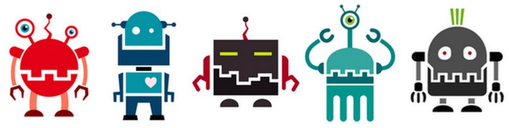
kaggle Facebook Challenge: Human or Robot ?

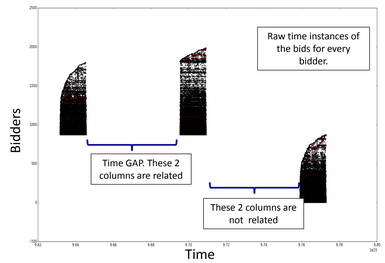
Introduction:

Penny auction sites are very important platform for all bidders around the world. The possibility of getting expensive and valuable products at low cost makes people to participate spending much time in front of the screen. But since the automatic bidding systems appearance, many users try to earn auctions without sweat on the chair and erode their fingers when pressing F5. The work exposed in this poster tries to detect these lazy users by exploring all bids made in a period of time to find any pattern that help to eliminate these undesired people of their existence



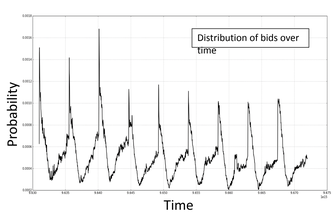
Preprocessing:

Since time instances were ofuscated, we take a look at them and try to obtain the original ones so that features have real meaning.



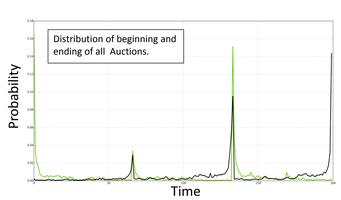
The 2 Gaps were removed. Also the first 2 columns were related Auctions and Bidders, but the third one was indepdent Auctions and Bidders so the bidders could be splitted into 2 groups.

In order to see it time instances were transformed by a non-decreasing function, we obtain the distribution of the bids time instances.



Due to the periodic distribution of the bids along time, we can conclude no significant transformation of the time has been made.

We also took a look at the sime instances where the Auctions began and finished in order to know it that is reliable information to construct features with.



Features:

Two data structures where created for extracting information from the data:

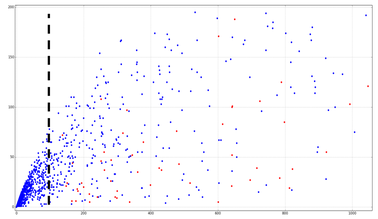
- Auctions Structure
- Bidders Structure

Numerical features

Integer features

- Number of auctions, bids, consecutive bids
- Number of coutries, IP's, phones

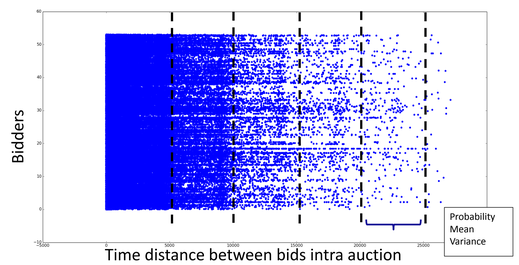
If not enough bids were made by a bidder, the system won't have enough staticts for taking him as bot. For users < 100 bids, 1300 were human, 7 bots (3 with just 1 bid)



Float features:

- Time distance between bids of the same user in the same auction and in any auction.
- Time distance to the last bid and distance to the first bid.
- Time distance between each own bid and the previous bid.

For every used feature, many of these measurements are obtained. Instead of just calculating statistics from these features, we divide the feature space into time slots and get statistics for every time slot.



For every time slot the following statistics have been obtained:

- Mean of time variables
- Varianve of time variables
- Probability of a bid belonging to the slot

Categorical features:

No relevant statistical information has been extracted from categorical features. Most of the time, users bids were for the same merchandise, so the mode of the variable do not provide information. Moreover, information provided by IP's and device types do not help to discriminate one class form another.

Manuel Montoya Catalá
Daniel Sierra Ramos

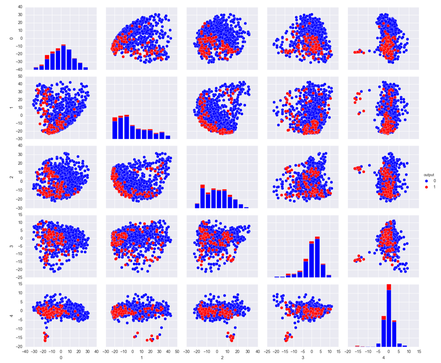
Feature Extraction:

After obtaining all features, is necessary to delete all possible collinearities. After several features extraction techniques, Kernel Partial Least Squares scheme has been employed.

After several test, 10 final features have been obtained as non-linear combinations of the original ones.

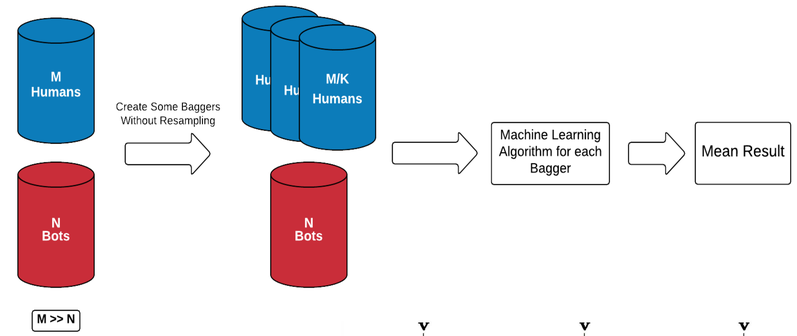
The right scatter matrix shows 5 of the 10 extracted features, where the blue points represent the humans, and the red ones represents bots.

As can be see, many of the 2D combinations show a good seperability for the two clases.

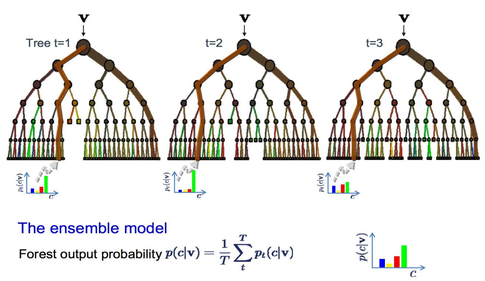


Detection System:

Due to classes are strongly unbalanced, is necessary to model a system that deals with this problem. So, a bagging model has been employed. In this model, all bot users are combined with the same number of randomly selected human users as shown in the next figure. Finally a set of N balanced set is obtained.



As machine learning algorithm, a Random Forest ensembling model has been employed. The non-linearity model provided by this method makes it suitable for this problem.



Results:

Next table shows some results obtained with this scheme

Training score	Validation score	Test score
0.98	0.94	0.89

Conclusion:

Given thist test score, probabily there is an overfitting issue with the solution presented by the RF classifier, even balanced dataset.