Human or Robot?

Based on the original Facebook hiring competition on Kaggle

TABLE OF CONTENTS

01 Background

02 Problem Statement

03 Dataset Analysis

04 Feature Engineering

05 Model Selection

06 Recommendations

01 BACKGROUND

Human bidders on the site 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. Their attempt at building a model to identify these bids using behavioral data, including bid frequency over short periods of time, has proven insufficient.

02 PROBLEM STATEMENT

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.

03

DATASET ANALYSIS

DATASET FEATURES

Merchandise

Auction

Country

Device

IP

Time

URL

Dataset contains 7656334 rows x 9 columns

04

FEATURE ENGINEERING

FEATURE ENGINEERING

Time

Most emphasized feature as in all online auctions, time is of the essence – a split second could be the deciding factor for a winning bid.

.diff() and .to_datetime() were two key methods used to further analyze the original time feature.

Time-related features that were generated:

- **instant_bids**: No. of simultaneous bids, i.e. multiple bids performed at the exact same time
- **8-hour, 6-hour, 4-hour timeframes**: Identify potential bidding patterns at a certain time of day
- **mean, std, min, max, 50%, iqr_range**: Various aggregation features through .describe()

FEATURE ENGINEERING

Auction

Unique identifier of each auction

Auction-related features that were generated:

- **first_bid, last_bid**: No. of times each bidder was the first or last bid of an action
- **num_bids**: Total no. of times each bidder made over the entire dataset timeframe
- **bids_per_auction, max_bids_per_auction**: Average and maximum no. of bids each bidder made per auction
- **ip_per_auction**: Average no. of IP addresses each bidder used (for bidding) per auction

FEATURE ENGINEERING

Device

Phone model used for the particular bid

Device-related features that were generated:

- **max_bids_per_device**: Maximum no. of devices each bidder used
- **mean_bids_per_device**: Average no. of devices each bidder used per auction

IP, URL

IP address and URL of the bidder

IP & URL-related features that were generated:

- **ip_per_bidder**: No. of unique IP addresses each bidder used
- **bid_per_ip**: No. of bids made per IP address
- **bid_per_url**: No. of bids made per URL

FEATURE ENGINEERING

Other Features

Miscellaneous features generated:

- No. of unique original features using .unique()

Merchandise

The category of the auction site campaign, which means the bidder might come to this site by way of searching for "home goods" but ended up bidding for "sporting goods" - and that leads to this field being "home goods".

Thoughts:
There was little to no significance of this feature, cross-checked with the .feature_importances_ attribute

05

MODEL SELECTION

MODEL SELECTION

Validation

Train-Test-Split
StratifiedKFold

Ensemble Modelling

XGBoost
RandomForestClassifier

Hyperparameter Tuning

GridSearchCV

AUC: 0.93822

Model performance equivalent to 12th place of the global Kaggle leaderboard

06

RECOMMENDATIONS

RECOMMENDATIONS

Ideas/potential features that were not implemented:

- Final-stretch bidding frequency
  - A bot that was designed to win as many auctions as possible, would likely be designed such that its bidding frequency would be much higher in the last stretch of each auction.
- Ensemble of ensembles