Facebook Bot-or-Not Classification

Names

- Alekya Pyreddy (903211489)
- Himadeep Reddy Reddivari (903208781)
- Kirthana Hampapur (903198301)

Project Goal

The goal of this project is to chase down robots for an online auction site. 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. To achieve this goal we develop prediction models for the outcome using various classifiers. We perform a detailed analysis of the model parameter selection and behavior of the tuning parameters is visualized. The models are validated and finally we implement a stacking approach to train a learning algorithm to combine the predictions of above classifiers, and predict the probability of a new bid being placed by a bot or not.

Description of the Data-set

The datasets are available at Kaggle ; courtesy of Facebook recruiting challenge. The most important initial step of this project is to extract valuable features for prediction. There are two types of data-sets. One is a bidder data-set that includes a list of bidder information, including their id, payment account, and address. The other is a bid data-set that includes 7.6 million bids on different auctions. The bidder and bid data-sets do not make any sense as the data as of now just contains auction ids, obfuscated payment accounts, mailing address and the outcome (0 or 1). They have very little valuable information. Thus feature extraction is important to learn the bidding behavior from the time of the bids, the auction, or the device. Also the raw data-set has to be cleaned and pre-processed to conduct any further data analysis. The data-set is from a real-time online auction website and the analysis will be meaningful and largely relevant.

Proposed methods

- Data Cleansing, Preprocessing and Feature Extraction
- Exploratory Data Analysis and Visualization
- Support Vector Machine and Neural Networks
- Random forest and Logistic regression
- Adaptive Boosting
- Optimal parameter selection and K-fold cross-validation
- Ensemble Stacking - constrained and other unconstrained
- Bayesian Hierarchical Model approach

Workload Allocation

Alekya	Exploratory Data Analysis and Visualization, Support Vector Machine and Neural Networks
Himadeep	Feature Extraction, Data Cleansing, Preprocessing, Adaptive Boosting, Bayesian Approach
Kirthana	Random forest and Logistic regression, Optimal parameter selection and K-fold cross-validation, Ensemble Stacking