Analysing the Santander Customer Satisfaction dataset

INTRODUCTION

In this project, I have tried to predict whether a certain customer is satisfied with the service or not, using the Santander Customer Satisfaction dataset. The raw data was taken from the corresponding competition on Kaggle. The goal of this project is to be able to predict the satisfaction of a certain customer, based on the data that we have on them, and to figure out what features of the data play the biggest role in making a customer satisfied. I have also used the data on the Kaggle page for this competition, to help figure out what certain features represented etc.

DATA

The train dataset consists of 370 different features, that are each tied to some customer (observation), of which there are 76020. The *TARGET* column is the value that I had to predict. It's entry is 1 for unsatisfied customers, and 0 for the happy ones. The test dataset consists of 75818 observations.

IMPORTANT FEATURES

- `var15` – one of the most important features – the age of the customer
- `var_num3` – the number of bank products that a customer has (had)
- `var3` – the nationality of the customer
- `var38` – the value of the mortgage that a customer has

DATA PREPROCESSING

The dataset had quite some challenges considering data processing, like the need to deal with removing constant features, removing duplicate features and dealing with some features manually.

I have removed the entries for the nationality of a customer, where there was no nationality specified, by replacing them with the most common nationality.

In the var38 feature, I have discovered that there are 14868 of the 117310.979016 entries. This value is very close to the mean of the column. Because of this uncommonly high number of this particular entry (117310.979016), compared to other entries, I logarithmically transformed the `var38` feature, to better the results.

MODELLING AND PREDICTING

I have used quite some modelling different techniques in this project, and the most successful of them all, based on the **receiver operating characteristic curve** accuracy score. I used this type of scoring

method, because it was the one required for the competition, and seemed rather interesting to me when I dove deeper into it.

RESULTS

The best result was achieved using the Gradient Boosting Classifier, which gave me an accuracy of 83%. Using those results I figured out t he importance of each feature. Here are the 20 most important ones: