CS 6375

Project Report

Facebook Recruiting: Human or Bot?

Number of free late days used: 1

Names of students in your group:

Jenila Rakesh Sanghvi JXS180012

Meet Jhatakia MKJ170000

Viral Doshi VPD180000

**Table of Contents**

[1. Introduction and Problem Description 4](#_Toc499810006)

[2. Related Work 7](#_Toc499810010)

[3. Dataset Description 11](#_Toc499810010)

[4. Pre-Processing Techniques 11](#_Toc499810011)

[5. Proposed Solutions and Methods 15](#_Toc499810017)

[6. Experimental Result and Analysis 17](#_Toc499810017)

[7. Conclusion 21](#_Toc499810022)

[8. Contribution of Team Members 22](#_Toc499810022)

[9. References 24](#_Toc499810024)

**Introduction and Problem Description**

Today the world has made immense advances in technology. The effects of technological advancement can both be positive and negative. Online auctions have become a vital part of e commerce website.

The major problem with having auction online is that one could not identify the bidder; i.e. whether the bidder is a human or a bot. This problem can be addressed as Auction Sniping which can be performed using biding robots.

Bidding robots are been used to make complex bidding decisions in split second time. Due to this, the human bidders don’t get a fair chance to win a bid which may get their back up and as a result would put an end to their usage of the auction website leading to loss of website’s core customer. Therefore it becomes important for the auction service provider sites to eliminate bots from bidding in their auction.

The goal of this project is to predict if an online bid is made by a machine (bot) or a human by using various machine learning techniques. Provided with some data about bidding information and information about bidders, we aim to identify the bots by recognizing pattern between the bids conducted by them using machine learning techniques. The important aspect of project includes task such as: feature extraction and selection, model selection, k-fold validation, learning and classification into bot or human and error analysis.

**Related Work**

As we were working on this project, we looked for references of related work and we found the following:

1. A machine learning framework to classify the Twitter users into bot, human, and cyborg. Author: Zi Chu, Steven Gianvecchio. Haining Wang and Sushil Jajodia

“Who is Tweeting on Twitter: Human, Bot, or Cyborg?” This paper classifies Twitter users into three categories: human, bot, and cyborg. Also the system consists of several components: the entropy component, the machine learning component, the account properties component, and the decision maker. This paper uses the following classifiers: Baseline Linear Classifier, Baseline Nearest Neighbour Classifier and Bayesian classification.

2. Using Machine Learning to Detect Fake Identities: Bots vs Humans

Author: Estée Van Der Walt and Jan Eloff.

This paper discuses about techniques and features to identify whether a twitter account is operated by a bot or by a human. It uses some of the features like ‘‘friend-to-followers ratio.’, ‘‘friend-count’’ and ‘‘follower-count,’’ to identify a bot or human user. They have created supervised learning models like support vector machines (SVMs), decisions trees, Naïve Bayes and neural networks.

**Dataset Description**

The datasets used for the project are from the Kaggle Competition website: <https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/data>

The dataset contains bid data, training data and test data.

Description about bid dataset:-

1. bid\_id - unique id for this bid.

2. bidder\_id – Unique identifier of a bidder.

3. auction – Unique identifier of an auction.

4. merchandise - The category of the auction site campaign.

5. device – Phone model of a visitor.

6. time - Time that the bid is made.

7. country - The country that the IP belongs to.

8. ip – IP address of a bidder.

9. url – url where the bidder was referred from.

Description about bidder dataset:-

1. bidder\_id – Unique identifier of a bidder.

2. payment\_account – Payment account associated with a bidder.

3. address – Mailing address of a bidder.

4. outcome – Label indicating whether a bidder is robot or not. Value 1 indicates a robot and 0 indicates a human.

The dataset contains 4700 data entries for test dataset, 2013 data entries for train dataset and 7.6 million data entries. Using this dataset attributes, we find ways to characterize bidding behaviour which can be used to generate features that can be inserted into a table. Then that table can be used to train a classification algorithm to distinguish between bots and humans.

**Pre-processing techniques**

Pre- processing techniques usually involves implementing the tasks such as:

1. Checking whether the dataset contains null or missing values.

2. Cleansing the dataset to remove any incorrect values.

3. Standardizing the features in the dataset.

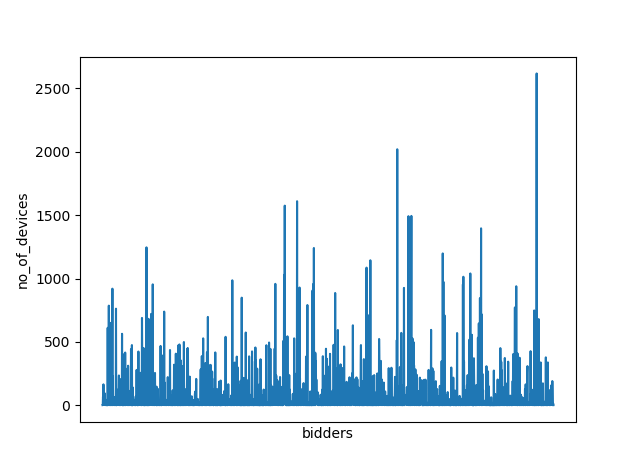
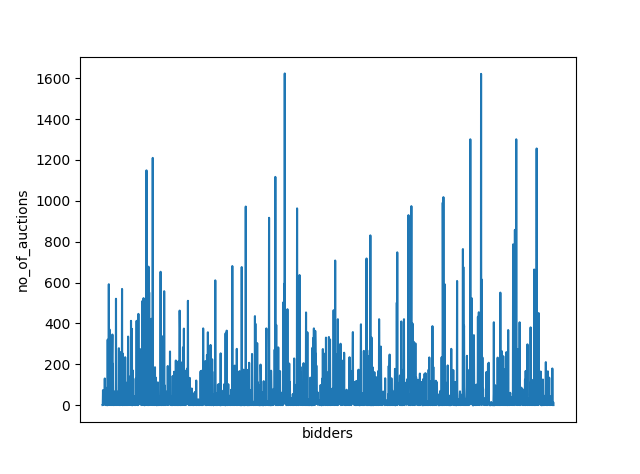
4. Converting any categorical values present in the dataset to numerical values.

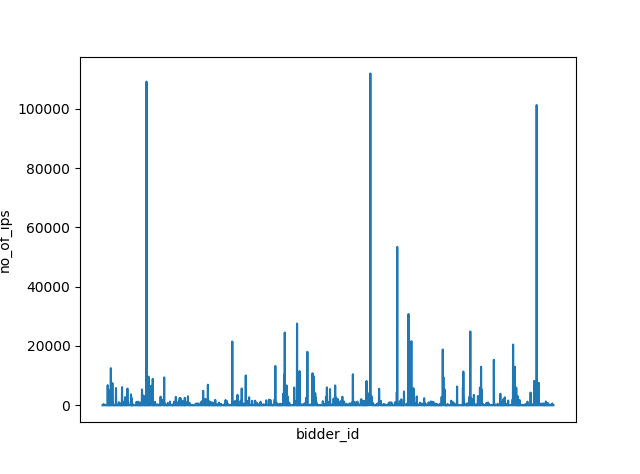
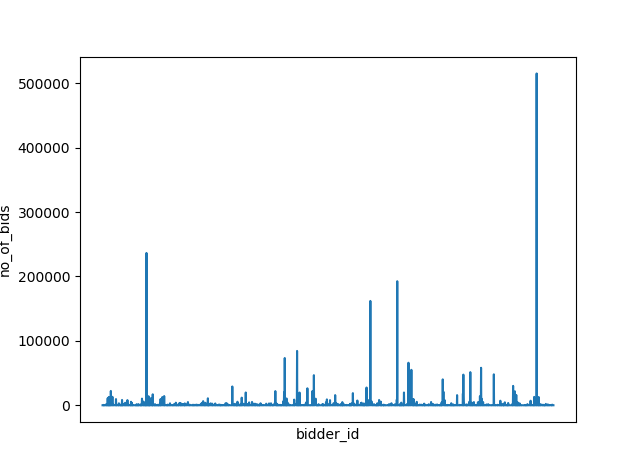
In order to perform the above stated pre-processing steps on the dataset we have used pandas and NumPy libraries as they provide various tools to handle the data and to standardize or normalize the features. In the training dataset we found out many missing values or null values which we then removed. As we already got the train and test dataset from the Kaggle website there was no need to split the bid dataset.

Using the pre-processed dataset we derived the features such as number of unique auctions a bidder has participated in, number of bids a particular bidder has raised, number of unique ip address through which the bidding requests come, and number of unique devices through which the bidding is done. All of the feature extraction from the dataset is done by applying aggregation function on the dataset.

By deriving these features from the given dataset, we can classify that whether the bot is bidding. For example consider the feature “number of unique ip”; this feature is used to identify the bots: if a bidder with a bidder\_id bids from different ip address frequently there is a high chance that the user bidding in auction is a bot and not a human. Same can be derived using the rest three features.

The graphs below plot bidders against number of devices, number of auctions and number of bids. There is also a plot between auctions and the number of bids in each auction. The number of bids vs. bidders plot shows the possibility of a bidder being a robot as a robot would bid a lot more times than a human. Also a robot would use multiple devices to bid and a bidder taking part in large number of auctions may also be robots.





**Proposed Solution and Methods**

For our project, we have used regression techniques like XGBoost and Random Forest in order to conduct the experimental analysis of the project. To conduct the analysis, models are created for machine to learn using the parameters such as learning rate, max\_depth, regression alpha and lambda along with n\_estimators which helps to identify bots.

**[A] Random Forest:**

1. Random forest is now the most commonly used algorithm in machine learning today.

2. As the name of the algorithm suggest it creates a random forest which is a group of decision tree.

3. Random forests are trained with the bagging method.

4. Random forest has equal number of hyperparameters as decision tree or a bagging classifier.

5. Random forest method searches for good features among a random subset of features, instead of searching for the most important features.

6. The advantage of using random forest is it can be used for both classification and regression tasks.

**[B] XGBoost:**

1. XGBoost which stands for stands foreXtreme Gradient Boosting is an optimized distributed library used for Gradient Boosting.

2. It implements machine learning algorithms under gradient boosting framework.

3. It provides distributed computing solution for faster training or larger datasets and several advanced features for model tuning, computing environments and algorithm enhancement

4. XGBoost can perform three kind of gradient boosting - Gradient Boosting, Stochastic Gradient Boosting and Regularized Gradient Boosting.

For the project to correctly classify human and bots we split the dataset as 80% training and 20% testing. A k-fold validation was applied on the training dataset to get validation done. The training dataset had 80% of the values which were used to perform the k-fold cross-validation on.

Then, using the models of XGBoost and Random Forest we achieved accuracy and ROC accuracy on training and testing dataset that was way better than the other models like decision tree, etc. Our project did not face the overfitting problem as the difference between the test and train accuracy was minimal.

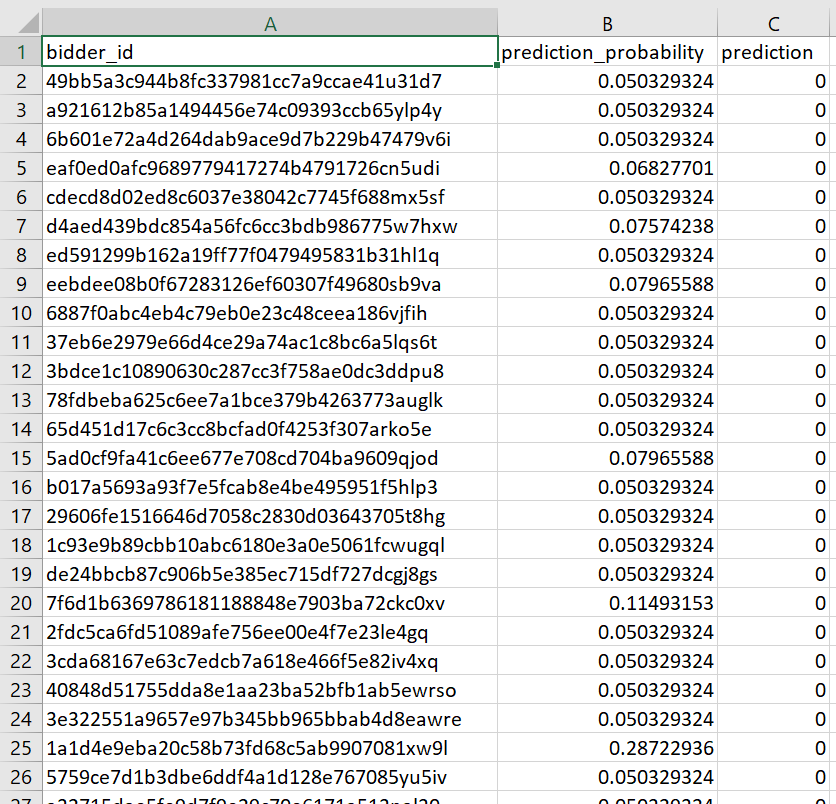
The computing resource and the programming environment used for the project are stted as follows: We have used Google Colaboratory for running the programs and we used the google drive link of the data sets to access the data sets into the Google Colaboratory environment. We have used the data sets provided with the competition question.

We have used machine learning libraries like Numpy, Pandas, Plotting libraries like Matplotlib, preprocessing libraries like sklearn.model\_selection for splitting the data and validation testing. We have also used pydrive, google.colab and oauth2client.client libraries for authentication to access the drive.

We used sklearn.metrics to calculate the accuracy and roc accuracy of the predictions. We have used xgboost for the XGBoost Classifier and sklearn.ensemble for RandomForestClassifier.

**Experimental Results and Analysis**

**Results depicting bidder id along with their probability of being a robot:**



**Log Table for Results:**

**XGBoost:**

|  |  |  |
| --- | --- | --- |
| **Experiment Number** | **Parameter Chosen** | **Results** |
| 1 | max\_depth = 10  learning\_rate = 0.012  objective = ‘binary: logistic’  n\_estimators = 280  reg\_alpha = 1.5 reg\_lambda = 1.0 | Train/Test Split = 80:20  Training Accuracy = 0.954  Test Accuracy = 0.968  Training ROC Accuracy = 0.910  Testing ROC Accuracy = 0.954 |
| 2 | max\_depth = 10  learning\_rate = 0.015  objective = ‘binary: logistic’  n\_estimators = 240  reg\_alpha = 1.5 reg\_lambda = 1.0 | Train/Test Split = 80:20  Training Accuracy = 0.958  Test Accuracy = 0.960  Training ROC Accuracy = 0.912  Testing ROC Accuracy = 0.944 |
| 3 | max\_depth = 6  learning\_rate = 0.01  objective = ‘binary: logistic’  n\_estimators = 240  reg\_alpha = 0 reg\_lambda = 0 | Train/Test Split = 80:20  Training Accuracy = 0.960  Test Accuracy = 0.948  Training ROC Accuracy = 0.917  Testing ROC Accuracy = 0.921 |
| 4 | max\_depth = 4  learning\_rate = 0.02  objective = ‘binary: logistic’  n\_estimators = 250  reg\_alpha = 1.0 reg\_lambda = 1.0 | Train/Test Split = 80:20  Training Accuracy = 0.952  Test Accuracy = 0.950  Training ROC Accuracy = 0.893  Testing ROC Accuracy = 0.926 |

**Random Forest:**

|  |  |  |
| --- | --- | --- |
| **Experiment Number** | **Parameter Chosen** | **Results** |
| 1 | max\_depth = 13  n\_estimators = 350  random\_state = None | Train/Test Split = 80:20  Training Accuracy = 0.947  Test Accuracy = 0.931  Training ROC Accuracy = 0.884  Testing ROC Accuracy = 0.887 |
| 2 | max\_depth = 3  n\_estimators = 230  random\_state = None | Train/Test Split = 80:20  Training Accuracy = 0.954  Test Accuracy = 0.923  Training ROC Accuracy = 0.889  Testing ROC Accuracy = 0.911 |
| 3 | max\_depth = 9  n\_estimators = 140  random\_state = None | Train/Test Split = 80:20  Training Accuracy = 0.950  Test Accuracy = 0.928  Training ROC Accuracy = 0.884  Testing ROC Accuracy = 0.895 |
| 4 | max\_depth = 30  n\_estimators = 410  random\_state = None | Train/Test Split = 80:20  Training Accuracy = 0.946  Test Accuracy = 0.928  Training ROC Accuracy = 0.878  Testing ROC Accuracy = 0.886 |

Experiment number 1 for XGBoost had the best parameter set as the Accuracy and ROC Accuracy was maximum. The prediction column in the result file has two values 0 and 1 where 0 indicates that the bidder\_id is a human bidder whereas the later indicates bot bidder.

**Conclusion**

We pre-processed and extracted various features from the Facebook Recruiting IV: Human or Robot dataset. All four extracted feature helps us to calculate the probability of the bidder being a Bot. Finally, we took the bidders with the highest “robot probabilities” to classify them as robot. After training the model and performing testing on the test dataset, we found 90 bidders that are bots and the rest 4610 bidders that are humans.

XGBoost model performed the best among the two models. It has the test accuracy rate of 0.968, whereas random forest has the test accuracy of 0.9478. Also the ROC accuracy value for XGBoost and Random Forest is 0.954 and 0.911 respectively.

After applying the above stated machine learning methods, the results generated would be useful for various auction sites like Listia, Ubid, ebay etc. For these auction sites it would be easy to identify whether a bidder is human or bot. This would prevent a bot to bid after a certain time by examining its initial bids. This would give a fair chance to legitimate users and customers.

Looking at the accuracy percentage, our machine learning model proved to be successful. The structure of the data helped us understand a lot about each individual bidder due to the high volume of bids dataset. We were able to find real correlations among the small bids, train and test dataset subset without overfitting the model and thus improving accuracy to 96.8%.

**Contribution of team members**

1. Project Selection and Brainstorming – Meet, Jenila, Viral

2. Research on the relevant tools and technologies – Jenila, Viral

3. Related work – Jenila, Viral, Meet

4. Dataset Preprocessing – Jenila, Meet

5. Proposed solution and method identification: Meet, Jenila, Viral

5.1 XGBoost– Meet, Jenila

5.2 Random Forest classification – Viral, Meet

6. Analysis of results - Meet, Jenila, Viral

7. Conclusion - Meet, Jenila, Viral

**References**

[1] <https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/>

[2] <https://www.analyticsvidhya.com/blog/2015/07/top-10-kaggle-fb-recruiting-competition/>

[3] <https://small-yellow-duck.github.io/auction.html>

[4] <https://en.wikipedia.org/wiki/Auction_sniping>

[5] <http://iwannadata.blogspot.com/2015/09/facebook-find-robot.html>

[6] <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

[7] <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>

[8] <https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7>