**Human Or Robot?**

Facebook Recruiting IV: Human or Robot?

## Abstract

*Our project aims to use data science and machine learning to solve a real-life problem of identifying robot bidders from human bidders. Our ML pipeline involves pre-processing, feature extraction, feature selection, model selection and classification tests. Experiments reveal different performance achieved by different models, our best performance achieves XXX and is between the Xth and Yth position of the competition.*

## Introduction

## Overview

The project is based on a past Facebook recruiting competition ("Facebook Recruiting IV: Human or Robot?”, 2015). The project aims at predicting for a specific bid whether the bidder is a robot or a human based on their behavior. The motivation is to facilitate a fair competition on the bidding website and therefore to improve the user’s satisfaction. The project is based on real-life dataset of an online auction website working with Facebook.

## Problem Statement & Description

The goal of this competition 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 auction activity.  
Our project is a supervised binary classification problem – determining one bidder between a human or a robot based on limited past bidding records. More formally the problem can be formulated: Given a set XBidders of N bidders, set of X’bids of M bids and set Y with 2 labels (human, robot), a machine classifier should predict for each sample of XBidders the appropriate corresponding label from Y.  
We use 2 evaluation criteria:

1. Area under the ROC curve (AU-ROC) which is used in the competition in order to be able to compare the results with the leaderboard.
2. Average Precision is used on the train data in order to assess the quality of the prediction because the data is highly imbalanced. Since the ratio of robot bids to human bids in the dataset is 3:20, if we label all bids as human we still could get 86.75% accuracy with high precision, however it will not be a good classification mechanism.

The main point of focus of the project is handling complex time-series based dataset.

## Related Work

There is a substantive amount of work in the field of online fraud prevention and in the field of Bots Identification which relates to our specific project.  
More specifically Online Auction Fraud Prevention and online auctions bots identification …..  
In regards of the specific contest. for consumers, the chances of landing a winning bid in online auctions has become increasingly more difficult with the abundance of “bidding robots”. Bidding robots such as “BidRobot” and “Auction Sniper” are pieces of software that are configured by the user to follow any number of auctions on different auction sites simultaneously, bidding in place of the user according to predefined settings and preferences.\

Chau & Faloutsos research tackle the problem of fraud detection in electronic auctions that use a Random Forest model leveraging price-based features to differentiate fraudsters from normal bidders. Packer & Huang tackled the Facebook competition and achieved 0.8747 AUROC using AdaBoost. Gu & Shi published a research based on the Facebook Competition in which they tested different models and achieved 0.94 AUROC in the private leaderboard.

## Dataset Description

The dataset includes basic account information and bidding records of all the bidders. There are no pre-built features at all, one most construct their own features with the bidding records provided. The dataset is comprised of the following two parts:

1. Bidder Information: Basic information regarding the bidder accounts and the appropriate labels ( labels of test bidders are not provided)
2. Bidding Information: Information regarding each bid in each auction

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | Bidders Dataset | | | | | # | Field | Detail | Comments | | 1 | Bidder\_Id | Unique bidder identifier |  | | 2 | Payment\_account | Bidder Payment account | Obfuscated | | 3 | Address | Mailing address of the bidder | Obfuscated | | 4 | Outcome | Tagging Label (1 for a robot and 0 for a human) |  | | |  |  |  |  | | --- | --- | --- | --- | | Bidding Dataset | | | | | # | Field | Detail | Comments | | 1 | bid id | Unique ID for the bids |  | | 2 | bidder id | Unique bidder identifier |  | | 3 | auction | Unique identifier of an auction |  | | 4 | merchandise | category of the auction site campaign |  | | 5 | Device | Phone model of a visitor |  | | 6 | Time | Time that the bid is made | Transformed | | 7 | Country | country that the IP belongs to |  | | 8 | Ip | Bid IP Address | Obfuscated | | 9 | url | URL the bidder was referred from |  | |

## Data Analysis & Visualization

We began by examining the data to understand important information regarding the dataset distribution and information regarding key behaviors that differentiate human and bot bidders.

1. General Dataset Information – There are 7.6M bids and 6614 unique bidders. The training set has 2087 bidders of which 1984 are labeled as human and 103 are labeled as robots. The test set had 4700 bidders, of which 4630 participated in bidding.
2. Statistics Analysis:

|  |  |  |
| --- | --- | --- |
| Statistic | Human | Robot |
| # of Bids | 1414 | 4004 |
| # of Bids Per Auction | 6 | 23 |
| # of Auctions Won | 6 | 18 |
| # Devices | 164 | 74 |
| # IP | 581 | 2388 |
| # URL | 335 | 545 |

The preliminary data analysis helped us to develop crucial insight regarding the data-science process and the differences between Robots and Human

1. The dataset is highly imbalanced – the ratio of bot to human bids in the train dataset 3:20. This potential raises overfitting threats if the test distribution is different then the train.
2. Robot have different bidding strategy than humans, they tend to be more active, bid more and win more auctions.
3. Robots tend to change the parameters they use more than humans – they use more IP addresses, devices and URLs.

After performing Time-Series analysis (our point of focus), we understood additional insights regarding the different between humans and robots (more information in paragraph 4.4)

1. Robots Bidding time patters are different that humans – less variance and dependency on time of the day.
2. Robots bid faster than humans (consecutive bids time differences)

## The Solution

## General Approach

Our approach to the problem displayed is based on the following stages:

1. Data Visualization & Analysis – understanding distinguishing patters between the robots and humans and develop hypotheses based on the differences.
2. Data Pre-processing – Cleaning that data and handling both NaN values and bidders without activity. Encoding the categorical variables (Merchandise, Payment Accounts, Address and etc) and correlating the bids and bidders.
3. Time-Series Analysis and Processing – analyzing the time-series data, creating time-based insights for the auction’s timelines and user activity and processing timestamps.
4. Feature Extraction – Perform Feature extraction based on our hypotheses
5. Model Selection & Evaluation
6. Compare, Improve and Iterate – Analyze specific samples missed or labeled incorrectly, create correlating features and properly tune the model parameters.

## Data Preprocess

## Data Cleaning

Initial data pre-process analysis raised several problems with noisy, inconsistent and missing values. There were 2 main problems to address regarding missing values:

1. Bidders without bids – 29 bidders were identified that didn’t have any bid in the bids dataset. All of the 29 bidders were labelled as human and therefore were dropped from the dataset without affected future classification.
2. Missing values – there were 2701 with inconsistent and missing value for the country the bid was performed from. These bids are 0.09% of the total amount of bids and below to different bidders and therefore these entries were dropped from the dataset.

## Outliers

Further analysis of the dataset raised several concerns regarding the consistency of the data, in particular with outliers in the data:

1. Number Of Bids Outliers – 5 robot had 1 bid while the average amount of bids per robot was 4004. Even though the data is imbalanced and skewed towards human bidders, we decided to drop the 5 robot bidders.

## Data Embedding & Representations

The dataset contains several categorical features that needs to be represented differently in order to be used in models:

1. Merchandise – contains information regards category of the auction site campaign leading to the specific auction, there are 9 different categories of merchandise. This data was represented with one-hot encoding.
2. Payment Account & Address – contain information regarding the payment accounts and address that the bidder had filled and does not depend on the specific bid. The field is of type String and is characterized by high variance between the bids. We didn’t perform One-hot-encoding for the field by rather used a ‘binning’ solution by representing the address by one of the following Boolean bins – rare, infrequent and frequent.

## Time-Series Analysis – Point of Focus

Our project main point of focus was a complex problem setting such a time-series analysis as presented below. We also believe we’ve also tackled non-trivial feature extraction both in regards for the time-series analysis and in regard to the fraud activity-based features (as mention in paragraph 4.5)

Our time-series analysis methodology and work can be broken into several stages:

1. Bid Time Analysis– The time-columns in the dataset was obfuscated to protect privacy but in a way that preserves the time-order of the different bids. By examining and visualizing the activity after discretizing the time information into time-slices a specific pattern occurs – 3 separate chunks (figure X) each chunk seems like a 3-day period (figure X+1).
2. Time Stamp Processing – The obfuscated time given has 1015 magnitude, these may cause numerical overflow problems when applying different models. We needed to processes the time information and transform it to a suitable magnitude. We transformed the time units into time intervals and then reassigned time-stamps, assuming peak activity is at 7pm.
3. Aggregating Time-Based Features- Based on the time-analysis and the time-activity related differentiation of the robots and humans we’ve developed hypothesis for further feature extraction. Our main hypothesizes were that the time distribution of bids, bidding speeds, parameters lifetime length and bid streaks will be different between robots and humans. We therefore developed Time-Series based feature extraction and described in the paragraph 4.5 below.

|  |  |
| --- | --- |
| figure X | figure X+1 |
|  |  |

## Feature Extraction

Following the dataset and time-series analysis described in sections 3.2 and 4.4 we extracted and hand-crafted a list of features. Each hand-crafted feature corresponds to one of the following 4 classes of hypothesis regarding difference between humans and robots:

1. User Activity Features – Aggregative features and statistics of the bidder’s activity in general. The underlying hypothesis is that human and robots differ in used parameters and activity.
2. Auction Activity - Aggregative features and statistics of the bidder’s activity per auction. The underlying hypothesis is that robots and human differ in their auction and bidding strategy.
3. Suspicious Activity – Features based on hand-crafted Rules and past robots activity analysis. The underlying hypothesis is that robots tend to share and re-use methods of operation.
4. Time-Series Activity – Features based on the time analysis. The underlying hypothesis is that robots and humans differ in regards to activity over time.

**Feature Extraction List-**

|  |  |  |
| --- | --- | --- |
|  | Feature | |
| # | Hypothesis Class | Feature |
| 1,2,3 | User Activity | Boolean – Rare & Infrequent IP, Address and Payment Account, |
| 4 | User Activity | Boolean- is payment account equals address |
| 5-11 | User Activity | Counts Per User - # Bids, Auction, Merchandise, Device, Country, IP, UR |
| 12-15 | User Activity | Average Number Of Bids Per Parameter - # Bids/Auction, Bids/Device, Bids/URL, Bids/Country |
| 27-31 | User Activity | Advanced Country Information Most Common Country & Fraction of Bids in Each Country Per User, Median/Max/Mean Number of Countries Per Auction |
| 16,26 | Auction Activity | Average Count Per User Per Auction - IP, Country |
| 32 | Auction Activity | Number of Auction Won |
| 34 | Auction Activity | Count Of Bids Per Use Per Price Percentile |
| 43-44 | Auction Activity | Count of Bids in first and last 10% of Won Auction |
| 45 | Auction Activity | Average Amount Of Bids in Won Auction |
| 46 | Auction Activity | Fraction of Bids per Price Percentile of Auction |
| 17 | Suspicious Activity | Fraction of Bids from Rare IP Address |
| 18,20,22,24 | Suspicious Activity | Parameter Used By a Robot in the past (Boolean Flag) – IP, Device, Country, URL |
| 19,21,23,25 | Suspicious Activity | Parameter Used By a Robot in the past (% from bids) – IP, Device, Country, URL |
| 33 | Time-Series Activity | Count Of Bids Per Use Per Time Percentile |
| 35-38 | Time-Series Activity | Number and Fraction of Bids in the First 10% and last 10% of the time of auction |
| 39-42 | Time-Series Activity | Average Min/Max and Global Min/Max of Bidder Consecutive Bids Per Auction |
| 43 | Time-Series Activity | Fraction of bids in each 6 hour time-frame window |
| 44-49 | Time-Series Activity | Min and Mean change time of parameter per bidder in consecutive bids – IP, Device, Country |
| 50-52 | Time-Series Activity | Max Bid count in 10/30/60 minutes timeframe |
| 53-54 | Time-Series Activity | Max and Mean bid streaks of consecutive bids with same parameters |

## Methods and Models

We use different kinds of models on the problem to see how the perform differently as depicted in figure XX in section 4.7.

First, we trained a baseline classifier which includes SVM (both linear and with RBF kernel). Then we use decision tree models- Random Forest and Gradient Boost Tree and XGBoost.

For each model we experiment with tuning hyper-parameters using random search and preform feature selection to select the best features for each model. We then train the models and compare the results.

## Experimental Results

## Results Pre-Feature Selection

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training | | | Cross Validation | | Test | | |
| AUC-ROC | Average Precision | AUC-ROC | | Average Precision | AUC-ROC | Average Precision |
| SVM Linear |  |  |  | |  | 0.7410 |  |
| Random Forest |  |  |  | |  | 0.93654 |  |
| Gradient Boost |  |  |  | |  | 0.91569 |  |
| XGBoost |  |  |  | |  | 0.92008 |  |

## Results Post-Feature Selection

## Model Improvement

## Resampling

## Analysis

## Conclusion

**Appendix 1-**

1. Bid Robot - <http://www.bidrobot.com/cool/>
2. Auction Sniper - [https://auctionsniper.com/](https://auctionsniper.com/\)
3. Fraud Detection in Electronic Auction -<http://www.cs.cmu.edu/~dchau/papers/chau_fraud_detection.pdf>
4. Human or Robot Xiuye Gu & Shuyand Shi - <http://cs229.stanford.edu/proj2017/final-reports/5161146.pdf>
5. Bid-War: Human or Robot? <https://pdfs.semanticscholar.org/167c/96db9b78e9629861ac699f105b87943c2bc6.pdf>

Appendix 2 - Hyper Parameters Tuning

## RBF

SVC(C=0.1, break\_ties=False, cache\_size=200, class\_weight=None, coef0=0.0, decision\_function\_shape='ovr', degree=3, gamma=1, kernel='rbf', max\_iter=-1,probability=True, random\_state=None, shrinking=True, tol=0.001, verbose=False)

## Random Forest

RandomForestClassifier(bootstrap=True, ccp\_alpha=0.0, class\_weight=None,

criterion='entropy', max\_depth=7, max\_features='auto',

max\_leaf\_nodes=None, max\_samples=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=800,

n\_jobs=None, oob\_score=False, random\_state=42, verbose=0,

warm\_start=False)

## Gradient Boost

GradientBoostingClassifier(ccp\_alpha=0.0, criterion='friedman\_mse', init=None,

learning\_rate=0.1, loss='deviance', max\_depth=4,

max\_features=None, max\_leaf\_nodes=8,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10,

n\_iter\_no\_change=None, presort='deprecated',

random\_state=None, subsample=1.0, tol=0.0001,

validation\_fraction=0.1, verbose=0,

warm\_start=False)