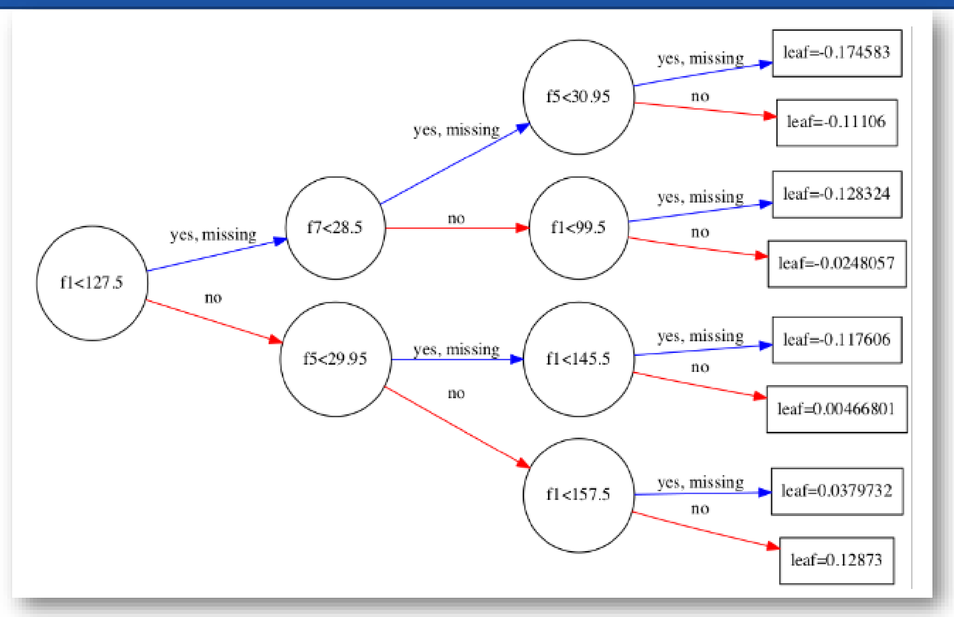
Analysing the Santander Customer Satisfaction dataset

L. Triska

Introduction 1

In the project I tried to predict if 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. I have also managed to figure out what features of the data play the
biggest role in making a customer satisfied.



Visualisation of a Gradient Boosting Decision Tree

Methodology 2

The methodology implies the of different classifiers on the dataset, and then comparing the results obtained from each classifier. I used the following ones: Decision Tree Classifier, Gaussian Naïve Bayes, Logistic Regression, Ada Boost Classifier, Random Forest Classifier, Bagging Classifier, Gradient Boosting Classifier

Data 3

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 equal to 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 4

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.
Here are some other specific challenges I dealt with:

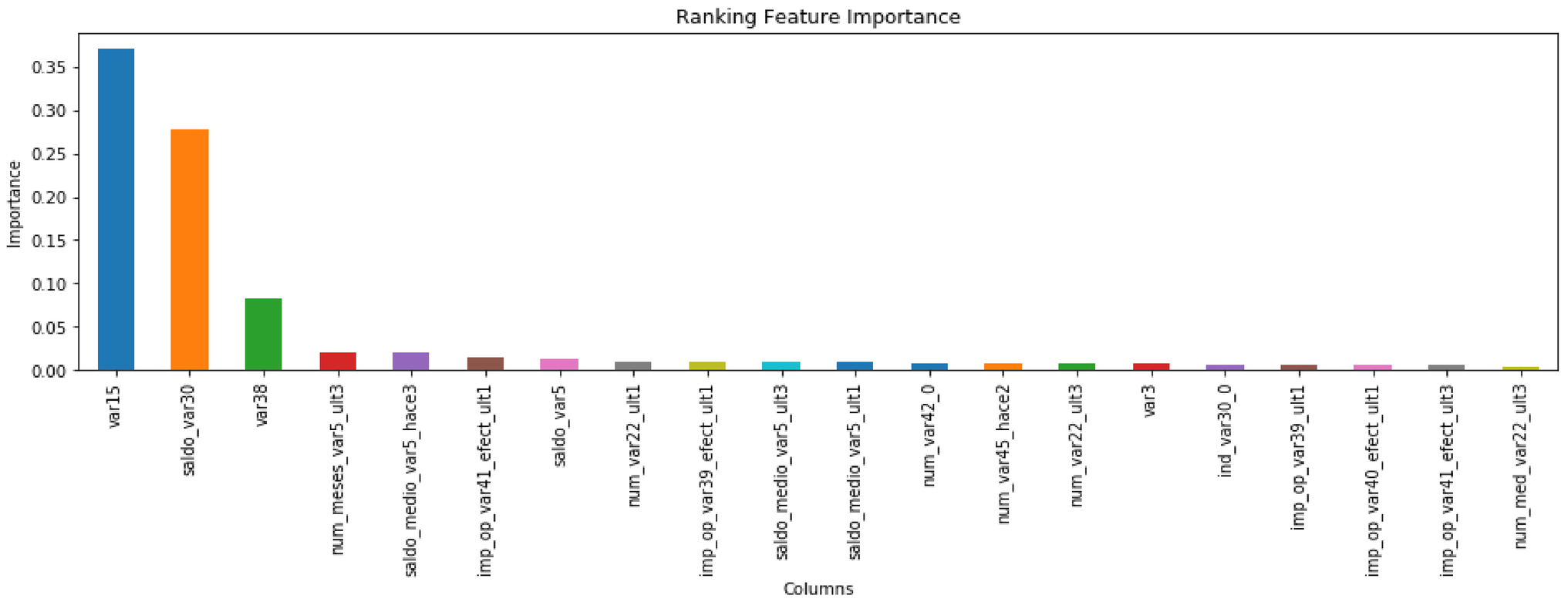
- I have removed the entries for the nationality of a customer, where there was no nationality specified, by replacing them with the most common nationality.
- In the var38 feature, I have discovered that there are 14868 of the **117310.979016** entries. This value is very close to the mean of the column. 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.

Results and Conclusions 5

The best result was achieved using the Gradient Boosting Classifier, which gave me an accuracy of 83%. Also, I think it's worth mentioning that the dataset is semi-anonymized, so it is unclear what a feature represents. I found the importance of each feature, the *most* important of which are:

1. var15 – the age of a person
2. saldo_var30 – a person's balance
3. var38 – the mortgage value

Here are the 20 most important features



Applied Sciences Faculty, Ukrainian Catholic University, Lviv, Ukraine
(triska@ucu.edu.ua)