Retail credit scoring using fine-grained payment data

Ellen Tobback[a] and David Martens[a]

[a]Department of Engineering Management, University of Antwerp

July 3, 2017

Abstract

Banks are continuously looking for novel ways to leverage their existing data assets. A major data source that has not yet been used to the full extent, is massive fine-grained payment data on the bank's customers. In this paper, a scalable and privacy-friendly design is proposed that builds predictive credit scoring models using a real-life data set of 183 million transactions made by 2.6 million customers. The data shows that amongst the clients that have transacted with at least one defaulter, 52% are defaulters themselves, while the overall default rate is 0.8%. We build upon this finding and create direct and implied networks from the payment data. The results show that our proposed design adds complementary predictive power to the current credit scoring models. Such improvement has a big impact on the overall working of the bank, from applicant scoring to minimum capital requirements.

1 Introduction

In this big data era, banks are looking for techniques and applications to leverage their existing data assets. Internally, banks have access to a broad range of customer data. Technological advancements such as mobile banking and contactless payments have substantially raised the number of registered transactions. As a result, next to socio-demographic data such as age, income and education, banks have data on purchasing and payment records which makes much of a person's behaviour visible. For example: the fact that a customer regularly transacts with other clients who have defaulted on a loan, combined with the fact that he makes regular payments at a casino and high-street shops can be telling of his default behaviour. Payment data has been used in the credit scoring literature to help banks better predict default and bankruptcy (Khandani et al. 2010, Bellotti and Crook 2013). Yet, this data source is not used to its full extent. Because payment data is too big or disorganized for traditional methods to handle, most studies have derived aggregated attributes from the fine-grained data, such as the average amount of transactions coming in and out, and thus discard important information. In this paper, we investigate how to

leverage this data source in its granular form and test both propositional and relational methods. We propose a scalable and privacy-friendly design that allows banks to include payment data in a non-aggregated manner. We test the results empirically using a data set from a European major commercial bank that contains 183 million checking account transactions made by 2.6 million clients holding a commercial loan.

This study is directly related to the recent credit turmoil that has shown the dangers of inaccurate credit risk modelling approaches. Focusing on payment data to enhance credit risk models has the advantage that it manages to combine interpretability with desired increased prediction performance. This comprehensibility aspect is a regulatory requirement, as a bank needs to be able to explain to a customer why credit has been denied (Martens et al. 2007).

Our study contributes to the credit scoring literature in several ways. (i) We provide empirical evidence that using fine-grained transaction data improves the accuracy of default predictions, (ii) we describe two methods (propositional and relational) and show that transaction data is best analysed in a relational manner, and (iii) building upon the empirical results, we offer advise to banks on their data collection and usage.

The outline of this paper is as follows. Section 2 reviews prior work on credit scoring and behavioural data. Section 3 describes the transformation of transaction data into default predictions. Next, Section 4 provides a detailed description of our experimental set-up and analyses the empirical results. The final section concludes the paper.

2 Credit scoring and behavioural data

There is a vast amount of research on credit scoring, covering statistical, operational research and machine learning methods. The first credit scoring models were created using discriminant (DA) (Durand et al. 1941) or regression analysis (Myers and Forgy 1963), however researchers quickly introduced logistic regression (Wiginton 1980), decision trees (Makowski 1985) and linear programming (Hardy and Adrian 1985) as alternative credit scoring methods. Since the 1990s the focus has shifted towards machine learning models such as Support Vector Machines (SVM) and neural networks. In a large benchmarking study, Baesens et al. (2003) show that neural networks and non-linear Least Square SVM report the highest performances for each data set considered in their study. However, the authors note that in terms of performance, logistic regression and DA are competitive with the non-linear classifiers.

In many countries, legislation requires financial institutions to explain why a certain credit was not granted. Applying non-linear, black box models decreases the comprehensibility of the credit scoring models. Even though methods exist that extract rules from black-box models (Martens et al. 2007), in a practical setting, credit scoring is still mainly based on simple classifiers such as logistic regression, DA and classification trees.

Most credit scoring research focuses on the modelling techniques and only to a lesser

extent on the input data. Traditionally, credit scoring models include socio-demographic data[1], data on the applicant's financial situation, employment and education data, and behavioural data (Van Gestel and Baesens 2009). Certain studies have included macroeconomic data to consider the market conditions at the time of application (Bellotti and Crook 2009, Bonfim 2009). Behavioural data describes the client's behaviour with regard to banking products. This can be credit card usage, transfer patterns on the transaction account or repayment behaviour on a different loan. Banks have behavioural data at their disposal if the applicant is an existing (credit) client. This data can be complemented with external data, e.g. from credit bureaus[2]. A number of studies have included behavioural data in their credit scoring models. Norden and Weber (2010) investigate the influence of credit line usage and the checking account balance on default risk of bank borrowers. They find that measures of account activity significantly enhance default predictions. Khandani et al. (2010) analyse patterns in consumer expenditures, savings and debt payments to predict credit card delinquencies. Bellotti and Crook (2013) use monthly account behavioural records to predict credit card defaults using dynamic models. However, all the above-mentioned studies transform the data to monthly aggregates, such as the transaction count per month, the monthly average account balance and the total inflow and outflow per month. In this paper, we employ behavioural data on checking account transactions in a fine-grained manner, using individual payments.

The use of behavioural data has proven to be successful in other domains as well, such as targeted advertising (Martens et al. 2016), fraud detection (Junqué de Fortuny et al. 2014) and customer retention (Verbeke et al. 2014). The nature of these large behavioural data sets requires a different modelling approach than the traditional, structured data sets. One option is to consider each action (i.e. payment) as a separate data entry and create a large matrix where each column is an interest, activity or entity. A different option is to create a network between nodes, with a node representing a client, where two nodes are linked through similar interests or activities: e.g. watching the same videos (Weber et al. 2013), visiting the same places (Provost et al. 2015) or liking the same pages on Facebook (De Cnudde et al. 2015). Behavioural data on checking account transactions has been used by Martens and Provost (Martens et al. 2016) to successfully target potential buyers of a financial product using a network structure, where two customers are linked if they have paid to the same entity. Within data mining we observe an increased use of social network data as input drivers for applications in marketing (Provost et al. 2009) and fraud detection (Hilas 2009). The main reason is the tremendous predictive power that is present in such relational data, with significant improvements compared to traditional

[1] In certain countries, legislation prohibits banks to discriminate based on certain socio-demographic information, such as age, gender, ethnic origin and religion. In the US, this is directly regulated by the Equal Credit Opportunity Act. In the EU this is indirectly regulated by Article 13 of the EC Treaty and translated into national legislation.

[2] In certain countries in continental Europe, there are no credit bureaus. Banks can collect information on existing credits from a national credit register.

approaches that only use individual customer data (Martens et al. 2016, De Cnudde et al. 2015). Network data can be seen more broadly than the typical social network data as data that defines any kind of relationship between entities. Two major categories of relational data can be distinguished: real network data and pseudo-network data. In a real network, two nodes are connected because a certain form of direct communication has taken place between them. In a pseudo-network, two nodes are connected because they have a common interest, activity or asset. The network is implied as there is no evidence that both nodes have ever communicated with each other. In this study, we build upon the proven success of network data and exploit the transaction data in a relational manner. Next to a direct network where consumers are linked if they made payments to each other, we create an implied or pseudo-social network where two consumers are linked if they made payments to the same entities.

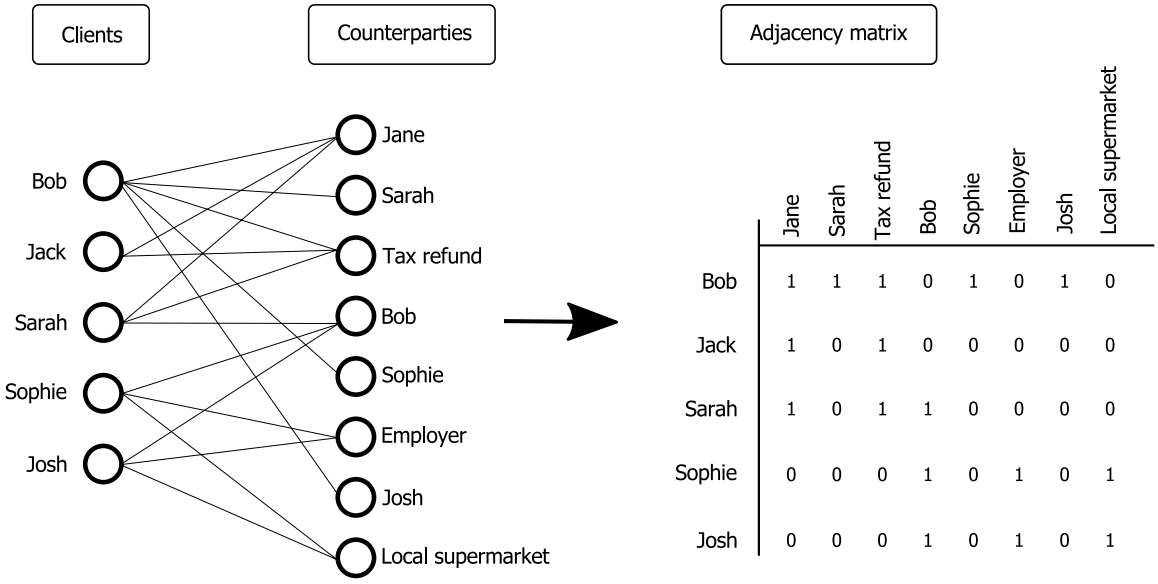
3 Transforming transactions into predictions

We investigate both propositional and relational models to use money transfer data of a client's transaction account. The propositional model follows the standard classification method and adds each unique account number that can be transacted with as a feature in the input space. This results in a large and sparse adjacency matrix B(m,n) that represents the behavioural data, with m the number of clients and n the number of counterparties (i.e. the unique set of account members that can be paid to). Each cell $c_{i,j}$ is a binary variable that denotes whether a transaction has taken place between client i and counterparty j. The matrix is created from the transaction log as illustrated by Figure 1. Bob, Jack, Sarah, Sophie and Josh all have an account and a commercial loan at the same bank. The graph on the left side represents their transactions over one month. The matrix on the right side is the adjacency matrix B(m,n).

The relational models use two types of network representations of the data: a direct network and an implied network. In the direct network, i.e. a unigraph with m nodes, two clients are linked if a transaction has taken place between them. In the implied network, which is a projection from a bipartite graph with m bottom nodes and n top nodes, two clients are linked if they have transferred money to or received money from the same entity. By creating both networks, we rely on the sociological concept of assortativity which states that people are more likely to form bonds with others who have similar characteristics such as values, beliefs and socio-economic status (McPherson et al. 2001). By creating a direct network, we build upon the theory of assortativity and assume that people of similar creditworthiness tend to cluster. The creation of the implied network is also justified by the assortativity concept, stating that similarity in one domain, such as transaction patterns, may indicate similarity in another domain, such as creditworthiness or willingness to redeem a loan.

Both networks are created from the transaction log as visualized in Figures 2 and 3. The

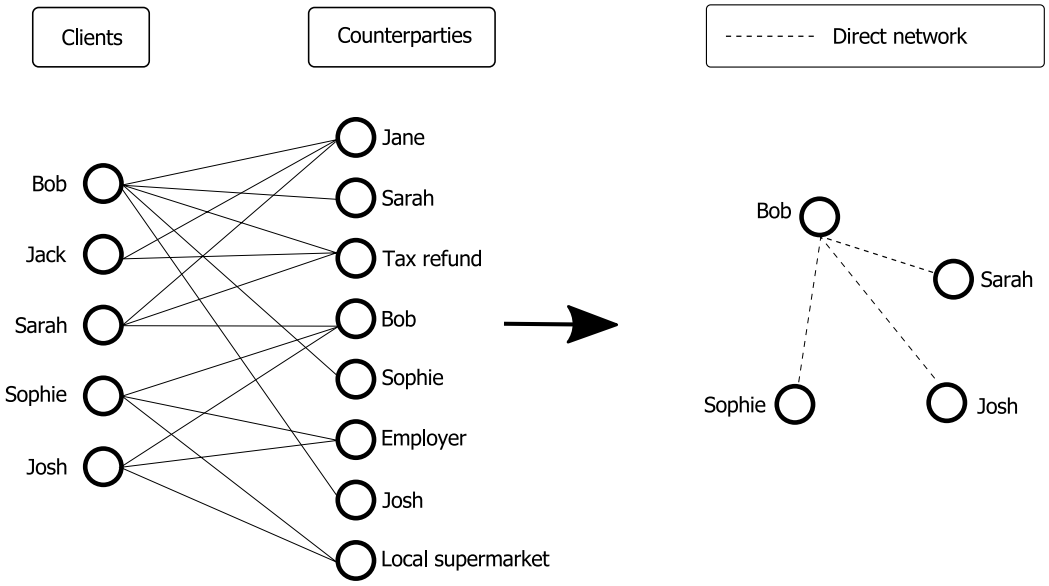
Figure 1: Matrix representation of the payment data: from the transaction log to an adjacency matrix.



graph on the right side of Figure 2 represents the direct network. It shows the connections between the clients that transacted with each other. These clients are both clients and counterparties in the transaction log. The debit transactions of Sarah, Sophie and Josh to Bob are listed as credit transactions on Bob's account. In the resulting network, Sarah, Sophie and Josh are directly connected to Bob.

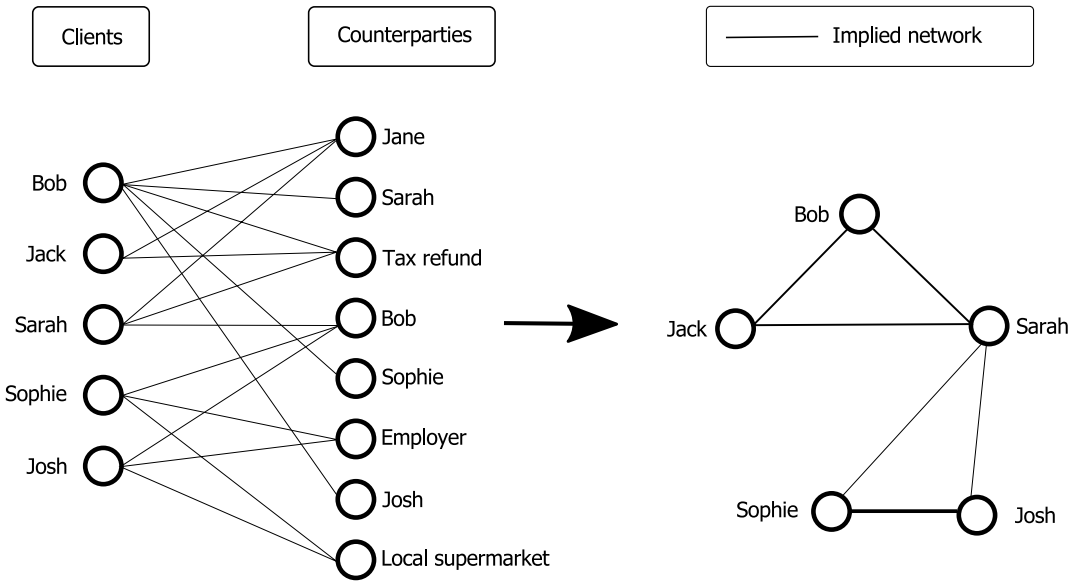
The implied network is a projection from the transaction log as illustrated by Figure 3. The network shows the connections between the clients that transacted with the same entity. Sarah, Sophie and Josh are connected in the implied network because they transferred money to the same account (i.e. Bob's account). The more entities they have in common, the stronger the connection. Sophie and Josh have a stronger connection than Josh and Sarah, because they have more transactions in common (i.e. Bob's account, their employer and the local supermarket). To create the implied network, we follow the three-step framework proposed by Stankova et al. (2015). In the first step, a weight is assigned to each counterparty according to the hyperbolic tangent of its inversed degree, with the degree equal to the number of clients that have made a payment to/received a payment from the respective counterparty. The hyperbolic tangent function downweighs entities that many clients have in common, as these are likely to be less distinctive for the target variable. Referring to the example of Figure 3, there will be more clients receiving a tax refund from the government tax agency (IRS or the country-equivalent) than clients paying their gro-

Figure 2: Direct network representation of the payment data: unigraph extracted from the transaction log.



ceries at a certain local supermarket. Hence, the supermarket should be assigned a larger weight than the tax agency as this implies higher similarity. Figure 4 shows the histogram of the number of clients per counterparty in our data set, accompanied by the corresponding weight. The Figure shows only the first part of the histogram. The actual maximum value of d, the number of clients per counterparty, is 1,388,977. Most counterparties have few transaction partners. However, there is a small number of counterparties that the majority of clients has transacted with. These are likely large companies or government organizations, such as energy and water suppliers and the tax agency. We obtained anonymised data, so no semantics could be extracted from the account numbers. The weights assigned to the counterparties (i.e. the top node weights S) are given by the black line, which is the hyperbolic tangent of the inverse degree d, the number of clients per counterparty. Clearly, the weighting scheme downweighs counterparties with many transaction partners more severely than entities that transacted with only a few clients. In the second step of the three-step framework, for each pair of linked clients, the edge weight is calculated by aggregating the weights of all shared counterparties. In the third step, classification is applied to the weighted network using a relational classifier. We further elaborate on the second and third step in Section 4.2.

Figure 3: Indirect (implied) network representation of the payment data: from a bipartite graph (transaction log) to a projected unigraph.



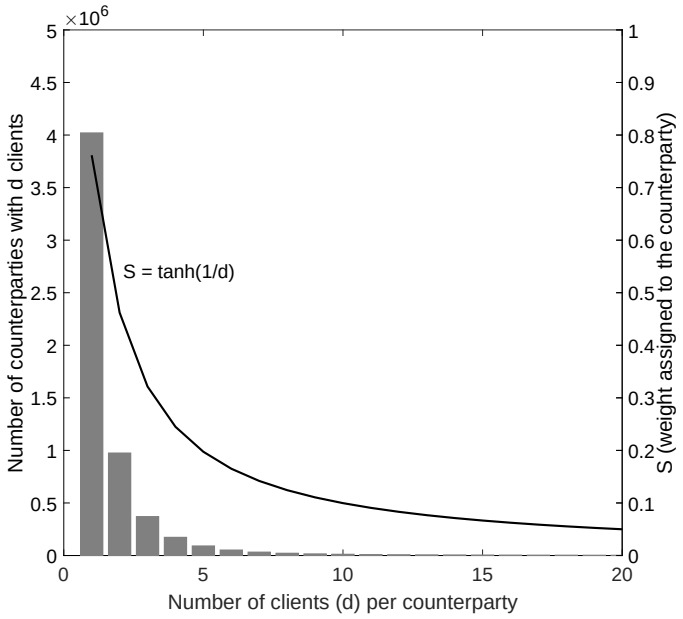
4 Experimental setup

4.1 Data

We received data from the transaction account, which includes an anonymised bank client indicator and a counterparty indicator for each transaction. We obtained 5 months of transaction data, containing over 180 million (debit and credit) transactions of 2.6 million bank accounts. Each bank account is linked to a consumer credit with a non-default status on 31 December 2014. The goal is to predict which loans will default in 2015. Transactions are marked as either point-of-sales (POS) transactions or transfers. The first category consists of payments made at a physical store with a debit card (credit card transactions are not included in the data), the latter category are electronic transfers from or to the client's account. Table 1 shows some relevant data characteristics. By adding more months, we increase the total number of transactions and counterparties. The number of bank accounts is kept stable and equal to those accounts that have made transactions on their account in December 2014.

Each bank account is accompanied by a rating score, determined by the bank's internal rating model that uses a 12-notch rating scale. This rating model is built using advanced modelling and includes socio-demographic and aggregated behaviour input variables. Due to confidentiality reasons, we are unable to describe the exact modelling procedure used. However, as the data is obtained from a large European bank, subject to regulatory oversight

Figure 4: Part of the histogram of the number of clients per counterparty (bars). The corresponding weight, the hyperbolic tangent of the inverse degree, is shown by the black line.



on its modelling, we can confidently state that the modelling procedure is in line with the state of the art modelling practices and the rating can therefore be considered as a good benchmark.

4.2 Study design

We estimate the performance of four models built using only transaction data: a propositional model, a direct network model, an implied network model and a linear ensemble model that combines the output scores of the direct and implied network. These performances are compared to the benchmark performance of the bank's own ratings. To test whether the payment data and the traditional data (represented by the ratings) are complementary, we create three additional linear ensemble models: one model that combines the output scores of the rating model with the scores of the direct network, one model that combines the scores of the rating model with the scores of the implied network and one model that combines the scores of the rating model with the scores of the direct and the scores of the implied network. The latter model is referred to as the 'full ensemble model'.

We use a ten-fold cross validation procedure, where 90% is used as training data and 10% as test data. The training data is further split in 80% to train and validate the

Table 1: Data characteristics

Number of months included	Number of clients	Number of counterparties	Number of transactions	Number of unique client-CP combinations
1 month (Dec)	2,585,227	4,141,402	42,865,861	28,278,074
2 months (+ Nov)	2,585,227	4,757,378	76,026,632	38,390,747
3 months (+ Oct)	2,585,227	5,254,154	114,043,910	48,401,003
4 months (+ Sep)	2,585,227	5,649,005	150,387,767	56,431,000
5 months (+ Aug)	2,585,227	5,945,217	182,645,116	63,042,608

classifiers using the transaction data and to train and validate the rating model, and 20% to train and validate the linear ensemble models.

To investigate the value of additional data, we start by incorporating only the most recent month (December 2014) and gradually increase the size of the data set by including the transactions further in the past. This allows us to see if it pays off to invest in the collection and storage of historic transactions.

Benchmark rating model The bank's 12-notch rating scale is used as a benchmark in this study. We could use the rating directly to test the performance, however, we decided to use the rating scales as input data in a simple linear prediction model using unary encoded dummies. In a later step, the output scores of the direct and implied network models can be added to this linear model, which allows us to precisely estimate the added predictive value of the payment data. This way we can preserve optimal flexibility as the weights of the rating dummies will be re-estimated on the training set for each new model, thereby allowing possible interactions between the individual ratings and the network variables. The rating scales are transformed into separate data entries using unary (thermometer) encoding in order to keep the ranking. Unknown ratings are replaced by the mode and are assigned a missing value flag. This results in 12 input variables (11 ratings and 1 missing value dummy).

As linear classifier, we apply a linear Support Vector Machine which solves the following optimization problem (Fan et al. 2008):

$$min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i max(1 - y_i\mathbf{w}^T\mathbf{x}_i, 0)^2 \tag{1}$$

With vector $\mathbf{w}$ the weights of the model and $\mathbf{x}_i$ and y_i representing the input vector and the label of the ith observation. $max(1-y_i\mathbf{w}^T\mathbf{x}_i, 0)^2$ is the squared hinge-loss function (L2). A ten-fold out-of-sample grid search is performed to find the optimal value of C, the regularization hyperparameter. We employ the LibLinear package from Fan et al. (2008) to run the SVM.

Propositional model The propositional model looks at the payment data from a standard classification perspective. Each counterparty in the adjacency matrix is an input feature of the propositional model. Depending on the number of months that are included in the data set, the number of variables in the model thus varies between 4 and 6 million. The model weights are calculated using linear SVM, using the same hyperparameter tuning as described above.

Direct and implied network models To create predictions for both the direct and implied networks, a relational learner is used. This learner is applied to the unigraph of the direct model and to the projected unigraph of the implied network. As a relational learner, we apply the weighted-vote Relational Neighbour (wvRN) classifier (Macskassy and Provost 2007). It is a simple, yet powerful classifier that uses the network structure to calculate a default probability score $P(L_i = c|N(i))$ for a company as a weighted average of its j neighbours' $(N(i))$ probability scores (see Equation 2). The classifier is based on the property of assortativity (McPherson et al. 2001), as it makes the assumption that the connected nodes are similar and therefore more likely to belong to the same class. We apply a smoothed version of wvRN that adds the default rate μ_c as smoothing factor.

$$\begin{aligned} P(L_i = c|N(i)) &= \frac{\sum_{j\in N(i)} w_{ij} P(L_j = c|N(j)) + 2\mu_c}{Z+2} \\ &\text{where the normalization factor Z is equal to } \sum_{j\in N(i)} w_{ij} \end{aligned} \tag{2}$$

Equation 2 calculates the probability that the label L of client i equals c, with c a binary indicator of default, given its neighbours $N(i)$ in the unigraph (projection). The resulting default probability score is the weighted sum of the default probabilities of a client's neighbours. In this study, the neighbour's default probability is set to either 0 or 1, depending on whether they defaulted or not. Referring back to our running example in Figure 3, we can find the default probability of Sophie using her implied network. Knowing that Sarah defaulted on her loan and Josh didn't, Sophie's default probability equals:

$$P(\text{Sophie defaults}|\text{Josh and Sarah}) = \frac{w_{\text{Sophie,Josh}} \times 0 + w_{\text{Sophie,Sarah}} \times 1 + 2\mu_c}{Z+2}$$

When estimating default probabilities, the traditional, unsmoothed, wvRN will assign boundary values to nodes with only one neighbour, i.e. one or zero depending on whether the neighbour has defaulted or not. Similarly, the method will assign boundary values when the node is surrounded by neighbours of only one type and zero when the node has no neighbours in the network. However, a client connected to no-one or to non-defaulted clients only still has a certain probability of default. To solve these problems, we calculate a smoothed version of the probability estimate using the concept of additive smoothing. Traditional additive smoothing starts from the prior assumption of equal probabilities for each class. This assumption is not valid for our credit scoring data set, therefore we replace the uniform probability of 0.5 by the default rate μ_c of the training set as prior. As a result, when using a smoothed wvRN, a client with no neighbours will receive the default rate μ_c.

The edge weight w_{ij} between client i and its neighbour j is different for the implied and direct network. The edge weights in the implied network are defined by Equation 3 and equal the sum of the top node weights S_k of all shared top nodes N_T in the bipartite graph.

$$w_{ij} = \sum_{k \in N_T(i) \cap N_T(j)} S_k \tag{3}$$

The top node weight S_k of node k is equal to the hyperbolic tangent of its degree d_k:

$$S_k = \tanh(\frac{1}{d_k}) \tag{4}$$

When calculating the weights of the top nodes, the test bottom nodes are not included. In our running example, this implies that the edge weight between Sophie and Josh is equal to $\tanh(\frac{1}{1}) + \tanh(\frac{1}{1}) = 1.5232$ and the weight between Sophie and Sarah to $\tanh(\frac{1}{2}) = 0.4621$.

For the direct network two different weighting schemes are considered. In the first scheme, the edge weight w_{ij} is a variable that denotes the number of months from the data set in which at least one transaction between both parties has taken place. When only one month of data (i.e. December) is considered, the edge weight is a binary variable. In the second scheme, the edge weight w_{ij} equals the number of transactions between both parties i and j.

Ensemble models As mentioned before, for each fold the training set is split into 80% to train the classifiers (training set 1) and 20% to train the ensemble models (training set 2). The direct and implied networks are built on the first training set and are used to estimate default probabilities for the clients in training set 2. These probability scores are then used as input features for the ensemble models, alongside the unary encoded rating dummies for those ensemble models that include the ratings. The ensemble models linearly combine the different variables with the weights estimated by a linear SVM.

4.3 Results

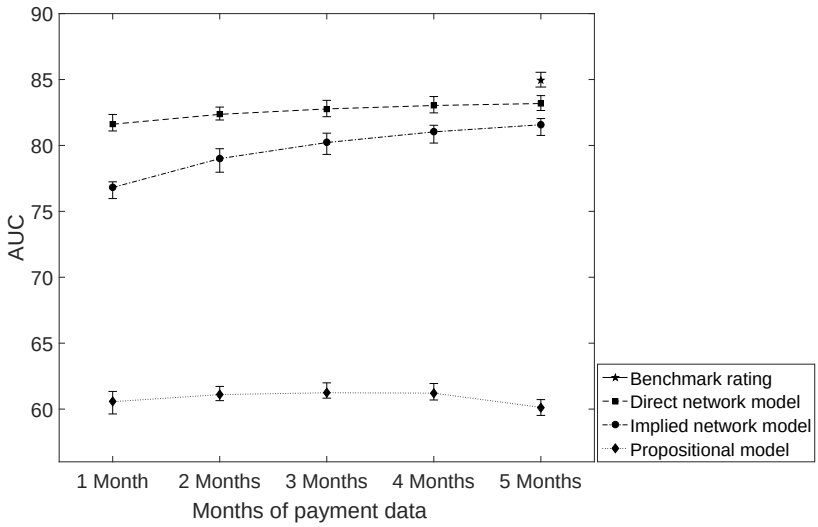
We compare the results of the four payment data models with the benchmark rating model and the rating ensemble models using the Area under the ROC-curve (AUC) characteristic Fawcett (2006) and the lift Berry and Linoff (2004) at 1% and 5% of the test set, averaged over the 10 folds.

Figure 5 plots the AUC performances for all models. For the sake of clarity, the results are spread over two graphs. Figure 5a plots the performances of the propositional model, the direct network model, the implied network model and the benchmark rating model. Figure 5b also plots the benchmark rating model (to facilitate comparison) and the four ensemble models. The direct network is created using the first weighting scheme, i.e. the number of months from the data set in which at least one transaction has taken place. The network models report high performances, however, they are still outperformed by the bank's own rating models. One exception is the ensemble model that combines the direct and implied network scores. The results show that a direct network has more predictive power than an implied network, indicating that your direct transaction circle is likely composed of people with similar creditworthiness. Figure 6 illustrates a set of default clusters (i.e. small networks that consist mainly out of defaulted clients) that are part of the direct network. The entire network is a collection of similar small default and non-default clusters. It motivates the intuition behind the relational model: if you are connected to numerous defaulters, you are likely also a defaulter. This intuition is also confirmed by Figure 7 which represents a client's default probability for increasing minima of defaulters (absolute or proportional) in its network. Remarkably, amongst the clients that are connected to at least 1 defaulter, 51.98% are defaulters themselves, compared to 0.77% in the complete test set.

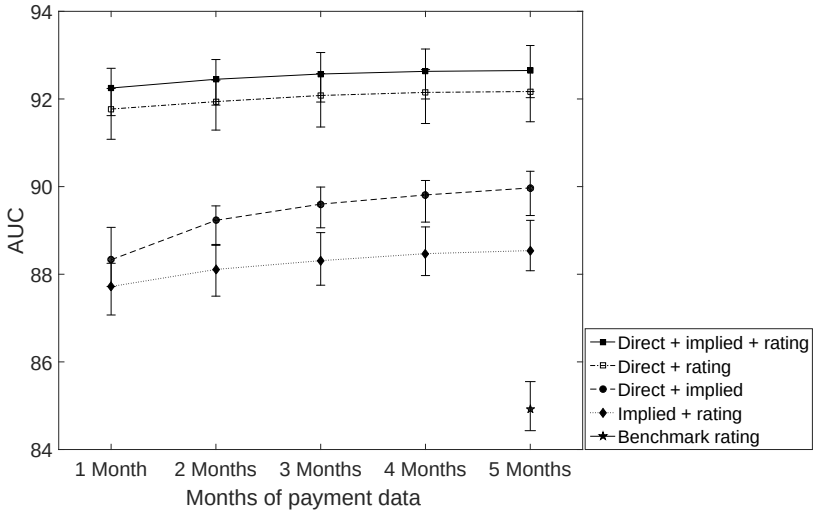
Best performances are reported for the ensemble models. Remarkably, the ensemble model that considers only payment data, i.e. the 'direct + implied' model, performs better than the ensemble model that combines the implied network with the ratings. The highest AUC values are found for the full ensemble model, closely followed by the model that combines the ratings with the scores of the direct network. The results show that traditional data and payment data have complementary predictive power in terms of AUC. The similarity-based network seems to add information beyond that already contained by the direct transactions network. Applying a propositional technique is clearly suboptimal for the fine-grained transaction data used in this study. Our results indicate that this data type should be exploited in a relational manner.

Figures 8 and 9 show the lifts at 1% and 5% percent averaged over the 10 folds. The highest lifts are reported for the full ensemble model, the 'direct + implied'-model and the 'direct + rating'-model, with a comparatively large gap in lift with the remaining five models. While the rating model scores better than the direct network model in terms of AUC, it performs worse in terms of the lift at the threshold of 1% : the direct network model reports a 115% $(= (\frac{39.6}{18.38} - 1) \times 100\%)$ higher average lift than the rating model when all five

Figure 5: Results in terms of out-of-sample AUC for the benchmarking and network models (a) and the ensemble models (b).



(a)



(b)

Figure 6: Graph representation of a sample of the direct network. Black nodes are defaulted clients and white nodes are non-defaulted clients.

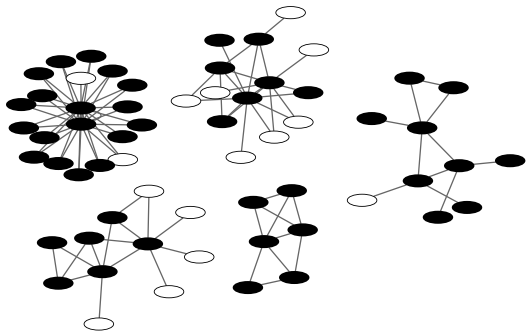
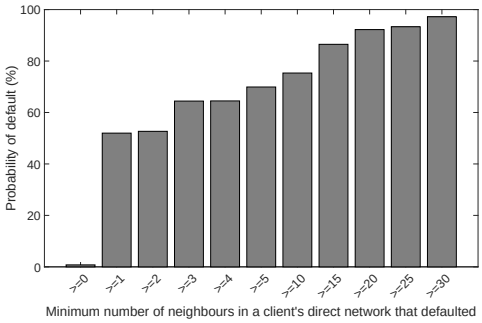


Figure 7: Probability of default for clients with increasing number and percentage of defaulted neighbours in their direct network.

(a) Number of neighbours



(b) Percentage of neighbours

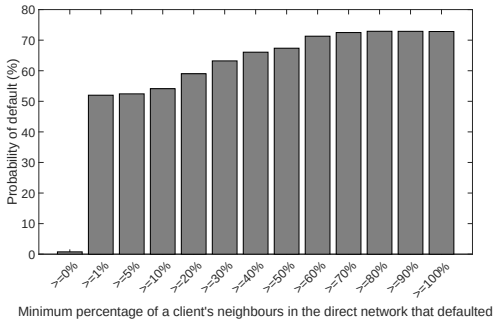
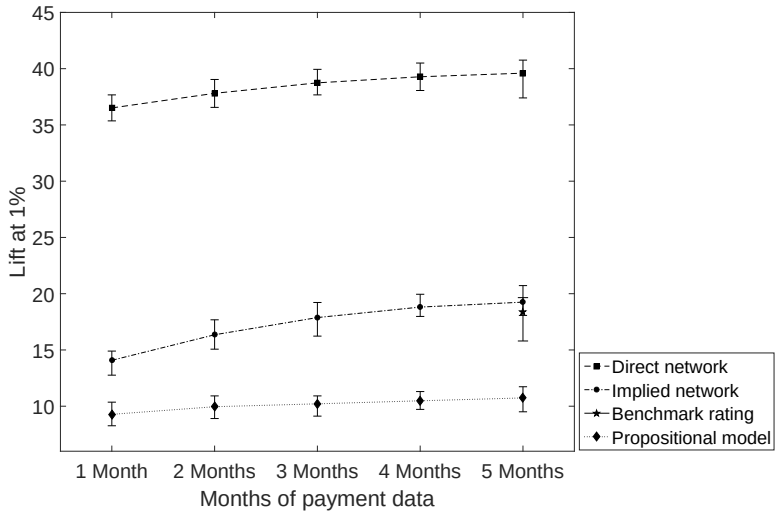
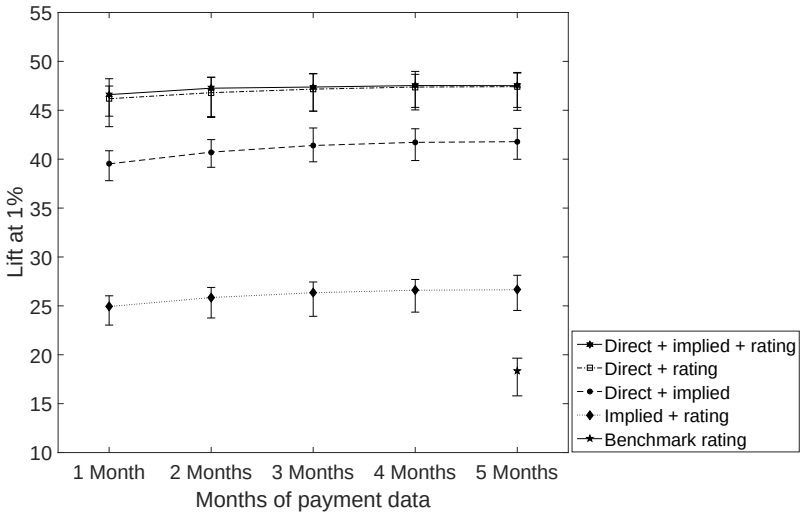


Figure 8: Results in terms of out-of-sample lift at 1 percent for the benchmark and network models (a) and the ensemble models (b).

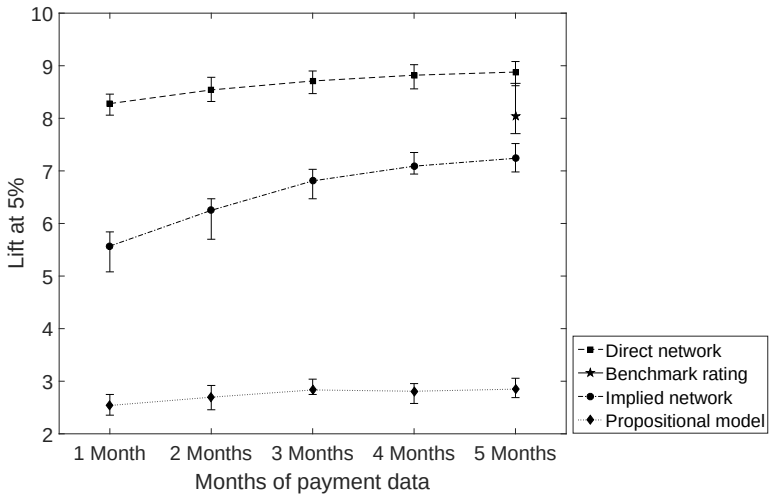


(a)

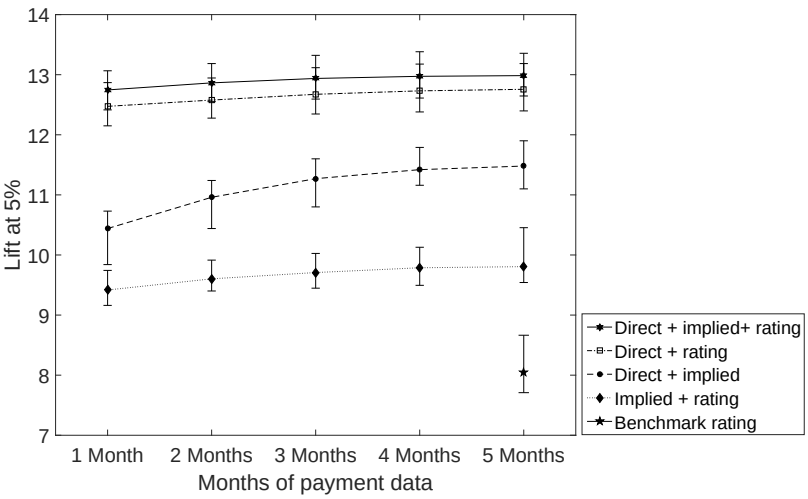


(b)

Figure 9: Results in terms of out-of-sample lift at 5 percent for the benchmark and network models (a) and the ensemble models (b).



(a)



(b)

months of transactions are included. However, this advantage over the rating model levels off at the 5% threshold, where the direct network model has only a 10% ($= (\frac{8.88}{8.05} - 1) \times 100\%$) higher lift than the rating model. In terms of lift it appears that adding the implied network score to the 'direct+rating'-model does not lead to higher performance: when using five months of data, the lift of the full ensemble model and the 'direct+rating' model overlap. The results are in line with other studies that use relational learners on fine-grained data: network data gives a boost to the model lift (Martens et al. 2016). In practical terms, this means that amongst the highest scores of the models that include payment data in a direct network there are more actual defaulters than amongst the highest scores of the traditional rating model. This 'boost' can also be seen in the ROC-curves in Figure 10. The direct model, the full ensemble model and the 'direct + implied'-model all have a cut-off at which the model detects more than 40% of the defaulters with almost zero misclassifications. The direct network model is surpassed by the rating model for lower cut-off values. Investigating the true positive (TP) and false positive (FP) rates of the direct network model's ROC-curve for the different cut-offs, shows that there is a sudden jump in the FP rate at a cut-off of 0.009. This score is the default rate of the training set and is assigned by the relational classifier to the bank's clients that have no known transactions with other clients of the bank: these are nodes with no links in the direct network, confirming the importance of more information. This finding shows that credit scoring and marketing can go hand in hand: (i) encouraging clients to increase their use of the bank's checking account will result in more information on a client's direct transaction network, and (ii) investing in positive worth-of-mouth marketing can lead to more links in a client's direct network if the people in its environment open an account at the bank.

Figures 5, 8 and 9 operate as learning curves (Perlich et al. 2003). In the first step only the most recent transactions (December 2014) are considered and at each step older transactions are included. Overall, we see that the performance increases with each extra month of transactions that is included in the features space, with the largest increase occurring in the beginning between 1 and 2 months. The model that seems to benefit the most from additional data is the implied network. We find the same conclusions as previous studies (Junqué de Fortuny et al. 2013) have found: when working with fine-grained data, bigger is better. However, the performance improvement caused by adding more transactions levels off after a certain point (Martens et al. 2016).

Regardless of the number of months included in the data set, the full ensemble model performs better than the other models. This is confirmed by the DeLong test of difference in AUC for correlated ROC-curves. The results of these comparison tests are reported in Table 2. The diagonal elements show the results for the model of the respective category. The rest of the matrix indicates the results of the different combinations of the corresponding data categories, i.e. the ensemble models. The full ensemble model, that uses all categories, is shown in the last row. For all cases, we find that the full ensemble model has a significantly higher AUC than the other models, with all p-values lower than 0.0001.

For the direct network, we consider two weighting schemes. The networks in Figures 5, 8

Figure 10: Receiver Operating Curve of one fold of out-of-sample predictions.

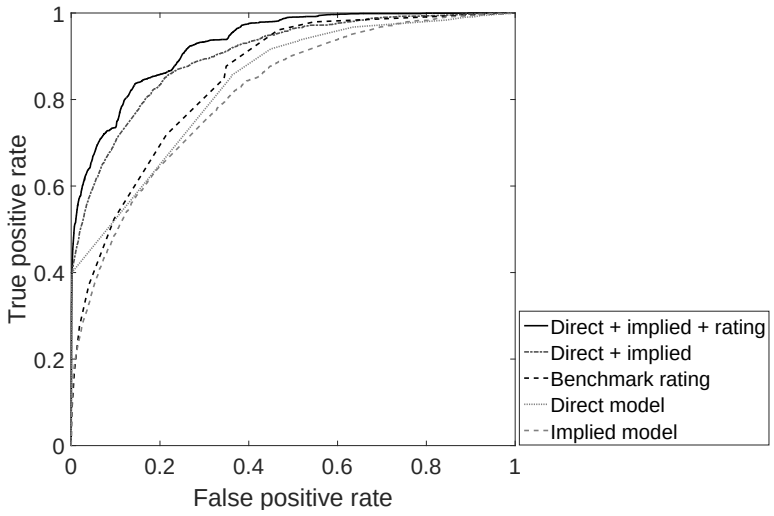
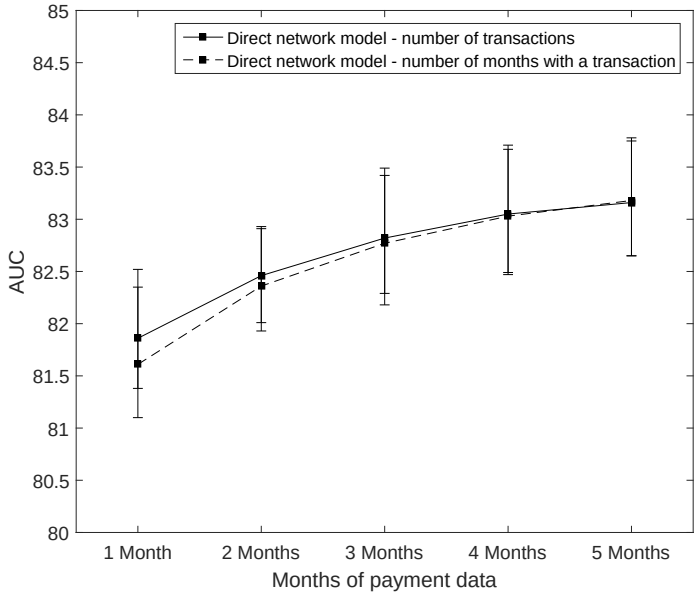


Table 2: Results in term of p-value of the DeLong test for the different models. The test compares the AUC performance of each model with the best performing model (the full ensemble model).

	Rating	Implied	Direct	Full ensemble
Rating	$<.0001$	$<.0001$	$<.0001$	-
Implied	$<.0001$	$<.0001$	$<.0001$	-
Direct	$<.0001$	$<.0001$	$<.0001$	-
Full ensemble	-	-	-	1.00

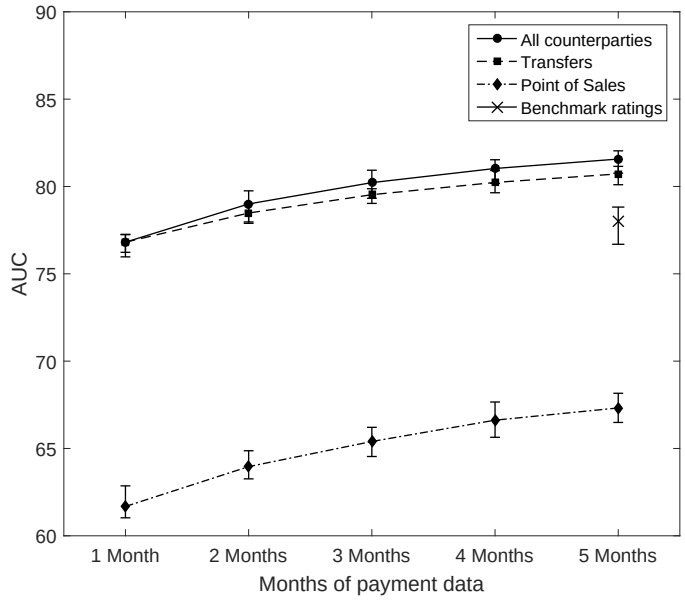
Figure 11: Results in AUC of the direct network model with two different weighting schemes.



and 9 apply the first type of edge weighting. The second type assigns the total number of transactions between both parties as weight to the edge. Figure 11 compares the AUC performances of the resulting networks using the two weighting schemes. The difference in performance between both networks is limited and levels off almost completely when all 5 months of transaction data are used. After the addition of the fourth month of transactions, the difference becomes insignificant (p-values > 0.59) as tested by the DeLong test. For banks it is thus sufficient to save only the unique transactions per month. The predictive value lies in the fact that a payment to a certain counterparty has been made, not in the amount or frequency.

As mentioned before, there are two types of counterparties in the data set: Point-of-sales and transfers. Figure 12 compares the AUC-results of the implied network when all counterparties are included with the network when only transfers or only POS are included. The results illustrate that most predictive power is included in the transfer transactions. The implied network created out of POS-transactions has limited predictive power and performs worse than the bank's own rating model. However, POS-transactions still add some complementary information to the transfer network, as the highest performance is found for the network created using all counterparties.

Figure 12: Results in AUC of the implied network for different counterparties.



4.4 Deployment

Introducing big data analytics in credit scoring may require a reallocation of banking resources. Traditional banks are often held back by legacy systems that are not adjusted to the task at hand (Capgemini, LinkedIn and Efma 2017). Simultaneously, banks can be reluctant to use external data sources for fear of defying customer's trust. This paper offers a big data application for banks that does not require large IT infrastructures and that uses internal data only. The data, in the form of transaction logs, are already available at banks. As the results show, with few months of transaction data, high accuracies can easily be obtained. The network scores can be integrated as a variable into the existing credit scoring models and can thus provide additional information without disrupting the entire credit scoring system. The smoothed wvRN classifier is a straightforward method with low computational time. On the largest data set, it took 48.96 seconds to run 10 folds for the implied network using Matlab and only 1.65 seconds for the direct network[3]. We should note that the data preparation process (extraction of the transaction log and transformation of the log to an adjacency matrix) is a time-intensive process. However, once extracted, the data can be utilized for purposes other than credit scoring as well, including targeted advertising and churn prediction.

An important issue to consider when using sensitive payment data is privacy. The design

[3]On an Intel Core i5-3470 CPU @ 3.20 GHz machine with 8Gb RAM.

we propose is privacy-friendly and is an example of privacy-by-design: privacy is embedded in the entire process. All data can be encrypted and only the encryption of the client IDs should be reversible. This decryption can be executed in a separate, protected environment that cannot be accessed by the modellers. We do not use the counterparties' semantics such as address, type of person, type of shop, thereby allowing full encryption. The results show that even the distinction between POS and transfer isn't necessary, as best performances are found when both types are included. At no point in the modelling process does the design require the modellers to look at the client's name or unencrypted payment data. The counterparties' IDs can thus be irreversibly hashed. However, irreversible hashing might be in conflict with the required interpretability of the model. Nevertheless, a high amount of privacy can still be maintained if the interpretation of the classifications is executed in a separate, secured IT environment The design does not require purchasing third-party information and is therefore less likely to face privacy and regulatory compliance issues, given that banks are transparent about the data used to build credit scores.

5 Conclusion

This paper investigates the use of transaction data for credit scoring. We examined propositional and relational methods to classify customers and find that transaction data should be modelled in a relational manner. We show that payment data adds complementary predictive power to the traditional credit scores. Best results are found when the default probability scores of a direct network (linking clients that transacted with each other) are combined with the scores of an implied network (linking clients if they transacted with the same entities) and the bank's own ratings. We find that electronic transfers are more predictive than point-of-sales transactions, though the model still benefits from the inclusion of both transaction types. Adding more information to the data set by including transactions further in the past increases the models' accuracies, though this increase appears to level off when all five months of transaction data are included.
In this study, we provide a big data application for credit risk assessment. The results confirm the large predictive value of behavioural data in credit scoring. The proposed design is easy to implement by financial institutions as it uses internal data and does not require a disruption of the existing IT infrastructure. Once the networks are created, they can be applied within the bank for different purposes other than credit scoring, such as churn prediction, fraud detection and targeted marketing. The design can also be extended to other credit scoring applications, including credit card default using credit card transactions and corporate default using corporate transactions.

References

Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research*

Society 54(6):627–635.

Bellotti T, Crook J (2009) Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society* 60(12):1699–1707.

Bellotti T, Crook J (2013) Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting* 29(4):563–574.

Berry MJ, Linoff GS (2004) *Data mining techniques: for marketing, sales, and customer relationship management* (John Wiley & Sons).

Bonfim D (2009) Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking & Finance* 33(2):281–299.

Capgemini, LinkedIn and Efma (2017) World fintech report 2017. Technical report, Capgemini, Linkedin and Efma, https://www.marsdd.com/wp-content/uploads/2015/02/CapGemini-World-FinTech-Report-2017.pdf.

De Cnudde S, Moeyersoms J, Stankova M, Tobback E, Javaly V, Martens D, et al. (2015) Who cares about your facebook friends? credit scoring for microfinance. Technical report, Univeristy of Antwerp Working Paper.

Durand D, et al. (1941) Risk elements in consumer instalment financing. *NBER Books* .

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9:1871–1874.

Fawcett T (2006) An introduction to roc analysis. *Pattern Recognition Letters* 27(8):861–874.

Hardy WE, Adrian JL (1985) A linear programming alternative to discriminant analysis in credit scoring. *Agribusiness* 1(4):285–292.

Hilas CS (2009) Designing an expert system for fraud detection in private telecommunications networks. *Expert Systems with applications* 36(9):11559–11569.

Junqué de Fortuny E, Martens D, Provost F (2013) Predictive modeling with big data: is bigger really better? *Big Data* 1(4):215–226.

Junqué de Fortuny E, Stankova M, Moeyersoms J, Minnaert B, Provost F, Martens D (2014) Corporate residence fraud detection. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1650–1659 (ACM).

Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34(11):2767–2787.

Macskassy SA, Provost F (2007) Classification in networked data: A toolkit and a univariate case study. *The Journal of Machine Learning Research* 8:935–983.

Makowski P (1985) Credit scoring branches out. *Credit World* 75(1):30–37.

Martens D, Baesens B, Van Gestel T, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research* 183(3):1466–1476.

Martens D, Provost F, Clark J, Junqué de Fortuny E (2016) Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly* 40(4):869–888.

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 415–444.

Myers JH, Forgy EW (1963) The development of numerical credit evaluation systems. *Journal of the American Statistical association* 58(303):799–806.

Norden L, Weber M (2010) Credit line usage, checking account activity, and default risk of bank borrowers. *Review of Financial Studies* 23(10):3665–3699.

Perlich C, Provost F, Simonoff JS (2003) Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research* 4(Jun):211–255.

Provost F, Dalessandro B, Hook R, Zhang X, Murray A (2009) Audience selection for on-line brand advertising: privacy-friendly social network targeting. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 707–716 (ACM).

Provost F, Martens D, Murray A (2015) Finding similar mobile consumers with a privacy-friendly geo-social design. *Information Systems Research* In Press.

Stankova M, Martens D, Provost F (2015) Classification over bipartite graphs through projection. Technical report, University of Antwerp Working Paper.

Van Gestel T, Baesens B (2009) *Credit Risk Management: Basic concepts: Financial risk components, Rating analysis, models, economic and regulatory capital* (Oxford University Press).

Verbeke W, Martens D, Baesens B (2014) Social network analysis for customer churn prediction. *Applied Soft Computing* 14, Part C(0):431 – 446, ISSN 1568-4946, URL `http://dx.doi.org/http://dx.doi.org/10.1016/j.asoc.2013.09.017`.

Weber I, Garimella VRK, Borra E (2013) Inferring audience partisanship for youtube videos. *Proceedings of the 22nd international conference on World Wide Web companion*, 43–44 (International World Wide Web Conferences Steering Committee).

Wiginton JC (1980) A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis* 15(03):757–770.