Predicting Financial Problems in Arrears Customers using Random Forests

(Bachelor's thesis)

Diederik van Krieken, s2009730, diederikvkrieken@gmail.com,
Marco Wiering[*], and Olav Aarts

July 29, 2014

Abstract

Many banks have felt the effects of the financial crisis through an increase in the debt of their customers. For the bank and customer it is surely preferred to reverse back into a positive balance as quick and long lasting as possible. This present research will be an attempt in determining predictability of which customers in arrears will have more difficulty in paying back their debts to the bank compared to others. This predictability will support the development of new approaching methods to these customers that are different than other clients. In this thesis we will focus to recognize customers with relevant financial problems. For this research a case study was done at one of the biggest Dutch banks. A sample group of customers whose accounts were liquidated was taken. By using random forests and different data techniques to cope with the imbalanced data set, features were found which allowed prediction of these customers in arrears. We found, among others, that the change in amount of transactions is a feature that predicts financial problems. These features will now be implemented so customers with these flags are identified earlier and will receive more appropriate treatment.

1 Introduction

Since the 2008 financial crisis, the number of people unable to repay their debt and/or loans to the banks has gone up. This change results in an enormous increase of clients being handed to the arrears departments which costs the bank substantial amounts of money. The arrears departments are responsible for the collection and recovery of the arrears of these clients.

Originally these arrears departments have had one single approach for all clients with arrears. They start with text messages, emails and letters. If the clients still fail to pay their debt they increase the attention by making phone calls and extra special treatment like removing fines. Since so many more clients have arrived in arrears management after the financial crisis these departments needed more efficient strategies and processes.

Predicting customer risk has been one of the main problems for banks (Thomas, 2000; Galindo and Tamayo, 2000). Many models have been used from logistic regression (Bolton, 2010) to neural networks (Scheurmann and Matthews, 2005; Chen and Huang, 2011; Ha, 2010) to others (Twala, 2010).

Currently a binary classifier that predicts whether a client will need extra attention is deployed in the mortgage arrears department of a bank (Sun, Wiering, and Petkov, 2014). This allows the client to be approached in a more efficient manner. Risky clients will immediately receive a phone call, while for the non-risky customers the standard procedure will be followed. The difference in approach means lower costs for the bank, and earlier contact, with possible better aid for the clients.

New insights in the arrears department have shown that clients should be divided in 4 separate groups. Clients can have financial problems, behavioral problems, both, or none. These 4 different groups require different approaches. Right now

[*]University of Groningen, Department of Artificial Intelligence

when a client is called, the employees will place the client in one of these groups to decide which action to take. This placement is done based on their own intuition. In the future an automated system might be possible to support the employees when calling.

In order to be able to predict financial problems independently from behavioral problems a model is built which will solely predict financial problems for customers in arrears.

In this thesis it is tried to build the model required above by building a binary classification model to predict whether new clients in arrears will have financial problems based on financial features. In order to do so a dataset of customers whose accounts were liquidated is obtained.

One of the requirements of the bank doing the research for is that it must be known which features are of importance in the model. This makes black box models like neural networks undesirable.

This dataset of liquidated accounts is preprocessed and a model is built using the random forests method (Breiman, 2001). Although different techniques can be used (Twala, 2010; Crook, Edelman, and Thomas, 2007), random forests consequently remain in the top classifiers (Caruana and Niculescu-Mizil, 2006; Cubiles-De-La-Vega, Blanco-Oliver, Pino-Mejías, and Lara-Rubio, 2013; Lariviére and den Poel, 2005) and are therefore preferred.

Different techniques are proposed and tested in order to cope with the imbalanced data set.

The research questions attempted to answer in this thesis are:

- Is it possible to predict financial problems for arrears customers using the random forests algorithm?

- Can the random forests algorithm be improved to cope with the imbalanced data set?

2 Method

The data required to predict financial problems in arrears are composed by finding all customers whose accounts were liquidated in two months of 2014. Exploratory data analysis has shown that 80% of all these customers, did not actually have financial problems but their arrear was a combination of neglect by these customers and unnecessary bank costs.

A common model used in bank classification is the random forests algorithm (Lariviére and den Poel, 2005). This method developed by Breiman (2001) is a combination of bagging with decision tree learning. In section 2.3 the working of this algorithm will be explained.

2.1 Data

A sample of the customers whose accounts were liquidated in the two months of 2014 was taken. Since we want to predict financial problems, this originally was only the group with financial problems. Another sample with an equal amount of accounts in arrears would be taken as customers without financial problems. Unfortunately this could not be done since the inflow of the financial problem customers was very unusual.

There was a substantial inflow of customers at a specific month mark. This very non-gaussian form would be difficult to recreate from people landing in arrears and probably is caused by procedure times. Unfortunately, the creation of a data set without financial problems was even made impossible since there was no history of customers in arrears during these months.

Further exploratory data analysis has shown that most of the customers did not really have financial problems. The main cause for these accounts to be liquidated was by the negligence of the customer.

Eventually a threshold of XXX Euro as written off value was taken in order to divide the dataset. Customers with a debt less than XXX Euros were seen as those without financial problems, those with more than XXX Euro debt had financial problems. This gave the dataset as seen in table 1.

In order to check the accuracy of the proposed models the dataset was split in a training and a test set. This was done by taking the accounts which were liquidated in month 1 for the training set and those who were liquidated in month 2 for the test set.

Table 1: Data distribution

	Financial Problem	No Financial Problem	Total
Month 1	5,23%	51,87%	57,10%
Month 2	6,89%	36,05%	42,90%
Total	12,08%	87,92%	100%

The only data used were those from before the last time the account landed in arrears. A customer could swing in and out of debt multiple times. The last time his account arrived in arrears, and did not have a single day that its balance was positive, was the last time he arrived in arrears.

2.2 Preprocessing

Computer limitations required us to decrease the data set. To handle the data and the different time series, each feature was decreased to a linear model. This was done with the data before the customer arrived in arrears. An example can be found in figure 1:

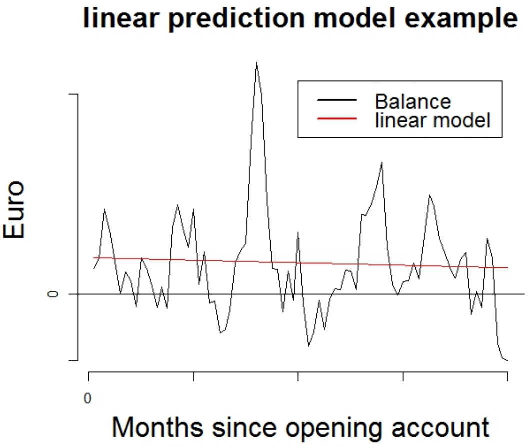

Figure 1: Example of linear model fit for Balance

This was done for the 6 financial variables Balance, True_balance, positive days, negative days, transactions and sustenance. These values were aggregated per month to cope with the data size. Finally we ended up with a dataset containing 12 features. Each variable has an intercept (β) and a slope (α): $f(x) = \alpha x + \beta$.

2.2.1 Balance

The balance of the account averaged per month.

2.2.2 True balance

Some customers had a quarterly limit meaning that they could legally be in debt. These customers were not in arrears when having a negative balance, so their allowed debt was added to their balance. A customer with a balance of -250 Euro but an allowed limit up to -1000 would have a "True balance" of +750. Important is that a quarterly limit customer who has not one positive balance day every three months automatically loses its limit allowance (its -1000 allowance gets revoked). This could mean that the customer goes from a "True balance" of 750 to -250 in a single day without any transactions.

2.2.3 Positive and negative days

Since we aggregate the data per month first, we can save the amount of positive and negative days per month. These values are not compensated for negative allowance, as described above, and take the balance value. Thus a person with an allowed debt will have negative days and might not be in arrears.

2.2.4 Amount of transaction

The amount of transactions each month.

2.2.5 Sustenance

The sustenance was calculated by adding all positive transactions and deposits per month.

2.3 Random forests

Random forests were introduced by Breiman (2001). This algorithm, shown in Algorithm 2.1, is an ensemble classification method which includes multiple decision trees. These trees are generated by using bootstrap subsets (with replacement) from the original sets, and random feature selection in the tree building process. Using these decision trees, a majority vote is used to predict the correct class.

As can be seen in the algorithm there are two variables that must be set: The amount of variables used and the amount of trees to be built: m_{try} and *ntrees*. Khoshgoftaar, Golawala, and Hulse (2007) suggest that a small amount of *ntrees* is sufficient, whereas the value of m_{try} should be calculated. On the other hand does Sun et al. (2014) suggest a *ntree* of 1000. For this model the optimal *ntree* value is found using built-in functions implemented

Algorithm 2.1 Random forests

```
Require: Data set $S = \{X_{i,1}, ..., X_{i,m}, Y_i\}$
  while $b = 1$ to $ntree$ (number of bootstrap samples): do
    Draw bootstrap sample $B^*$ of size $N$ from training data
    (Grow random-forests tree $T_b$:)
    while Number of node sizes $< n_{min}$ do
      Select $m_{try}$ variables at random from $p$ variables
      Pick best split-point
      Split node into daughter nodes
    end while
    Output ensemble of trees $\{T_b\}_1^M$
  end while
```

by Breiman, Cutler, Liaw, and Wiener (2003). As shown in section 3.4 the optimal amount of trees depends on whether the balanced, under-sampled or default random forests algorithm is used.

Breiman (2001) suggests three possible values for m_{try}: $\frac{1}{2}\sqrt{m_{try}}$, $\sqrt{m_{try}}$, and $2\sqrt{m_{try}}$. Calculations show that a m_{try} value of 3 is preferred. See figure 2. This corresponds with the $\sqrt{12}$ as suggested by Breiman (2001).

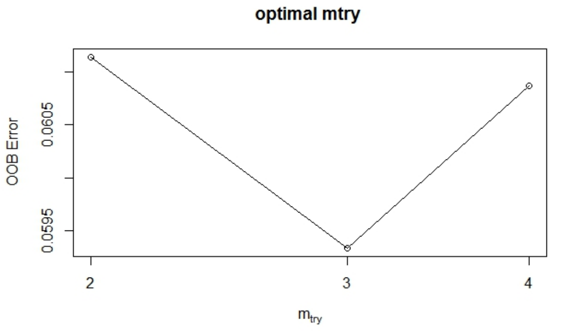

Figure 2: Error rate for different m_{try} values

2.4 Imbalanced data

As shown in section 2.2 the data used are imbalanced. The cases with no financial problem outnumber the financial problem cases 9:1. There are different methods (He and Garcia, 2009) to deal with imbalanced data sets. For this research we have decided on only two methods. Pre-sampling the data (Liu, Wu, and Zhou, 2009; Hulse, Khoshgoftaar, and Napolitano, 2007) and balanced random forests (Xie, Li, Ngai, and Ying, 2009; Chen, Liaw, and Breiman, 2004). These options for imbalanced datasets are discussed below.

2.4.1 Pre-sampling

Liu et al. (2009) shows that under- and oversampling are the preferred pre-sampling methods before training the model.

When under-sampling the amount of samples for the majority class is reduced to a 1:1 relation. For example, the majority class outnumbers the data 10:1. The new sample model will decrease the majority class to be of equal size as the minority class by selecting random samples.

When oversampling, the data set is increased by taking each sample from the minority class multiple times until a 1:1 ratio is reached.

2.4.2 Balanced random forests

As explained before, random forests use bagging methods in order to create subsample sets. Balanced random forests use these bagging methods to under-sample or oversample (Xie et al., 2009). Basically each time a subset is built, the balanced random forests algorithm is required to take a subset where the amount of both classes is equal. The algorithm is shown in Algorithm 2.2.

2.5 Linear model comparison

Although we explained the highest results are usually obtained using random forests, a general linearised model (Nelder and Baker, 1972) is built to check whether this statement holds. The optimal threshold is found by calculating the upper left corner of the receiver operating characteristic (ROC) curve based on the training set.

2.6 Fitness of the model

2.6.1 Confusion matrix

The confusion matrix (table 2) contains the amount of true positive, true negative, false positive and false negative of the predicted data. By using this table one can say something about the sensitivity

Algorithm 2.2 Balanced random forests

```
Require: Data set $S = \{X_{i,1}, ..., X_{i,m}, Y_i\}$
  while $b = 1$ to $M$ (number of bootstrap samples): do
    Draw minority class bootstrap sample from training data
    Draw equal amount of cases with replacement from majority class.
    (Grow random-forests $treeT_{b*}$):
    while Number of node sizes $< n_{min}$ do
      Select $m_{try}$ variables at random from $p$ variables
      Pick best split-point
      Split node into daughter nodes
    end while
    Output ensemble of trees $\{T_b\}_1^M$
  end while
```

and accuracy. Due to the imbalanced data set, accuracy is not desirable since both classifications errors are taken to be equally important and a misclassified financial problem (predicted as not a financial problem, is a financial problem (false negative)) is a lot worse than a false positive.

Table 2: Confusion matrix

		Predict class	
		Financial Problem	No Financial Problem
Actual Class	Financial Problem	True Positive (TP)	False Negative(FN)
	No Financial Problem	False Positive (FP)	True Negative(TN)

Thus two methods are used to show the fitness of the model.

2.6.2 ROC

ROC, as described in Fawcett (2006), is a method to evaluate a learning model which takes the predicted true positive rate (sensitivity) and plots it against the predicted false positive rate. This is dependent on which threshold value is taken.

$$TruePositive_{rate} = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$TrueNegative_{rate} = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

After calculating this curve, an area under the curve (AUC) can be obtained. A perfect model would have an AUC of 1.0, perfect classification, or 0.0, which is a perfect misclassification. A random model has an AUC value of 0.5.

To classify the fitness of the model the AUC is used. Although (Hand, 2009) advises against using this method, and while Hand's solution is the H-measure, a more appropiate solution is found is using the Expected Misclassification Cost (EMC) as proposed by West (2000).

2.6.3 Expected misclassification cost

As explained before, the costs associated with a False Negative (Customer predicted to have no financial problem, but has a financial problem) is higher than vice versa (false positive). According to West (2000) (as a rule of thumb) the relative ratio of misclassification must be 1:5 and is defined as followed where $C_{12} = 5$ and $C_{21} = 1$. $P_{12} = n_2/N_2$ is the false positive rate and $P_{21} = n_1/N_1$ the false negative rate. π_1 is the prior probability of a not financial problem customer and is $\pi_1 = 0.88$. π_2 is the prior probability of a financial problem customer and is $\pi_2 = 0.12$.

$$EMC = C_{21}P_{21}\pi_1 + C_{12}P_{12}\pi_2$$

3 Results

As explained in section 2, 3 different random forests models were built (the linear model will be discussed later). One using the original data set, and 2 others using different ways of coping with the imbalanced data set. The ROC curves for these three methods can be seen in figure 3

3.1 Basic random forests

The first random forests model already has an AUC of 0.882. Its confusion matrix (table 3) shows a very high accuracy, but as predicted many false negative errors. This is due to the high imbalanced dataset.

Table 3: Predictions basic random forests model

		Predicted	
		Financial Problem	No Financial Problem
Actual Class	Financial Problem	5,86%	10,10%
	No Financial Problem	0,67%	83,37%

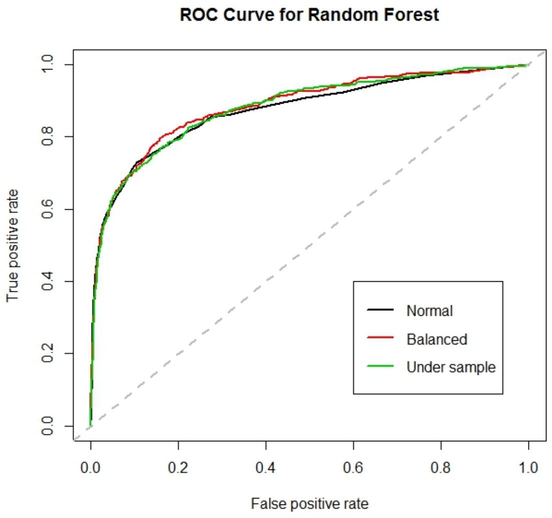

Figure 3: ROC curve of the different random forests models

Based on these results we can conclude that the basic random forests already is a very successful model.

3.2 Random forests under-sample

Using the under- and over sampling methods the following results are obtained. The oversampling method has an AUC of 0.875 where the under-sample has a minimal but predicted increase to 0.88 (Liu et al., 2009; Hulse et al., 2007; Barandela, Valdovinos, Sánchez, and Ferri, 2004). For further comparison only the under-sampled random forests will be used. The confusion matrix (table 4) shows the result with a lot less false negative errors. As explained in section 2.6.1 false negative errors weight heavier than false positive. The bank prefers calling more people with possible financial problems than neglecting them.

Table 4: Predictions basic random forests model trained using under-sampled data

		Predicted	
		Financial Problem	No Financial Problem
Actual Class	Financial Problem	12,05%	3,91%
	No Financial Problem	12,44%	71,60%

3.3 Balanced random forests

When using the balanced random forests method an AUC of 0.884 is obtained. This is higher than the under-sample random forests, but again, more false negative errors are made (table 5).

Table 5: Predictions balanced random forests model

		Predicted	
		Financial Problem	No Financial Problem
Actual Class	Financial Problem	10,88%	5,08%
	No Financial Problem	6,58%	77,46%

3.4 Optimal *ntree*

As can be seen in figure 4 and as shown in Khoshgoftaar et al. (2007) 100 trees are already sufficient to build a reliable random forests model. In these plots the estimated error rate (accuracy) is shown against the amount of trees in our random forests. Although in section 2.6.1 it was explained that accuracy is an undesirable way of measuring fitness, it is preferred for quick impressions to determine the optimal *ntrees* (Breiman et al., 2003).

Interesting though, and as shown in figure 5, more trees are needed when building a balanced random forests model than the basic random forests. This, however, is still not as many trees as Sun et al. (2014) obtained the optimal result with. For the under-sampling method even more trees are needed to obtain a balanced result as shown in figure 6.

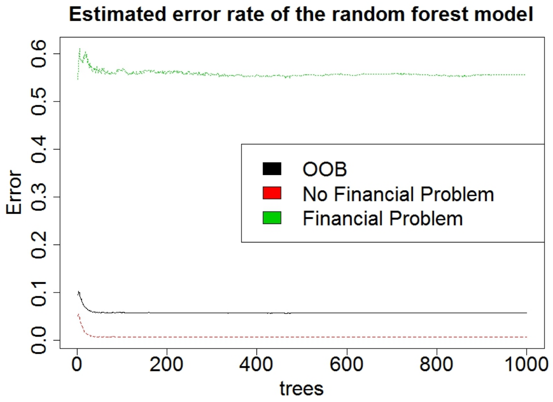

Figure 4: Accuracy indication of basic random forests model

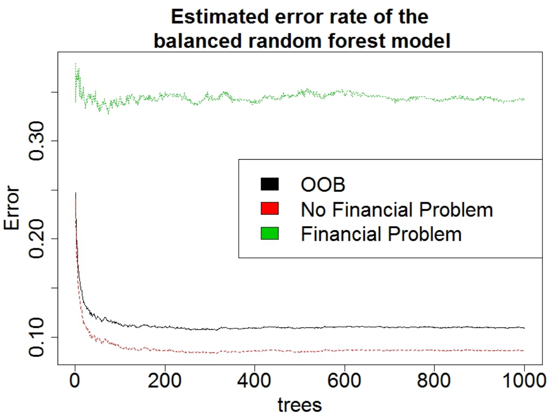

Figure 5: Accuracy indication of balanced random forests model

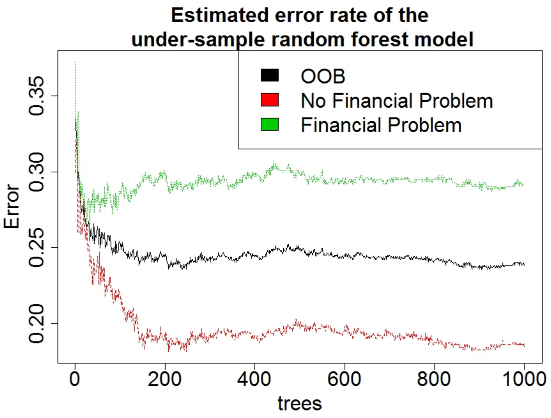

Figure 6: Accuracy indication of under-sample random forests model

3.5 Expected misclassification cost

As explained in section 2.6.2 the EMC value should be used when showing the fitness of the model, as West (2000) proposed. The results can be seen in table 6. For full model comparison the AUC and EMC values are plotted in figure 7.

Table 6: EMC values per model

	Random forests	Under sample	Balanced
EMC value	0,56179	0,30445	0,32703

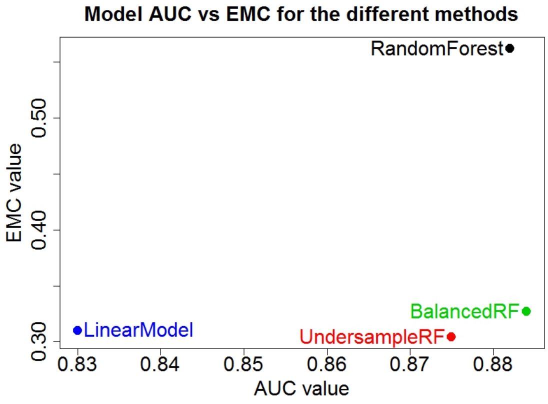

Figure 7: AUC vs EMC for the different methods

3.6 Linear model

Our linear model has an AUC of 0.83 and is indeed not as good a predictor as the random forests model. The optimal threshold, as found by finding the upper left value of the ROC curve, gives the confusion matrix as shown in table 7. Interestingly though does this confusion matrix correspond with an EMC value of 0.32. So although our model has a lower AUC value, the optimal linear model has an EMC value similar to the other models.

Table 7: Predictions linear model

		Predicted	
		Financial Problem	No Financial Problem
Actual Class	Financial Problem	12,18%	3,78%
	No Financial Problem	15,41%	68,63%

3.7 Feature importance

For all three models, the feature importances are quite different. The results can be seen in figures 8 to figures 10. Intersect is the β value, the slope is represented by the α.

They do have one constant similarity: the change in transactions. In each model, one of the main predictors is the Transaction α.

4 Conclusion and Discussion

The high AUC value for our model could not have been predicted for the little amount of variables

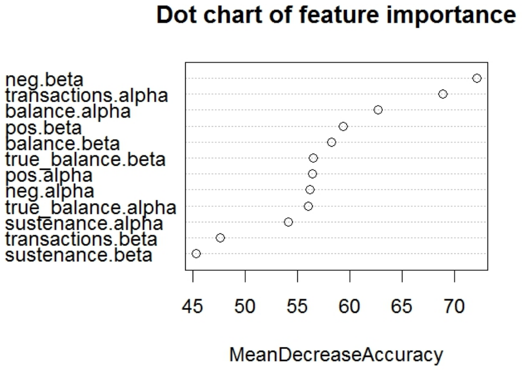

Figure 8: Feature importance for the basic random forests model

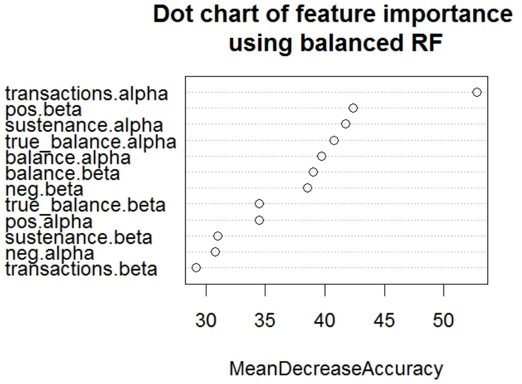

Figure 9: Feature importance for the balanced random forests model

used in the model and might be an indication of a flaw. One possibility could be that the linear function for the balance already predicts the final balance. This however is incorrect: the only data used was that before the last time in arrears (as explained in section 2.2) and the final balance was not included in the data. Thus, the linear balance model did not predict the final balance. Furthermore as can be seen in section 3.7 it is not one of the main features.

Unfortunately the data set only consisted of customers whose accounts were liquidated. For further analysis a sample dataset with both customers that were in arrears without liquidated accounts and customers with liquidated accounts should be taken.

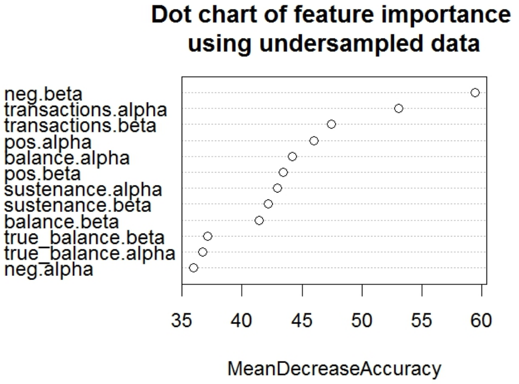

Figure 10: Feature importance for the underbalanced random forests model

Although our models have very similar AUC values, balanced and under-sampling random forests are our preferred final models. This is because they have the lowest EMC values. This immediately shows the importance of the EMC. Although all three models nearly have the same AUC values, their predictions are all quite different as also can be seen in the error plots: figure 4, 5, and 6.

4.1 Feature importance

From the results it can be concluded that a change in the amount of transactions is one of the most important indicators for a financial problem. This new valuable insight will now be used in the final model that is to be implemented in the arrears department.

The high differences between the feature importance for the different models, suggests a high correlation between the features. This is usually the case when each feature is easily replaced by another feature. This, however, is not the case as can be seen in the correlation plot of figure 11. What can be seen in the correlation plot is of course the correlation between positive/negative days and the α and β of the different feature predictors.

4.2 Research questions

The research questions attempted to answer in this thesis were:

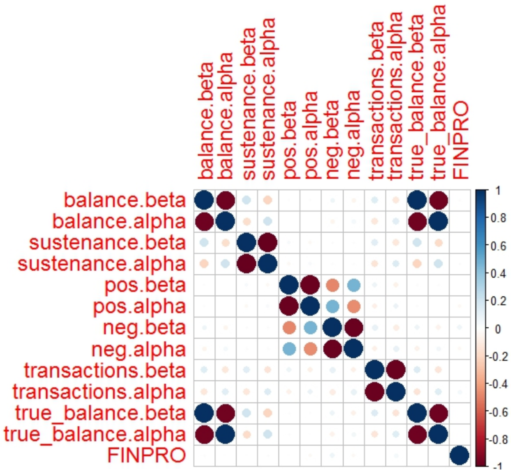

Figure 11: Variable correlations

- Is it possible to predict financial problems for arrears customers using the random forests algorithm?
- Can the random forests algorithm be improved to cope with the imbalanced data set?

Both questions can be answered affirmatively: Yes, financial problems can be very accurately predicted for arrears customers using random forests, and yes the different ways to cope with the imbalanced data set improve the results.

4.3 Future work

In this thesis the out-of-box confusion tables were used whereas it is possible to change the threshold which results in different confusion matrices. This can be done for instance by taking the upper left point on a ROC, or by minimizing the EMC value. A quick attempt showed however that it might not automatically improve the results since an EMC of 0.22 could be obtained on the training set using the balanced random forests but this resulted on an EMC of 0.67 on the test set. Hence there is room for future work.

References

Ricardo Barandela, Rosa M Valdovinos, J Salvador Sánchez, and Francesc J Ferri. The imbalanced training sample problem: Under or over sampling? In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 806–814. Springer, 2004.

Christine Bolton. *Logistic regression and its application in credit scoring.* PhD thesis, University of Pretoria, 2010.

L Breiman, A Cutler, A Liaw, and M Wiener. randomforest: Breimans random forest for classification and regression. *Fortran original by Leo Breiman and Adele Cutler and R port by Andy Liaw and Matthew Wiener. R package*, pages 4–3, 2003.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 10/01 2001. URL `http://dx.doi.org/10.1023/A%3A1010933404324`. J2: Machine Learning.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 2004.

S. C. Chen and M. Y. Huang. Constructing credit auditing and control & management model with data mining technique. *Expert Systems with Applications*, 38(5):5359–5365, 5 2011.

Jonathan N. Crook, David B. Edelman, and Lyn C. Thomas. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465, 12/16 2007.

María-Dolores Cubiles-De-La-Vega, Antonio Blanco-Oliver, Rafael Pino-Mejías, and Juan Lara-Rubio. Improving the management of microfinance institutions by using credit scoring models based on statistical learning techniques. *Expert Syst.Appl.*, 40(17):6910–6917, dec 2013. URL http://dx.doi.org/10.1016/j.eswa.2013.06.031.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

J. Galindo and P. Tamayo. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1-2):107–143, 04/01 2000. URL http://dx.doi.org/10.1023/A%3A1008699112516. J2: Computational Economics.

Sung Ho Ha. Behavioral assessment of recoverable credit of retailers customers. *Information Sciences*, 180(19):3703–3717, 10/1 2010.

David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942. ACM, 2007.

T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse. An empirical study of learning from imbalanced data using random forest. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 310–317, Oct 2007. ISBN 1082-3409.

Bart Lariviére and Dirk Van den Poel. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2):472–484, 8 2005.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):539–550, 2009. ID: 1.

John A. Nelder and RJ Baker. *Generalized linear models*. Wiley Online Library, 1972.

Esther Scheurmann and Chris Matthews. *Neural network classifers in arrears management*, pages 325–330. Artificial Neural Networks: Formal Models and Their Applications ICANN 2005. Springer, 2005.

Zhe Sun, M. A. Wiering, and Nicolai Petkov. Classification system for mortgage arrear management. In *IEEE Computational Intelligence for Financial Engineering and Economics*, 2014.

Lyn C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172, 2000.

Bhekisipho Twala. Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 37(4):3326–3336, 4 2010.

David West. Neural network credit scoring models. *Computers & Operations Research*, 27(11):1131–1152, 2000.

Yaya Xie, Xiu Li, E. W. T. Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Syst.Appl.*, 36(3):5445–5449, apr 2009. URL http://dx.doi.org/10.1016/j.eswa.2008.06.121.