Data-Science report

Santander
CONSUMER BANK

Table of content

1 Introduction

1.1 Problem

For organisations and management customer satisfaction is off course a key measure of success. Unhappy customers will leave your company and unhappy customers don't explain and will suddenly just leave.

In this case we will analyse the Santander bank data and identify which customers are dissatisfied early in the relationship. This enables Santander bank to take pro-active steps to improve a customers happiness, before its too late. The dataset contains hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with being a client of the bank.

In addition to this main goal they also want to extract maximum value from the data.

1.2 Goal

The main goal of this case is to predict which costumers will be happy and which will be unhappy early in the relationship enabling Santander to take proactive steps. In addition we will extract maximum possible value from the data. The valuable insight we'll get from the data will be presented at the end of the report as actionable management insights.

1.3 Management questions

Q1: Which customers are satisfied and which are dissatisfied?
Q2: What are the most important predictors?
Q3: Which valuable insights can we get from the data?

1.4 Points of attention

We will describe a few points of attention to mitigate the risks attached to data-science projects.

1.4.1 Profitability of the Data-Science project

In this case we've kept a lean/agile way of working. That's the reason we started with a fast GBM prediction. To deliver value and a benchmark result as soon as possible. From there on we were focused on keep improving predictions(value optimization) until we've reached our management goal. From there on we focus on the next goals.

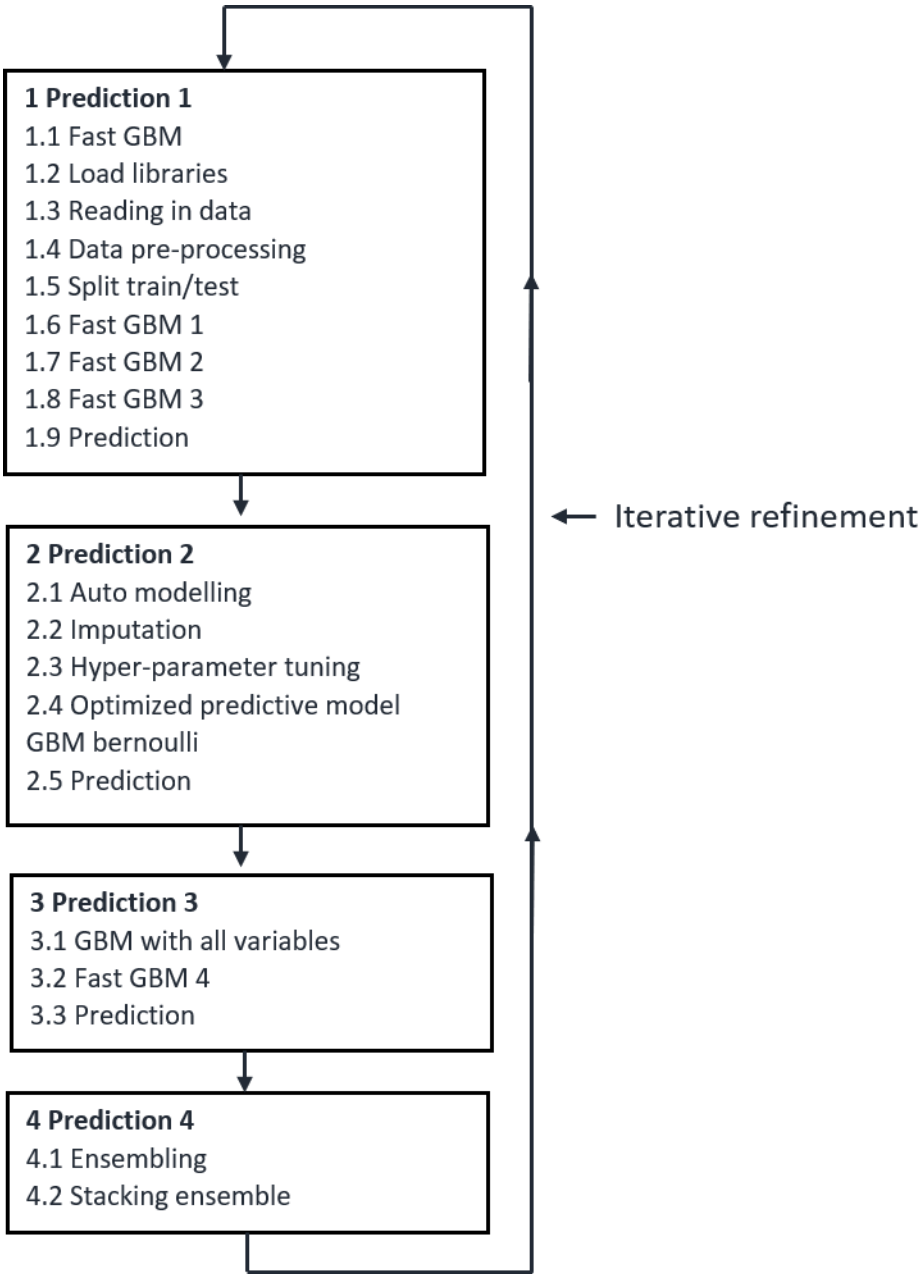
1.4.2 Understanding/Trust of the management

To increase understanding and Trust of the management we'll describe the full process in detail in section 2. (The detailed process description will follow soon).

1.4.3 Data quality

The data was downloaded from Kaggle.com. A very trustworthy source.

2 Procedure

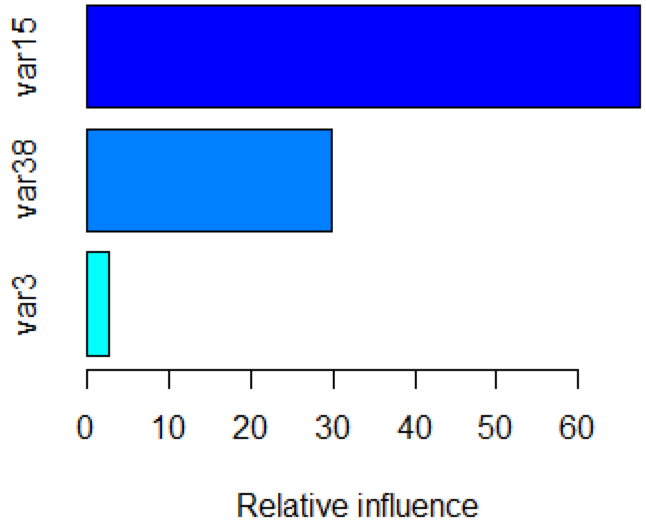


1 Prediction 1
1.1 Fast GBM
1.2 Load libraries
1.3 Reading in data
1.4 Data pre-processing
1.5 Split train/test
1.6 Fast GBM 1
1.7 Fast GBM 2
1.8 Fast GBM 3
1.9 Prediction
2 Prediction 2
2.1 Auto modelling
2.2 Imputation
2.3 Hyper-parameter tuning
2.4 Optimized predictive model GBM bernoulli
2.5 Prediction
3 Prediction 3
3.1 GBM with all variables
3.2 Fast GBM 4
3.3 Prediction
4 Prediction 4
4.1 Ensembling
4.2 Stacking ensemble
Iterative refinement

2.1 Prediction 1

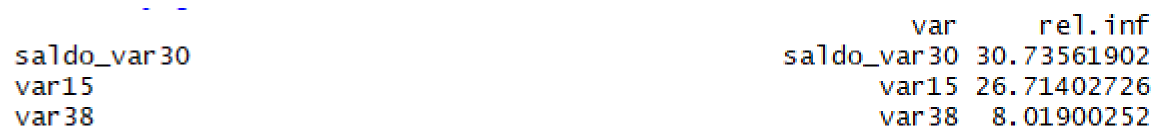
For prediction 1 we applied a fast GBM model to get a first important insight in the data. Therefore first we load the R libraries(R functions) we need to be able to perform the analysis. Then we read in the dataset from the C drive and perform several pre-processing steps, such as taking a subset of the data to speed up the analysis. Then we split the data in a train and test set. In the first GBM we inspected the 3 variables that stood out/differed from the other variables. The result of GBM 1 is shown in figure 1. Var 15 is the most important variable. Then var 38 and then var3(Although var 3 doesn't have a very high score).

Figure 1 – Number of satisfied(1) and unsatisfied(2) customers.



For the second GBM we inspected all the variables. The results show a long list of variables. However 3 features/variables are most important. Again saldo_var30 and var15. This means these 2 variables explain a large part of the variance in Y=TARGET(customer satisfaction). Var 38 is third most important.

Table 1 – Most important features



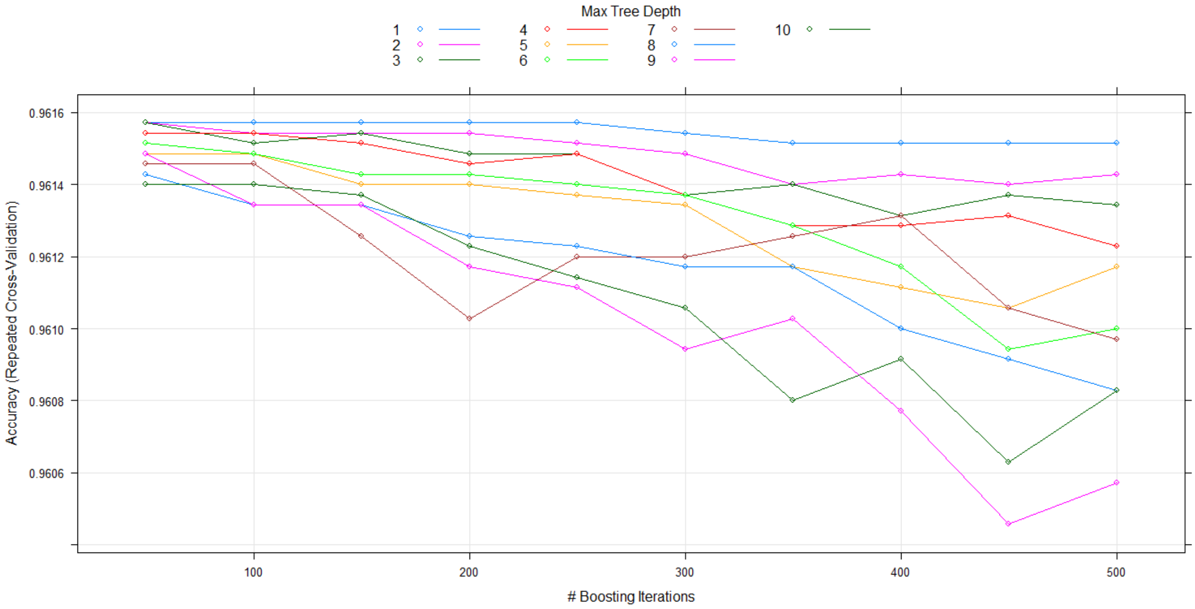
```
                                                                var      rel.inf
saldo_var30                                             saldo_var30  30.73561902
var15                                                         var15  26.71402726
var38                                                         var38   8.01900252
```

Therefore in the third GBM we input these 3 variables. Then with this model we make the pr edictions and we reach an AUC score of 0.83. This is a good score (See section 3.1.2).

2.2 Prediction 2

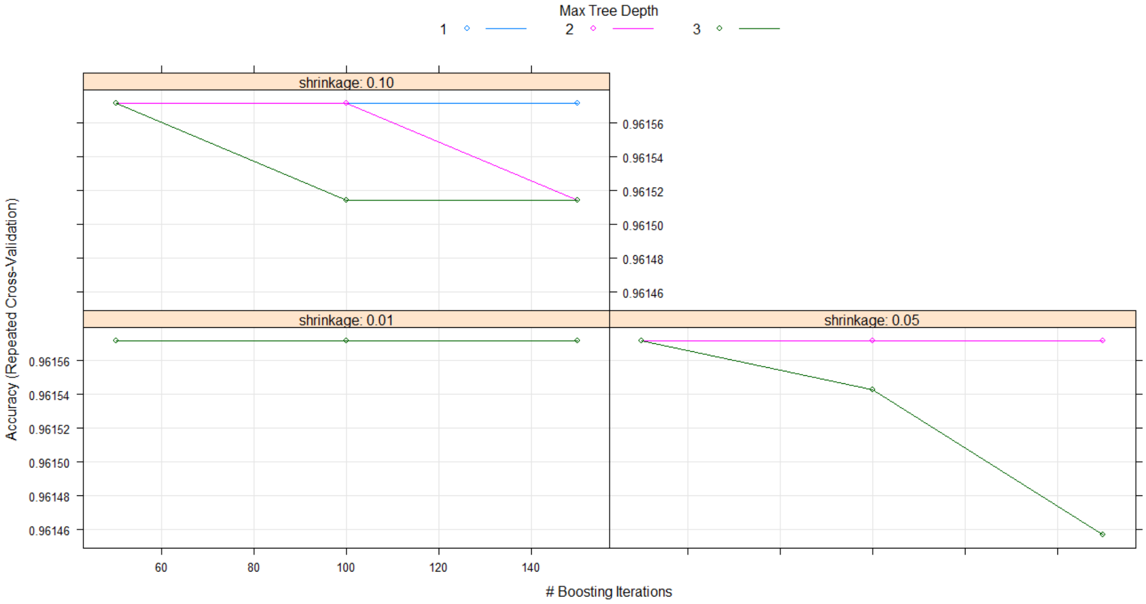
In the second prediction we've used some automated R functions. First we did a full automated grid search.

Figure 2 – Full automated hyperparameter search



Based on the full automated grid search results we selected the hyperparameter inputs for the next partial automated grid search.

Figure 3 – Partial automated hyperparameter search



The optimal hyperparameters of this grid search were used as input hyperparameter to make our prediction with the GBM model.

2.3 Prediction 3

This section follows soon.

2.4 Prediction 4

This section follows soon.

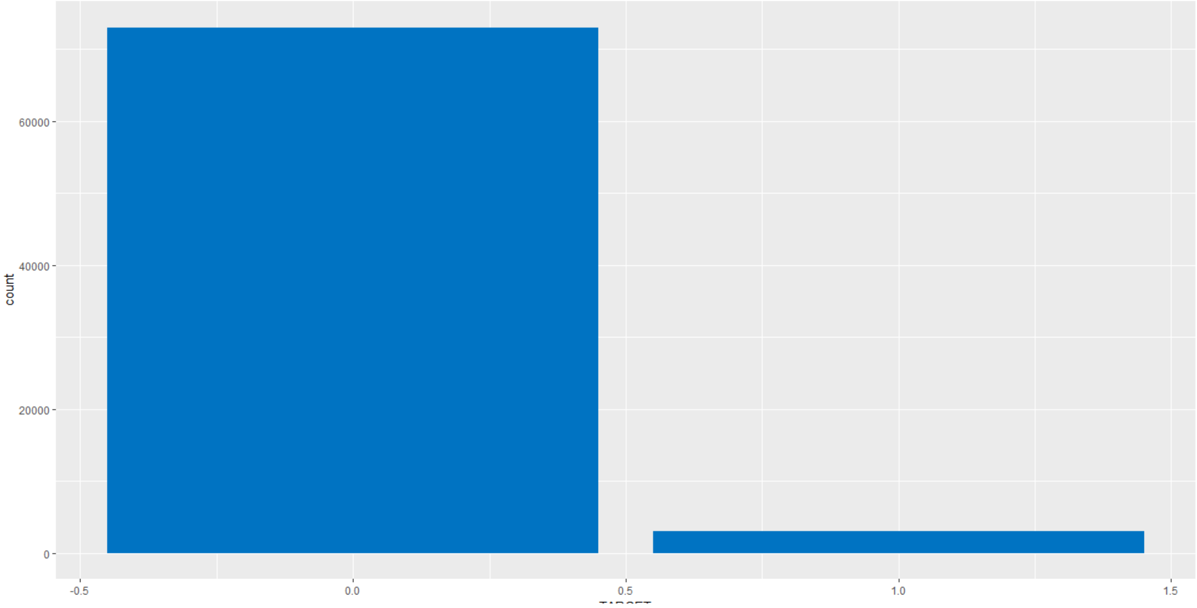
3 Results

3.1 Prediction 1

3.1.1. The dependent variable

The dependent variable consists of 73.012 0 (satisfied) values and 3008 1 (unsatisfied) values. 96% of the customers are satisfied and 4% are unsatisfied. This can graphically be seen in figure 1.

Figure 4 – Number of satisfied(1) and unsatisfied(2) customers.



3.1.2 AUC

The AUC score of the first prediction is 0.8281295.The AUC (Area under the curve) score is an important metric for classification problems. Probably the second most important metric after accuracy. An AUC score of 1 is a perfect test and a score of 0.5 is the lowest score. A general guide for classifying the accuracy of the test;

.90-1 = excellent (A)
.80-.90 = good (B)
.70-.80 = fair (C)
.60-.70 = poor (D)
.50-.60 = fail (F)

The AUC score of 0.83 means a good (B class) accurate score.

3.2 Prediction 2

3.2.1 Accuracy

The accuracy is 0.96 (whether a customer is satisfied with his/her banking experience.

3.2.2 AUC

The AUC is

3.3 Prediction 3

3.3.1 AUC

The AUC is

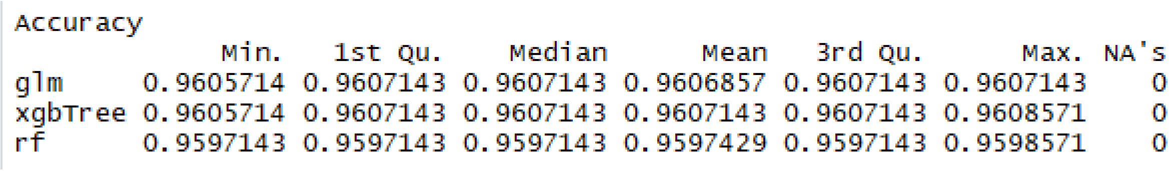
3.4 Prediction 4

For prediction 4 we used a stacking ensemble with the models logistic regression, XGboost and Random forest as 3 base models.

3.4.1 Independent model accuracy

The mean accuracy scores are shown in table 2. The XG boost model scores slightly better than the logistic regression, however the score is the same 96% accuracy score that we achieved with the second prediction.

Table 2 – Model accuracy 3 base ensemble models

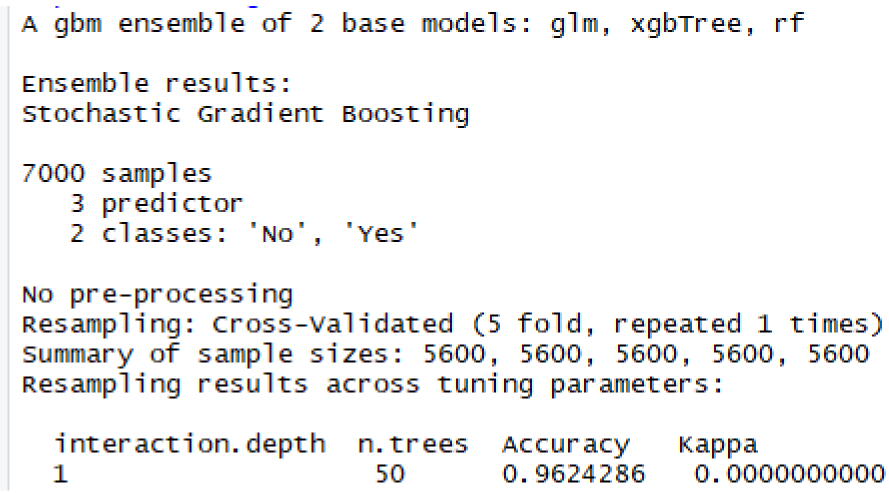


```
Accuracy
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
glm     0.9605714 0.9607143 0.9607143 0.9606857 0.9607143 0.9607143    0
xgbTree 0.9605714 0.9607143 0.9607143 0.9607143 0.9607143 0.9608571    0
rf      0.9597143 0.9597143 0.9597143 0.9597429 0.9597143 0.9598571    0
```

3.3.2 Stacking ensemble accuracy

Next we'll do a stacked ensemble. Figure 3 shows the accuracy of the stacked ensemble is 0.96.

Figure 5 – Stacking Ensemble results



```
A gbm ensemble of 2 base models: glm, xgbTree, rf

Ensemble results:
Stochastic Gradient Boosting

7000 samples
   3 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 1 times)
Summary of sample sizes: 5600, 5600, 5600, 5600, 5600
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                  50       0.9624286  0.0000000000
```

4 Conclusion

4.1 General conclusion

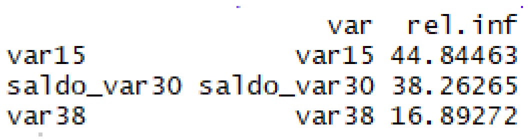
Q1: Which customers are satisfied and which are dissatisfied?

The GBM models we've used predicted the satisfaction scores on the test set with 96% accuracy.

Q2: What are the most important predictors?

With the following 3 predictors we were able to predict accuracy in (dis)satisfaction with 96%.

Table 4 – 3 most important variables



```
                    var   rel.inf
var15             var15 44.84463
saldo_var30 saldo_var30 38.26265
var38             var38 16.89272
```