Kateryna Hrebeniuk

SANTANDER CUSTOMER SATISFACTION

KAGGLE COMPETITION

General Assembly DS-BOS-08 April 25, 2016

PROBLEM STATEMENT

From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around. What's more, unhappy customers rarely voice their dissatisfaction before leaving.

Santander Bank is asking to help them identify dissatisfied customers early in their relationship. Doing so would allow Santander to take proactive steps to improve a customer's happiness before it's too late.

The competition provides dataset with hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience.

It is supervised learning binary classification task. My goal is to learn training data and predict the categorical class labels of new instances based on past observations.

DATA

Train and test datasets are provided by Kaggle. An anonymized datasets contain large number of numeric variables. "ID" column shows the customer's identifier. The "TARGET" column is the variable to predict. It equals 1 for unsatisfied customer and 0 for satisfied customer.

File descriptions:

- train.csv - the training set including the target (76020 rows x 371 columns)
- test.csv - the test set without the target (75818 rows x 370 columns)

OUTCOME

Output file should contain two columns – "ID" and "TARGET". "ID" column should show the customer's identifier from test dataset and "TARGET" column should show probabilities for each observation.

There are **5,125 teams** that participate this competition. The best public AUC score, that is calculated on approximately 50% of data, is **0.844134** and the worst AUC score is **0.296686**. The goal is to find the best score using different algorithms but try not to overfit the model.

DATA EXPLORATION

Train dataset

Shape of dataset: 76020 x 371

There are 0 row(s) out of 76020 have missing values.

Counts by satisfaction:

Satisfied customers: 73012 (96.0 %)

Unsatisfied customers: 3008 (4.0 %)

Test dataset:

Shape of dataset: 75818 x 370

There are 0 row(s) out of 75818 have missing values.

DATA CLEANING

Descriptive statistics

34 columns out of 371 are empty and 29 columns are duplicated!

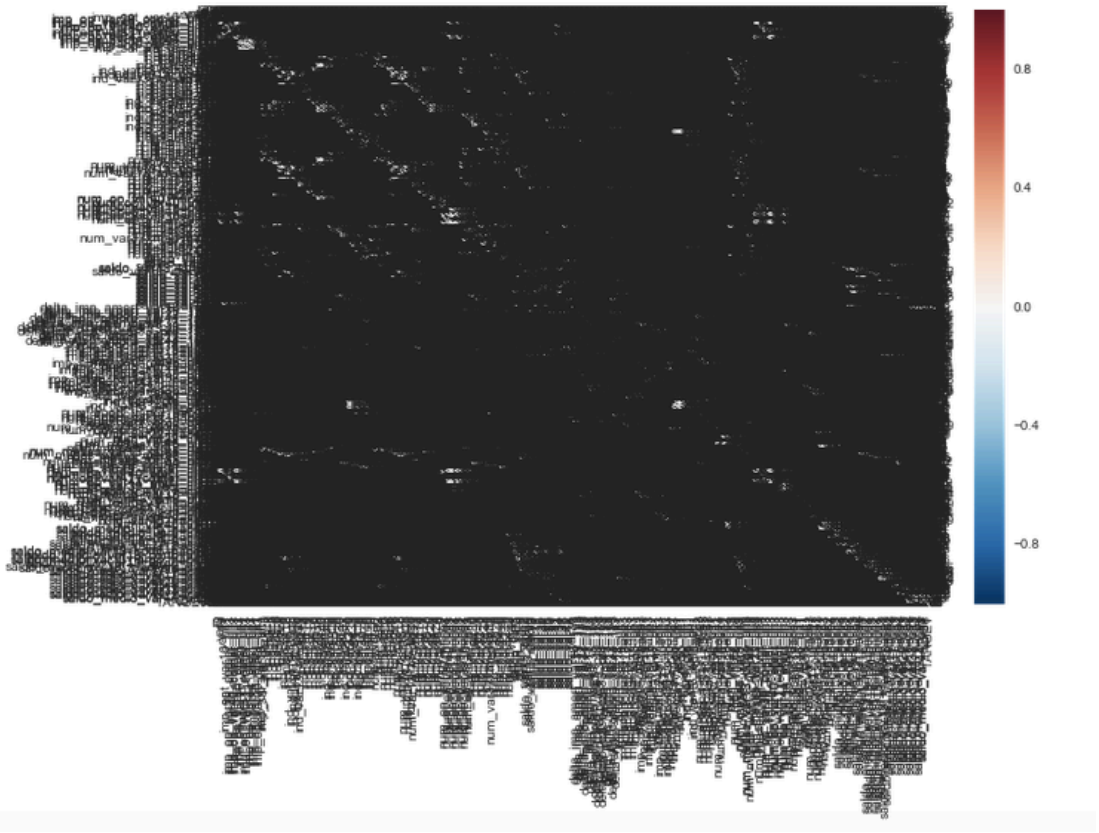
	ID	var3	var15
count	76020.000000	76020.000000	76020.000(
mean	75964.050723	-1523.199277	33.212865
std	43781.947379	39033.462364	12.956486
min	1.000000	-999999.000000	.000000
25%	38104.750000	2.000000	23.000000
50%	76043.000000	2.000000	28.000000
75%	113748.750000	2.000000	40.000000
max	151838.000000	238.000000	105.00000(

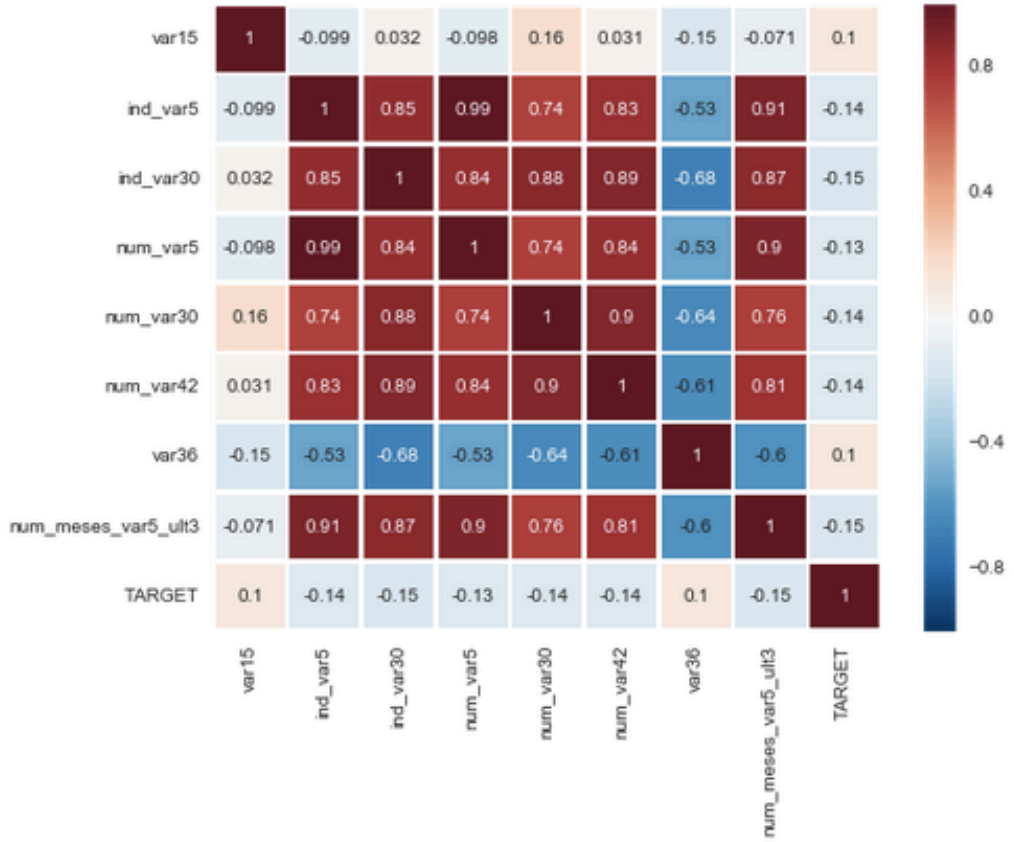
ind_var1	ind_var2_0	ind_var2	ind_var5_(
76020.00000(	76020	76020	76020.000
0.003762	0	0	0.958024
0.061221	0	0	0.200535
0.000000	0	0	0.000000
0.000000	0	0	1.000000
0.000000	0	0	1.000000
0.000000	0	0	1.000000
1.000000	0	0	1.000000

DATA MINING

Explore correlations

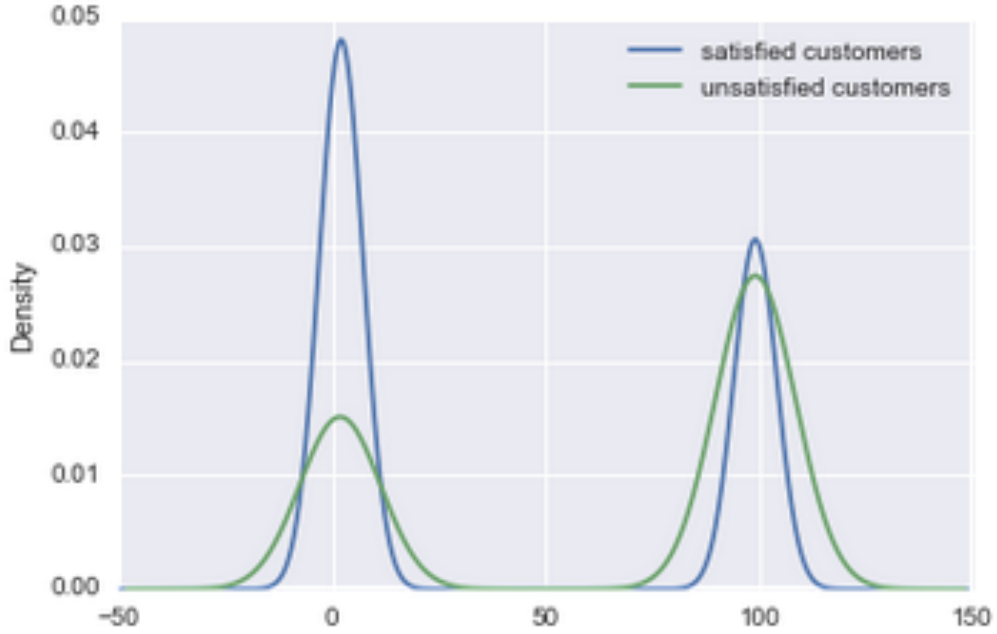
all feaures vs most correlated with target





DATA MINING

var36 exploration



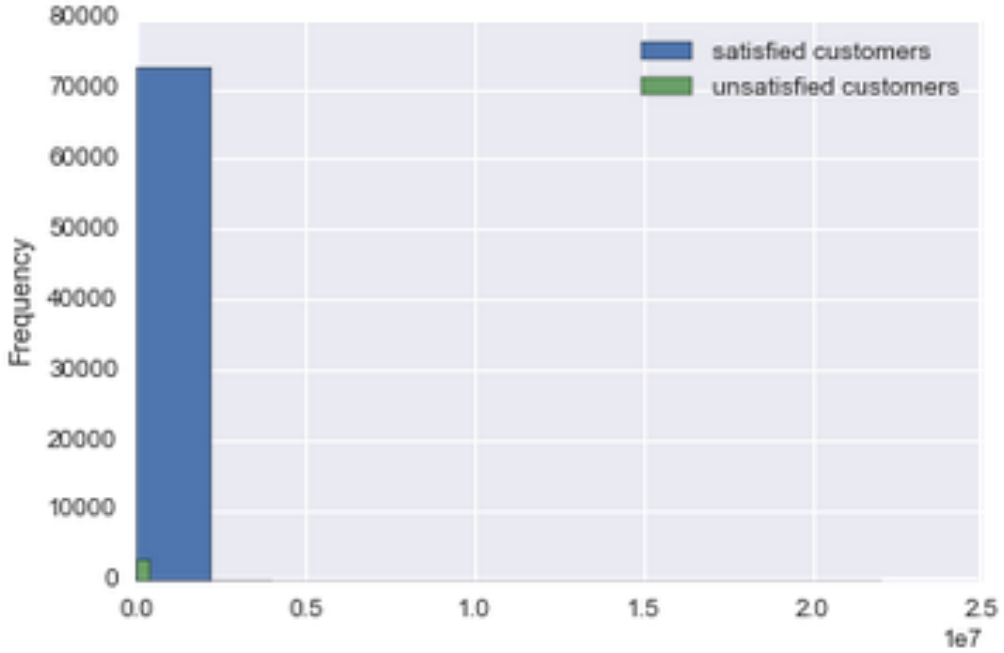

```
count    76020.000000
mean        40.449079
std         47.362719
min          0.000000
25%          2.000000
50%          3.000000
75%         99.000000
max         99.000000
Name: var36, dtype: float64
```

According to density chart above customers with values of 0,1,2 or 3 in var36 are less unhappy than those who has 99.

DATA MINING

var38 exploration

It seems that var38 is customer's account balance. And I wanted to see if there is relationship between customer's satisfaction and account balance.

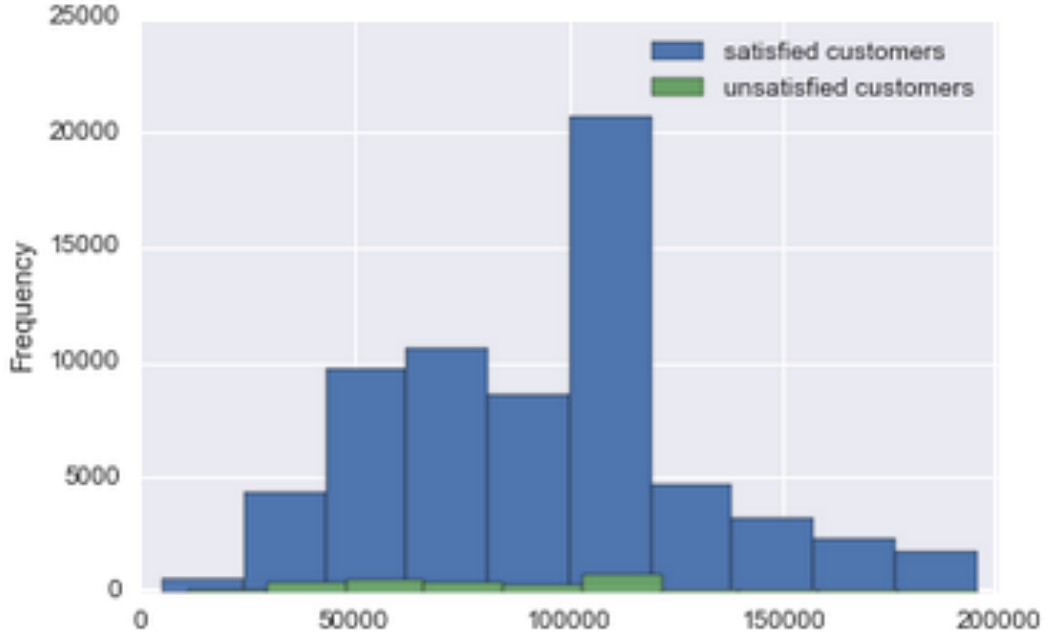


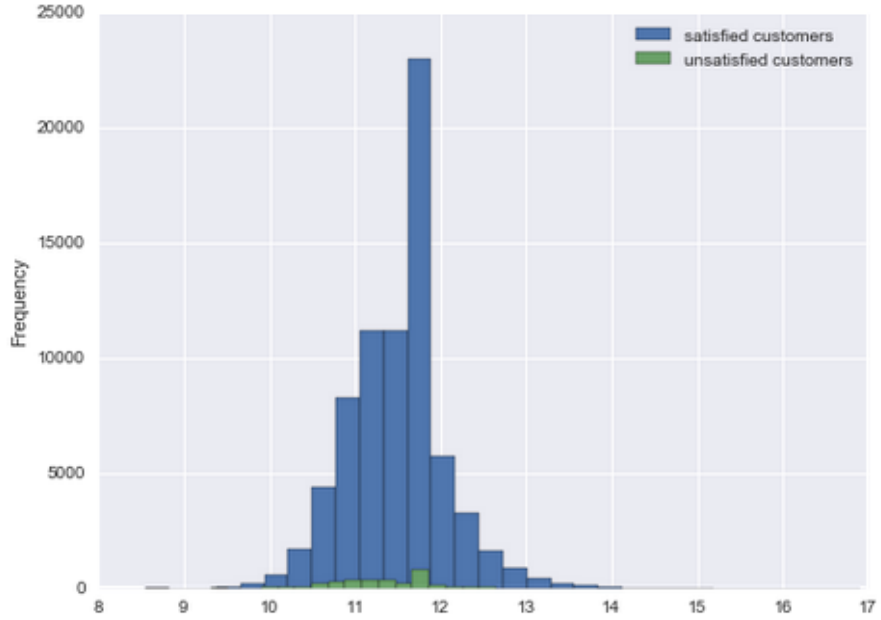

```
count         76020.000000
mean         117235.809430
std          182664.598503
min            5163.750000
25%           67870.612500
50%          106409.160000
75%          118756.252500
max        22034738.760000
Name: var38, dtype: float64
```

DATA MINING

var38 exploration

removed outliers vs logarithmic

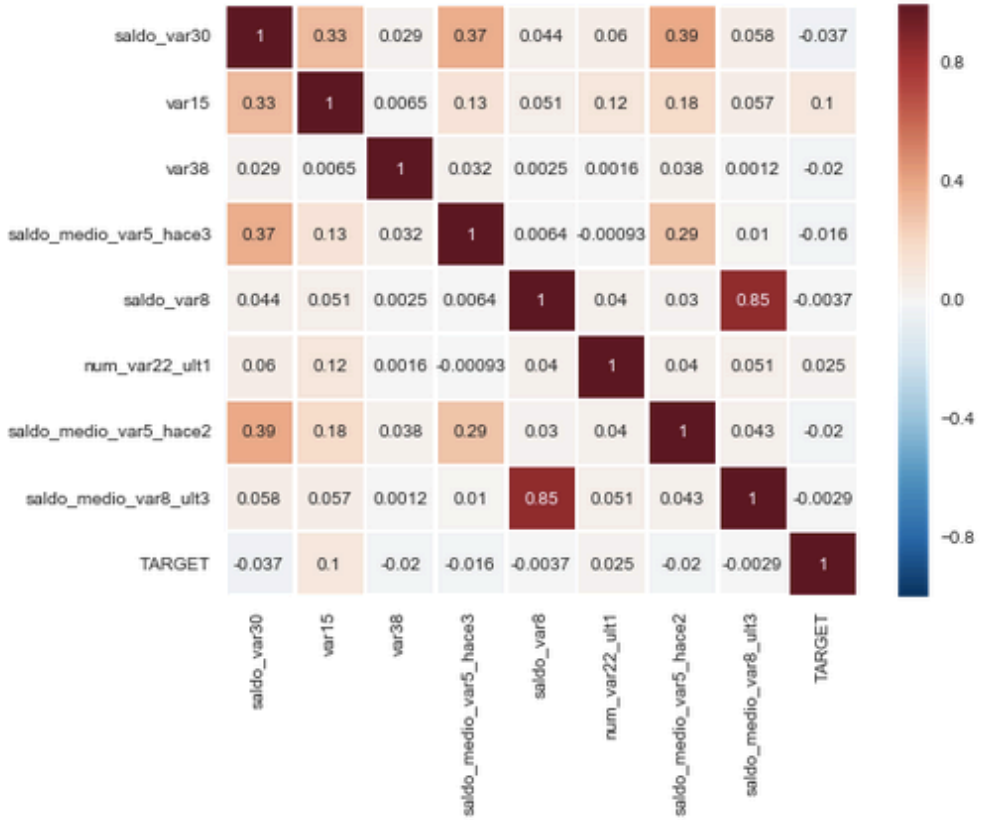




FEATURE SELECTION

I used **GradientBoostingClassifier()** to find important features

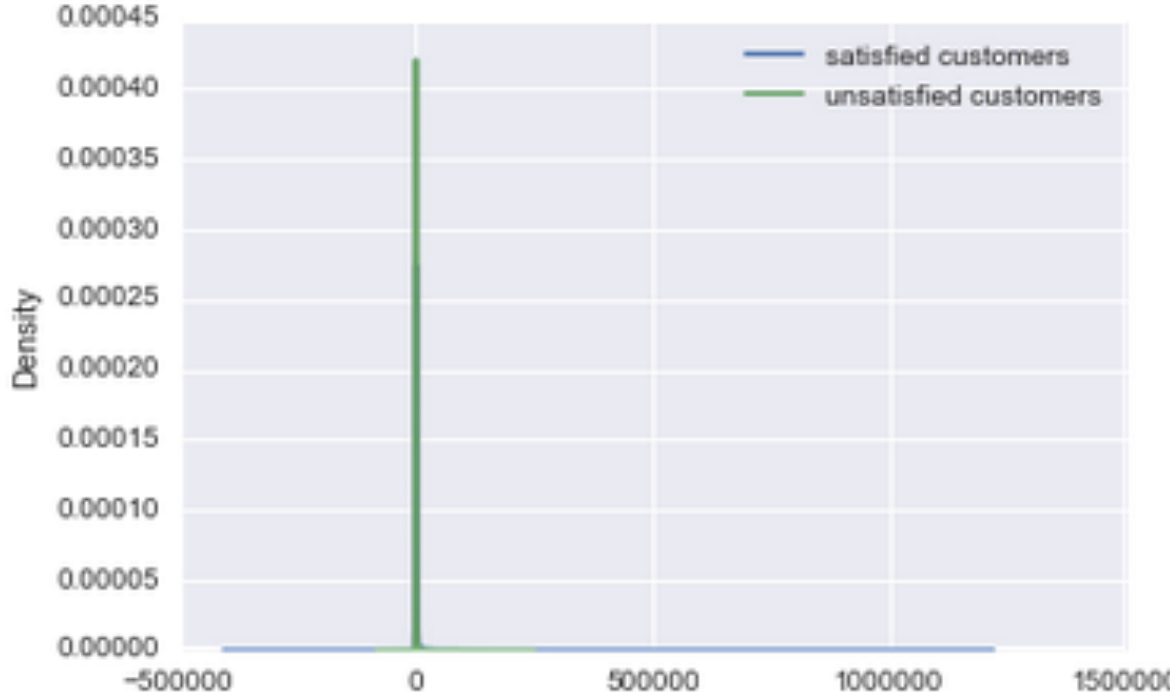
features correlated with target	the most important features
var15	saldo_var30
ind_var5	var15
ind_var30	var38
num_var5	saldo_medio_var5_hace3
num_var30	saldo_var8
num_var42	num_var22_ult1
var36	saldo_medio_var5_hace2
num_meses_var5_ult3	saldo_medio_var8_ult3

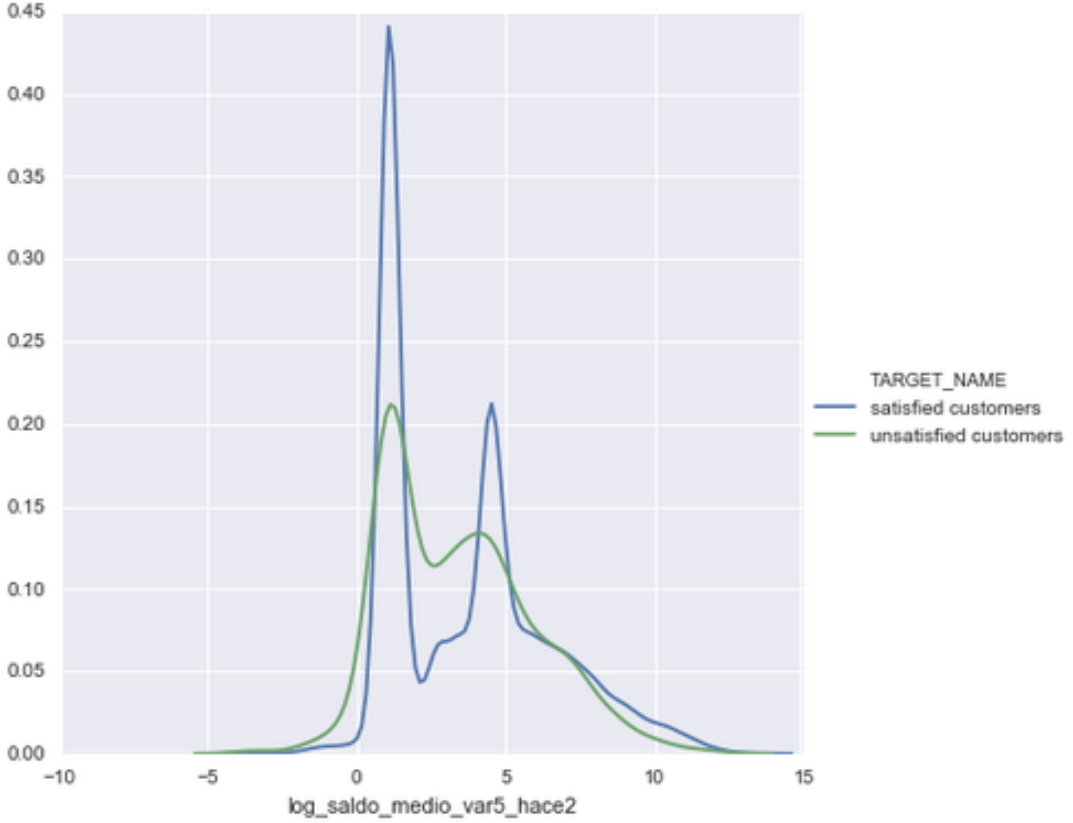


FEATURE SELECTION

saldo_medio_var5_hace2 exploration

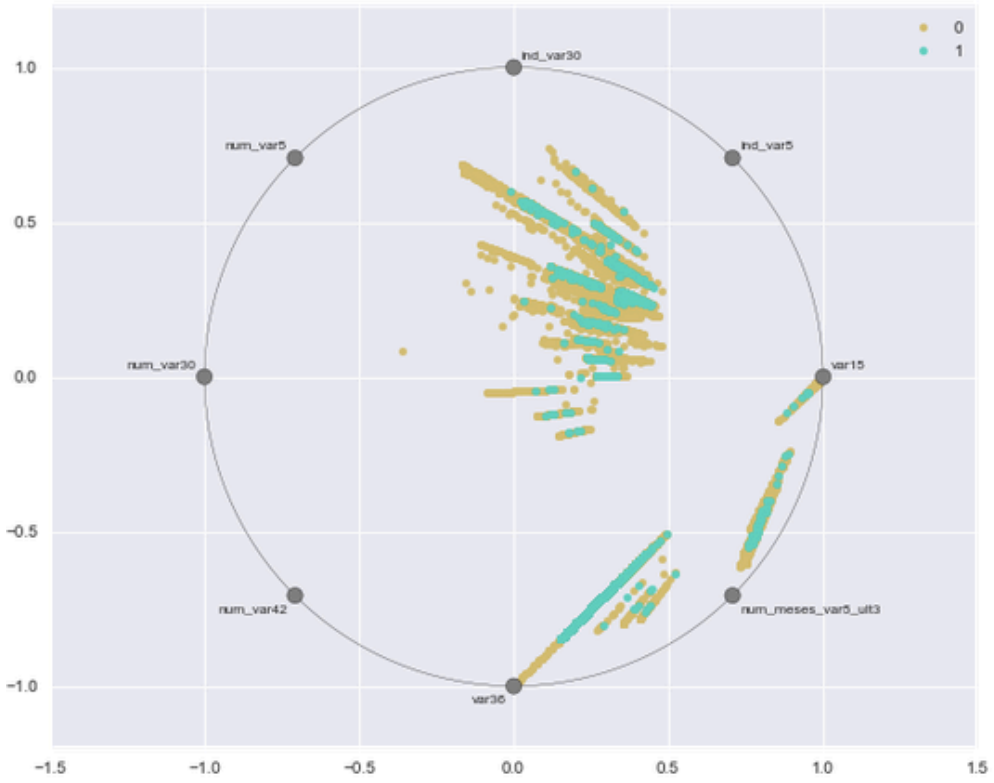
original vs logarithmic



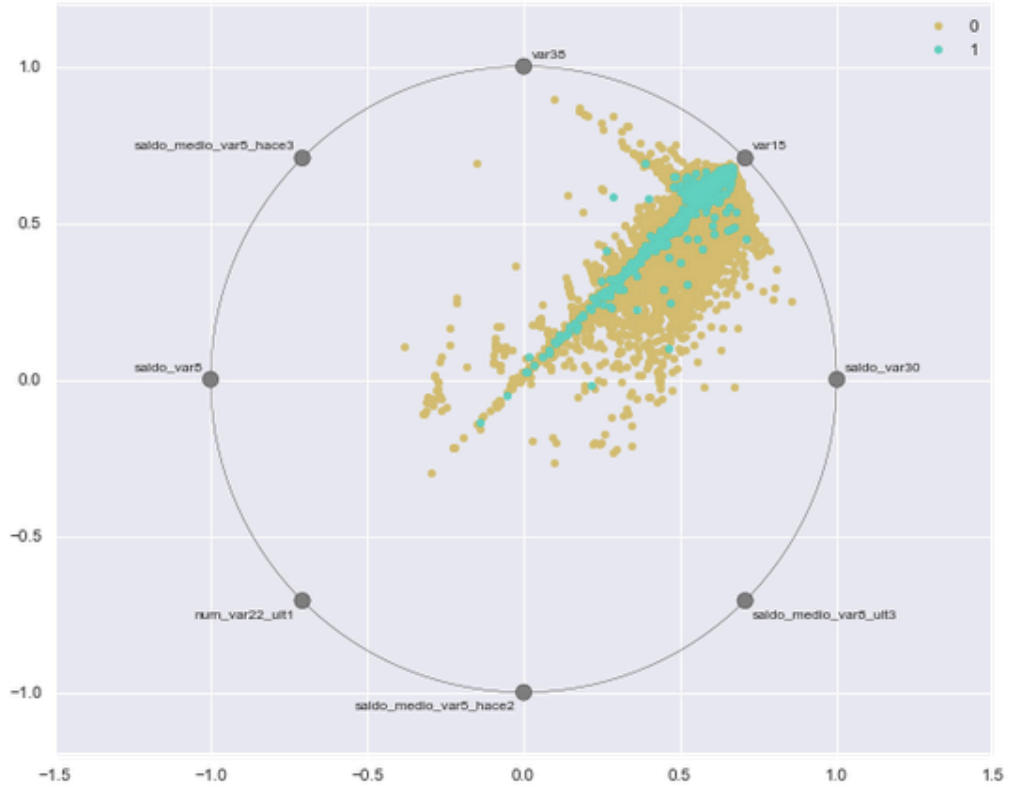


RADVIZ VIZUALIZATION

features correlated with target

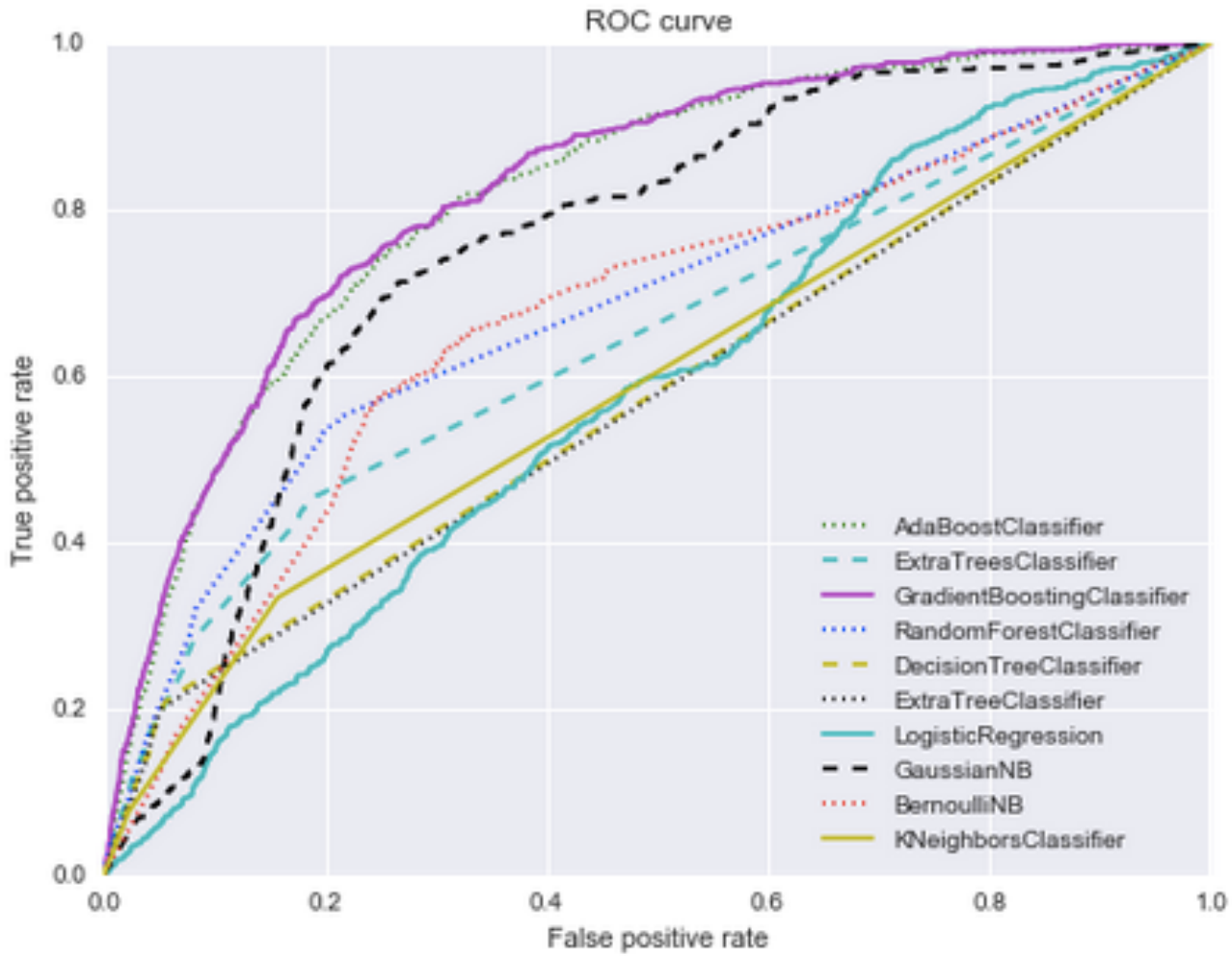


the most important features



RadViz is a way of visualizing multi-variate data. It is based on a simple spring tension minimization algorithm. Basically you set up a bunch of points in a plane. In our case they are equally spaced on a unit circle. Depending on which class that sample belongs it will be colored differently.

MODELING



	ROC AUC
ExtraTreeClassifier	0.5736
DecisionTreeClassifier	0.5767
LogisticRegression	0.5797
KNeighborsClassifier	0.5902
ExtraTreesClassifier	0.6412
BernoulliNB	0.6686
RandomForestClassifier	0.6804
GaussianNB	0.7520
AdaBoostClassifier	0.8158
GradientBoostingClassifier	0.8230

RESULTS

Public Leaderboard

(calculated on approximately 50% of the test data)

logistic regression model with top 8 selected features	**0.537662**
decision tree model with top 8 selected features	**0.767169**
decision tree model with all features	**0.817955**
gradient boosting model with all features	**0.833733**

CONCLUSIONS

It was a very interesting project and experience. The data was pretty clean and I spent my time exploring the features and comparing different prediction algorithms. From the data exploration, I have found that satisfied and dissatisfied customer data is mostly overlapped. I have tried various models and found that the Gradient Boosting ensemble model from scikit-learn package works the best for this data set.

My AUC score on public board is **0.833** and the best score in the Kaggle competition is **0.844**. According to Kaggle forum the highest scores were obtained using the XGboosting (Extreme Gradient Boosting) algorithm.

Another way to get better performance can be achieved by using automated parameter tuning.

Kateryna Hrebeniuk

THANK YOU!

General Assembly DS-BOS-08 April 25, 2016