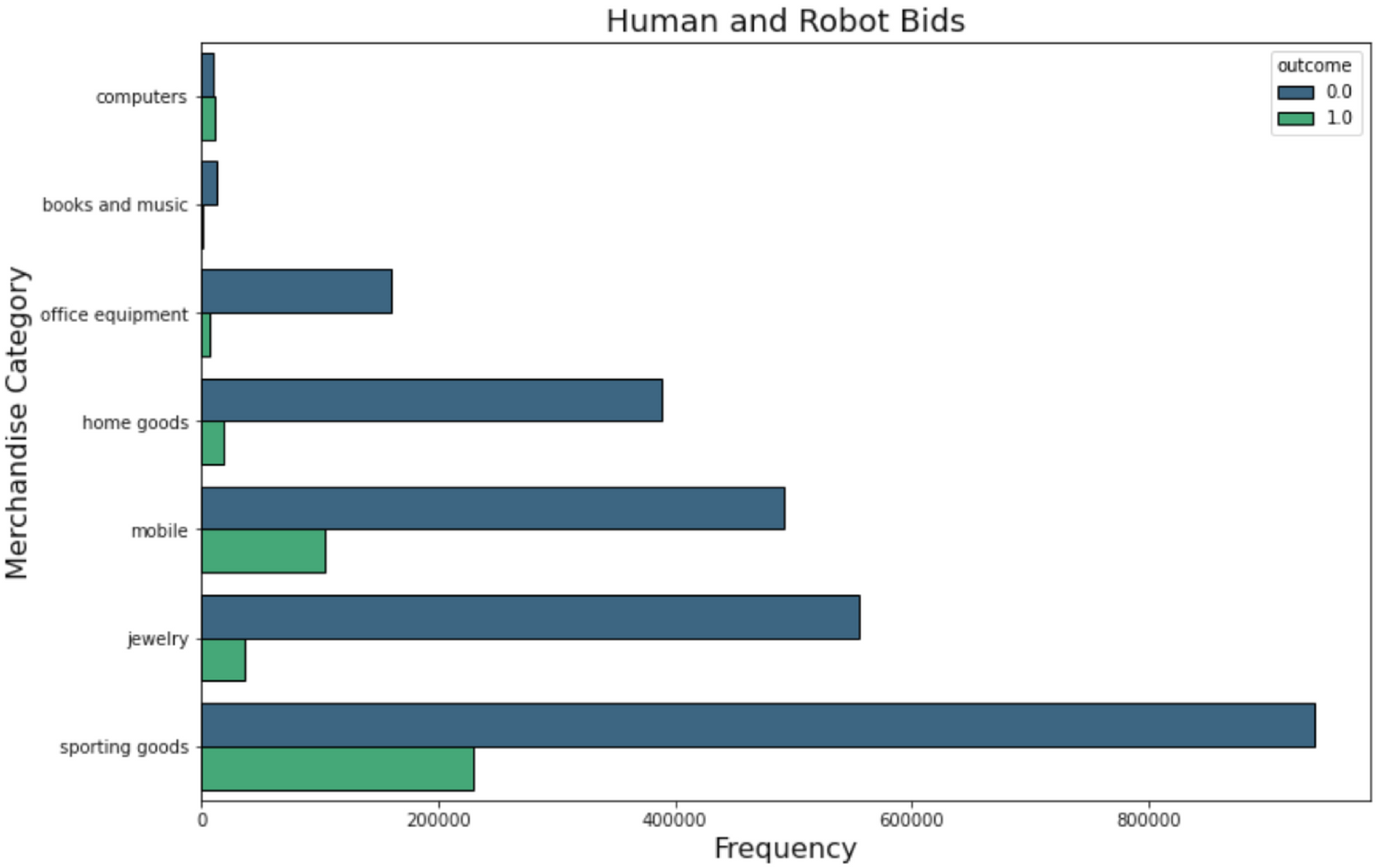
- ROBOT
  - Mean = 4,004, Median = 716.
- HUMAN
  - Mean = 1443, Median = 14

IP Addresses

- ROBOT
  - Mean = 2,388, Median = 290.
- HUMAN
  - Mean = 581, Median = 11



New features will highlight these differences in behaviors.



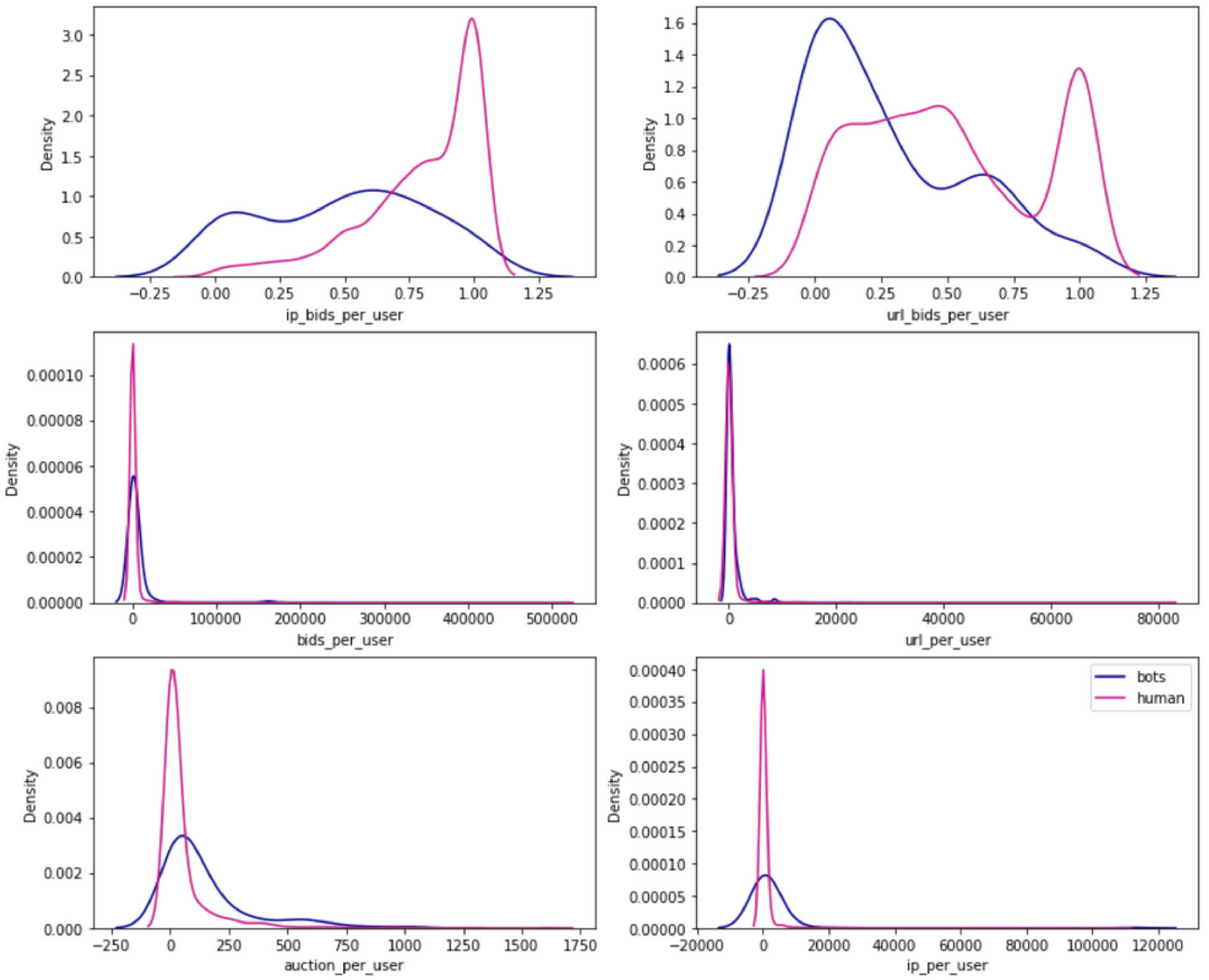
New features will highlight these differences in behaviors.

Feature Engineering

16 Features Created

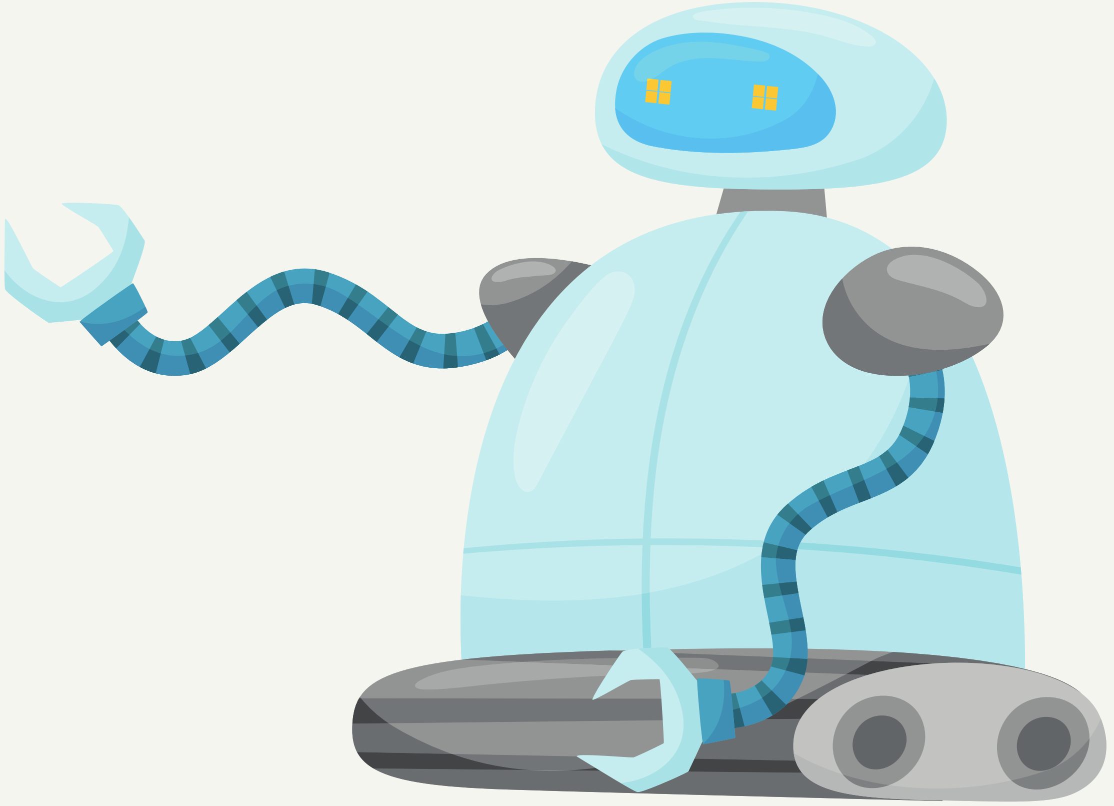
Some of the Features:

- 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 number of auctions for each country per user
- Mean/median number of IP addresses per auction per user



Advanced Analytics & Insights

Machine Learning



Binary Classification

Predicting a class: "human" or "robot"

Evaluation Metrics

- Recall
- AUC

Models applied

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier

Models Summary

Accuracy is 0.94;

- True Negatives: 372
- False Positives: 2
- False Negatives: 22
- True Positives: 1

LOGISTIC REGRESSION

Accuracy is 0.95;

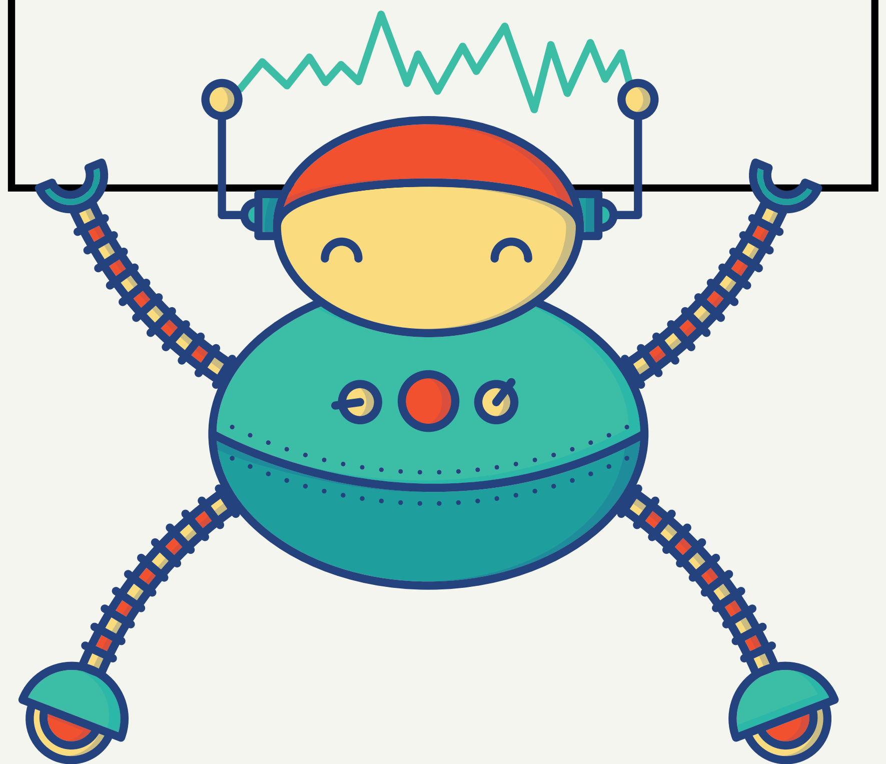
- True Negatives: 374
- False Positives: 0
- False Negatives: 19
- True Positives: 4

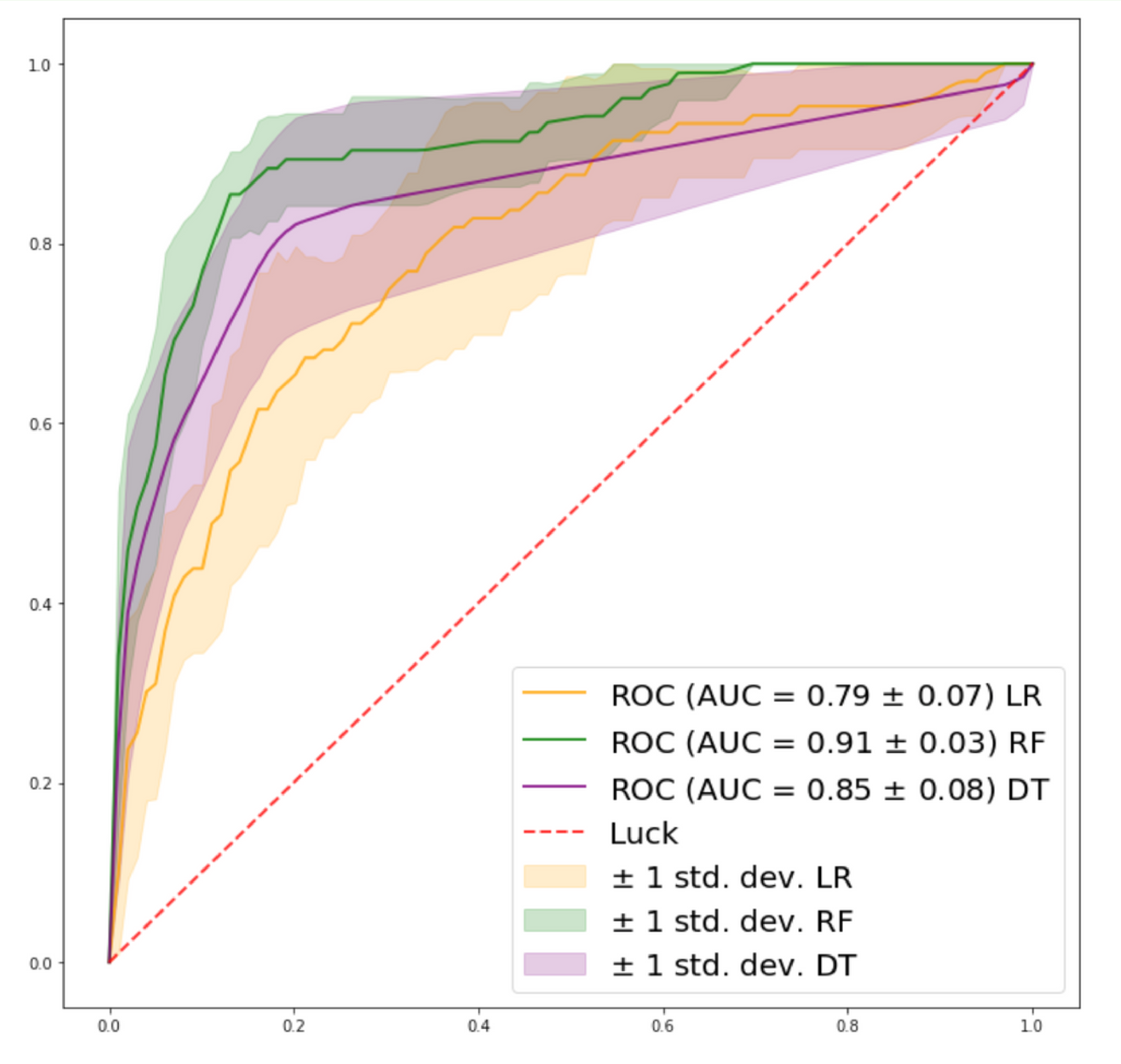
RANDOM FOREST CLASSIFIER

Accuracy is 0.94;

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

DECISION TREE CLASSIFIER





Evaluation

ROC - AUC Curve

- Random Forest has the highest AUC value:
    - 0.91 +/- 0.03

Misclassification

Type I error - False Positive (Human classified as robot) is better than type II error - False Negative (Robot classified as human).

Possible Solution

CAPTCHA is one approach to manage with type I error.

Conclusion &

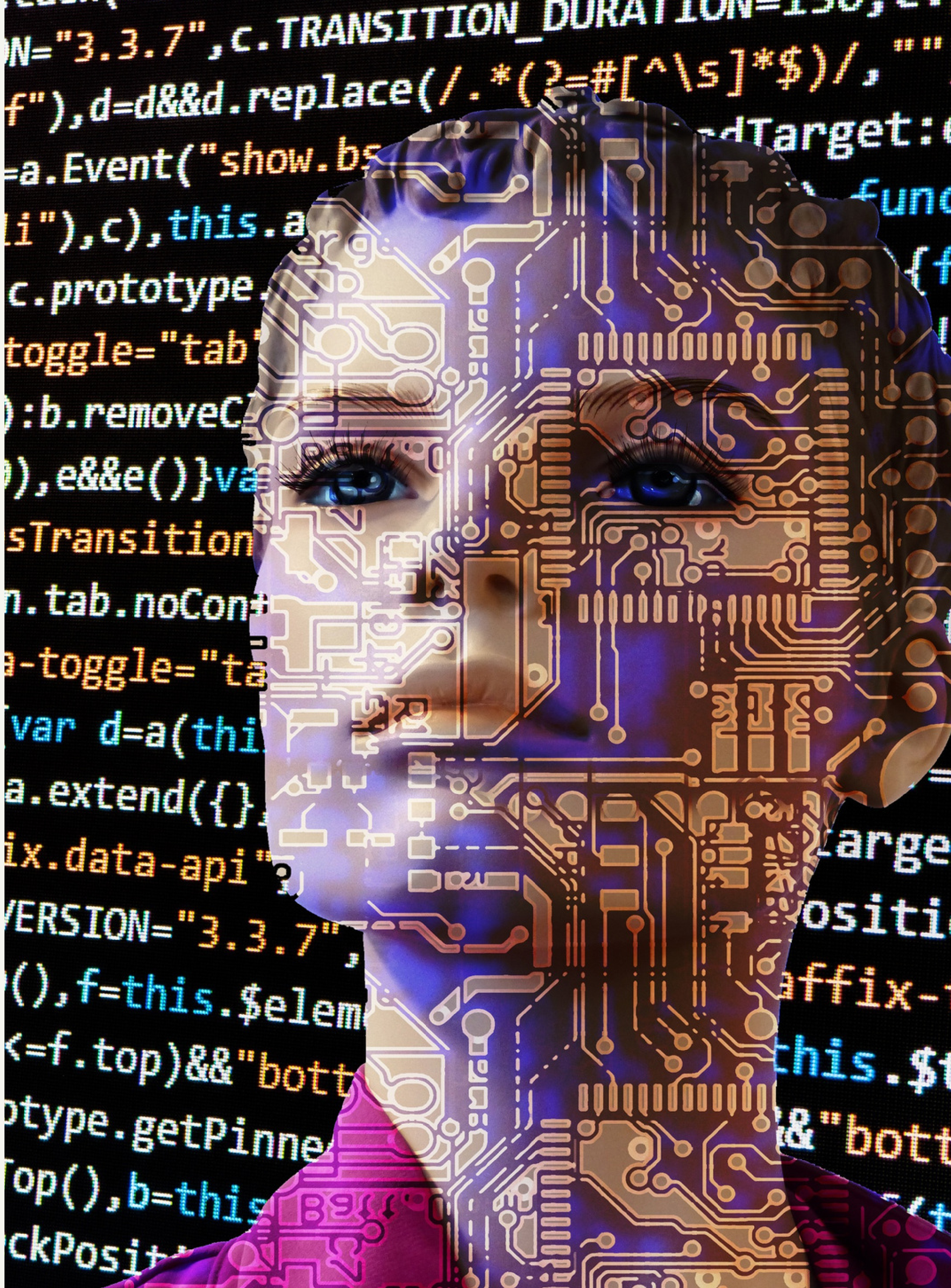
Suggestions for future improvement

Conclusion

1. Decision Tree classifier has the highest recall.
2. Random Forest classifier has the highest AUC.
3. Neither has a high success rate identifying robots.
4. Models are not ready to go into production.
5. Small sample size (Test set has 397 users, 23 labeled robots).

Next Steps

1. Create new features from the time column like:
    a. Maximum number of bids made within a 20 minute span;
    b. Median time between a user's bid and that user's previous bid;
    c. Number of simultaneous bids
2. Try XGBoost, Support Vector Machine and Naive Bayes.
3. Require more data.



Questions?

Thank you!