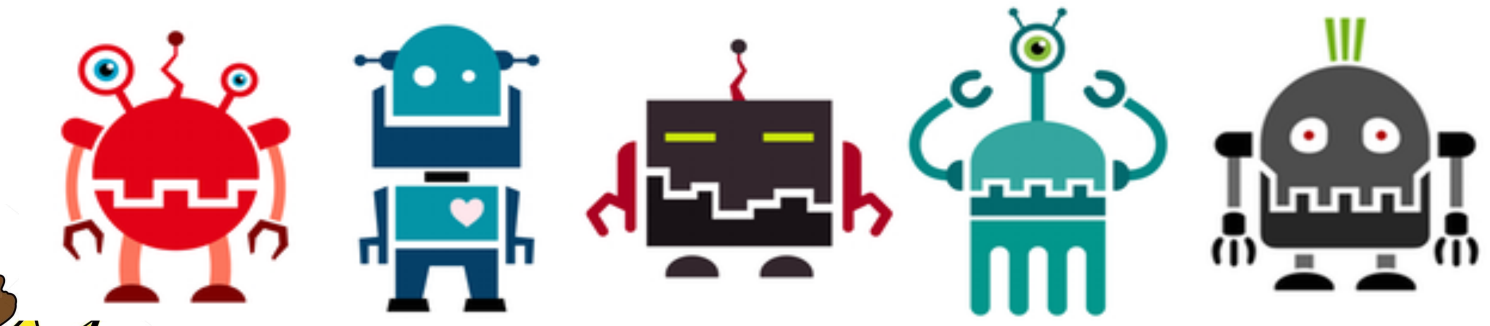
Human or Robot?



Xiuye Gu, Shuyang Shi

{xiuyegu, bsnsk}@stanford.edu

Summary

On online auction sites, bidders participate in auctions by bidding certain objects they want. Due to the existence of software-controlled bidders, i.e. robots, human bidders on the sites are becoming frustrated with their inability to win auctions, and therefore the core customer base of that site can be plummeting. In order to improve customer experience, platforms need to recognize robot bidders and eliminate their bidding from auctions.

Our project follows a Kaggle competition, aiming to classify human bidders and robot bidders based on their bidding behaviors.

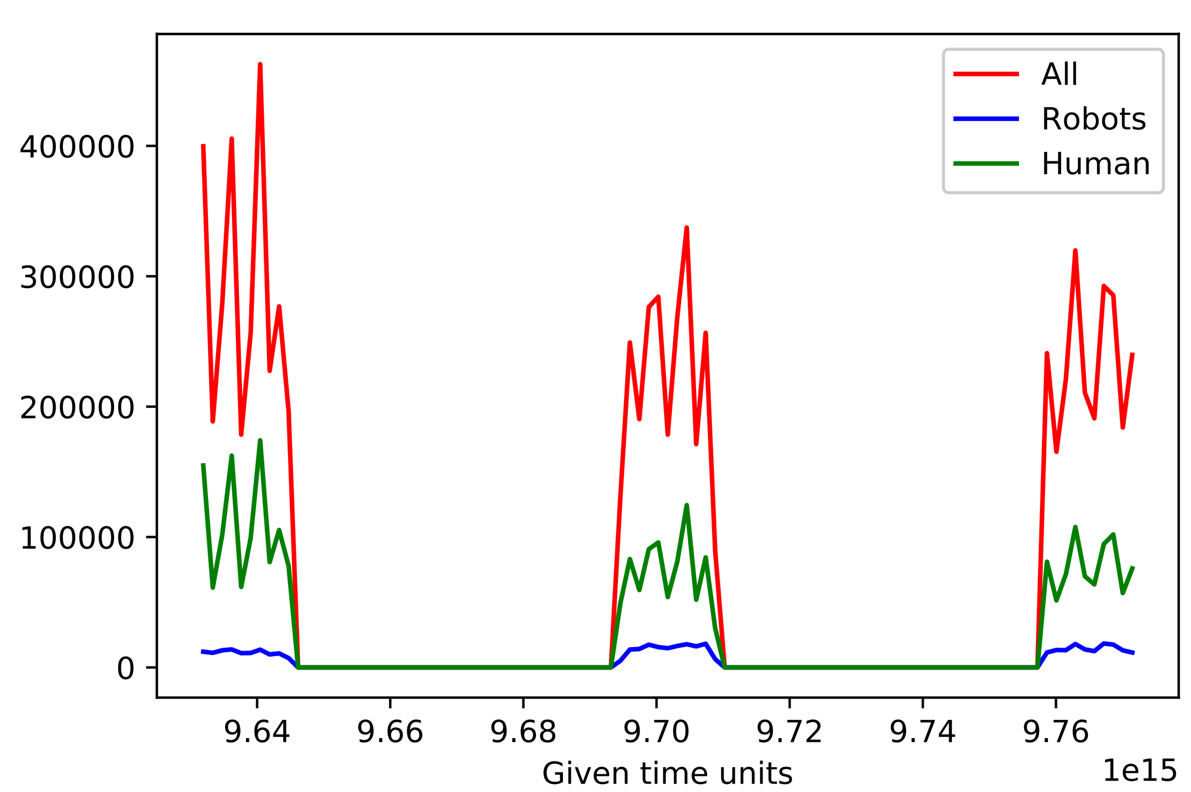
Our best test performance achieve **0.94** AUROC, which is between the 4th and 5th position on the private leader board.

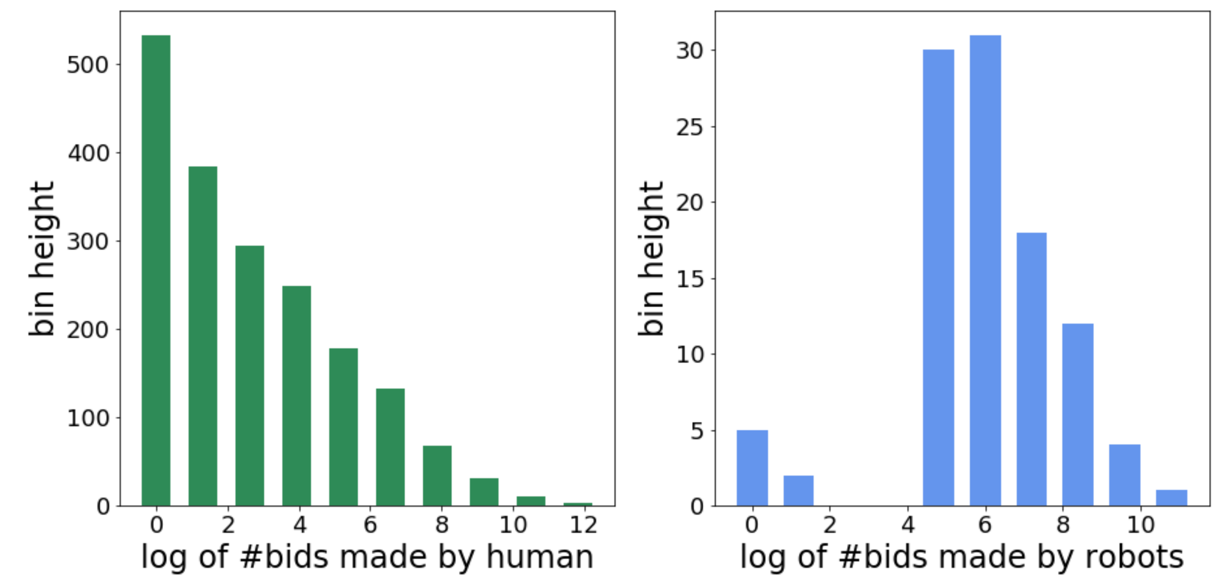
Data

We use the dataset provided in the competition, which is consists of a bidder dataset and a bid dataset. The bidder dataset mainly provide bidder's ID and their labels. The training set has **1984** human and **103** robots, and the test set has 4700 bidders. Notice the how SMALL and UNBALANCED is the training set!

The bid dataset contains each bid's auction, merchandise category, device, time, country, IP, and URL.

Features





Dense Features
of bids made
mean bids made per auction
of auctions participated
of country the bidder went
of device/IP/URL used
of auctions won
Time difference between consecutive bids made by the same user
Response time
Bids' price
Avg changing IP time
Log entropy of IPs/URLs used
Sparse Features
of bids made in each small time interval
Percentage of bids made in each country
Merchandise category (one-hot encoding)

Models

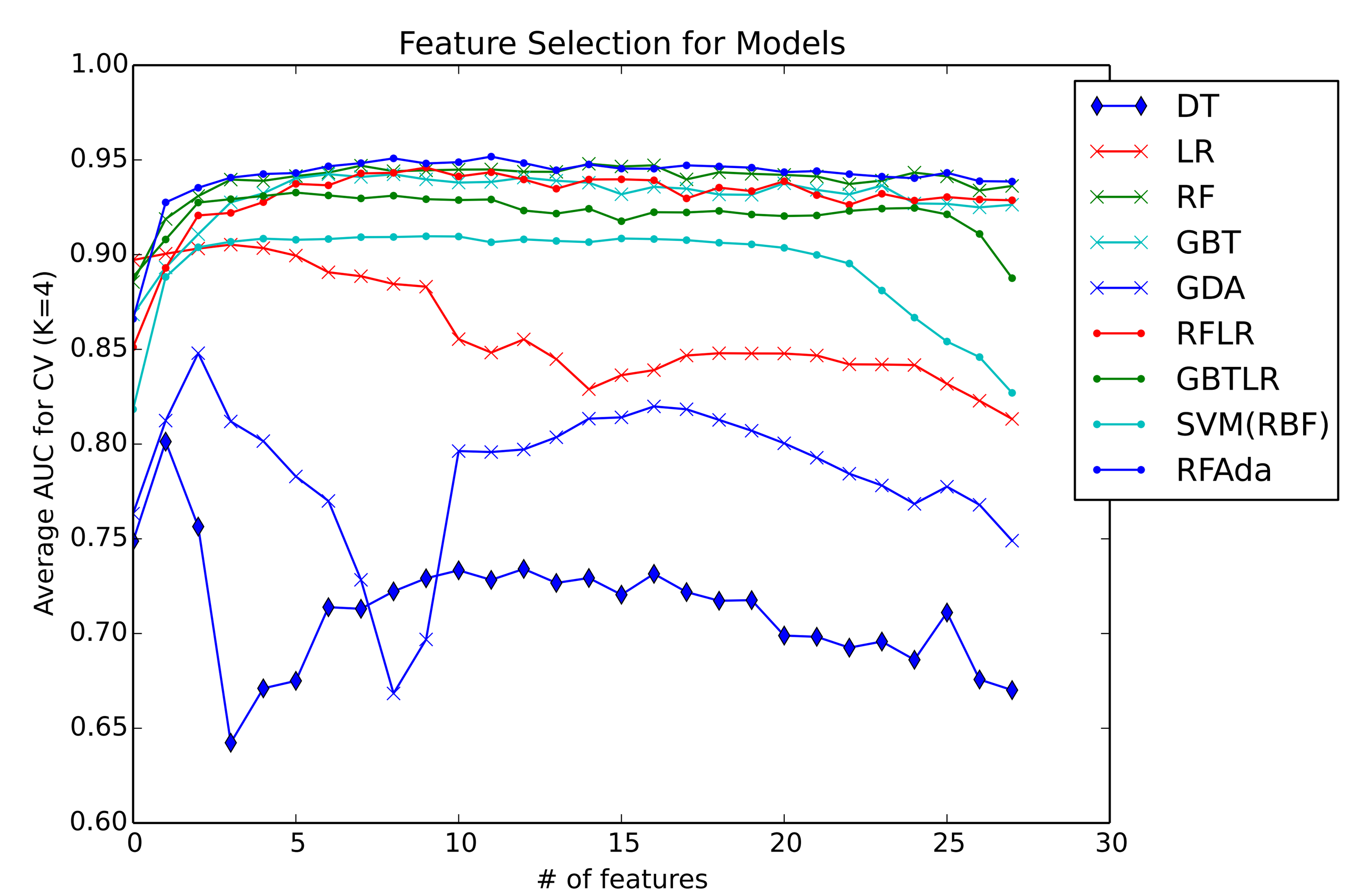
Basic	Tree-based	Composite	Miscellaneous
Linear Regression	Random Forest	RF LR	GDA
SVM Linear/RBF	Gradient Boost Tree	GBT LR	DNN
Decision Tree	Ada boost Random Forest		

AUC on predefined feature set

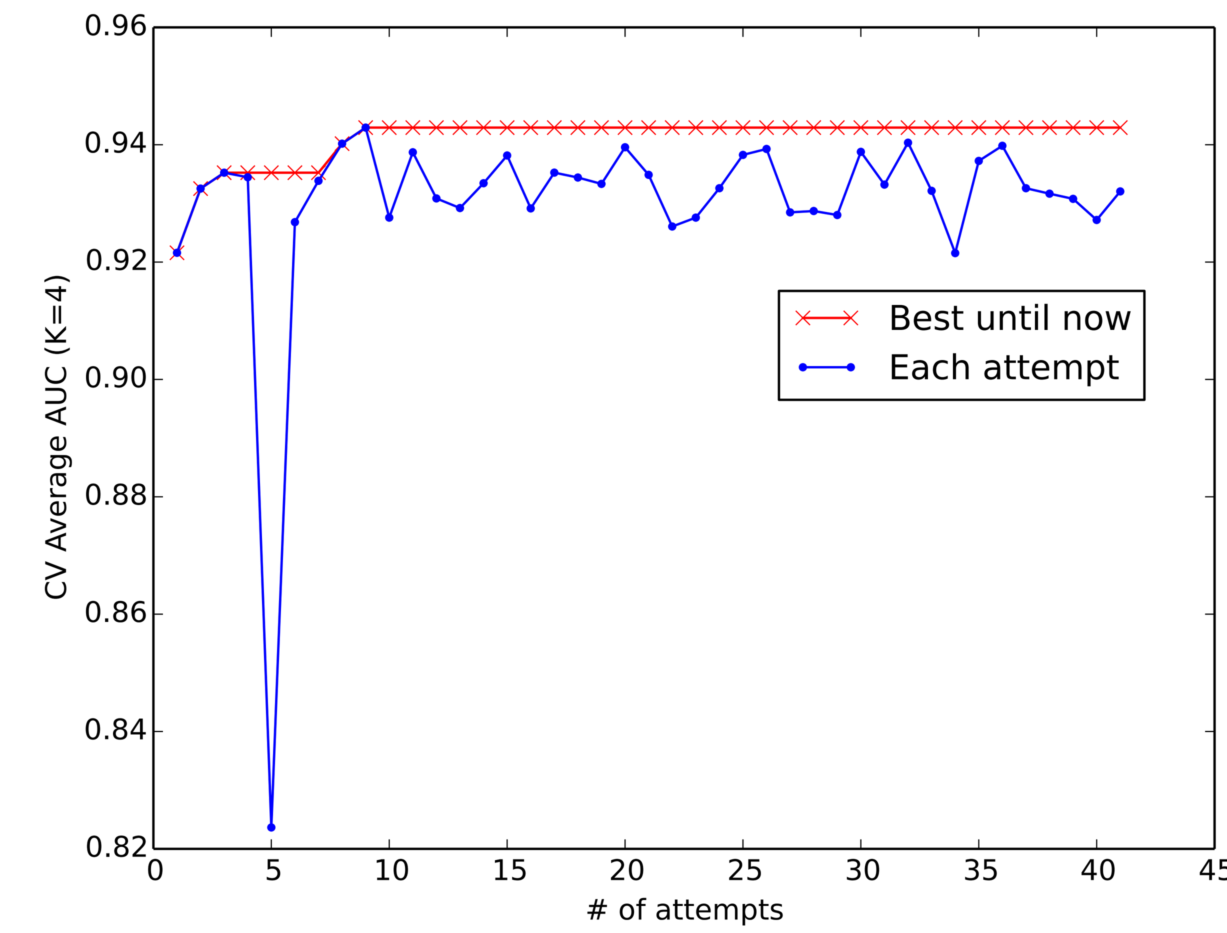
Models	Training	CV	Test
LR	0.8732	0.8082	0.8121
SVM (RBF)	1.0	0.8874	0.8217
SVM (Linear)	0.9981	0.6655	0.5834
DT	1.0	0.6711	0.6050
RF	1.0000	**0.9404**	0.9330
RFAda	1.0	0.9373	**0.9370**
GBT	1.0	0.9180	0.9204
GDA	0.7521	0.7468	0.7369
RFLR	0.9937	0.9255	0.9144
GBTLR	0.9531	0.9125	0.8534
DNN(TF)	0.9655	NA	0.8441
DNN(sklearn)	0.9999	0.8384	0.8235
DNN(unnormalized)	0.5	0.5	0.5025

AUC after feature selection

Models	Training	CV	Test
LR	0.9054	0.9053	0.8901
SVM (RBF)	0.9982	0.9097	0.8479
DT	0.8679	0.8013	0.7934
RF	0.9997	0.9403	**0.9259**
RFAda	1.0	**0.9455**	0.9220
GBT	1.0	0.9341	0.9069
GDA	0.8491	0.8480	0.8247
RFLR	0.9888	0.9327	0.9138
GBTLR	0.9376	0.9121	0.8743



Random search hyper parameter



Ablative analysis on RF

Feature	CV AUC
Full feature sets	0.9416
median of tdiff	0.9448
# of device	0.9347
min of price	0.9300
std of price	0.9315
min of response	0.9232
log entropy of ip	0.9250
# of country	0.9269
# of bids	0.9255
min of tdiff	0.9178
Avg changing IP time	0.9119
# of auction	0.8988
log entropy of url	0.8897
mean of tdiff	0.8523

Conclusion

In this real-life problem, different machine learning models have very different results. By feature extraction, feature selection, model selection, and ablative analysis, we understand that tree-based models have good accuracies in such problems, while models like GDA make poor options, and we explored the way to find the best possible method in such problems.