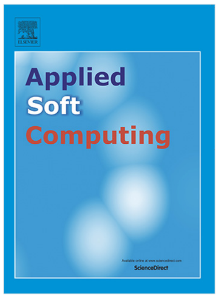
Accepted Manuscript



The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics

María Óskarsdóttir, Cristián Bravo, Carlos Sarraute, Jan Vanthienen, Bart Baesens

PII: S1568-4946(18)30560-X
DOI: https://doi.org/10.1016/j.asoc.2018.10.004
Reference: ASOC 5121

To appear in: *Applied Soft Computing Journal*

Received date : 3 April 2018
Revised date : 2 July 2018
Accepted date : 1 October 2018

Please cite this article as: M. Óskarsdóttir, et al., The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics, *Applied Soft Computing Journal* (2018), https://doi.org/10.1016/j.asoc.2018.10.004

The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion using Mobile Phone Data and Social Network Analytics

María Óskarsdóttir[a,*], Cristián Bravo[b], Carlos Sarraute[c], Jan Vanthienen[a], Bart Baesens[a,b]

[a]*Dept. of Decision Sciences and Information Management, KU Leuven, Belgium.*
[b]*Dept. of Decision Analytics and Risk, University of Southampton, United Kingdom.*
[c]*Grandata Labs, 550 15th Street, San Francisco, California 94103, USA.*

Abstract

Credit scoring is without a doubt one of the oldest applications of analytics. In recent years, a multitude of sophisticated classification techniques have been developed to improve the statistical performance of credit scoring models. Instead of focusing on the techniques themselves, this paper leverages alternative data sources to enhance both statistical and economic model performance. The study demonstrates how including call networks, in the context of positive credit information, as a new Big Data source has added value in terms of profit by applying a profit measure and profit-based feature selection. A unique combination of datasets, including call-detail records, credit and debit account information of customers is used to create scorecards for credit card applicants. Call-detail records are used to build call networks and advanced social network analytics techniques are applied to propagate influence from prior defaulters throughout the network to produce influence scores. The results show that combining call-detail records with traditional data in credit scoring models significantly increases their performance when measured in AUC. In terms of profit, the best model is the one built with only calling behavior features. In addition, the calling behavior features are the most predictive in other models, both in terms of statistical and economic performance. The results have an impact in terms of ethical use of call-detail records, regulatory implications, financial inclusion, as well as data sharing and privacy.

Keywords: Credit Scoring, Social Network Analysis, Profit Measure, Mobile Phone Data, Decision Support

---

*Corresponding author

Email addresses: maria.oskarsdottir@kuleuven.be (María Óskarsdóttir), C.Bravo@soton.ac.uk (Cristián Bravo), charles@grandata.com (Carlos Sarraute), jan.vanthienen@kuleuven.be (Jan Vanthienen), bart.baesens@kuleuven.be (Bart Baesens)

1. Introduction

Credit scoring is undoubtedly one of the oldest applications of analytics where lenders and financial institutions perform statistical analysis to assess the creditworthiness of potential borrowers to help them decide whether or not to grant credit [1]. Fair Isaac was founded in 1956 as one of the first analytical companies offering retail credit scoring services in the US. Its well-known FICO score (ranging between 300 and 850) has been used as a key decision instrument by financial institutions, insurers, utilities companies and even employers [2]. The first corporate credit scoring models date back to the late sixties with Edward Altman developing his well-known z-score model for bankruptcy prediction, which is still used to this day in Bloomberg reports as a default risk benchmark [3]. Originally, these models were built using limited data–consisting of only a few hundred observations–and were based on simple classification techniques such as linear programming, discriminant analysis and logistic regression, which is the current industry standard given its high interpretability [2]. The importance of these retail and corporate credit scoring models further increased due to various regulatory compliance guidelines such as the Basel Accords and IFRS 9 which clearly stipulate the inputs and outputs of a credit scoring model together with how these models can be used to calculate provisions and capital buffers.

The emergence of more sophisticated classification techniques such as neural networks, support vector machines and random forests led to various extensive benchmarking studies aimed at improving credit scoring models in terms of their statistical performance (e.g., in terms of area under the ROC curve or classification accuracy) [4, 5]. Many of these studies concluded that traditional credit scoring models based on, e.g., simple logistic regression models, performed very well and newer classification techniques could only offer marginal performance gains. In other words, research on developing high-performing credit scoring models has more or less stalled. We believe the best investment in better credit scoring models is not to turn the attention to newer classification techniques but to leverage innovative Big Data sources instead.

While these new sources of data present the opportunity to profile potential borrowers using a wider representation of behavior, they also present an ethical challenge. Mobile phone data, e.g., in the form of call-detail records (CDR), allows constructing a very detailed social network, and using this information to profile repayment behavior can be seen as unfair to borrowers that could be punished for their mobile cell phone behavior. Recently, the use of positive information has been put forward as a necessary source of data that should be included in scoring models [6]. Positive information is defined as all information that represents the good financial behavior, providing a clearer definition of the factors that make a good borrower. Barron and Staten [7] show, for example, that not using positive information leads to a decrease of up to 47.5% in credit availability.

This paper introduces mobile phone data as a new Big Data source for credit scoring and shows that while it is a powerful source of information, it should be used strictly in a positive framework to increase the access to financing to borrowers who would otherwise be out of options until a much later stage. To motivate the use of this information in financial institutions, its potential is studied in both statistical and profit terms.

Big Data is typically defined in terms of its 5 V's: Volume, Variety, Velocity, Veracity and Value. Recent special issues of Information Systems Research [8] and MIS Quarterly [9] indicate the explosion of interest in Big Data within the IS community. The use of mobile phone data for credit scoring is a fitting example of this since it comes in huge volumes (Volume), has not been explored before (Variety), is generated on a continuous daily basis (Velocity) and is usually stored using a well-defined call-detail record log format (Veracity). In this paper, its Value is also quantified by focusing both on its statistical performance (e.g., in terms of area under the ROC curve) and on its bottom line impact in terms of profit. Additionally, an evaluation of the qualitative performance of the data in terms of positive information for enhanced financial inclusion is provided. This study is based on a unique data set combining banking, sociodemographic and CDR data. CDR are logs of all phone calls between the customers of a telecommunications provider (telco), see Table 1. More specifically, the data set includes all CDR of the bank's customers, the CDR of the people they are in contact with and the banking history of these customers. Overall, it adds up to a year and a half of banking history of over two million bank customers and the calling activity of almost 90 million unique phone numbers spanning five months. This unique combination of data gives the opportunity to explore the potential of enriching traditional credit scoring models with social network effects reflecting calling behavior. The three key research questions are:

Q1 What is the added value (in terms of both AUC and profit) of including call data for credit scoring?

Q2 Can call data replace traditional data used for credit scoring?

Q3 How does default behavior propagate in the call network?

To the best of our knowledge, these questions have not been researched before. Each of the questions will be answered from both a statistical as well as a profit perspective, which is another key contribution of this paper. Furthermore, the implications for financial inclusion are evaluated.

The impact of this research is manifold. A successful application of boosting the performance of credit scoring models using call data would improve credit decision-making and pricing. The insights could also facilitate access to credit for borrowers with little or no credit history. This is the case for young borrowers, lenders exploring new markets or in developing countries with young credit markets. In all these cases, the borrowers are not expected to have a credit history, but they do have mobile phone records. Knowing how default behavior propagates in a call network also has regulatory implications. For example, the Basel Accords try to capture default correlation in order to better protect a financial institution against unexpected losses [2]. The research can shed new light on how default behavior is correlated. This could lead to better provisioning and capital buffering strategies, thereby improving the resilience of the financial system against shocks and macroeconomic downturns. Knowing how default behavior propagates in a call network also has other regulatory implications. If CDR data is indeed useful for credit prediction, then banks and credit bureaus have a strong economic incentive to collaborate with telecommunications companies to share data in order to perform this type of analyses.

In the next section, a review on the literature on Big Data in credit scoring as well as previous research on call networks is provided. In section 3, the theoretical background and methodology applied in the case study is described, with the experimental setup detailed in section 4. The results are presented in section 5, followed by a discussion on their various implications in sections 6 and 7. The paper concludes with a summary of the contributions and discussion on possibilities for future work.

2. Related Work

Many analytical modeling exercises start from a flat data set, build a predictive model for a target measure of interest (e.g., churn, fraud, default) and evaluate it on an independent out-of-sample data set. An assumption which is (oftentimes) tacitly made is that the data is independent and identically distributed. Recent research questioned this assumption and analyzed how customers can influence each other through the different social networks that connect them [10]. Various types of social behavior can be observed. One is homophily, which states that people have a strong tendency to associate with others whom they perceive as being similar to themselves in some way. Social influence occurs when people's behavior is affected by others with whom they interact [11, 12]. Some of the social behavior can also be attributed to other (e.g., external) confounding factors [13]. The idea of network learning is to embed social behavior patterns in the predictive models to successfully leverage the impact of joint customer actions [14]. A key input to any social network learning exercise is the network itself, which consist of nodes and edges. In certain settings, the definition of these networks is relatively straightforward. As an example, consider churn prediction in telco where the network can obviously be constructed based upon data stored in the CDR. Earlier research found significant social network effects for predicting churn in telco [15]. Another example is credit card fraud detection where a network can be defined by connecting credit cards to merchants. Also in this setting, strong social network effects have been found [16].

In credit scoring, there is a firm belief amongst both researchers and practitioners that default behavior of borrowers is correlated. To illustrate this, the Basel Accord models default correlation by means of an asset correlation term, which is set to 15% for residential mortgages and 4% for qualifying revolving exposures. However, both these numbers have been set in a rather arbitrary way, or based upon some empirical but not published procedure [17]. This interdependency has been proven to be a significant factor amongst small and medium-sized enterprises [18]. One of the key challenges in understanding network effects or default propagation in credit scoring concerns the definition of the network itself. Preliminary attempts have been made to build networks between customers in online peer-to-peer lending. For example, Lin et al. [19] illustrated that online friendships with non-defaulters increases the credit score. These findings were also confirmed by Freedman and Jin [20], with an additional caution that online ties on their own may not reveal true information about creditworthiness and may also be manipulated [21]. De Cnudde et al. [22] developed credit scoring models for microfinance using social media network information extracted from Facebook accounts. Their results suggest that explicit networks of friends who interact are more predictive than of friends who do not, but implicit networks of people with similar behavior are better than both explicit friendship networks. In industry, social networks are already being exploited to assess creditworthiness, by technology companies such as Lenddo, that make use of social media connections to analyse people's default risk [23].

More recently, the interest in using call networks as a new Big Data source for credit scoring has gained traction, e.g., with Wei et al. [21] formulating the potential value of credit scores obtained with networks–for example, based on social media or calls–and how strategic tie-formation might affect these scores. Although especially interesting in relation to the Chinese government's plan for a social credit system [24], the study is only theoretical and is missing an important empirical evaluation of the proposed models [21]. Moreover, recent press coverage on specialized smartphone applications that evaluate people's creditworthiness using the huge amount of data generated by their handsets indicates the potential of call networks as an alternative data source for credit scoring [25, 23]. Most of these studies have focused on the use of social networks in the context of social media, or have discussed the potential of CDR-induced social networks in credit scoring.

The literature on the analysis of CDR is rich [26]. The idea of using CDR data for credit scoring stems from the fact that the way people use their phone is assumed to be a good proxy for their lifestyle and economic activity. Previous research confirmed that using CDR data to build call networks by linking together individuals who are in contact with each other, results in social networks that can be used in both descriptive and predictive studies on age, gender, ethnicity, language, economic factors, geography, urbanization, and epidemics [27, 28, 29, 30, 31]. For example, Leo et al. [31] confirm the presence of homophily in terms of economic behavior using call networks. More specifically, they show that wealth and debt are unevenly distributed and that people are better connected with those that share their socioeconomic class. Furthermore, Haenlein [32] investigated the distribution of customer revenue within a call network and demonstrated that high revenue customers are primarily related to other high revenue customers and the same for low revenue customers.

3. Methodology

This paper contributes to the literature by investigating the use of CDR data for credit scoring in terms of value. Here, the proposed methodology for extracting appropriate information from the CDR data by means of social networks and influence propagation is detailed. Furthermore, techniques for evaluating model and feature performance in terms of profit are presented.

3.1. Call Networks: Featurization and Propagation

A call network is a network where the nodes $\mathcal{V} = \{v_1, \dots v_n\}$ are people present in a CDR log. These logs are kept for billing purposes and include information about every phone call made by the

Table 1: An example of a CDR log. In the actual dataset the phone numbers are encrypted.

Call Start Date	Call Start Time	Call Duration (sec)	From Number	To Number
01MAY2017	14:51:14	715	(202) 555-0116	(701) 555-0191
02MAY2017	14:34:37	29	(803) 555-0129	(202) 555-0116
01MAY2017	20:34:14	9	(803) 555-011[illegible]	(406) 555-0137
02MAY2017	20:03:38	89	(701) 555-01[illegible]	(803) 555-0129

customers of a telecommunications operator, including the encrypted phone numbers of the customers that made and received the phone call as well as timing and length. An example of such a log can be seen in Table 1. Information from CDR about time and duration of phone calls (or text messages) can be used to connect the people in the network to create the edges, $e_{i,j} \in \mathcal{E}$. The edges are either undirected, such as when two customers share a phone call but it is irrelevant which customer made the call; or directed, in which case we distinguish between outgoing and incoming edges (i.e., all phone calls made by and received by a person, respectively). The edges are represented by an n by n binary matrix, called adjacency matrix A, where a non-zero entry denotes an existing edge between node v_i and v_j in an undirected network and from/to v_i to/from v_j in a directed network with outgoing/incoming edges. The edges can also carry weights to indicate the intensity of the relationship between two people, for example the number or duration of phone calls they share in a given time period. The weights are denoted by the weight matrix $W = (w_{i,j})$, where $w_{i,j} \in \mathbb{R}^+ \cup \{0\}$. The first order neighborhood of a node v_i is the collection of nodes v_j that share an edge with v_i, that is

$$N_i^1 = \{v_j | e_{i,j} \in \mathcal{E}, j = 1, \ldots, n\}.$$

In some networks, the nodes can be labelled, or assigned to a class that is later used in a predictive analytics framework. In this application, there are two types of labels. The first type of label regards default, in which case the customers in the call network belong to one of two classes: they are either defaulters, who have been in arrears for more than 90 days within a twelve month period (bad customers); or they are non-defaulters (good customers). [1] When building a credit scoring model, the goal is to assign one of these two classes to each customer of interest and it is the target variable of the classification problem in this study. In the call network there are also customers who, during the timespan of the network, have run into payment arrears for one or two months in addition to defaulters with three months of payment arrears. For clarity, all these customers are referred to as delinquent

[1] We use the Basel definition of default. [2]

customers and it is the second type of label. The delinquent customers have the possibility to influence others in the network to also run into payment arrears–also referred to as default influence–and they are used when generating features as explained below.

In order to use the information that is contained in the call networks for building credit scoring models, network features are extracted for each node in the network by aggregating information about its position within the network and connectivity to other nodes. As in similar studies, a distinction is made between direct network features, which are derived from the node's first order neighborhood, and indirect network features that take into account the whole network structure [33]. As stated earlier, the aim is to study how delinquent customers may influence others with whom they are connected. Therefore, by assuming there is prior knowledge about some delinquent customers in the network (i.e., having a subset of nodes with known labels) that knowledge can be incorporated in the network features by exploiting social ties. To this end, both direct and indirect network features are extracted as illustrated in Figure 1. The direct network features represent the presence and number of delinquent customers in a node's first order neighborhood. They are easy to extract and provide a representative overview of people's social connections [34]. However, the influence of payment arrears is likely to reach further than just the first order neighborhood. This effect is modeled using two distinct propagation methods that have been effective in previous research and are designed to simulate real-life behavior: Personalized PageRank (PR) and Spreading Activation (SPA). The results of both methods are exposure scores which are categorized as indirect network features. Although other propagation methods exist, such as Gibbs sampling and relaxation labelling, these were not applied here because they have been shown to be less effective for prediction in call networks [35], are less scalable and as such did not fulfil the requirements of this study. The features resulting from these three approaches will be used as input features when building credit scoring models, but first a more detailed explanation is provided.

3.1.1. Link Based Features

Lu and Getoor [34] presented a framework for inferring labels for nodes in a network based on labels of neighboring nodes. They defined three features that can be extracted from the neighborhood of a node: count-link, mode-link and binary-link. These represent, respectively, the frequency of classes in the neighborhood, their mode, and a binary indicator for each class. Futhermore, using a logistic regression model, Lu and Getoor [34] showed that these features are very predictive for the

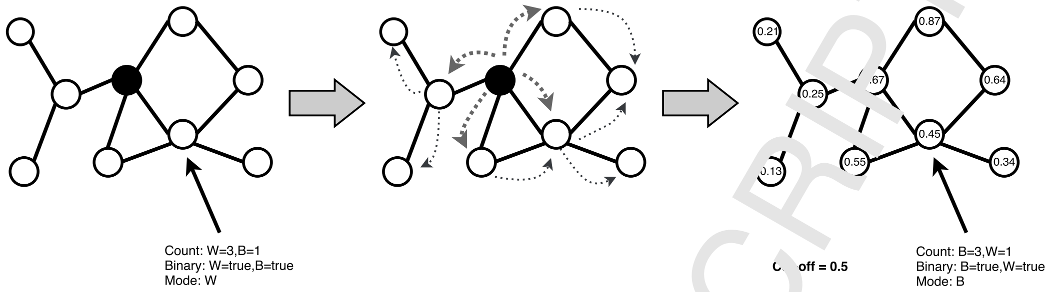


Figure 1: The figure demonstrates the computation of link-based measures, before and after a propagation method is applied to the network. The figure on the left shows a network with one black node and eight white nodes. The link-based features of the node to which the arrow points are summarized, where B means black and W means white. The figure in the middle demonstrates the application of the propagation method with the resulting exposure scores shown for each node in the figure on the right. After the exposure scores have been computed, a cut-off point is set at 0.5 and nodes with a score that is higher than the threshold are labeled black (B), and white (W) otherwise. Subsequently, link-based exposure features are extracted for the node to which the arrow points.

class of the node itself.

Extraction of link-based features is based on the presence of delinquent customers with varying number of payment arrears.

3.1.2. Personalized PageRank

The propagation method Personalized PageRank (PR) was developed for search engines (e.g., Google) to rank webpages while also taking into account an initial source of information, such as frequently visited web pages [36] but can also be used for different kinds of linked data [16, 33]. For the nodes in a network with weight matrix W, the method iteratively computes exposure scores ξ_{k+1} based on the exposure scores in the node's neighborhood ξ_k and a random jump to other nodes in the network, determined by the information source z – also called restart vector – using the equation

$$\xi_{k+1} = \alpha W \xi_k + (1-\alpha) z,$$

where $1-\alpha$, the damping factor, denotes the probability of a random jump and k is the iteration step. As a result of the initial information source, exposure scores of nodes closer to the source nodes are higher. Here, the delinquent customers are the information source.

3.1.3. Spreading Activation

The propagation method Spreading Activation (SPA) originates from cognitive psychology and simulates how information, or energy, spreads through the network from a set of source nodes. It is used to model a 'word-of-mouth' scenario, where influence–in this case from delinquent customers–spreads through the network. 'Word-of-mouth' has been shown to be effective in social networks [37, 38]. Before the method begins, a set of active nodes $V^A \subset V$ possesses the energy $E^0(V_A)$. In each step k of this iterative method, a part d of an active node's energy $E^k(V_A)$ is spread to the nodes in its neighborhood while the rest of the energy remains. The part that is transferred, is distributed according to the relative weights of the links to neighboring nodes, expressed by the transfer function

$$E_{transfer} = \frac{d \cdot w_{i,j}}{\sum_{w_{i,s} \in W} w_{i,s}} E^k(V_i^A).$$

The method stops when no more nodes are being affected and the changes in energy of the already affected nodes are smaller than a given threshold value. The total energy always remains the same, but spreads throughout the network.

3.1.4. Link-Based Exposure Features

After a propagation method, such as PR or SPA, has been applied to a network, each node possesses an exposure score that can be viewed as the relative ranking of the node compared to the rest of the network. The score can be used as a feature directly or by determining a cut-off value. Nodes with an exposure score lower than the cut-off are defined as low-risk nodes and those with an exposure score above the cut-off as high-risk nodes [33]. Then, based on this re-labelling of the network, new link-based features can be extracted. This is demonstrated in Figure 1.

3.2. The Expected Maximum Profit Measure

Model selection highly depends on how the performance is measured. Traditional measures for credit scoring models include AUC, Gini coefficient and the KS statistic that either assess the discriminative ability of the models or the correctness of the categorical predictions [5]. The recently proposed Expected Maximum Profit (EMP) measure has an advantage over these traditional measures because it considers the expected losses and operational income generated by the loan, and is tailored towards the business goal of credit scoring [39]. Most importantly, when applied to credit scoring models it facilitates computing the models' value, the fifth V of Big Data. The measure is based on the expected maximum profit measure, originally developed for customer churn prediction [40], and is expressed for credit scoring by

$$EMP = \int_{b_0}\int_{c_1} P(T(\Theta); b_0, c_1, c^*) \cdot h(b_0, c_1) dc_1 db_0$$

where

$$P(t; b_0, c_1, c^*) = (b_0 - c^*)\pi_0 F_0(t) - (c_1 + c^*)\pi_1 F_1(t)$$

is the average classification profit per borrower given the prior probabilities of being a defaulter (non-defaulter), π_0 (π_1), and the cumulative density functions of defaulters (non-defaulters), $F_0(s)$ ($F_1(s)$). Furthermore, b_0 is the benefit of correctly identifying a defaulter, c_1 the cost of incorrectly classifying a non-defaulter as a defaulter, c^* the cost of the action, $\theta = \frac{c_1+c^*}{b_0-c^*}$ the cost/benefit ratio and $h(b_0, c_1)$ the joint probability density function of the classification costs [39]. The maximum profit is achieved by optimizing the cut-off dependent average classification profit where the optimal cut-off value is

$$T = argmax_{\forall t} P(t; b_0, c_1, c^*).$$

As a result, the measure clearly defines an optimal fraction, expressed as

$$\bar{\eta}_{EMP} = \int_{b_0}\int_{c_1} [\pi_0 F_0(T(\Theta)) + \pi_1 F_1(T(\Theta))] \cdot h(b_0, c_1) dc_1 db_0,$$

representing the fraction of applications that should be rejected to receive maximum profit. Verbraken et al. [40] showed that the EMP corresponds to integrating over the range of the ROC curve that would be considered in a real application, discarding the segment that has a very high, unreasonable cost, and that it is an upper bound of the profit a company could achieve by applying the respective classifier.

When deriving the parameters b_0, c_1 and c^* and the probability distribution $h(c_1, b_0)$, Verbraken et al. [39] rely on the profit framework discussed in Bravo et al. [41]. Thus, b_0 is specified as the fraction of the loan amount that is lost after default or

$$b_0 = \frac{LGD \cdot EAD}{A} =: \lambda, \tag{1}$$

where LGD is the loss given default, EAD is the exposure at default and A the loan amount. Furthermore, c_1 equals the return on investment (ROI) of the loan and $c^* = 0$ since rejecting a customer does not generate any costs. It only remains to determine $h(b_0, c_1)$ where $ROI(c_1)$ is assumed to be constant but $\lambda(b_0)$ needs to be estimated for each dataset because it is more uncertain with a multitude of possible distributions.

Table 2: Confusion matrix for computing model profit

		Predicted class	
		Non-default	Default
Actual class	Non-default	$ROI \cdot A$	$-ROI \cdot A$
	Default	$-LGD \cdot EAD$	0

3.2.1. *Model Profit*

The EMP fraction can subsequently be used to compute the profit of a given model. First, it is translated into a cut-off value, which depends on the number of instances in the test set. The instances are labelled as defaulters or non-defaulters depending on whether their predicted score is higher or lower than the cut-off. Then for each customer in the test set, the confusion matrix in Table 2 is used to compute the loss or gain produced by the customer. The model profit is finally computed by aggregating the profit of all customers.

3.2.2. *Feature Importance in Terms of Profit*

When a credit scoring model is built using the random forest algorithm, its properties can be used to measure the profit impact of each feature in the model. Assuming a random forest model RF was built using N trees $(T_i)_{i=1}^{N}$ and M features $(F_j)_{j=1}^{M}$, the feature importance in terms of profit can be computed in the following way.

1. Apply the random forest model RF to the test set and extract class predictions for each tree $T_i \in RF$.

2. For each tree T_i compute the profit $P(T_i)$ using the confusion matrix in Table 2.

3. For each feature F_j in the test set, compute the mean decrease in profit. This is defined as the difference between the average profit of trees where $F_j \in T_i$ and the average profit of trees where $F_j \notin T_i$, given by the equation

$$P(F_j) = \frac{\sum\limits_{i, F_j \in T_i} P(T_i)}{|\{T_i : F_j \in T_i\}|} - \frac{\sum\limits_{i, F_j \notin T_i} P(T_i)}{|\{T_i : F_j \notin T_i\}|}$$

   where $F_j \in T_i$ means that feature F_j is in tree T_i.

4. Sort $P(F_j)$: the features with the highest values are those with the greatest mean decrease in profit.

The result of this method is a ranking of the features in terms of the importance with respect to profit.

4. Experimental Design

4.1. Data Description

The data used in this study originates from a telecommunications operator and a commercial bank that both operate in the same country. The datasets are anonymized and do not contain personal information such as the name and address of customers. The telco data contains five consecutive months of CDR data of almost 90 million unique cell phone numbers as described in Table 1. The data from the bank includes over two million customers and it consists of three parts, namely sociodemographic information, such as age, marital status and postcode; debit account activity, including timing and amount of payments; and credit card activity. Both sociodemographic and debit account activity span three months and conform the historic part of the dataset. For the credit card activity, there is information about when the cards were issued, the total credit limit, monthly values of how much of the credit remains and how often the customers have failed to repay their debt until twelve months after receiving the card. The credit card transactions serve as the key input to the credit scoring application because the data provides information about monthly payment arrears. This is used to predict the creditworthiness of the customers. The knowledge about the credit limit and remaining credit on the cards also allows the computation of the EMP.

4.2. Experimental Setup

The credit scoring models are built for customers who received a credit card within a three-month period in 2015 and they are referred to as subjects. An overview of the experimental setup can be seen in Figure 2. The credit card data contains information which enables the labeling of the subjects as defaulters or non-defaulters by counting how many late payments they have in the year after signing up for the card. As previously noted, the Basel definition is used where having three or more late payments implies default. The label or target vector is denoted by $y_{Default}$.

To create the bank component of the dataset, both the sociodemographic and debit account data is used. More precisely, sociodemographic features such as age, marital status and residency as reported at the time of the credit card application are extracted. Furthermore, debit account activity in the month prior to receiving the credit card is considered and used to extract features representing spending behavior, as can be seen in Table 3. Based on Singh et al. [42], two types of temporal-behavioral features

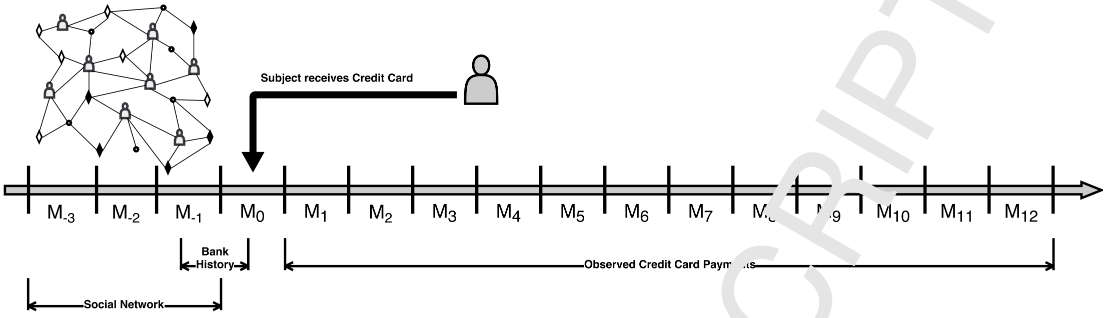


Figure 2: Experimental setup for one timeframe.

that have been shown to correlate with financial well-being and consumption are included. The first one, diversity, measures how customers spread their transactions over various bins, represented by the days of the week in this case. For each customer i and each bin j, the fraction of transactions p_{ij} that fall within bin j is computed. The temporal diversity of customer i is then defined as the normalized entropy of all transactions counted in all seven bins with M being the number of non-empty bins, or

$$D_i = \frac{-\sum_{j=1}^{7} p_{ij} \log p_{ij}}{logM}.$$

In addition, the loyalty of a customer is defined as

$$L_i = \frac{f_i}{\sum_{j=1}^{7} p_{ij}}$$

where f_i is the fraction of all transactions of customer i that happen in their k most frequently used bins. In this case, loyalty characterizes the percentage of transactions that take place during a customer's three most active days. The collection of both sociodemographic and debit account features is called 'sociodemographic' features and denoted with x_{SD}.

The telco data is used as the key input for the social network part of the analysis. As mentioned before, the subjects received their credit cards within a period of three months and the subjects are considered in each month separately, which results in three timeframes t_1, t_2 and t_3. To build a call network for each timeframe, the CDR of three whole months prior to the card acquisition month is aggregated and people that have shared a phone call during this period are linked together after discarding any phone calls lasting less than five seconds. Thus, there are three call networks spanning three months each. Each network consists of all subjects that received a credit card in the month succeeding the last month in the network, everyone they shared a phone call with and all phone calls between everyone in the network. In addition to the subjects, there are also other types of people in the

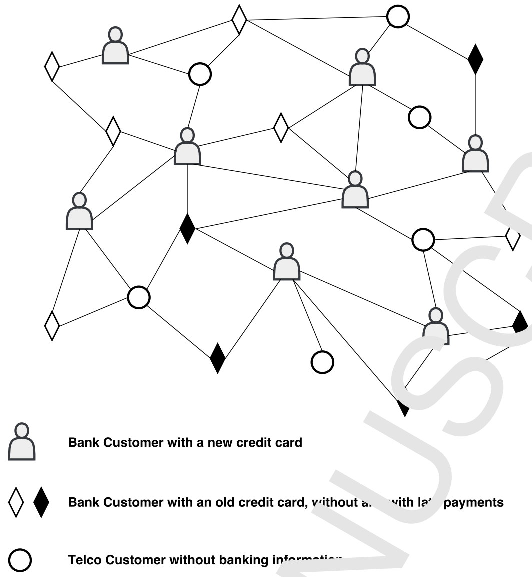


Figure 3: The figure demonstrates the various types of people that are present in the call network.

network, as Figure 3 shows. The people-shaped entities are the subjects, whereas the diamond-shaped entities denote other bank customers (i.e., people who did not receive a credit card during the three months). They may, however, already possess a card, and those that are known to have had payments arrears are colored black. These are the delinquent customers in the network, as described in section 3.1. Bank customers without payment arrears are colored white. The circular entities in the network are people who are customers of the telco but not of the bank.

For all subjects in each of the three timeframes, four types of network features, both direct and indirect, are extracted. First, features representing the calling behavior of the subjects. Thus, the number and duration of incoming, outgoing and undirected phone calls taking place during the day and night and on different days of the week are computed. These features are denoted by x_{CB}. As described in subsection 3.1, information about delinquent customers in the network–the black diamonds–is used and they are labeled with respect to three distinct criteria: having one or more late payments, having two or more late payments and, having three or more late payments. This gives the opportunity to distinguish the severity of their financial situation in relation to the influence they spread. These three label vectors serve as the information source z and active nodes V^A when applying PR and SPA, respectively. Based on these labellings the extraction of link-based features, computation of PR and SPA exposure scores together with link-based exposure features as described in subsection 3.1 is performed. To construct the weight matrix W, edges in all networks are weighted by the number of phone calls and both incoming and outgoing edges as well as undirected networks are considered. The parameters in the propagation algorithms are set to the default values $\alpha = 0.85$ (for PR) and $d = 0.85$ (for SPA), based on exploration of the data which showed robust results. For the link-based exposure features, the cut-off point is defined as the minimum exposure score of the delinquent customers with at least three late payments, since having at least three late payments defines default. All the link-based features are viewed as one group of features denoted by x_{LB}. Finally, the feature groups x_{PR} and x_{SPA} are respectively composed of the exposure scores of PR and SPA together with the corresponding link-based exposure features.

The result of the featurization process is a dataset of the form

$$x = \{x_{SD}, x_{CB}, x_{LB}, x_{PR}, x_{SPA}\} \qquad y = \{y_{Default}\}$$

Table 3 describes some of these features. After combining the two data sources, extracting all the features described above and cleaning up the dataset, 22,000 observations remain and over 300 features. The fraction of defaulters is 0.0449 or just under 5% default rate.

With the datasets featurized, credit scoring models are built using binary classifiers with a 70%/30% split into training and test set. Before building the models, highly correlated variables are removed and undersampling of the training set conducted to reduce class imbalance, as is common when applying analytics techniques [43]. Final model performance is evaluated using the test set. The binary classifiers logistic regression, decision trees and random forests are used for the empirical analysis. Logistic regression is the industry standard for building credit scoring models [2]. Decision trees are included since they are more powerful than logistic regression, while at the same time guaranteeing interpretability of the model. They are implemented using recursive partitioning with ten-fold cross validation on the training set to tune and prune the trees. Both the logistic regression and decision tree models are compared against random forests which are an ensemble method that constructs multiple decision trees that jointly decide upon the credit score. Random forests are considered to be a very powerful, black box analytical modeling technique. As a result of parameter tuning, 500 trees were used to build each forest.

Table 3: Descriptions of some of the features that were extracted from the data sources. In the table IN, OUT and UD stand for networks with incoming, outgoing and undirected edges, respectively. The number in the brackets (x) indicates how delinquent customers were defined with respect to the number of payment arrears in each case.

Feature Group	Notation	Number	Feature	Description
Socio demo graphic	SD	35	Age	Current age of the customer
			Amount Spent	Total amount spent in the month before receiving the credit card
			Mean Spent p. Day	Average amount spent per day during the month before receiving the credit card
			Diversity-NE Value	Diversity of value spent over non-empty bins during the month prior to receiving the credit card
			Diversity-ALL Number	Diversity of number of purchases over all [illegible] bins during the month prior to receiving the credit card
			Loyalty-Number	Loyalty of number of purchases in the top three bins during the month prior to receiving a credit card
Calling Behavior	CB	72	Count IN	Total number of phone calls received during the three months of the social network
			Weekend Duration OUT	Aggregated duration of all phone calls made on weekends during the three months of the social network
			Tuesday Duration UD	Aggregated duration of all phone calls made and received on Tuesdays during the three months of the social network
Link-Based	LB	36	Binary (0) IN	Binary indicator of having neighbors with no late payments, in a network with incoming edges
			Binary (1) OUT	Binary indicator of having neighbors with one late payment, in a network with outgoing edges
			Binary (2) UD	Binary indicator of having neighbors with two late payments, in a network with undirected edges
			Binary (3) UD	Binary indicator of having neighbors with three late payments, in a network with undirected edges
			Count (0) IN	Number of neighbors with no late payments, in a network with incoming edges
			Count (1) OUT	Number of neighbors with one late payment, in a network with outgoing edges
			Count (2) OUT	Number of neighbors with two late payments, in a network with outgoing edges
			Count (3) UD	Number of neighbors with three late payments, in a network with undirected edges
Persona-lized PageRank	PR	54	Exposure (1) IN	Exposure score after applying PR on a network with incoming edges and delinquent customers with one or more late payments.
			Exposure (2) OUT	Exposure score after applying PR on a network with outgoing edges and delinquent customers with two or more late payments.
			Exposure (3) UD	Exposure score after applying PR on a network with undirected edges and delinquent customers with three or more late payments.
			Binary High Risk (1) IN	Binary indicator of having neighbors with high exposure scores after applying PR on a network with incoming edges and delinquent customers with one or more late payments.
			Binary High Risk (2) OUT	Binary indicator of having neighbors with high exposure scores after applying PR on a network with outgoing edges and delinquent customers with two or more late payments.
			Count High Risk (3) IN	Number of neighbors with high exposure scores after applying PR on a network with incoming edges and delinquent customers with three or more late payments.
Spreading Activation	SPA	54	Exposure (1) IN	Exposure score after applying SPA on a network with incoming edges and delinquent customers with one or more late payments.
			Exposure (2) OUT	Exposure score after applying SPA on a network with outgoing edges and delinquent customers with two or more late payments.
			Exposure (3) UD	Exposure score after applying SPA on a network with undirected edges and delinquent customers with three or more late payments.
			Binary High Risk (1) IN	Binary indicator of having neighbors with high exposure scores after applying SPA on a network with incoming edges and delinquent customers with one or more late payments.
			Count High Risk (1) UD	Number of neighbors with high exposure scores after applying SPA on a network with undirected edges and delinquent customers with one or more late payments.
			Count High Risk (3) IN	Number of neighbors with high exposure scores after applying SPA on a network with incoming edges and delinquent customers with three or more late payments.

Table 4: Statistical Model Performance (AUC).

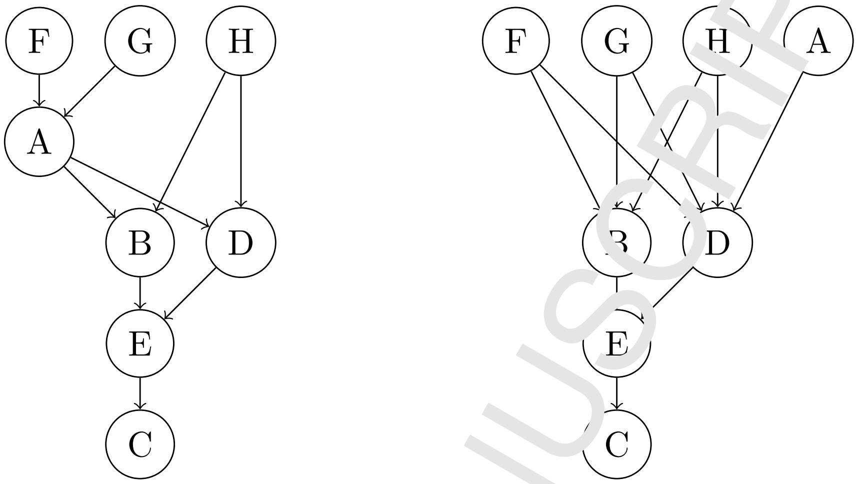
	Model		Classifier	
Model ID	Feature Groups	Logistic Regression	Decision Trees	Random Forest
A	SD	0.5869	0.7004	0.8993
B	CB	0.5351	0.704[illegible]	0.8700
C	LB	0.5485	0.7429	0.7697
D	PR	0.5163	0.7611	0.8339
E	SPA	0.5281	0.7188	0.8063
F	SD,CB	0.6115	0.712[illegible]	0.9227
G	CB,LB,PR,SPA	0.5182	0.7307	0.9154
H	SD,CB,LB,PR,SPA	0.6121	0.7263	0.9224

5. Results

The results are organized in three parts starting with empirical tests to establish the networks' relational dependency. Subsequently, the results of the proposed methodology are detailed, first in terms of statistical performance and then in terms of economic performance.

5.1. Homophily amongst Defaulters

A network is homophilic if nodes with a certain label are to a larger extent connected to other nodes with the same label. In the default networks, homophily is present if the fraction of edges between defaulters and non-defaulters is significantly smaller than the expected fraction of such edges in the network. A one-tailed proportion test with a normal approximation for homophily amongst defaulters resulted in a p-value of less than 0.0001, which means that there is evidence of homophily [44]. Furthermore, homophily in networks can also be measured with dyadicity and heterophilicity, that is, the connectedness between nodes with the same label and of different labels, respectively, compared to what is expected in a random network [44]. The networks analyzed here, have a dyadicity amongst defaulters of 0.8689 while the heterophilicity is 0.8137. This means that the networks are not dyadic, as defaulters are not more connected amongst themselves, but they are heterophilic, i.e., there are less connections between defaulters and non-defaulters. Based on these results, there is foundation for applying social network analytic techniques to predict default in the call networks.



(a) 95% Confidence

(b) 99% Confidence

Figure 4: Domination graphs.

5.2. Statistical Model Performance

Credit scoring models are built with the features in each feature group separately, as well as three models with a combination of feature groups, as seen in Table 4. The first five models *A*, *B*, *C*, *D* and *E* study the main effects of each feature group. Model *F* combines the sociodemographic features with the calling behavior features, model *G* includes all feature groups except the sociodemographic features and in model *H* we consider all feature groups. Other combinations of feature groups were tried, but they did not provide more significant results than the ones shown. As is common practice in credit scoring, statistical model performance is measured by the area under the receiver operating curve (AUC). The AUC summarizes the trade-off between model sensitivity and specificity in a single number between 0 and 1 with higher values meaning better performance.

From Table 4, it is clear that the performance with respect to the three classifiers varies substantially. Overall, the logistic regression models perform the worst, of which models including sociodemographic features (models *A*, *F*, *H*) perform best. Logistic regression models do not yield a better performance when using network-related features. This hints at a non-linear behavior that cannot be properly captured by a generalized linear model.

The random forests produce the best-performing models and the remaining discussion will therefore be focus on them. First, the test of DeLong et al. [45] is applied to the receiver operating curves

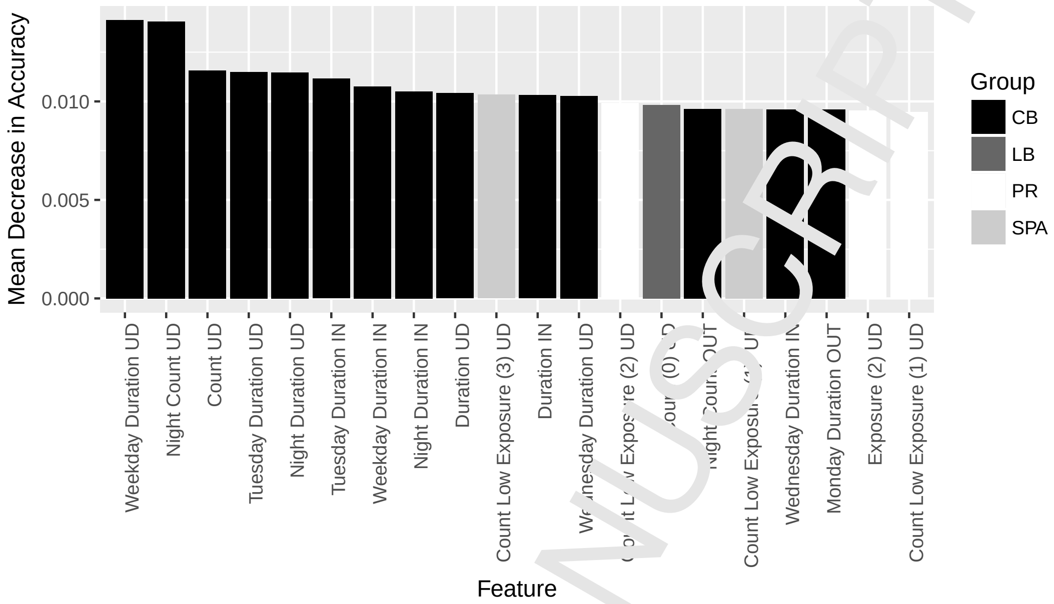
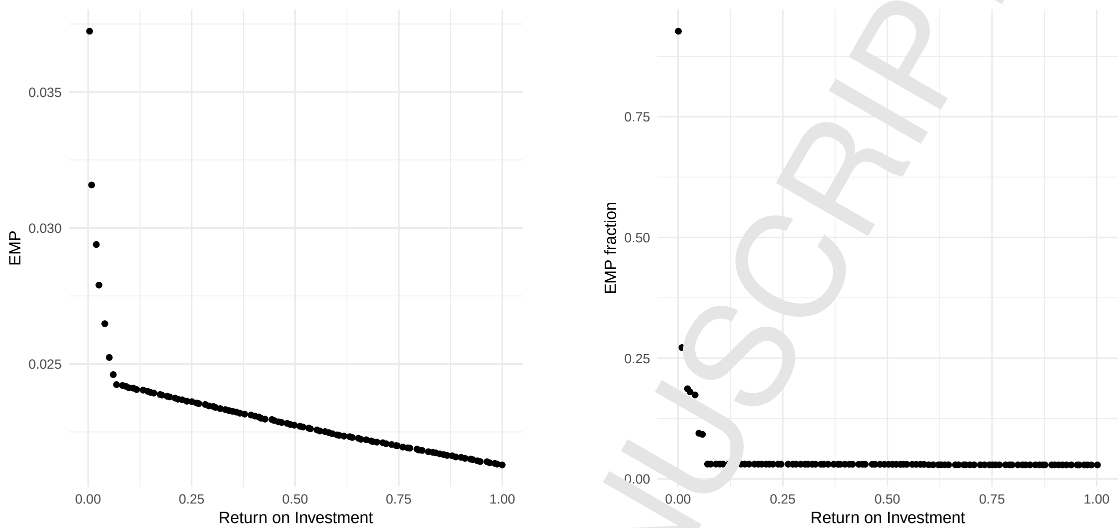


Figure 5: Feature importance: Mean decrease in accuracy.

(ROC) of each pair of random forest models to compare their performance. The results can be seen in the domination graphs in Figure 4. The best performing models are at the top and models that perform worse are lower down. The arrows indicate a significant improvement in statistical performance at 95% and 99% confidence level on the left and right, respectively. The figure on the left demonstrates that there is not a significant difference in the performance of the three models with a combination of features (F, G, H), but models with only one type of features (A, B, C, D, E) perform significantly worse with the link-based features (C) performing worst overall. Secondly, the importance of the features in model H is explored to determine their ability to predict default and rank the usefulness of the features. This is displayed in Figure 5 for the mean decrease in accuracy for the 20 most important variables. The mean decrease in accuracy of a particular feature measures how much the accuracy of the resulting model decreases when that feature is left out of the model, and as a result, gives a score of how important it is in the model. Figure 5 demonstrates that the calling behavior features are ranked the highest, followed by PR features and SPA features, and a single LB feature.

5.3. Economic Model Performance

The previous subsection showed that the statistical performance of more complex credit scoring models with a combination of feature groups is significantly better than models with only one feature



(a) EMP (b) EMP fraction

Figure 6: Sensitivity Analysis for ROI.

group, and even better than that of models with sociodemographic features alone. Here the economic performance of the models is evaluated and the importance of features in terms of profit by applying the EMP to the random forest models. For the EMP (see section 3.2) various parameters need to be specified. To compute the benefit of correctly identifying a defaulter, λ (see Equation 1), the credit card limit is used as the principal (*A*) and the drawn amount on the card at the time of default as exposure at default (*EAD*). The two remaining parameters: loss given default (*LGD*) and return on investment (*ROI*), are domain specific and not obtainable from the data directly. Therefore, an exploration of their effect on the EMP is provided. An analysis of the variation in EMP as a function of *LGD* shows substantial robustness, which means that the economic performance of the models does not greatly depend on *LGD*. Considering this, and based on expert judgement, this parameter is set to 0.8. In contrast, EMP decreases when *ROI* increases as is evident from Figure 6, which shows the EMP and its implied cutoff (EMP fraction) as a function of *ROI* when *LGD* is set at 0.8. The value for *ROI* is determined based on the 'elbow' in these figures and set to 0.05. The inflection point is the point where the *ROI* becomes the biggest influence (thus the linear behavior) and so it is appropriate to choose a value that balances profits for the rest of the analyses. Subsequently, the distribution of λ can be estimated, see Figure 7. As in Verbraken et al. [39] there are two peaks in the distribution, one at each end of the unit interval and with the assumption that λ follows a uniform distribution in

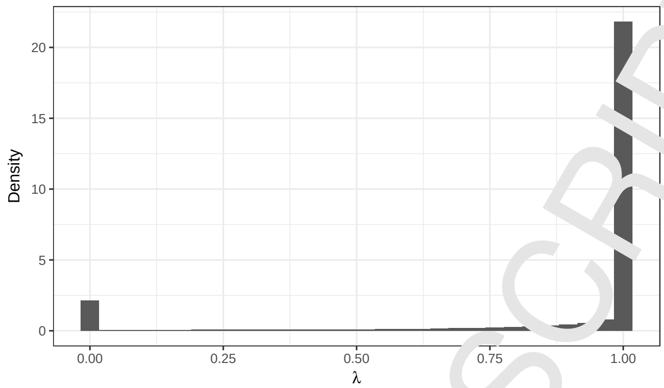
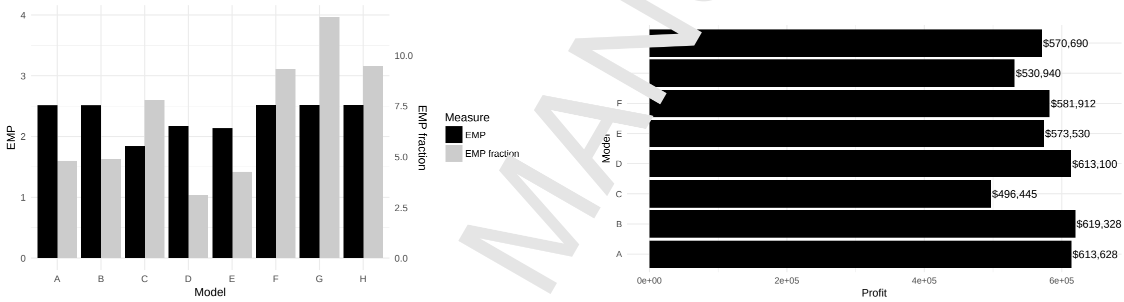


Figure 7: Distribution of λ.



(a) EMP and EMP fraction

(b) Profit

Figure 8: Economic Model Performance.

between. The peak at 0 represents credit card holders who have had payment arrears and have paid back fully, whereas the peak at 1 indicates those that never paid back their debt. This distribution is used to determine the values for p_0 and p_1, see Verbraken et al. [39].

With all parameters estimated, the next step is to compute the expected maximum profit, the profit maximizing fraction of rejected loans and the model profit (as described in Section 3.2.1) for the random forest models in Table 4. The results can be seen in Figures 8. The value for EMP is expressed as a percentage of the total loan amount and measures the incremental profit relative to not building a credit scoring model. The ranking of the values for the expected maximum profit is consistent with the ranking of the AUC values in Table 4, and again models A, F, G and H are considered best and C the worst. The EMP fraction values vary, however, and therefore so do the model profits. The profit maximizing fraction represents the fraction of credit card applications that should be rejected in order to obtain the maximum profit. The fact that the fraction for model G is

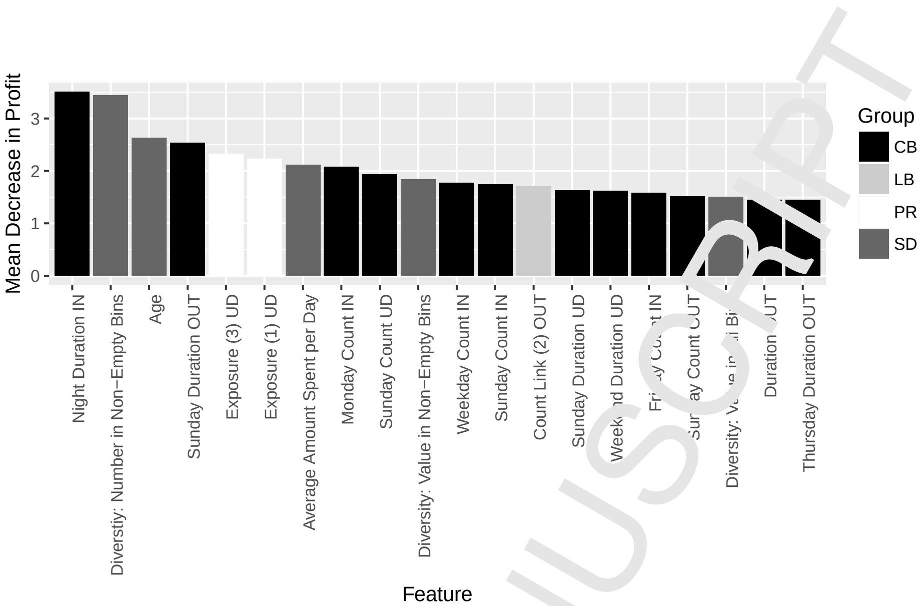


Figure 9: Feature Importance: Mean decrease in Profit.

so much higher than the rest with the profit remaining the same, indicates that the model focuses on the most profitable customers. Regarding the model profits, there is a substantial increase compared to not using a model. Model *B* has the highest profit, followed by models *A* and *D*. Models *F* and *H* also produce decent profits, whereas *G* does not, at least not when compared to the rest. Again, model *C* performs the worst. Of the models with only one data source, model *B* (built with calling behavior variables) brings the higher singular results. As no history is available for these borrowers, a possible explanation is that their socioeconomic standing can be deduced from their immediate network. Note however that this difference is marginal, as model *A* (sociodemographic variables) follows it. Of the combined models, models *F* (with both sociodemographic variables and calling behavior) and *H*, with all available variables, produce the best results in terms of profits.

Figure 9 shows the mean decrease in profit for the 20 most important features in model *H* computed using the technique described Section 3.2.2. This profit perspective shows more variation in groups of features than the statistical one. As for mean decrease in accuracy, more than half of the features are calling behavior features, but in contrast to Figure 5, a quarter of the features are sociodemographic features, which in this case are features that measure consumption. This is consistent with the result indicating that, of the combined models, model *F* was associated to a larger profit. To test for correlation among the ranking of features according to the two measures we computed the Spearman's ρ, Kendall's τ and Goodman and Kruskal's γ correlation coefficients. The resulting values did not indicate a correlation among the rankings.

It is interesting to see that having only network features allows us to discriminate potentially better customers. That would mean that we can look, using the network connections, beyond simple socioeconomic and sociodemographic traits, and actually profile more profitable customers. In the models, network features help discriminate different customers, who cannot be captured by common features, and it happens that these customers bring a lot of profit.

6. Discussion

Based on these results, the three research questions in section 1 can be addressed.

The first research question Q1 assesses the value that call data adds to credit scoring models in term of AUC and profit. For statistical performance, models including all features performed best, with the AUC value increasing by 0.023 points in the best model when compared to the sociodemographic model *A*. The economic performance of the models in terms of EMP, EMP fraction and profit can be seen in Figure 8. The model with the highest profit is model *B* and it is slightly better than the traditional model *A*. Models with a combination of feature groups (*F* ,*G* and *H*) produce lower profit but their EMP values are the highest. The reason for the lower profit is the high EMP fraction, which indicates that these models are more conservative and exclude a higher proportion of the defaulters. These results indicate that the CDR data complements the conventional data and there is added value when including the CDR data in credit scoring models and even when used without the traditional features.

The results also provide an answer to the second research question Q2: Can call data replace traditional data used for credit scoring? In terms of both statistical and economic performance, the results indicate that the predictive power of call data is just as good or might be even better than traditional data for these borrowers. This is clear from the high performance of model *B*. In addition, the importance of the calling behavior features shows that these are very predictive, much more so than the traditional features. This result demonstrates the merit of this research. Given the high predictive power of the call data, borrowers without enough bank information can benefit from the approach by giving access to their call records to obtain credit.

Finally, the last research question Q3 about how default behavior propagates in the network, can be addressed. The results of the homophily test in section 5.1 showed a lower fraction of connections amongst defaulters and also between defaulters and non-defaulters. This might partially be a consequence of the low number of defaulters in the network overall. Furthermore, insights about the propagation of default behavior can be obtained from the importance of the features. Firstly, for mean decrease in accuracy, see Figure 5, a few PageRank and Spreading Activation features are important. These are predominantly 'Count Low Exposure' features which represent the number of neighbors with low exposure score. This indicates that not having a high risk neighbor is predictive of non-default. The PageRank feature 'Exposure (2)' is also among the 10 most important features. It represents the PageRank exposure score when the influence comes from delinquent customers with two or more late payments and indicates a propagation effect of default behavior. Second, for the mean decrease in profit in Figure 9 there are two PageRank exposure scores, based on one and three late payments of delinquent customers. From these observations we can say that, in terms of propagation of default influence, Personalized PageRank is more effective than Spreading Activation. A more thorough analysis of how default propagates is needed to better understand the effect of each of these features.

7. Impact of Research

The research findings presented in this paper have possible impact at various levels. This section identifies three different levels and provides a discussion of the implications of each one.

7.1. Regulatory Impact

The Basel Accords model unexpected losses using a Merton single-factor model where the asset value of an obligor depends upon a systematic (e.g., the macroeconomy) and an idiosyncratic (e.g., obligor-specific) risk component [2]. Asset correlations are then also factored in to see how default behavior is correlated and, as such, model system risk. A key concern relates to the exact values of these asset correlations. For corporates, the assets can be quantified by inspecting balance sheets, and various financial models have been introduced to quantify corporate asset correlations.

For retail exposures (e.g., credit cards, mortgages, installment loans), it becomes considerably more difficult as the assets are less tangible. Retail asset correlations have been specified in the Basel Accords using some empirical, but not published, procedure reflecting a combination of supervisory judgment and empirical evidence. As such, they are fixed at 4% for qualifying revolving exposures (e.g., credit cards) and 15% for mortgages. Given their impact on capital calculation, it would be desirable that these asset correlations are sustained by a solid theoretical framework and accompanying empirical validation. In this research, we illustrated how default behavior on credit cards propagates in a call network. These insights pave the way for additional research aimed at quantifying asset and default correlation for retail exposures in a more sound and solid way. This can then lead to better regulatory asset correlation values which in turn leads to a better protection of the financial system.

7.2. Financial Inclusion

The results may also have a societal impact that affect borrowers in developed and developing countries in different ways. In the former case, people who are joining the financial market for the first time, such as young people and immigrants, face trouble when applying for loans because they do not have a credit history. Instead, they need to spend time and effort to build their credit history before financial institutions can assess whether they are creditworthy.

In developing countries where historical financial data is often nonexistent, the impact is even greater. As reported by the World Bank, over two billion adults worldwide do not have a basic account which makes up more than 20% of the adult population in some countries [46]. The benefits of behavior-based microfinance in these countries are evident, as having access to small credits has a social impact on communities, helping to fight poverty and enhancing economic development [47]. In contrast to the lack of banking history, the high(er) availability of call data in these countries provides an alternative for credit scoring, thereby facilitating credit access to a wider segment of the population. According to the results, features extracted from these untraditional data sources are good predictors of credit behavior (e.g., models B, G and H). In addition, the numerous smartphone applications that are already being deployed in some developing countries are a prime example of the success of these methods. They offer immediate small loans, that are repaid within a short period of between three weeks and six months and have lower interest rates, ranging between 6% and 12% as opposed to the 25% interest rate in traditional microlending [25].

7.3. Privacy and Ethical Concerns

The results of this study are furthermore affected by privacy regulations because the implementation of some of the models depends on different parties sharing the data. Since there are no worldwide applicable standards for data-sharing of that kind, we illustrate how this might occur by studying the reality of the US and the EU in what follows.

In the US, there is no single federal law regulating data transfer between affiliates. The transfer of financial information between a bank and a telco is protected under the Gramm-Leach-Bliley Act

[48]. This legislation allows transfer of personally identifiable information originating from a financial service provider to a third party if the parties design a contract that disallows disclosure and use of information outside the project. In general, such a contractual framework should satisfy most other pieces of regulation that might indirectly apply to the sharing of data in the other direction (i.e., from the telco to the bank or credit bureau).

In the European Union, in contrast, there is a strong body of legislation regulating data sharing. Given that CDR data are a form of communication, and the objective of the model is to process it along with banking data in an automated way, two pieces of legislation apply: Regulation 2016/679 [49] regarding the protection of privacy for natural persons, best known as the General Data Protection Regulation (GDPR), and the "ePrivacy Regulation" [50] regarding the processing of personal data in the electronics communications sector.

The ePrivacy Regulation deals with if, and how, communications data at a disaggregated level can be used. Article 30 in particular mandates that a service such as a financial score, which is not only for billing or providing the mobile service, is a "value added service", and thus requires explicit authorization from the user. This authorization might be given in the contract, for example, or ex post to the signing of the contract via electronic authorization. In case none of these provisions can be set in place, then the sharing of CDR data cannot occur unless the data is anonymized.

The key challenge is how to make the data available to the other party, so defaulters can be correctly identified. Fortunately, there are methods that can provide privacy-preserving data linkage [51] that can be followed in order to join the data securely without compromising the individual on either side of the sharing process. Methods such as Privacy Preserving Probabilistic Record Linkage [P3RL 52], that are in use in the medical sciences, allow secure data-sharing between partners. The secure transfer of data is also very simple to satisfy, following proper encryption and secure access protocols. The GDPR has additional provisions on data storage, forcing companies to store data only for the time necessary to provide the service, so the party receiving the linked data only for the purposes of model development must ensure proper disposal of the data after the model development. Finally, note that the model itself is considered aggregated data. Article 22 of the GDPR allows the safe use of statistical models when these are required to establish a contract with the counterparty (the financial institution), which is the case when a loan is granted.

There is as well an ethical concern in using data that depends on the social network of the borrowers to restrict funding to them. This is of course not a practice that should be recommended from the results of this model, as it would constitute unfair discrimination. However, when borrowers do not have any past behavior information that allows institutions to make a decision, or they have not accumulated enough additional information to profile them correctly, then CDR information can clearly contribute to increase financial inclusion. Thus, we propose that the use of this data be done in strict positive terms. This can be easily done when constructing a credit score: it is common practice to discretize continuous variables and give a score based on the Weight of Evidence for each of the segments [2]. An ethical use of this information would simply assign the neutral score to those segment which would unfairly punish the borrowers, leaving the positive segments that would provide easier access to funds.

8. Conclusion

This study presents the statistical and economic advantages of exploiting Big Data and social network analytics for credit scoring applications. We use phone call logs are used to build call networks and social network analytics applied to enhance the performance of models that predict creditworthiness of credit cards applicants. We do this from both a statistical and profit perspective and demonstrate how incorporating telco data has the potential of increasing the Value of credit scoring models. Furthermore, we identify which features are most important for this predictive task, both in terms of statistical performance and profit. According to the results, models that are built with features that represent calling behavior perform best, both when performance is measured in AUC and profit. We also show that these features dominate other features in terms of importance. This is an interesting result because it means that how people use their phones can be used as the sole data source when deciding whether they should be given a loan or not. Thus we propose that the data should be used in strict positive terms, to facilitate financial inclusion for people that lack enough information for correct profiling.

The main limitation of our this is the data itself. The scorecards that were built are for the applications of credit cards, and it is unclear how the results would generalize for other types of credits such as microloans or mortgages. In industry, numerous applications for granting microloans via smartphones by analyzing user's behavior exist. According to various reports, behavioral features are important in these applications as well, but that is difficult to verify without published scientific results. Similar data could be obtained from peer-to-peer lending platforms, or through agreements between telcos and banks/credit bureaus, where there is access to both default status of users as well as behavioral features. Behavioral data similar to the mobile phone data shown in this work could also be gathered from social media platforms such as Twitter. The data in this study originates from a single country where a telco and a bank have a special agreement to share the data. Therefore, an analysis of similar data from other countries or data for other types of credits would strengthen the external validity of the presented results. In practice, lenders use credit bureau variables, such as FICO scores, when assessing creditworthiness, and unfortunately they were not available for these analyses, but would be an interesting extension of our work.

It is already clear that the mobile phone data used in this study is big in the sense of 'Volume', 'Velocity', 'Veracity' and 'Variety'. Our analysis of the data and the resulting well-performing models show that it also has a positive effect for financial inclusion and on model profit, and as such is also important for 'Value': the fifth V of Big Data!

Acknowledgments

The authors would like to thank Ariel Berenstein for his contribution to this research.

the authors acknowledge the support of a large Belgian bank that wishes to remain anonymous.

References

[1] Thomas LC. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International journal of forecasting 2000;16(2):149–72.

[2] Baesens B, Scheule H, Rösch D. Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. John Wiley & Sons; 2016.

[3] Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance 1968;23(4):589–609.

[4] Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J. Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the operational research society 2003;54(6):627–35.

[5] Lessmann S, Baesens B, Seow HV, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. Eur J Oper Res 2015;247(1):124–36.

[6] World Bank. General principles for credit reporting. 2011. URL: http://documents.worldbank.org/curated/en/662161468147557554/pdf/70193-2014-CR-General-Principles-Web-Ready.pdf.

[7] Barron JM, Staten M. The value of comprehensive credit reports: Lessons from the us experience. Credit reporting systems and the international economy 2003;8:273–310.

[8] Agarwal R, Dhar V. Editorial—-Big data, data science, and analytics: The opportunity and challenge for IS research. 2014.

[9] Baesens B, Bapna R, Marsden JR, Vanthienen J, Zhao JL. Transformational issues of big data and analytics in networked business. MIS quarterly 2014;38(2):629–31.

[10] Sundararajan A, Provost F, Oestreicher-Singer G, Aral S. Research commentary —- information in digital, economic, and social networks. Inform Systems Res 2013;24(4):883–905.

[11] Newman M. Networks: an introduction. OUP Oxford; 2010.

[12] Lee GM, Qiu L, Whinston AB. A friend like me: modeling network formation in a location-based social network. Journal of Management Information Systems 2016;33(4):1008–33.

[13] Aral S, Walker D. Tie strength, embeddedness, and social influence: A large-scale networked experiment. Management Sci 2014;60(6):1352–70.

[14] Macskassy SA, Provost F. Classification in networked data: A toolkit and a univariate case study. Journal of Machine Learning Research 2007;8(May):935–83.

[15] Verbeke W, Martens D, Baesens B. Social network analysis for customer churn prediction. Applied Soft Computing 2014;14:431–46.

[16] Van Vlasselaer V, Bravo C, Caelen O, Eliassi-Rad T, Akoglu L, Snoeck M, et al. Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems 2015;75:38–48.

[17] Gordy MB. A risk-factor model foundation for ratings-based bank capital rules. Journal of financial intermediation 2003;12(3):199–232.

[18] Calabrese R, Andreeva G, Ansell J. "Birds of a feather" fail together: Exploring the nature of dependency in SME defaults. Risk Analysis 2017;.

[19] Lin M, Prabhala NR, Viswanathan S. Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending. Management Sci 2013;59(1):17–35.

[20] Freedman S, Jin GZ. The information value of online social networks: lessons from peer-to-peer lending. Tech. Rep.; National Bureau of Economic Research; 2014.

[21] Wei Y, Yildirim P, Van den Bulte C, Dellarocas C. Credit scoring with social network data. Marketing Sci 2015;35(2):234–58.

[22] De Cnudde S, Moeyersoms J, Stankova M, Tobback E, Javaly V, Martens D. What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance. Journal of the Operational Research Society 2018;:1–11.

[23] Kharif O. No credit history? no problem. lenders are looking at your phone data. 2016. URL: `https://www.bloomberg.com/news/articles/2016-11-25/no-credit-history-no-problem-lenders-now-peering-at-phone-data`.

[24] Chin J, Wong G. China's new tool for social control: A credit rating for everything. The Wall Street Journal 2016;.

[25] Dwoskin E. Lending startups look at borrowers' phone usage to assess creditworthiness. The Wall Street Journal 2015;.

[26] Naboulsi D, Fiore M, Ribot S, Stanica R. Large-scale mobile traffic analysis: a survey. IEEE Communications Surveys & Tutorials 2016;18(1):124–61.

[27] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008;2008(10):P10008.

[28] Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA. Geographic constraints on social network groups. PLoS one 2011;6(4):e16939.

[29] Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. Science 2012;338(6104):267–70.

[30] Sarraute C, Blanc P, Burroni J. A study of age and gender seen through mobile phone usage patterns in mexico. In: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE; 2014, p. 836–43.

[31] Leo Y, Fleury E, Alvarez-Hamelin JI, Sarraute C, Karsai M. Socioeconomic correlations and stratification in social-communication networks. Journal of The Royal Society Interface 2016;13(125):20160598.

[32] Haenlein M. A social network analysis of customer-level revenue distribution. Marketing letters 2011;22(1):15–29.

[33] Van Vlasselaer V, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B. Gotcha! network-based fraud detection for social security fraud. Management Science 2016;63(9):3090–110.

[34] Lu Q, Getoor L. Link-based classification. In: ICML; vol. 3. 2003, p. 496–503.

[35] Óskarsdóttir M, Bravo C, Verbeke W, Sarraute C, Baesens B, Vanthienen J. Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. Expert Systems with Applications 2017;85:204–20.

[36] Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Tech. Rep.; Stanford InfoLab; 1999.

[37] Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, et al. Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th international conference on Extending database technology: Advances in database technology. ACM; 2008, p. 668–77.

[38] Backiel A, Verbinnen Y, Baesens B, Claeskens G. Combining local and social network classifiers to improve churn prediction. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ACM; 2015, p. 651–8.

[39] Verbraken T, Bravo C, Weber R, Baesens B. Development and application of consumer credit scoring models using profit-based classification measures. Eur J Oper Res 2014;238(2):505–13.

[40] Verbraken T, Verbeke W, Baesens B. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE transactions on knowledge and data engineering 2013;25(5):961–73.

[41] Bravo C, Maldonado S, Weber R. Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. European Journal of Operational Research 2013;227(2):358–66.

[42] Singh VK, Bozkaya B, Pentland A. Money walks: implicit mobility behavior and financial well-being. PloS one 2015;10(8):e0136628.

[43] Baesens B. Analytics in a big data world: The essential guide to data science and its applications. John Wiley & Sons; 2014.

[44] Baesens B, Van Vlasselaer V, Verbeke W. Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons; 2015.

[45] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;:837–45.

[46] World Bank . Financial inclusion. 2017. URL: `http://www.worldbank.org/en/topic/financialinclusion/overview#1`.

[47] Copestake J. Mainstreaming microfinance: Social performance management or mission drift? World Development 2007;35(10):1721–38.

[48] Cuaresma JC. The Gramm-Leach-Bliley Act. Berkeley Technology Law Journal 2002;17(1):497–517.

[49] European Union . 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). Legislation; 2016.

[50] European Union . Directive 2002/58/ec concerning the processing of personal data and the protection of privacy in the electronic communications sector. Legislation; 2002.

[51] Clifton C, Kantarcioğlu M, Doan A, Schadow G, Vaidya J, Elmagarmid A, et al. Privacy-preserving data integration and sharing. In: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM; 2004, p. 19–26.

[52] Schmidlin K, Clough-Gorr KM, Spoerri A. Privacy preserving probabilistic record linkage (p3rl): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC medical research methodology 2015;15(1):46.

Vitae

María Óskarsdóttir is a PhD candidate at the Faculty of Economics and Business (FEB) at KU Leuven. She has obtained a Master degree in mathematics at Leibniz Universität Hannover. Her current research topics include social network analytics and analysis of mobile phone data with applications in churn prediction and credit scoring. She has published papers in international peer-reviewed journals like Expert Systems with Applications and Business & Information Systems Engineering. She will complete here doctoral studies in September 2018.

Dr. Cristián Bravo is Associate Professor of Business Analytics at Department of Decision Analytics and Risk, University of Southampton. His research focus on the development and application of data science methodologies in the context of credit risk analytics, in areas such as deep learning, text analytics, and social network analysis. He also serves as editorial board member in Applied Soft Computing and in the new Journal of Business Analytics of the UK's OR Society. He is the co-author of the book "Profit Driven Business Analytics", published by Wiley in English and Chinese.

Carlos Sarraute holds a M.Sc. degree in Mathematics from the University of Buenos Aires and a Ph.D. in Informatics Engineering from ITBA. He currently is director of research at Grandata, where he works in understanding and predicting Human Dynamics, by analyzing big data coming from the mobile phone and banking industries. He participates in scientific collaborations with MIT, City College of New York, ENS Lyon and UBA among others. He has presented his work in data science and social network analysis conferences, such as ASONAM, KDD and NetMob. He has published in several journals including Nature Communications.

Jan Vanthienen received his PhD degree in Applied Economics from KU Leuven, Belgium. He is a full professor of Information Systems at the Department of Decision Sciences and Information Management, KU Leuven and (co-)authored more than 200 full papers in international journals and conference proceedings. His research interests include information and knowledge management, business rules, decisions and processes. and business analysis and analytics. He received an IBM Faculty Award on smart decisions, and the Belgian Francqui Chair at FUNDP. Currently he is department chair at the Department of Decision Sciences and Information Management of KU Leuven.

Professor Bart Baesens is a professor of Big Data and Analytics at KU Leuven (Belgium) and a lecturer at the University of Southampton (United Kingdom). He has done extensive research on Big Data & Analytics, Credit Risk Modeling, Fraud Detection and Marketing Analytics. He has written more than 200 scientific papers, presented at top international conferences and co-authored 9 books. He has received various best paper and best speaker awards.

ACCEPTED MANUSCRIPT

***Highlights (for review)**

- A novel approach for credit scoring using mobile phone data and social networks
- Implies enhanced financial inclusion in the context of positive credit information
- Incorporating mobile phone data increases statistical performance
- The best model in terms of profit includes only calling behavior features
- Individual calling behavior is most predictive of creditworthiness