# Fraud detection

## Exploration & missing value handling

After data loading, exploratory data analysis was conducted.

The missing values were assumed to be meant by “?” symbol and everything else was assumed to be categories with information value (e.g. categorical values within a numerical variable). Since there was no information about the variables available, and the reason behind the missing values was unknown, the missing values were marked with a neutral category in the categorical variables.

The categorical labels from the numerical variables were extracted into separate variables, and then filled with the neutral category as placeholders.

There were too many missing values to delete the records, and imputation would have introduced noise if the missing values didn't belong to either available category.

Since we are talking about financial transactions, the class distribution needed to be checked. This revealed that there was a class imbalance within the dataset. This was solved by sub-sampling. Several methods with the rf algorithm of the Caret package were compared, and at the end the SMOTE method worked the best:

**Up-sampling:**

* Accuracy : 0.6939
* 95% CI : (0.651, 0.7344)

**Down-sampling:**

* Accuracy : 0.7245
* 95% CI : (0.6826, 0.7636)

**SMOTE:**

* Accuracy : 0.7653
* 95% CI : (0.7252, 0.8022)

**ROSE:**

* Accuracy : 0.5286
* 95% CI : (0.4833, 0.5735)

## Training & tuning

The caret package was used to train various algorithms on the sub-sampled data using repeated cross-validation (5 number and 3 repeats) and a grid search of 10. After training, the ones with either Specificity or Sensitivity above 0.6 and an overall accuracy of at least 0.7 were selected.

The selected models were then tuned to their best performance considering the available data.

**KKNN:**

* Accuracy : 0.8939
* 95% CI : (0.8632, 0.9197)
* Sensitivity : 0.9435
* Specificity : 0.7857

**Greedy Prototype Selection:**

* Accuracy : 0.8163
* 95% CI : (0.7791, 0.8496)
* Sensitivity : 0.8958
* Specificity : 0.6429

**L2 Regularized Linear Support Vector Machines with Class Weights:**

* Accuracy : 0.7327
* 95% CI : (0.6911, 0.7714)
* Sensitivity : 0.7411
* Specificity : 0.7143

After the model tuning, the models were combined into an ensemble model, where each model had a weight of 1, in other words their predictions were averaged.

Confusion Matrix and Statistics

Reference

Prediction bad good

bad 309 26

good 27 128

Accuracy : 0.8918

95% CI : (0.8609, 0.9179)

No Information Rate : 0.6857

P-Value [Acc > NIR] : <2e-16

Kappa : 0.7495

Mcnemar's Test P-Value : 1

Sensitivity : 0.9196

Specificity : 0.8312

Pos Pred Value : 0.9224

Neg Pred Value : 0.8258

Prevalence : 0.6857

Detection Rate : 0.6306

Detection Prevalence : 0.6837

Balanced Accuracy : 0.8754

'Positive' Class : bad

## Evaluation of the results

The results show that only 27 bad transactions were misclassified, while 26 good transactions were classified as bad. The good part of the result is that over 91% of the bad transactions were identified. The not so good part is, that in order to do this, 17% of the good transactions had to be misclassified.

Because of the limited resources available (time, computational power, amount of data, meta-data), the following improvements could be made:

* Feature engineering, if meta-data could be obtained.
* More data would most likely improve the model performance.
* Trying out various other models from the caret package and reworking the ensemble with it.