
Introducing
a bank for your
communities.
Santander
Introducing
a bank for your
communities.
REDUCE YOUR
DAILY EXPLETIVES.
Trilet

Hi there! I’m a software engineer based in NY and working for the UN, and an aspiring data scientist at Flatiron Academy

Barto Molina, Flatiron DS, April 2019 cohort

the problem
methodology

- Predict customer satisfaction
- Dataset preparation
- Machine learning algorithms

features
types

numerical features

Related to user account. i.e. balance.

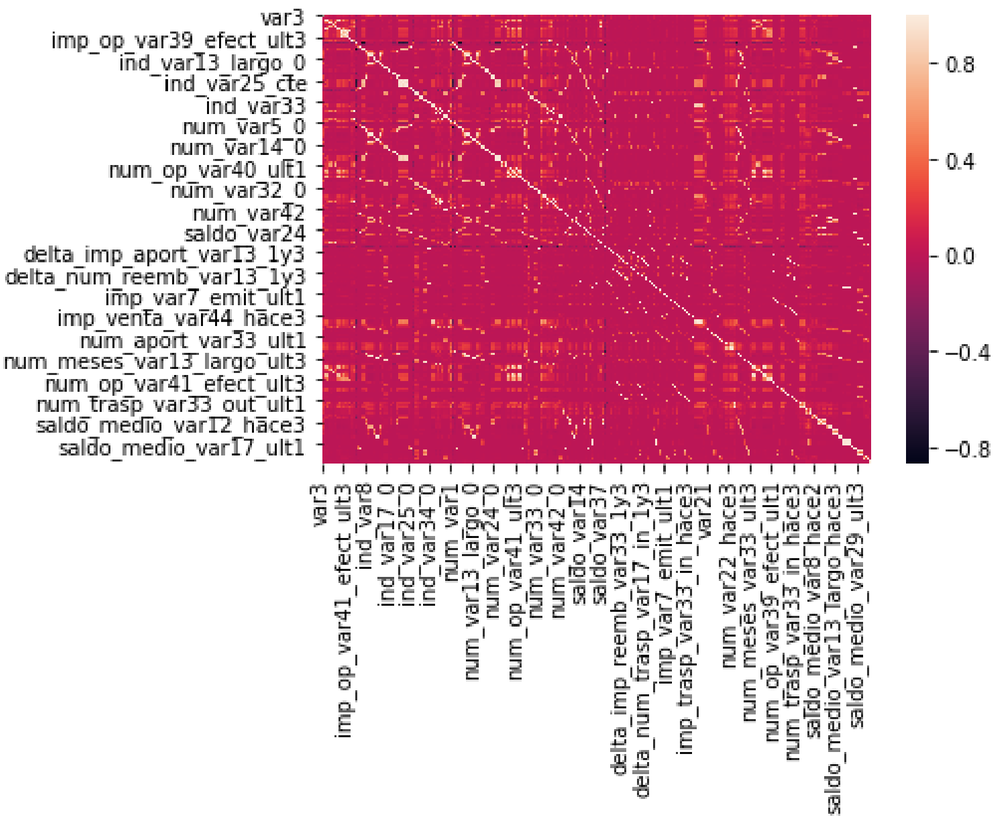
categorical features

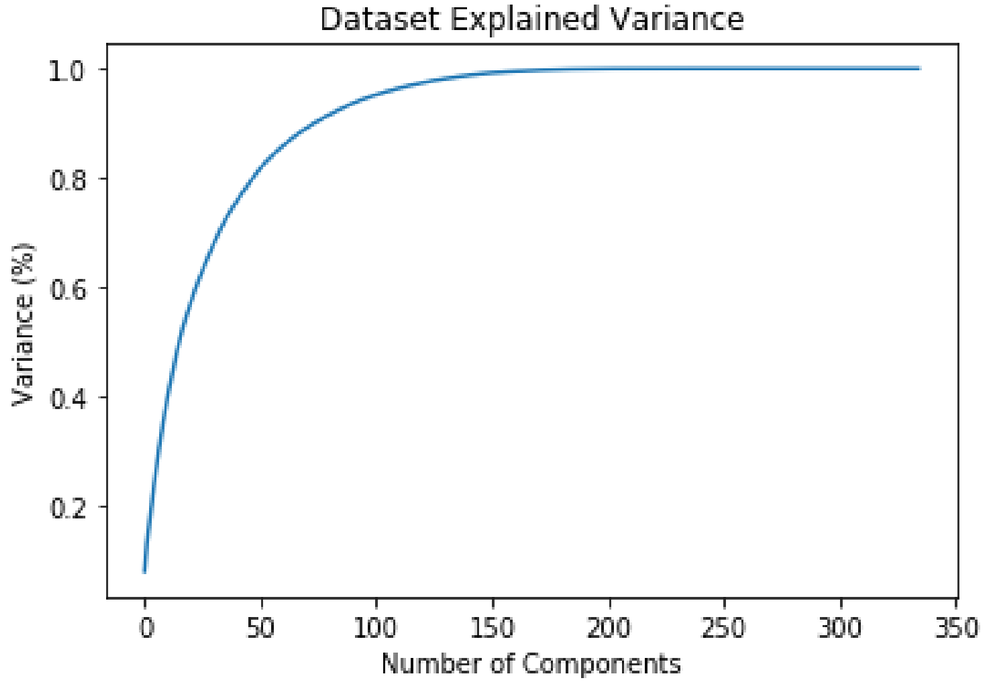
Like the different products associated to the users.

other features

Country, Age, value of the Mortgage.

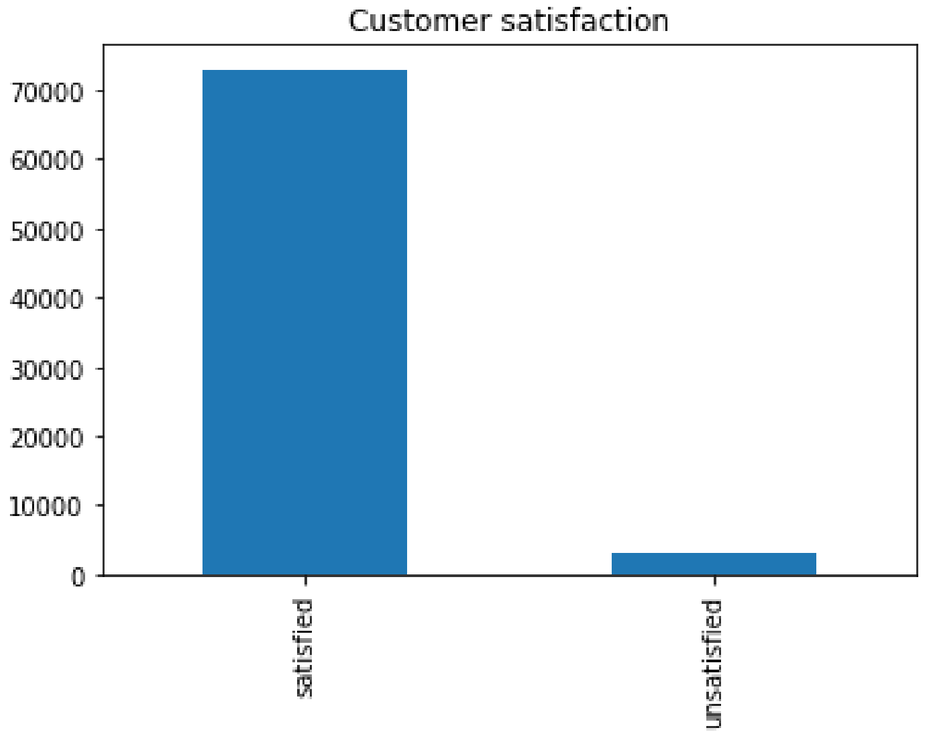
features
PCA





target
unbalanced

4% unsatisfied customers



target
resampling

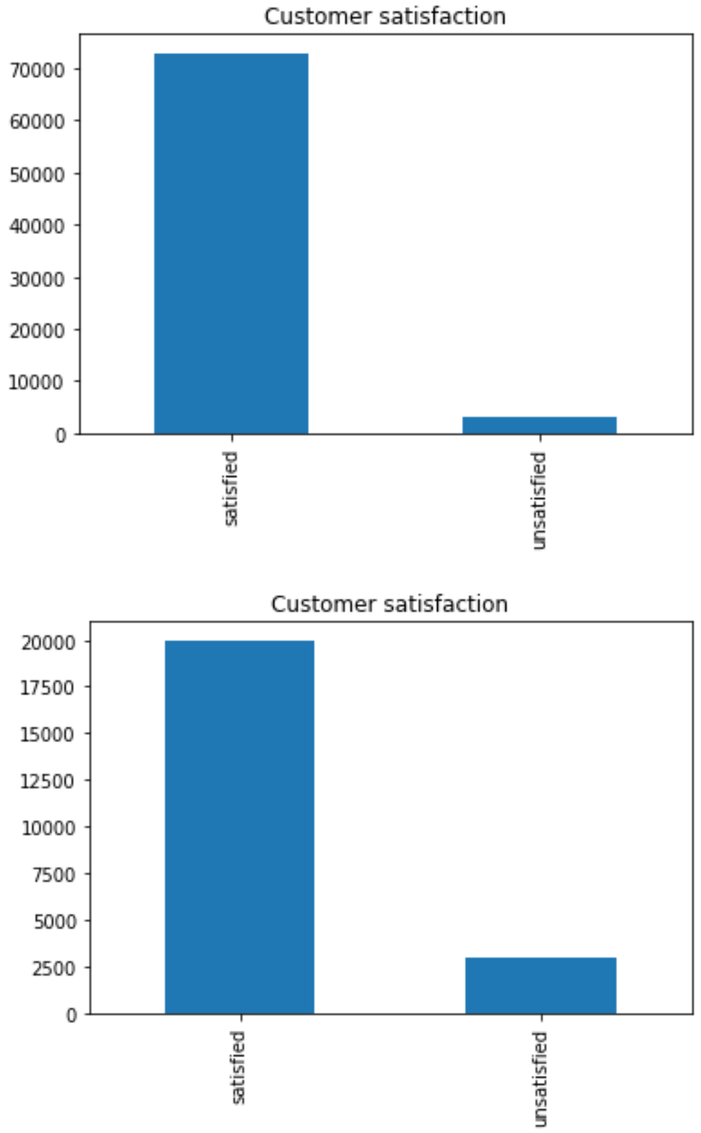
undersampling

- Tomek links
- Cluster centroids

oversampling

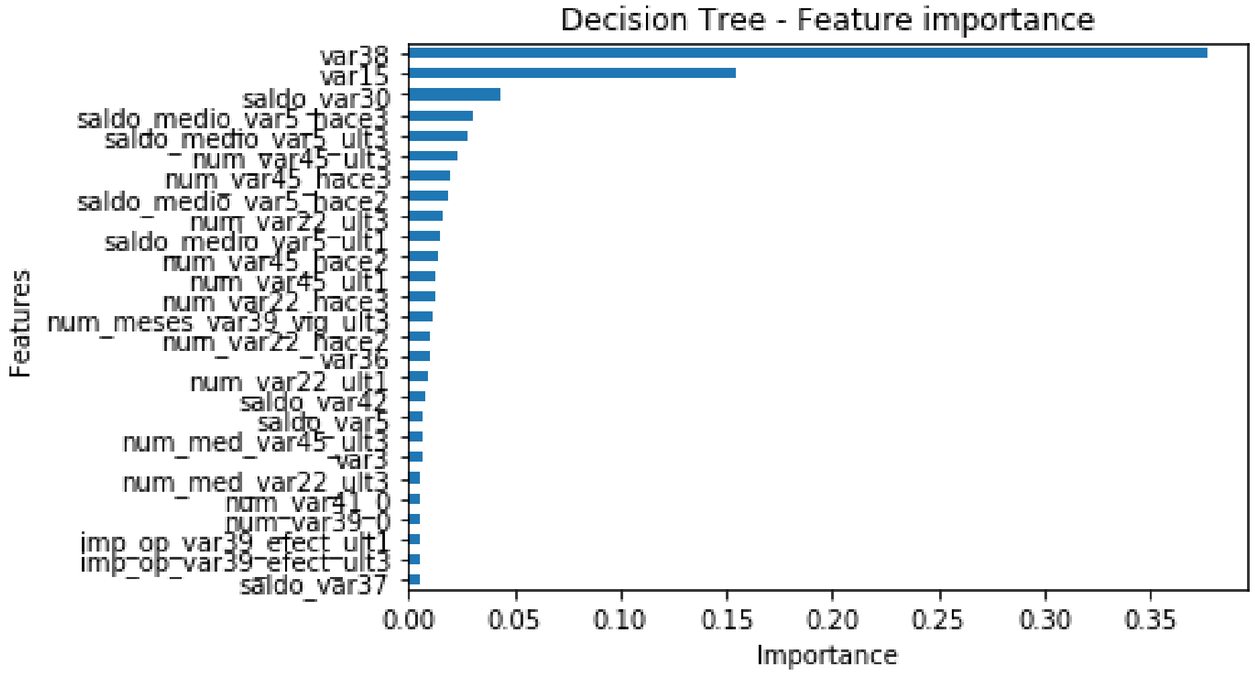
- SMOTE

Random undersampling

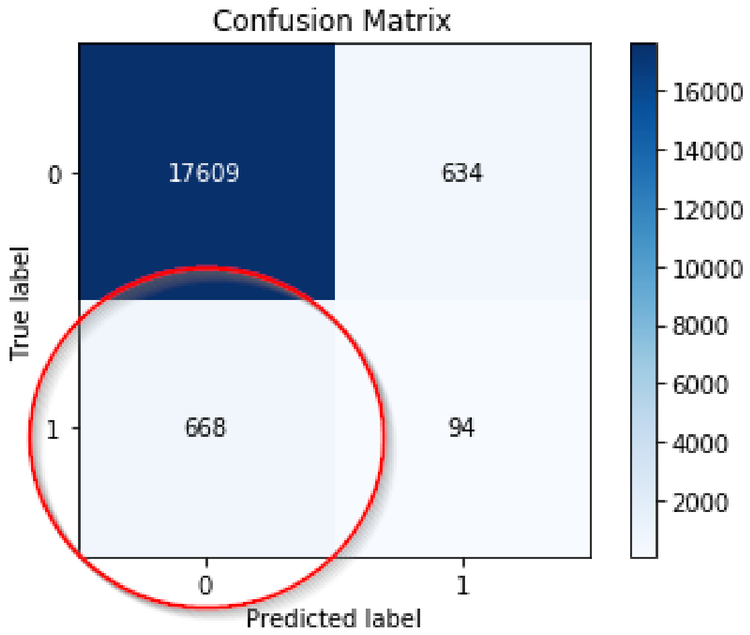


random forest
feature importance

- mortgage
- age
- account balance



random forest
unbalanced dataset



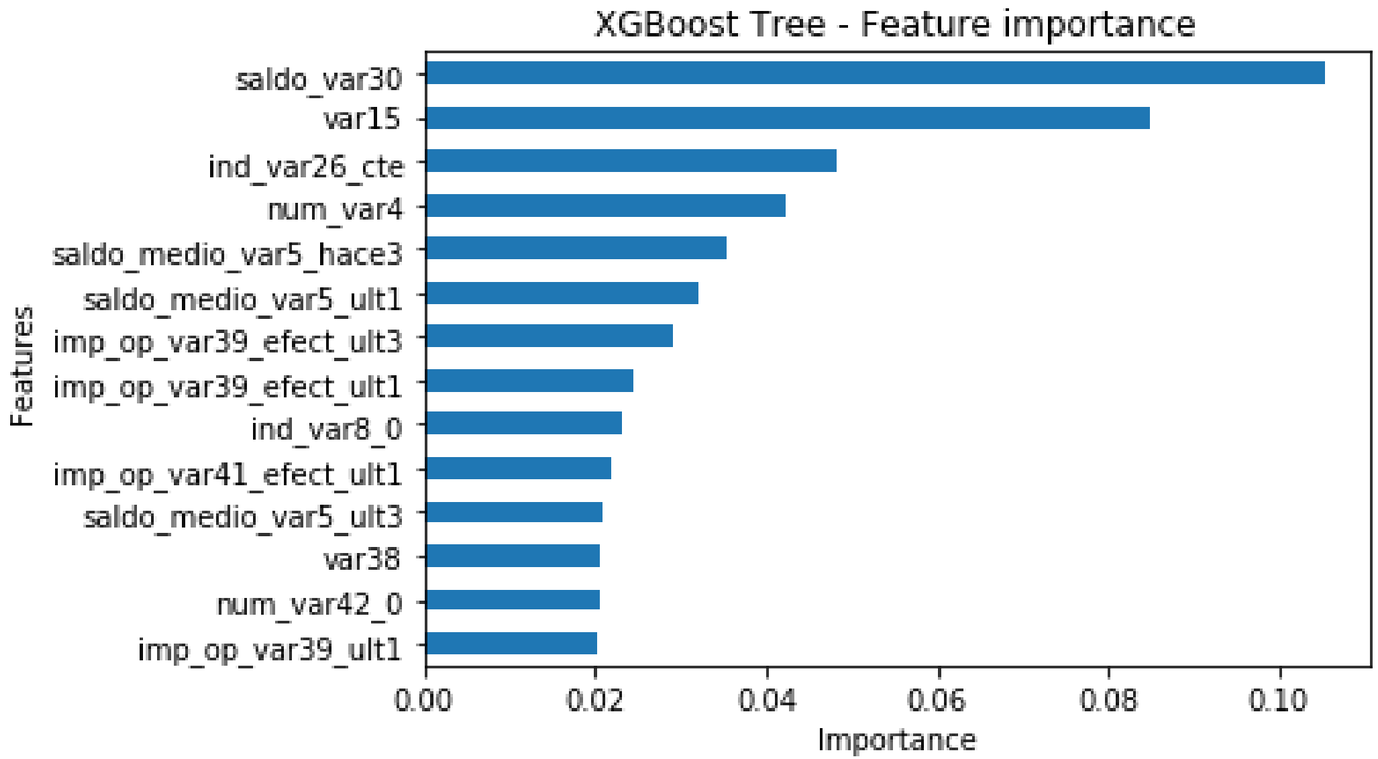
	precision	recall	f1-score	support
0	0.96	0.97	0.96	18243
1	0.13	0.12	0.13	762
accuracy			0.93	19005
macro avg	0.55	0.54	0.55	19005
weighted avg	0.93	0.93	0.93	19005

random forest
comparison

Model	data	accuracy	f1 avg	f1 weighted
Base Model (Decision Tree)	standarized	.93	.54	.93
Base Model (Decision Tree)	resampled	.81	.60	.81
Random Forest	resampled	.85	.62	.84
GridSearch Random Forest	resampled	.87	.60	.84
Random Forest	standard PCA	.95	.53	.94

XGBoost
feature importance

- account balance
- age
- ind_var26_cte



conclusions
lessons learned

missing / incomplete data
difficult to interpret.

high dimensional data
computational power required.

Unbalanced datasets
sampling, hyperparameter tuning.

conclusions
future work

unbalanced data

Further work on the sampling / hyperparameter tuning.

additional ML algorithms

SVM, AdaBoost.

pipelines

Streamline the machine learning process.

thank you.