Retail credit scoring using fine-grained payment data

Retail credit scoring using fine-grained payment data

Ellen Tobback[a] and David Martens[a]

[a]Department of Engineering Management, University of Antwerp

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 design is proposed that builds predictive credit scoring models using the fine-grained payment data. Using a real-life data set of 183 million transactions made by 2.6 million customers, we show that the scalable implementation that is put forward leads to a significant improvement in AUC, with only seconds of computation needed. When investigating the 1% riskiest customers, twice as many defaulters are detected when using the payment data. Such an 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, like any other large company, are looking for ways to leverage their existing data assets. Internally, banks have access to a broad base 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 sociodemographic data such as age, income and education, banks have data on the purchasing and payment records which makes much of a person's behaviour visible. 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 while only rarely receiving money from an employer can be telling for his default behaviour. Payment data has been used in the credit scoring literature to help banks better predict default and bankruptcy. 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 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[1]. 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.

The recent credit turmoil 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 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 [18].

Our study contributes to the credit scoring literature in several ways. 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 our initial findings, we offer advice to the banks on their data collection and analyses to perform.

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 artificial intelligence methods. The first credit scoring models were created using discriminant (DA) [7] or regression analysis [22], however researchers quickly introduced logistic regression [32], decision trees [17] and linear programming [11] as alternative credit scoring methods. Since the 1990's the research focus has shifted towards techniques such as Support Vector Machines (SVM) and Artificial Neural Networks (ANNs). In a large bench-

[1]Whether the bank is allowed to explicitly use this payment data in such a manner in the context of the European General Data Protection Regulation (GDPR) [24] is part of a larger discussion, which is why we propose the privacy-friendly design as discussed in Section 4.4

marking study, Baesens et al. [1] show that neural networks and non-linear Least Square SVM report the highest performance 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 credit scoring. Even though methods exist that extract rules from black-box models [18], 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[2], data on the applicant's financial situation, employment and education data, and behavioural data. Certain studies have included macroeconomic data to consider the market conditions at the time of application [2, 5]. Following the definition by Shmueli [28], big behavioral data captures human behavior through the actions and/or interactions of people. These form a record of a person's behavior captured as fine-grained features. In this setting, behavioural data can be the fine-grained 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[3]. This data can be complemented with external data, e.g. from credit bureaus.[4] A number of studies have included behavioural data in their credit scoring models. Norden and Weber [23] investigated 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. [15] analyse patterns in consumer expenditures, savings and debt payments to predict credit card delinquencies. Bellotti and Crook [3] use monthly account behavioural records to predict credit card defaults using dynamic models.

[2]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.

[3]In Europe, a new European PSD2 directive has come into effect in January 2018 [8], which encourages the mobility of consumers' payment data. This implies that payment data becomes available from other banks, at the request of the customer. The implications in this setting are elaborated on in Section 4.4.

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

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 use behavioural data on checking account transactions in a fine-grained manner, using the individual payments.

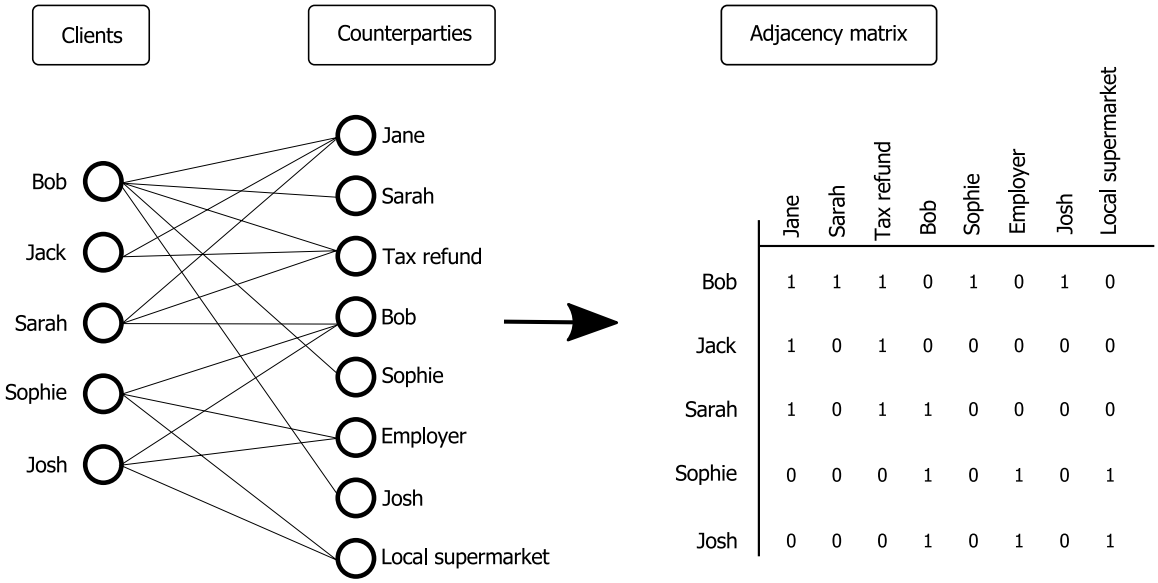
The use of behavioural data has proven to be successful in other domains such as as targeted advertising [20], fraud detection [14] and customer retention [30]. The nature of these large behavioural data sets requires a different modelling approach than the traditional, structured data sets. One option is to consider each action 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 all nodes where two nodes are linked through similar interests or activities: e.g. watching the same videos [31], visiting the same places [27] or liking the same pages on Facebook. Behavioural data on checking account transactions has been used by Martens and Provost [20] 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 [26] and fraud detection [12]. The main reason is the tremendous predictive power that is present in models built on such relational data, with significant improvements compared to traditional approaches that only use individual customer data. Network data can be seen more broadly than the typical friendship relationship among persons as data that defines relationships 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 transaction 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 $B(m, n)$.

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

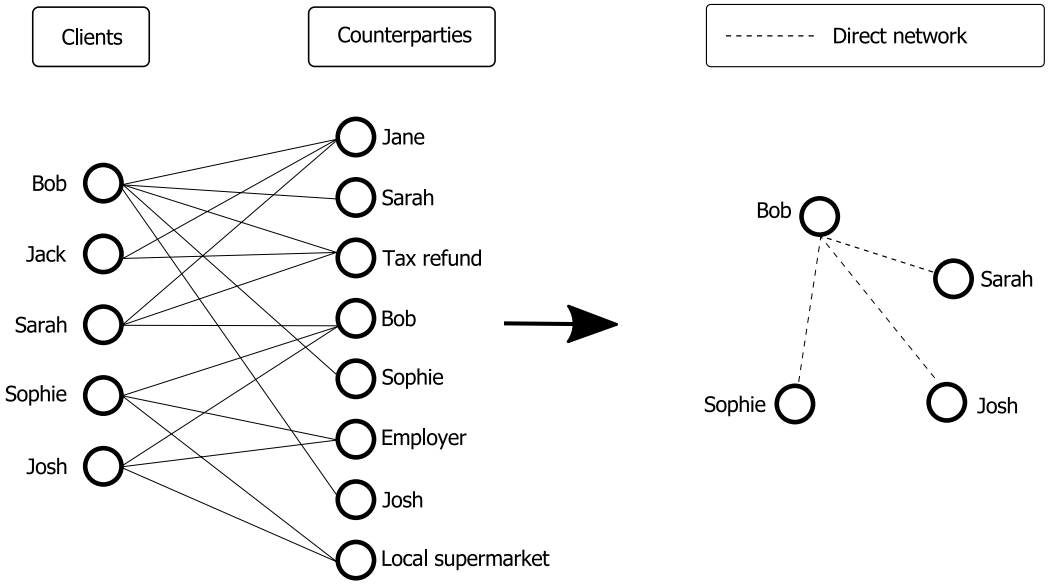


The relational models use two types of network representations of the data: a direct network and an implied network. In the direct network, i.e. a unigraph with m nodes, two clients are linked if a transaction has taken place between them. In the implied network, which is a projection from a bipartite 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, socio-economic status [21]. By creating a direct network, we build upon the theory of assortativity and assume that people of similar creditworthiness tend to cluster [20].

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

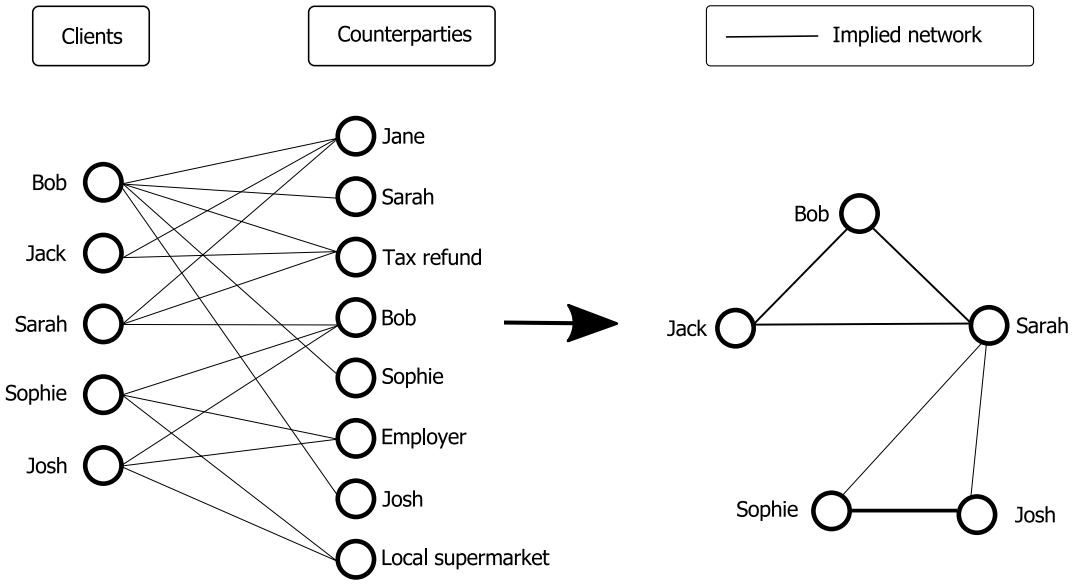
Figure 2: Direct network representation of the payment data: unigraph extracted from the transaction log.



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 and their employer).

To create the implied network, we follow the three-step framework proposed by Stankova

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

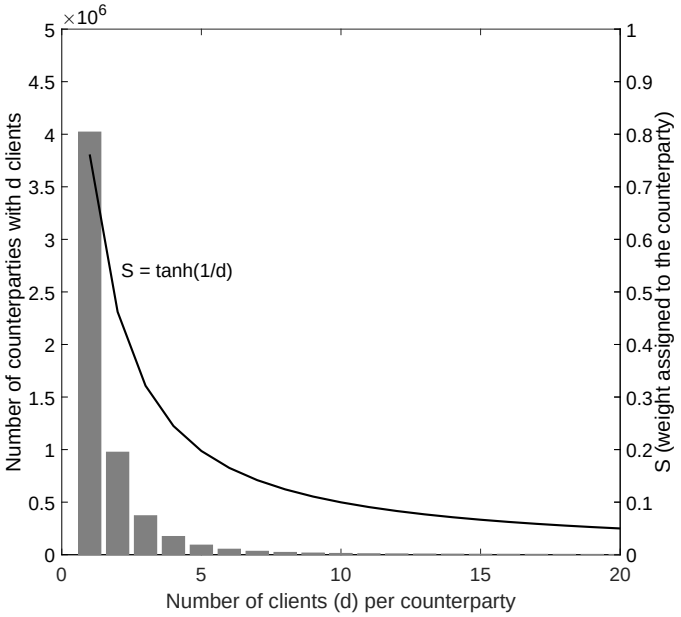


et al. [29]. To each counterparty a weight is assigned according to the hyperbolic tangent of its inversed degree, with the degree equal to the number of clients that have made a payment to/received a payment from the respective counterparty. The hyperbolic tangent function downweighs entities that many 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 (Internal Revenue Service or the country-equivalent) than clients buying their groceries at a certain local supermarket. Hence, the supermarket should be assigned a larger weight than the tax agency. Figure 4 shows the histogram of the number of clients per counterparty in our data set, accompanies by the corresponding weight. Most counterparties have few transaction partners. However, there is a small number of counterparties that the majority clients has transacted with. These are likely to be large companies or government organizations, such as energy and water suppliers and the tax agency. 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. The weighting scheme downweighs counterparties with many transaction partners more severely than entities that transacted

with only a few clients[5]. In the second step of the three-step framework the weights of all shared entities are aggregated.

The implied network is named Pseudo-Social Network (PSN): as in a true social network, strongly connected consumers demonstrate a strong similarity, at the very least in the particular merchants with which they transact. It is a pseudo-social network because, by and large, the linked consumers probably have no true social relationship with one another [20].

Figure 4: 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.



[5]The reader might wonder if this will not cause overfitting. In previous work [29], we looked at this risk by also including a tunable beta function as the top node weighting function. This beta function can take different forms, with a shape similar to the one of Figure 4, but also the possibility to downweigh the very infrequent counterparts. Interestingly, the optimal form is still the one we observe in the tangens hyperbolicum function (moreover, the tangens hyperbolicum is the one recommended in this study).

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 2.5 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 point-of-sales (POS) transactions or transfers. The first category consists of payments made at a physical store with a debit card (credit cards 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 sociodemographic 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 on its modelling, we can confidently state that the modelling procedure is in line with the state of the art modelling practices.

4.2 Study design

We estimate the performance of 4 models built using only transaction data: a propositional model, a direct network model, an implied (PSN) network model and a linear ensemble model that combines the output scores of the direct and implied network. The performance of these models is 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

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 (December)	2,585,227	4,141,402	42,865,861	28,278,074
2 months (+ November)	2,585,227	4,757,378	76,026,632	38,390,747
3 months (+ October)	2,585,227	5,254,154	114,043,910	48,401,003
4 months (+ September)	2,585,227	5,649,005	150,387,767	56,431,000
5 months (+ August)	2,585,227	5,945,217	182,645,116	63,042,608

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 classifiers using the transaction data and to train and validate the rating model, and 20% to train and validate the linear ensemble models. Ideally, the study should be performed completely out-of-time. However, due to data restrictions we are limited to the use of an out-of-sample testing framework.

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 historical 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. 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. The rating scales are transformed into separate data entries using unary encoding. 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 Machines which solves the following optimization problem [9]:

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

With vector $\mathbf{w}$ representing the weights of the model and $\mathbf{x}_i$ and y_i representing the input vector and the label of the i_{th} observation. $max(1 - y_i \mathbf{w}^T \mathbf{x}_i, 0)^2$ is the squared (L2) hinge-loss function. A ten-fold out-of-sample grid search was performed to find the optimal value of C, the regularization hyperparameter. We employed the LibLinear package from Fan et al. [9] 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 amount of months that are included in the data set, the amount of variables in the model thus varies between 4 and 6 million. The model weights are calculated using linear SVM.

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 PSN network. As a relational learner, we apply the weighted-vote Relational Neighbour (wvRN) classifier [16]. It is a simple, yet powerful classifier that uses the network structure to calculate a credit default probability $P(y_i = c|N(i))$ 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 [21], as it makes the assumption that the connected nodes are similar and therefore more likely to belong to the same class. We applied a smoothed version of wvRN that adds the default rate μ_c as smoothing factor.

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

Equation 2 calculates the probability that the label y of client i equals c, with c being a binary indicator of default, given its neighbours $N(i)$ in the unigraph (projection). The resulting credit default probability 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.

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 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:

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

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 characteristic [10] and the lift [4] at 1 and 5% of the test set, averaged over the 10 folds.

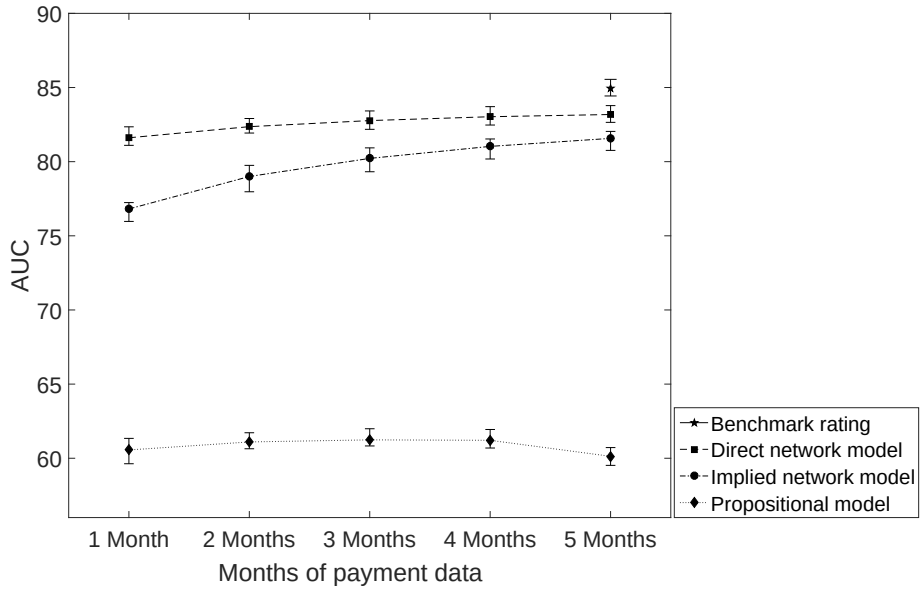
Figure 5 plots the AUC performance for all models. For sake of clarity, the results are spread over two graphs. Figure 5a plots the performance of the propositional model, the direct network model, the implied network model and the benchmark rating model. Figure 5b also plots the performance of the benchmark rating model (to facilitate easy 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 performance, however, they are still outperformed by the bank's own rating models. The only 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 to be composed of people with similar creditworthiness. Figure 6 illustrates a set of default clusters 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 to be a defaulter too. 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 (in the training set), 51.98% are defaulters themselves, compared to 0.77% in the complete test set. In other words: for every 100 clients in the test set that are linked to at least 1

defaulter in the training set, 52 of them will default.

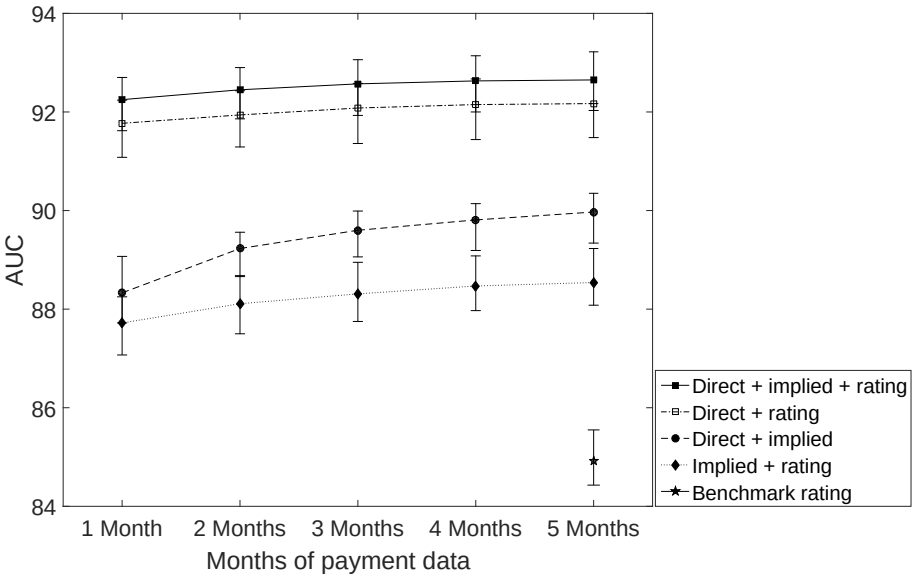
The best performing models are the ensemble models. The ensemble model that considers only payment data, i.e. the 'direct + implied' model, performs better than the ensemble model that combines the implied network with the ratings. The highest AUC values are found for the full ensemble model, closely 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 above 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% higher average lift than the rating model. However, this advantage over the rating model levels off at the 5% threshold, where the direct network model has only a 10% 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 [20]. In practical terms, this means that amongst the highest (worst) 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 thus nodes with no links in the direct network, confirming the importance

Figure 5: Results in terms of out-of-sample AUC.



(a)



(b)

Figure 6: Graph representation of a sample of the direct network.

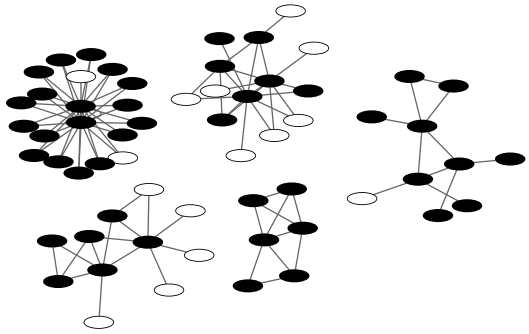
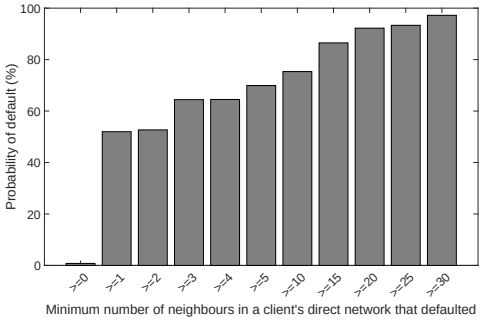


Figure 7: Probability of default for clients with increasing number and percentage of defaulted neighbours in their direct network.

(a) Number of neighbours



(b) Percentage of neighbours

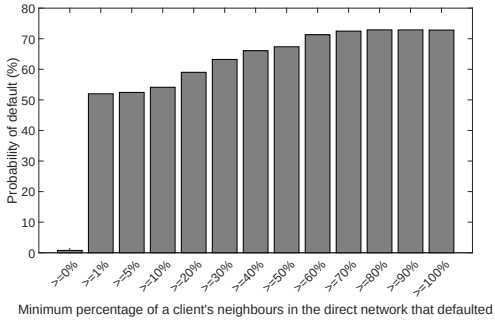
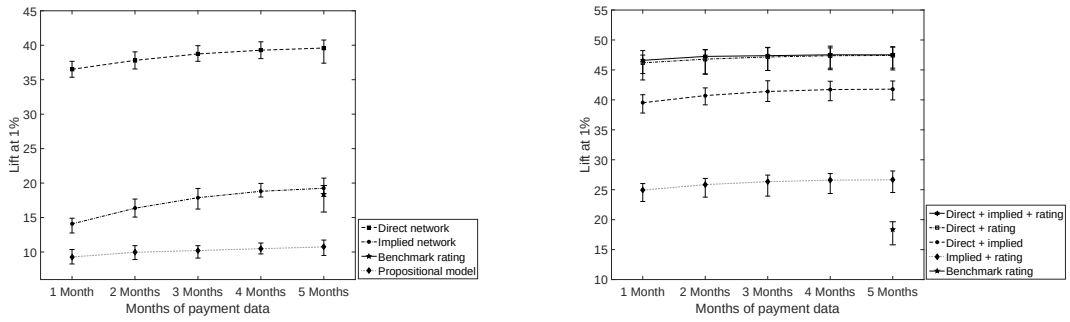
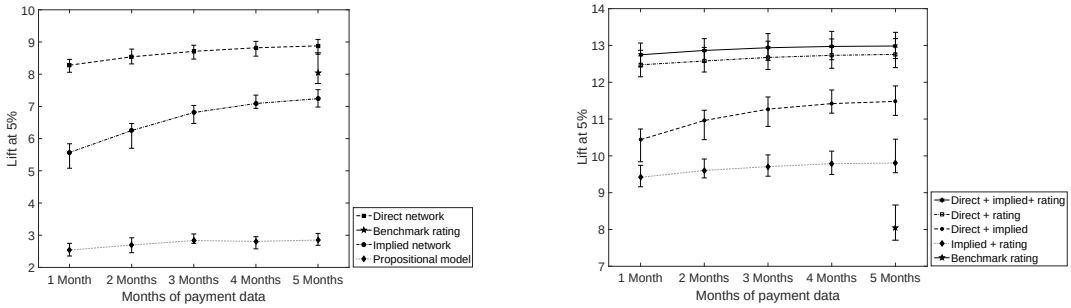


Figure 8: Results in terms of out-of-sample lift at 1 percent.



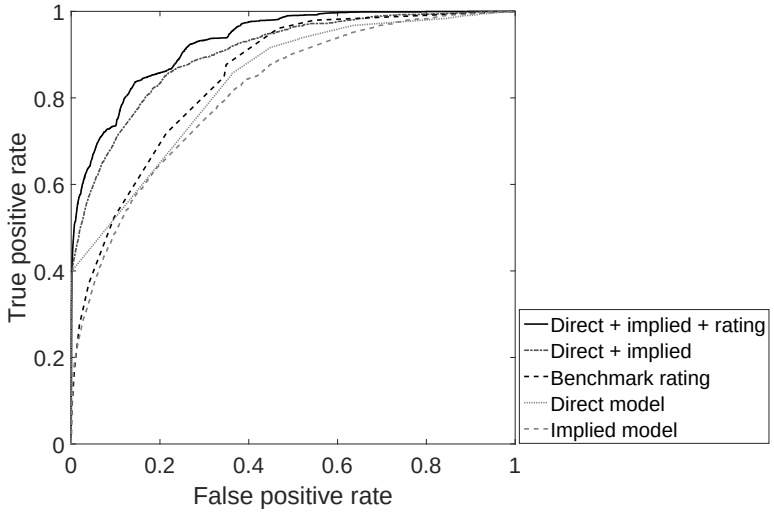
(a) **(b)**

Figure 9: Results in terms of out-of-sample lift at 5 percent.



(a) **(b)**

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



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 word-of-mouth marketing can lead to more links in a client's direct network if the people in his or her environment open an account at the bank.

Figures 5, 8 and 9 operate as learning curves [25]. 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 (PSN) network. Contrary to previous studies [13], we do not increase the number of clients, but the number of counterparties and known transactions per client. Nonetheless, we find the same conclusions: when working with fine-grained data, bigger is better [13, 20].

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 a set of one-sided paired t-tests over all ten folds. 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

Table 2: P-values of a one-sided paired t-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	7.77e-14	5.81e-14	2.96e-08	-
Implied	5.81e-14	3.39e-14	6.95e-10	-
Direct	2.96e-08	6.95e-10	2.27e-15	-
Best model	-	-	-	1.00

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 1e-07.

For the direct network, we consider two weighting schemes. The networks in Figures 5, 8 and 9 applied 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 performance values 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 > 0.31) as tested by a one-sided paired t-test. We would recommend banks reproduce such learning curves on their own data. With our results in mind (on the given dataset for the given prediction task), it is sufficient to save only the unique transactions per month, as the predictive value seems to lie 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 PSN 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 11: AUC of the direct network model with two different weighting schemes.

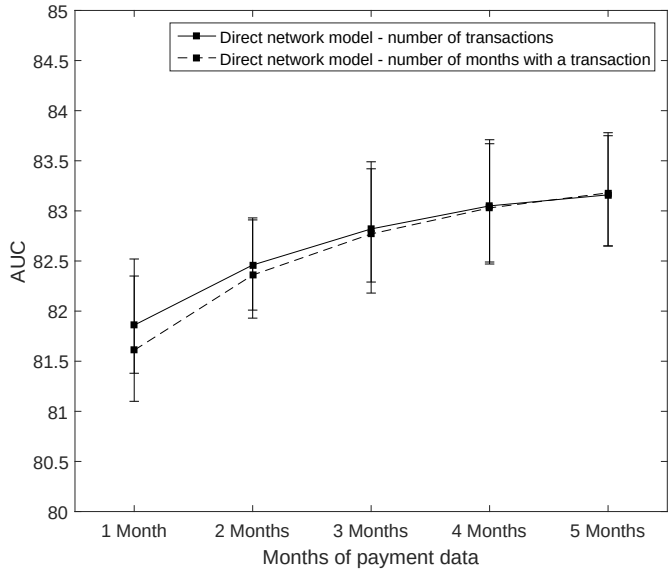
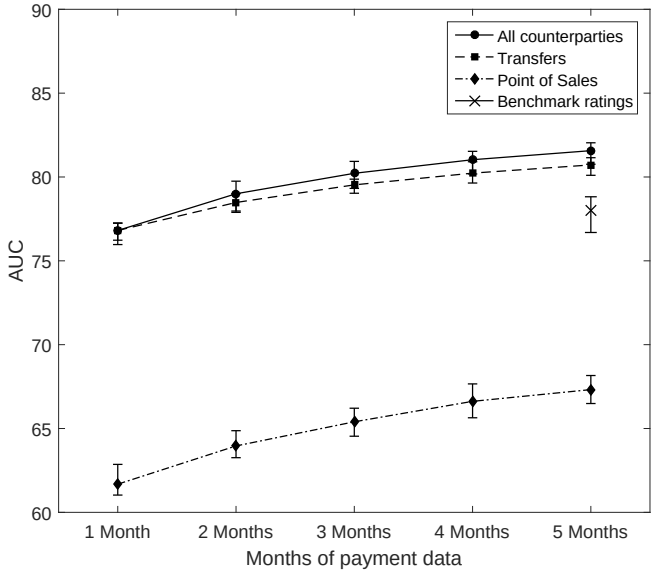


Figure 12: AUC of the implied network for different counterparties.



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 [6]. 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, a high accuracy can easily be obtained. 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 [6]. 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. This process can also be done in an incremental manner, so once the first batch has been processed, the weekly or monthly processing to add the new transactions would be quite fast. Finally, 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.

In Europe, a new European PSD2 directive has come into effect in January 2018, which encourages the mobility of consumers' payment data [8]. This implies that customers can ask their current bank to send all financial transaction data to another bank. With an increased digitalisation in the bank sector, there is a growing need for customers to be allowed to smoothly apply for a (consumer) loan through the banking app. This is a huge opportunity for our design: even for non-existing customers, a credit score will be computable if the customer provides the bank with its financial transactions. This implies that a person is no longer limited to its current bank to apply digitally and smoothly for a loan. A potential threat that looms for current major banks, is the implication that also startups or digitally native companies such as Google, Amazon, Facebook or Apple can start gathering payment data from other banks (of course only at the explicit request of the customers) and with the use of our design can build or improve their existing (marketing, credit, fraud) scoring models.

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

An important issue to consider when using sensitive payment data is privacy. The design 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 the best performance is found when both types are included. The counterparties' IDs can thus be irreversibly hashed. 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 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.

Another ethical implication for the use of fine-grained payment data is the explainability of such models. Global explanation methods such as investigating the coefficients in a linear model or rule extraction [18, 19] are difficult, as there are millions of features (the account numbers of the counterparties). For such high-dimensional, sparse data, instance-level explanation methods have been proposed recently, such as the EDC method [19] (stands for EviDence Counterfactual, as well as Explaining Document Classification). A single (e.g. default) prediction is explained by the minimal set of transactions made by this person such that if those transactions would not have occurred, the customer would no longer be predicted to be a defaulter. Although this is an interesting approach to understand the decisions made, and potentially to improve the model (by explaining false positives), such explanations are not provided to the customers. The score based from our design would mainly be used as a complement to already available models (which is also the recommendation when considering the predictive performance). Upon request for more explanation by the customer, the information provided is limited to the loan evaluation process, meaning that the decision has taken into account aspects as budgetary analysis, credit register information, and credit scoring. Most banks already use payment data in an aggregated manner, the explanation hence does not change with the inclusion of our design. There is however an additional ethical issue, where we need to avoid that customers, becoming aware of the use of their payment data for credit scoring, avoid transacting with economically vulnerable persons with assumed low credit score. On the other hand, this same economically vulnerable group benefits from this design, as they would be able to

apply for credit at more banks (not just its own bank), and as such the design also promotes financial inclusion.

5 Conclusion

This paper investigates the use of transaction data for credit scoring. Our first contribution is the examination of both propositional and relational methods to classify customers and the finding that transaction data should be modelled in a relational manner. Our second contribution is the demonstration that payment data adds complementary predictive power to the traditional credit scores. Thirdly, we offer guidelines to banks on the data to use and analyses to consider. The 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 PSN 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' accuracy, 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 be extended to other credit scoring applications, including credit card default using credit card transactions and corporate default using corporate transactions. Additionally, the European PSD2 directive also implies that banks (but also startups or digitally native companies such as Google, Amazon, Facebook or Apple) can start gathering payment data from other banks (of course only at the explicit request of the customers) to improve their existing scoring models.

As we found that bigger data leads to better predictions, the question arises at what point more data will not be beneficial anymore? More data has two dimensions: more transactions for a given customer (in time for example), or having transaction data over more customers. Another interesting issue for future research is the inclusion of domain knowledge or the detection of potential bias in the scoring models.

References

[1] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

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

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

[4] M. J. Berry and G. S. Linoff. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2004.

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

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

[7] D. Durand et al. Risk elements in consumer instalment financing. *NBER Books*, 1941.

[8] European Commission. Payment services (psd 2) - directive (eu) 2015/2366. Technical report, European Commission, 2015.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

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

[11] W. E. Hardy and J. L. Adrian. A linear programming alternative to discriminant analysis in credit scoring. *Agribusiness*, 1(4):285–292, 1985.

[12] C. S. Hilas. Designing an expert system for fraud detection in private telecommunications networks. *Expert Systems with applications*, 36(9):11559–11569, 2009.

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

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

[15] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

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

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

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

[19] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–100, Mar. 2014.

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

[21] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.

[22] J. H. Myers and E. W. Forgy. The development of numerical credit evaluation systems. *Journal of the American Statistical association*, 58(303):799–806, 1963.

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

[24] Official Journal of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Technical report, European Parliament and Council, 2016.

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

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

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

[28] G. Shmueli. Analyzing behavioral big data: methodological, practical, ethical, and moral issues. *Quality Engineering*, 29:57–74, 2016.

[29] M. Stankova, D. Martens, and F. Provost. Classification over bipartite graphs through projection. Technical report, University of Antwerp Working Paper, 2015.

[30] W. Verbeke, D. Martens, and B. Baesens. Social network analysis for customer churn prediction. *Applied Soft Computing*, 14, Part C(0):431 – 446, 2014.

[31] I. Weber, V. R. K. Garimella, and E. Borra. Inferring audience partisanship for youtube videos. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 43–44. International World Wide Web Conferences Steering Committee, 2013.

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