Solution for Santander Customer Satisfaction competition, 3rd place

Lucas Silva, Gilberto Titericz, Dmitry Efimov, Ikki Tanaka, Darius Barušauskas, Marios Michailidis, Mathias Müller, Davut Polat, Stanislav Semenov, Dmitry Altukhov

April 28, 2016

Outline

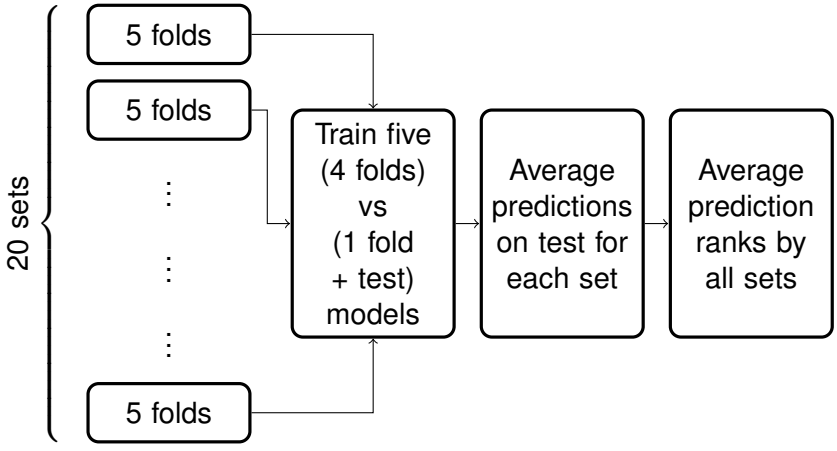
Cross validation scheme

Feature engineering

Models

Ensembling

Cross validation scheme



Feature preprocessing

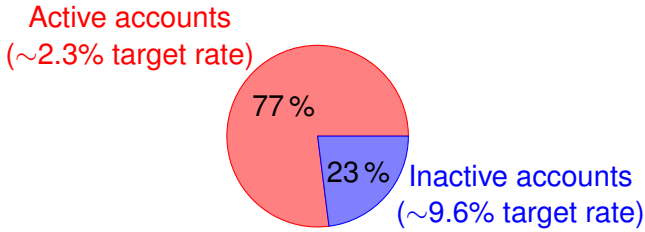
- removing constant features
- removing identical variables
- removing indicator features
- scaling features
- var38 = 117310.979016494 $\Rightarrow$ var38 = NA

Feature engineering

- **Sum of Zeros**
- **t-SNE**: mapping of feature space to lower feature space such that joint distribution of distances between points stays the same
- **PCA**: principal component analysis
- **K-means**: data clustering with number of cluster as a feature
- **Likelihoods**: the probability of target 1 within the subset of samples
- **Special features**

Special features

- Binary feature to determine if account is “active” or “inactive”: inactive accounts contain 18 non-constant features only



- Percentile rank of var38 (income?) within each var15 (age?) group
- Binary mod 3 features

Models

- **XGBoost**
- **RGF**
- **Neural network**
- **FTRL**
- **Random Forest**
- **Adaboost**
- **ExtraTrees**
- **KNN**
- **Lasso**
- **SVM**

Ensembling

- For each model we used bagging technique (train model on different subsets and average predictions) to reduce the noise
- The final prediction is a linear combination of predictions from all models
- AUC metric is not differentiable function, so we cannot use gradient descent
- To optimize AUC and find the best coefficients for the linear combination we used Nelder-Mead method (function `optim` in `R`)

Some lessons

- Build CV scheme correctly
- Trust your CV
- Choose the simplest model
- Working in big team is a very nice experience
- Search for something unusual in the data (special features?)

Thank you! Questions?

Dmitry Efimov

diefimov@gmail.com

kaggle.com/efimov

github.com/diefimov