What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance

What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance

Abstract

Microfinance has known a large increase in popularity, yet the scoring of such credit still remains a difficult challenge. Credit scoring traditionally uses sociodemographic and credit data, which we complement in an innovative manner with data from Facebook. A distinction is made between the relationships that the available data imply: (1) LALs are persons who resemble one another in some manner, (2) friends have a clearly articulated friendship relationship on Facebook, and (3) BFFs are friends that interact with one another. Our analyses show two interesting conclusions for this emerging application: the BFFs have a higher predictive value then the person's friends and secondly, the interest-based data that define LALs, yield better results than the social network data. Moreover, the model built on interest data is not significantly worse than the model that uses all available data, hence demonstrating the potential of Facebook data in a microfinance setting.

Keywords: Data mining, Decision support systems, Microcredit, Credit scoring, Networks and graphs, Default prediction

1 Introduction

"The first thing [in credit] is character.
Before money or property or anything else."
– J.P. Morgan

In microfinance, where credit history data is often lacking, character is considered an important predictor for loan repayment (Schreiner 2003). Manual screening of the applicants by the loan officer is used to gather information about their trustworthiness. Though effective, this is a timely and costly process. Attempts to replace the credit screening process with automated credit scoring have shown that the use of traditional socio-demographic and credit data is insufficient (Schreiner 2000; Van Gool et al. 2012). These types of data are unable to capture the unwillingness to repay the loan, one of the main causes of low repayment rates. Microfinance comes with a social mission of alleviating poverty, enhancing economic development and achieving social impact in the community (Copestake 2007). The creditworthiness decisions should be in line with this social mission. Investing in improved credit scoring models helps microfinance lenders to distinguish the risky population from the target population.

We obtained data from Lenddo, a company specialized in social authentication and scoring technology [1]. Lenddo uses alternative data to provide credit scoring and verification for the emerging middle class in developing markets. The company has developed patented technology to collect and process billions of data points, and uses advanced machine learning techniques to build predictive algorithms. Lenddo has multiple algorithms which draw upon a wide array of data from Facebook, Twitter, LinkedIn, Gmail, Yahoo, Android, IOS, machine fingerprinting, etc. Its LenddoScore product is currently being used by banks and lending institutions worldwide to reduce risk, reach new customers and improve customer service. Lenddo's technology is designed to service thin-file and new-to-credit consumers, such as the upcoming middle class who is "underbanked" and in need of small loans and other financial services. The borrowers often lack an established credit history, making commercial banks reluctant to grant them credit but are often active users of social networks, enabling Lenddo to provide unique insights about their creditworthiness. For the purpose of this paper, only a small anonymised subset of Lenddo's data was shared and analysed. The analysis and methodology presented in this article are similar in concept to the approaches used by Lenddo, however they do not describe any of the algorithms and scoring solutions currently or previously used by Lenddo in its business.

The predictive modelling task that we consider is identifying the risky loan applicants that would not fully repay their loan. For the analysis, we use data from Facebook [2] and categorise it as follows: socio-demographic data, interest data and social network data. The socio-demographic data includes traditional features such as age, place of residence and education level. The interest data captures fine-grained data related to for example the pages a user likes or the companies he worked for. Finally, the social network data consists of friendship connections between borrowers on Facebook. We use and combine this data in an innovative manner for credit scoring purposes as these define different relationships: look-a-likes, friends and BFFs (see Fig. 1).

[1] http://partners.lenddo.com

[2] The data was obtained through informed consent by the users: when downloading the app, the users need to opt-in for providing the available Facebook data, meaning that the customer gave explicit permission for Lenddo to fetch and use the data.

Look-a-likes (LAL) refer to people that are similar to one another. In this case this can be interpreted as persons either demonstrating similarities regarding certain sociodemographic characteristics, liking the same pages on Facebook, having a Facebook-friend in common or commenting on the same status. Clearly, this does not say anything about any real connections between those persons. That is, these individuals are not necessarily connected in real life, in fact they most likely have never met at all. However, the information included in these similarities can be an important predictor for default behavior since similar behavior in one domain (e.g. preferences) might imply similarities in other domains (e.g. default) as well (Martens and Provost 2011; Moeyersoms and Martens 2015; Provost, Martens, and Murray 2015; Raeder et al. 2012). Additional Facebook data is available as explicitly stated *Friends*. The last category of data implies relationships of the form *Best Friends Forever (BFFs)*. These are Facebook friends that interact with one another, be it being tagged together in a picture, commenting on each others' status, etc. Note that we do not distinguish in strength regarding the BFF relation, i.e. two persons are considered BFFs both in the case where one interaction occurs and in the case where multiple interactions are recorded.

The contributions of this paper are three-fold, as illustrated in Fig. 1. To our knowledge, we are the first to investigate the use of Facebook data for credit scoring for microfinance. The potential of such an automated credit scoring process is innovative and has large implications for the widespread use of microfinance and the potential economic growth of developing countries. Secondly, whereas previous studies that use Facebook data for predictive modeling focus on either the social network data or the interest data, we explicitly assess the combination of both. Finally, within the area of social network Facebook data, we further investigate the difference in predictive power of different levels of closeness, i.e. friends versus BFFs.

2 Related Work

2.1 Credit scoring for microfinance

Up to now, the use of interest-based and social network Facebook data to predict creditworthiness has not been investigated. Research on credit scoring mainly focuses on the use of structured data, such as sociodemographic factors (Banasik, Crook, and Thomas 2003; Hand, Sohn, and Kim 2005) and balance sheets (Emel et al. 2003; Min and Jeong 2009), thereby ignoring the high-quality information available in other data formats. In microfinance, the applicant's selection is often judgmental, i.e. the loan officer assesses the risk based on its own prior experience, his opinion on the applicant and the loan conditions (Schreiner 2003). In many cases the loan officer communicates with the local community of the client to get an idea about the client's trustworthiness (Morduch 1999). In literature, this type of lending is called relationship-based lending where the lender gains information about the borrower during the course of their relationship. A second type of microfinance lending is group-based lending, in which social capital is created and used to alleviate the problem of asymmetric information and moral hazard (Hermes and Lensink 2007). Social capital - defined by Putnam (Putnam 1995) as "features of social organization such as networks, norms, and social trust that facilitate cooperation and coordination" - operates under the form of peer-pressure in these joint liability groups.

Research on microfinance credit scoring is limited. Zeller (Zeller 1998) and Sharma and Zeller (Sharma and Zeller 1997) used group, community and lender or program characteristics to describe credit risk of joint liability groups. Schreiner (Schreiner 2003) remarked that statistical scoring will probably not work well for group-based lending, since there is no data on individual risk. Group risk appears to be much less strongly linked to group characteristics than individual risk to individual characteristics. Van Gool et al (Van Gool et al. 2012) investigated whether traditional credit scoring is applicable to microfinance lending. Using borrower, loan and lender characteristics they built a credit scoring model for a Bosnian microlender. They found that their credit scoring models are not able to fully replace the traditional credit process of manual screening. These findings confirm the conclusion of Schreiner (Schreiner 2000) whose study revealed that automated credit scoring complements, but does not replace the judgment of a loan officer based on qualitative, informal knowledge about the character of the applicant.

What the above mentioned studies have in common, is that they only use structured data in their credit scoring models. This data includes loan characteristics (purpose of the loan, duration of the loan), borrower characteristics (age, gender, education) and credit history (repayment of previous loans) and therefore does not differ much from the credit scoring models used in traditional lending. The complex nature of microfinance necessitates an assessment of character. Schreiner (Schreiner 2003) advises microlenders to search for personal character traits that are predictive of repayment behavior. Recently, Wei et al (Wei et al. 2014) showed in a theoretical framework how network data can improve the accuracy of customer credit scores. Their framework is based upon the assumption of homophily, the notion that linked entities are more likely to have the same characteristics.

Furthermore, Facebook has patented technology to assess creditworthiness of users based on credit ratings of people present in the users' social network (Facebook Inc 2014). Although not deployed yet, Facebook's interest in this data corroborates the possible value that lies in the use of alternative data for credit scoring purposes.

2.2 Interest-based vs social network data

Different types of data are commonly used for predictive modeling in a retail setting (Van Gestel, Baesens, and Martens 2015). Except for the conventional sociodemographic data, social network and interest data can be considered as well. Social network data represents real relationships between customers, while interest data refers to the often fine-grained observed interests and preferences of persons.

A seminal paper that uses social network data is that of Hill et al. (Hill, Provost, and Volinsky 2006), which uses the social relationships observed in calling behavior to predict

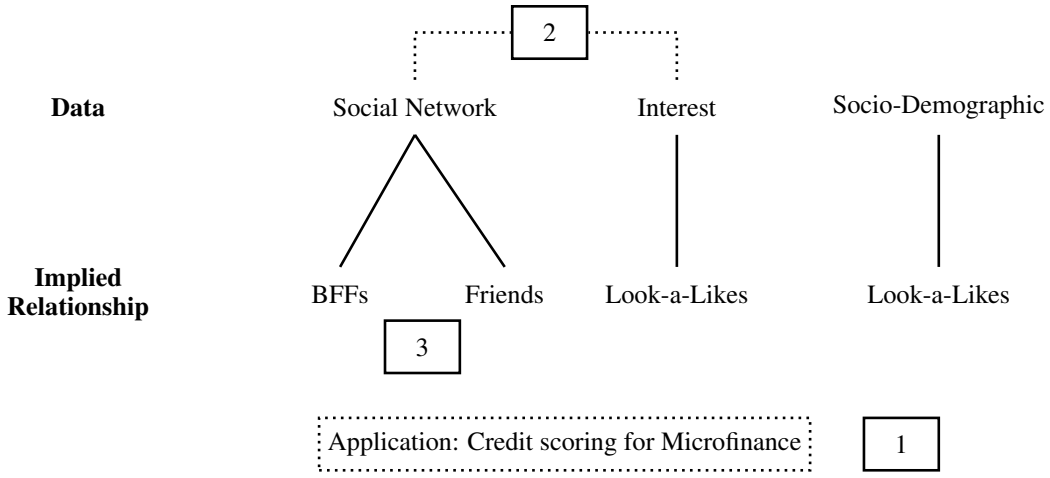

Figure 1: Contributions.

product/service adoption in a telecommunications setting. Other studies have looked at call behavior as well to predict churn (Verbeke, Martens, and Baesens 2014) and social network data for viral marketing (Domingos 2005). However, often no real network data is available and other characteristics which are beyond the traditional socio-demographics data, can be used to detect similarities between people. For instance, Kosinski et al. (Kosinski, Stillwell, and Graepel 2013) and Junque de Fortuny et al. (Junqué de Fortuny, Martens, and Provost 2013) looked at predicting different personality traits from a dataset of users liking Facebook pages. The studies of Goel et al. (Goel, Hofman, and Sirer 2012) and Hu et al. (Hu et al. 2007) predict demographic attributes and Raeder et al. (Raeder et al. 2012) predict brand interest from people's browsing history. Weber et al. (Weber, Garimella, and Borra 2013) reveal political views from history of videos watched on YouTube. For financial applications, Martens et al. (Martens and Provost 2011) predict interest in financial products from transactional datasets of consumers making payments to merchants and Provost et al. (Provost, Martens, and Murray 2015) consider geolocation data to connect people if they visited the same places with the goal of predicting brand interest.

To the best of our knowledge, no study has included both social network and fine-grained interest-based data in order to predict default in microfinance settings. In this work, both data types are combined so that potential differences in predictive power between the data sources can be observed.

3 Data

A balanced sample is made available to us of 4,985 loan applications made by 4,512 users. As stated previously and visualized in Fig. 1, we employ three data categories which we use to distinguish three levels of relations in terms of look-a-likes, friends and BFFs. We use Fig. 2 to illustrate these. Note that any names or personally identifiable information shown in this paper are fictitious and do not relate to names or information of actual Lenddo members.

Figure 2: Illustration of look-a-likes, friends and BFFs.

Table 1: Overview of the constructed data matrices indicating when people are connected in the network, the category of the resulting relation, the number of features (m), the number of active elements (N) and the sparsity (ρ) defined as $\rho = N/(m \times n)$, with n the number of data instances ($n = 4,985$).

Name	Represented data	Category	m	N	ρ
Sociodemo	Socio-demographic attributes of a person	Socio-demographic data	29	111,989	83 %
LAL_Likes_Item	Persons liking a page on Facebook	Interest-based LAL	48,701	127,241	0.052%
LAL_Likescat_Item	Persons liking a category of a page on Facebook	Interest-based LAL	238	53,441	4.504%
LAL_Groups_Item	Persons joined in a group on Facebook	Interest-based LAL	38,037	55,399	0.029%
LAL_Education_Item	Persons going to specific educational institutions	Interest-based LAL	4,620	11,015	0.048%
LAL_Employers_Item	Persons working for employers	Interest-based LAL	5,190	13,173	0.051%
LAL_Position_Item	Persons holding employment positions or business titles	Interest-based LAL	3,393	9,983	0.059%
LAL_Comments_Items	Persons commenting on a status	Relational LAL	2,141,630	1,763,453	0.017%
LAL_Photos_Items	Persons mentioned in a picture	Relational LAL	293,155	404,896	0.028%
LAL_Links_Items	Persons mentioned in a link	Relational LAL	297,410	407,358	0.028%
LAL_Status_Items	Persons mentioned in a status	Relational LAL	667,298	806,411	0.024%
LAL_Videos_Items	Persons mentioned in a video	Relational LAL	27,442	33,602	0.024%
LAL_Likes_Items	Persons liking an item (video/status/photo/comment)	Relational LAL	4,122,418	2,846,613	0.014%
LAL_Comments_All	Persons giving/receiving comments to/from each other	Relational LAL	896,164	1,217,744	0.027%
LAL_Photos_All	Persons mentioning one another in one of their photos	Relational LAL	731,574	235,645	0.007%
LAL_Links_All	Persons mentioning one another in one of their links	Relational LAL	2,627,614	1,051,770	0.008%
LAL_Status_All	Persons mentioning one another in one of their statuses	Relational LAL	630,749	490,942	0.016%
LAL_Videos_All	Persons mentioning one another in one of their videos	Relational LAL	46,078	30,899	0.014%
LAL_Likes_All	Persons liking each other's video/status/photo/comment	Relational LAL	1,817,619	2,692,752	0.030%
LAL_Comments_Borrowers	Borrowers giving/receiving comments to/from each other	Relational LAL	4,985	20,301	0.081%
LAL_Photos_Borrowers	Borrowers mentioning one another in one of their photos	Relational LAL	4,985	9,199	0.037%
LAL_Links_Borrowers	Borrowers mentioning one another in one of their links	Relational LAL	4,985	14,318	0.057%
LAL_Status_Borrowers	Borrowers mentioning one another in one of their statuses	Relational LAL	4,985	9,949	0.040%
LAL_Videos_Borrowers	Borrowers mentioning one another in one of their videos	Relational LAL	4,985	1,496	0.006%
LAL_Likes_Borrowers	Borrowers liking each other's video/status/photo/comment	Relational LAL	4,985	29,814	0.120%
FRI_FBFriends	Borrowers befriending one another	Friends	4,985	30,347	0.122%
BFF_Comments	Friends giving/receiving comments to/from one another	BFF	4,985	18,391	0.074%
BFF_Photos	Friends mentioning one another in one of their photos	BFF	4,985	8,609	0.035%
BFF_Links	Friends mentioning one another in one of their links	BFF	4,985	13,072	0.053%
BFF_Status	Friends mentioning one another in one of their statuses	BFF	4,985	9,469	0.038%
BFF_Videos	Friends mentioning one another in one of their videos	BFF	4,985	1,438	0.006%
BFF_Likes	Friends liking each other's video/status/photo/comment	BFF	4,985	22,606	0.091%
BFF_All	Friends having any kind of interaction	BFF	4,985	24,243	0.098%

Table 1 shows a list of the constructed data structures, to which we will refer in the following subsections, along with some relevant data characteristics.

3.1 Socio-demographic data

The socio-demographic data originates from mandatory and optional information the user provides both Lenddo and Facebook. Such variables include date of birth, hometown, religion and school level. A total of 29 socio-demographic characteristics are used in the constructed Sociodemo matrix. The number of missing values is approximately 16.65%. Note that a missing value might denote data intentionally left blank by users, which is also modeled in the input data.

3.2 Interest data

In addition to traditionally available socio-demographic characteristics, we also have fine-grained interest characteristics, which let us determine look-a-likes.

First, there are interests which manifest themselves immediately which we use to define interest-based look-a-likes (Interest-based LALs in Table 1). Liking a Facebook page or joining a Facebook group are direct testimonies of an interest. We also use schools visited, employers worked for and employment positions held to define an interest. Note that borrowers are not required to provide this information. In Fig. 2, both Scout and Jane like the page of Harper Lee and therefore are look-a-likes. These manifest interests result in the LAL_*_Item matrices which model in a binary manner persons (rows) and their interest (page or category of that page), group, school, employer or employment position (columns). Based on these structures, people with common interests can be found. Fig. 3 displays the degree distributions for look-a-likes based on similar interests (pages and the categories of these pages) and groups. The distributions illustrate that many of the Facebook pages, Facebook page categories or groups are likely to have a small number of likes or memberships respectively, and only a non-negligible number of them are connected to many users, which is in line with previous research (Ugander et al. 2011).

Secondly, interests can also become clear by looking at interactions between users which we define as relational look-a-likes (Relational LAL in Table 1). In order to delimit the space of interactions considered in this study, we refer to interactions on Facebook belonging to one of these: (1) Interacting with a person using plain text, links, photos or videos (here, both *sharing* of the text, link, photo or video on someone's wall and *tagging* are included), (2) Commenting on text, links, photos or videos, and (3) Liking text, links, photos or videos. If two users comment on a status or like a status of the same person, this may imply a common interest. In Fig. 2, Sherlock and Jane are look-a-likes as both of them comment on Lizzie's status. Eric and Jo are not friends, but both might be members of the Data Science group on Facebook which implies a common interest, making them look-a-likes.

Three types of data matrices are constructed to model interaction-based look-a-likes in the network. First, the LAL_*_Borrowers matrix of size 4,985 x 4,985 represents

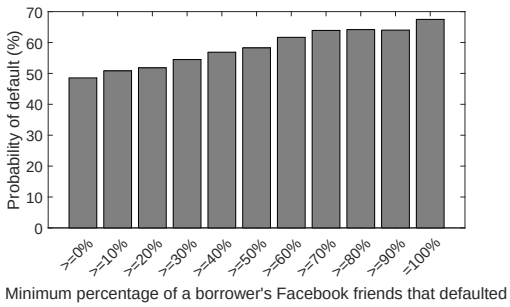

Figure 4: Probability of default as a function of the proportion of neighbours in the friends network that defaulted.

borrowers directly interacting with one another through comments, photos, links, statuses, videos or likes. Since these interactions do not imply the users being friends, this matrix clearly represents look-a-likes. The second matrix, LAL_*_All, extends the previous one by also including interactions with Facebook users that are non-borrowers. Lastly, LAL_*_Items attempts to add even more information by representing an interaction between users (rows) and items (columns). Including the specific item commented on for example may add more detailed information with respect to the look-a-like relation.

3.3 Social network data

Social network data is used to distinguish plain friends from BFFs. Two users are referred to as friends if they befriended one another on Facebook. In the first interaction in Fig. 2, Jane and Lizzie become friends. This information is modeled in the FRI_FBFriends matrix. Fig. 4 plots the probability of default against the proportion of neighbours in the friends network that defaulted. The default probability increases along with the proportion of defaulted friends and reaches a maximum of 67.49% for borrowers that are connected to defaulted friends only.

Two Facebook friends that actually interact with one another by e.g. liking one another's status, makes them BFFs. When Jane comments on Lizzie's status in the second interaction of Fig. 2, Lizzie and Jane change from being just friends to being BFFs. Supposing Jo, Scout, Elinor and Lizzie befriended one another in the past, Lizzie tagging them in her status update, makes all of them BFFs. This data is modeled in the BFF_* matrices by combining the direct interactions in LAL_*_Borrowers with the friends in FRI_FBFriends. Fig. 5 plots the default probability as a function of the proportion of defaulted neighbours in a BFF network, where two friends are classified as BFFs if they have interacted through photos. Comparing this Figure with Figure 4 shows that the BFF network is more discriminative than the general friends network. The default probability reaches a maximum of 100% for borrowers with at least 90% defaulted neighbours in the BFFs network. Fig. 3 displays the degree distribution of the friends and the BFFs on a log-log scale. Both distributions are monotonically decreas-

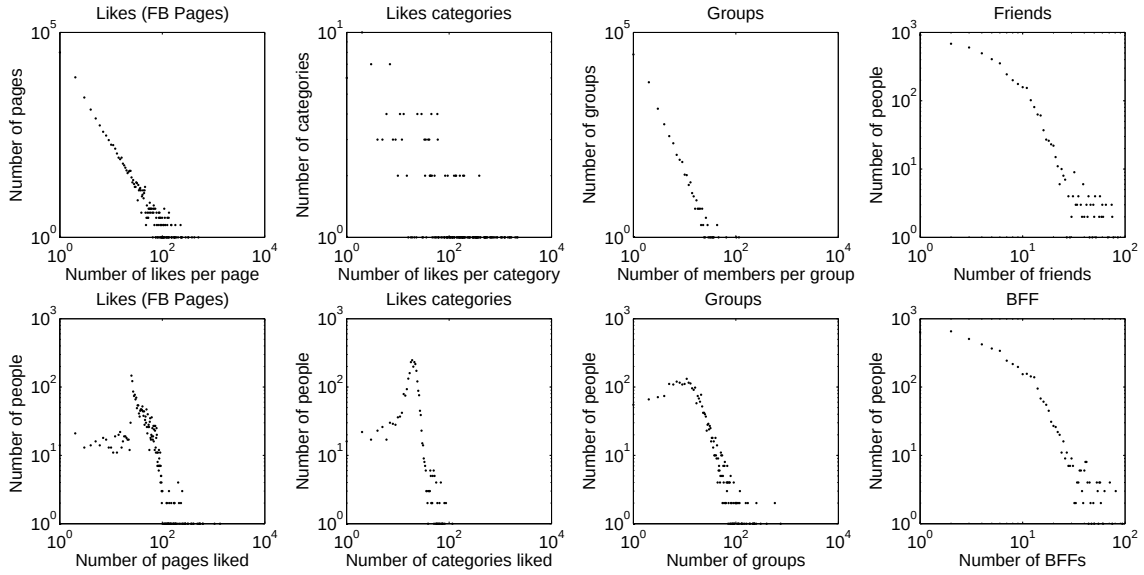

Figure 3: Degree distributions for the pages, the categories of the pages, the groups, the friends and the BFFs.

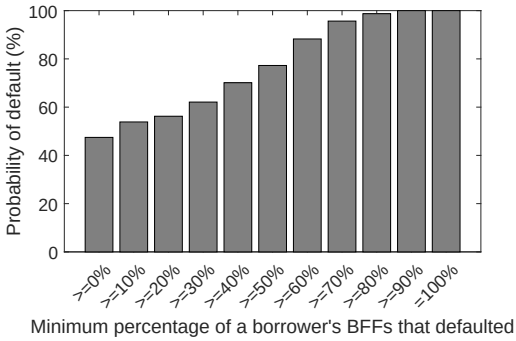

Figure 5: Probability of default as a function of the proportion of defaulted neighbours in the BFF network of friends interacting through photos.

ing and very similar to that of the entire Facebook community (Ugander et al. 2011), where most users have a moderate number of friends and only a few users have a high degree.

4 Methodology and Results

4.1 Methodology

For each of the data categories we use specifically tailored techniques that we describe in more detail in the following section.

The *look-a-like data* (LAL_* matrices) can be modelled as bipartite graphs (bigraphs), which are graphs with two types of nodes and edges exist only between nodes of different types. In our case, one set of the bigraph nodes represents the loan applicants and the other set refers to their items of interest. We use the proposed three-step framework for node classification within bigraphs by Stankova et al. (Stankova, Martens, and Provost 2015) to create a weighted projection of the bigraph and then apply a unigraph relational learner. The projection is created by connecting the persons that have at least one shared interest and weighted in the following manner. Based on the empirical results from the study, we apply the tangens hyperbolicum function to weight the items of interest, by assigning a lower score to the very popular items as being less informative for the target variable. In the following step, we calculate the strength w_{ij} between two persons i and j in the projection by summing the weights of their shared items of interest. To this weighted unigraph representation of the bigraph, we apply the network-only Link-Based classifier (nLB) (Lu and Getoor 2003), which is a powerful relational learner that is able to capture complex network patterns (Stankova, Martens, and Provost 2015). The nLB classifier builds a class vector $CV(i)$ for every training instance (i.e. node) i in the network which contains the probability estimates (scores) that the node under study has a class label default or non-default (see Equation 1). From the formula, one can see that the probability estimate of a node i belonging to a certain class (c), is calculated as a weighted average of the scores of its neighbouring nodes ($j \in N(i)$). Subsequently, nLB creates a logistic regression model based on these class vectors (see Equation 2).

$$CV(i)_c = \frac{\sum_{j \in N(i)} w_{ij} \cdot P(l_j = c)}{\sum_{j \in N(i)} w_{ij}} \quad (1)$$

$$P(l_i = c|N(i)) = \frac{1}{1 + e^{-\beta_0 - \beta CV(i)}} \quad (2)$$

As an alternative to this network based approach, we also look at this from a standard classification perspective, where

we apply a state-of-the-art discriminative learner on the matrix representation of the data (Wu et al. 2008). More specifically, we employ a linear SVM from the package LibLinear (Fan et al. 2008) to the sparse, high-dimensional feature data. In a similar manner, the *social network data* (FRI_FBFriends and BFF_* matrices) can be modelled as graphs with only one type of nodes (unigraphs), where the persons are connected to their Facebook friends or BFFs. For this type of data, we again apply the linear SVM to the adjacency matrix and the nLB classifier directly on the unweighed unigraphs. Additionally, we also build a baseline SVM model with the 29 *socio-demographic* variables (Sociodemo matrix) available for each loan applicant. The categorical variables are included in the model by dummy encoding them.

Finally, we incorporate all the pieces of information into two ensemble models, where the socio-demographic data is combined with the scores from the SVM (first ensemble model) and nLB (second ensemble model) respectively, applied over the interest-based and the social network data. As a classification technique for the ensembles we use a linear SVM, since we need to be able to understand the decisions made by the classifier. Comprehensibility is an important issue in credit scoring for legal and regulatory reasons. National legislations often demand that financial institutions explain why a particular credit is denied and regulations such as the Basel Accords and International Financial Reporting Standards dictate that financial institutions must understand their credit risk models and predictions. For the experimental setting we use a 10-fold cross-validation procedure where (i) 40% of the data is used for training and validation of the classifiers used with the interest based and the social network data, (ii) 40% is for training, 10% for validation and 10% for testing the ensemble model. As explained by Moeyersoms and Martens (Moeyersoms and Martens 2015), it is paramount that we carefully calculate the scores for the interest-based and the social network data on a separate subset of the data that is not used for building the ensemble in order to avoid overfitting.

4.2 Results

As can be seen in Figure 1, we have different types of data available which each imply a different type of relationship. For each of these three types (look-a-likes, friends and BFFs) and for their combination (ensemble model), we will first assess their predictive value in evaluating creditworthiness. We start by analysing and comparing their predictive performance in terms of AUC reached by the techniques described above. Next, we perform a statistical significance test to interpret the predictive performance in a sound manner. For BFFs and look-a-likes respectively, we measure the performance of each individual data category as well as the performance of a stacked model that is a linear model of the scores of the individual data categories. E.g. the stacked look-a-like model is built by applying a linear SVM on the different LAL features.

The predictive results for all different data sources are given in Figures 6 and 7 for the SVM-models and Figures 8 and 9 for the nLB-models. To make the distinction between BFF and LAL clear, the results per technique are divided over two Figures: a Figure that contains the performance of the different BFF features (i.e. Figure 6 for SVM and 8 for nLB) and a Figure that contains the performance of the different LAL features (i.e. Figure 7 for SVM and 9 for nLB). Performances of the friends data, socio-demographic data and ensemble model are indicated in all Figures to allow easy comparison. Please note that the performance of the socio-demographic model is the same for both the SVM and nLB Figures, since this model does not contain network data and is always modelled using a linear SVM. The ensemble model is different in both cases. In Figures 6 and 7 the ensemble model is a linear model that includes the socio-demographic features and the *SVM-scores* of all the different BFF, LAL and friends features, while in Figures 8 and 9 the ensemble model combines the socio-demographic features with the *nLB-scores* of all the different BFF, LAL and friends features. Prediction performance is measured in terms of AUC, a widely-used performance evaluation metric in the machine learning community that represents the probability that a randomly chosen positive instance is ranked higher by the classification technique than a randomly chosen negative instance (Fawcett 2006). The X-axis shows the AUC whereas the Y-axis denotes the different data features. The reported AUC is the mean AUC over the ten folds. In the Figures, the error bars and the dotted lines for the friends, socio-demographic and ensemble models represent the 95% confidence intervals. The first observation is that the look-a-likes data, especially Likes and Likes categories have the most predictive value as compared to other data sources. Interestingly, for both methods the look-a-likes data performs better as compared to BFFs and friends data. That is, it appears from the results that similarities in interests or behavior includes more information than the real social network of a person with respect to default prediction. Likes and Likes categories perform better than the socio-demographic model and thus appear to have more predictive information than the demographic data that is traditionally used. Moreover, the confidence interval of the stacked look-a-likes model even crosses the confidence interval of the ensemble model that contains all data features, indicating that it might be sufficient to collect only the look-a-likes data. Especially noteworthy is the fact that the three best performing look-a-likes data features (likes pages, likes categories and groups) are interest-based look-a-likes and not look-a-likes through direct links (e.g. being tagged in the same picture) or a mutual connection (e.g. responding to a status of a mutual friend).

The baseline socio-demographic model appears to have a large predictive performance as well, thereby performing better than BFFs, friends and even most of the look-a-likes data. When comparing BFFs with friends data, it can be seen that there is no major difference between BFFs and friends when applying the SVM. The nLB on the other hand, shows that most of the BFF data has higher predictive value as compared to friends. This could indicate that real, active friendships are more predictive than merely being connected on Facebook. However, it is important to note that the number of connections in the BFF networks is rather low. It is therefore possible that the networks are too small to make reliable

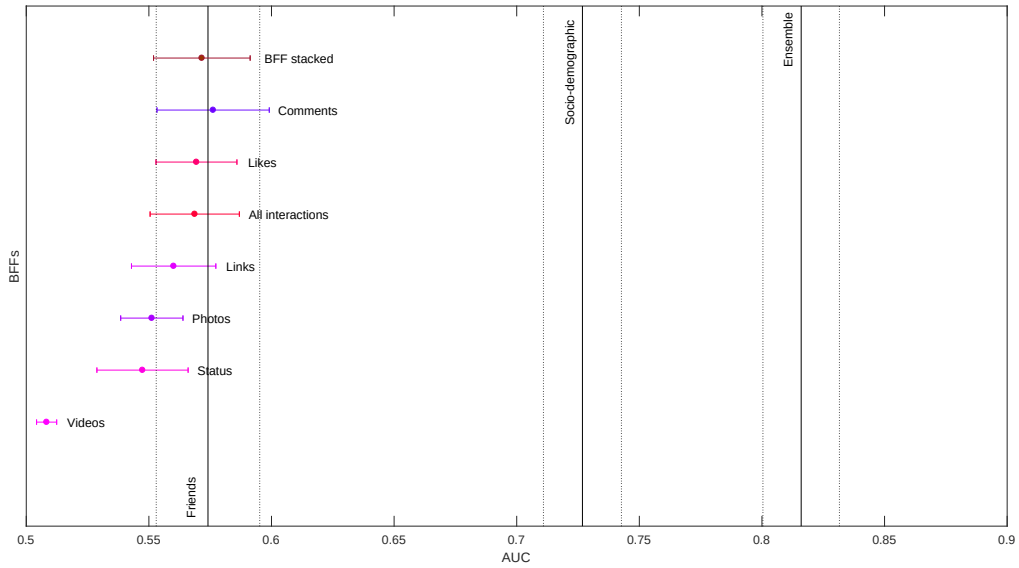

Figure 6: AUC results for the different BFF data categories when using a linear SVM.

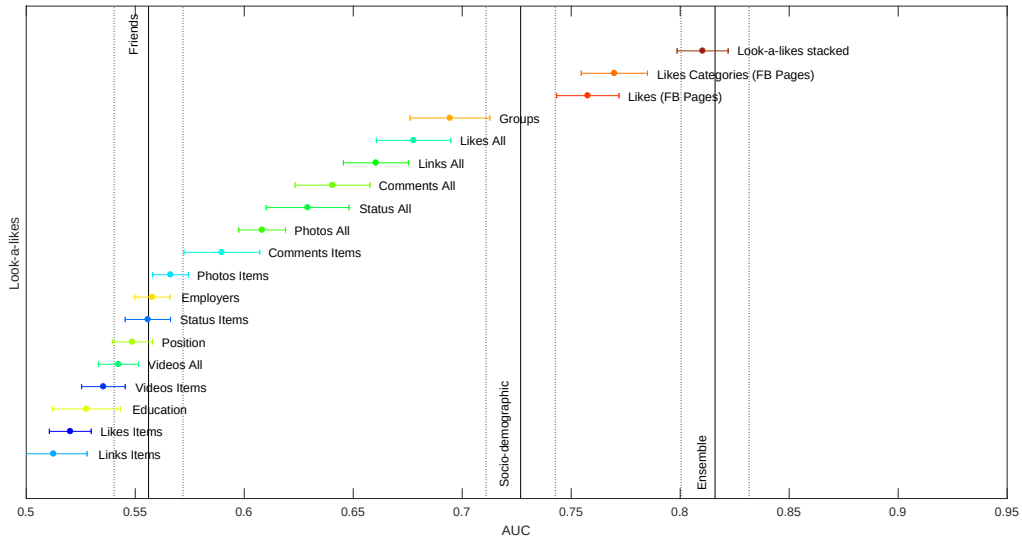

Figure 7: AUC results for the different look-a-likes data categories when using a linear SVM.

conclusions.

Lastly, we compared the performance of the ensemble model with models that include different combinations of the data types. Specifically, we compare the AUC of the ensemble model with the AUCs of linear models that contain (i) only the variables of one specific data type (such as socio-demographic data or look-a-likes) and (ii) that combine two different data types (such as look-a-likes *and* friends). The significance test results of comparisons are reported in Table 2. Here, p-values of different models which include different combinations of the data types are statistically tested in terms of AUC, by using a Wilcoxon signed-rank test. Every model is tested as compared to the top-performing model which is, as can be seen from the previous results, the ensemble model. Table 2 reports the significance results for the models that combine the SVM scores of the data types. The diagonal elements show the model where all features of the respective data type are included. The rest of the matrix indicates the results of the combinations of the corresponding data categories, by using a linear SVM. As an example; in the matrix the intersection between SD and FRI data shows a p-value of 0.002. This indicates that the AUC of the lin-

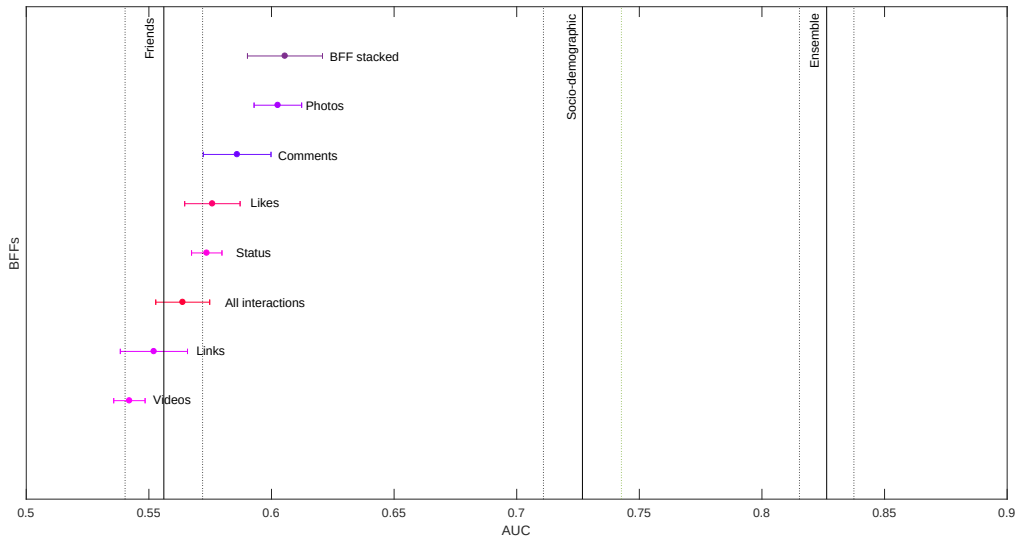

Figure 8: AUC results for the different BFF data categories when using the network-only Link-Based classifier (nLB).

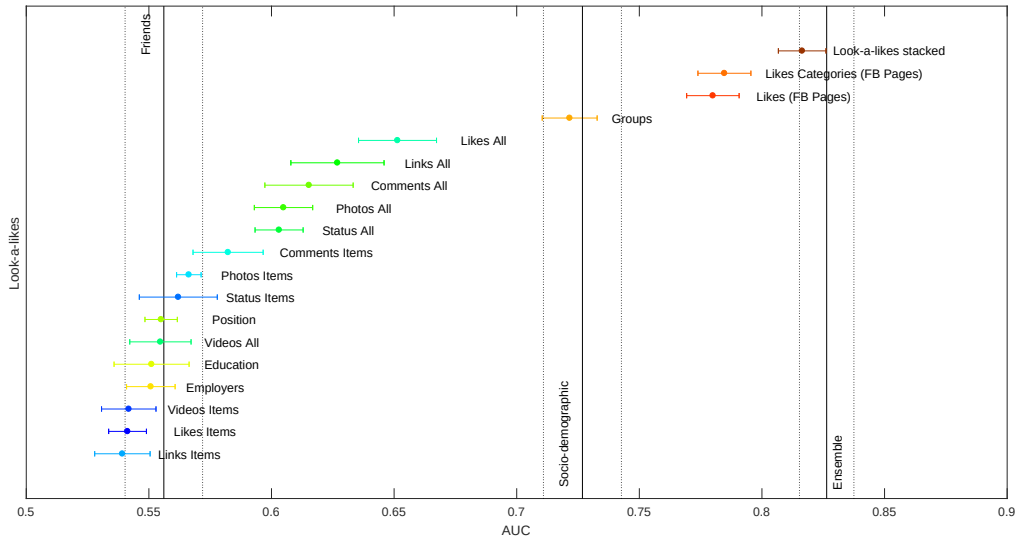

Figure 9: AUC results for the different look-a-likes data categories when using the network-only Link-Based classifier (nLB).

ear SVM that includes both socio-demographic and friends data is significantly worse (at the 1% level) as compared to the ensemble model. The ensemble model, that uses all the data, is shown in the last row. Performances that are not significantly different at the 5% level from the top performance (ensemble model) with respect to a Wilcoxon signed-rank test are tabulated in bold face. Statistically significant underperformances at the 1% level as compared to the ensemble model are emphasized in italics. From this table, one can conclude that although the ensemble model performs best, the performance of the model which includes the look-a-likes data is not significantly worse as compared to the ensemble model (p-value of 0.2324). The same can be seen for other combinations of data which include the look-a-likes data. Similar conclusions can be made for the linear models that combine the nLB-scores of the different data features. Again, this confirms our previous finding that interest data gives more information than the social network data. Moreover, this implies that in this case, using one source of data (look-a-likes) is sufficient to build the predictive model and

Table 2: Significance test for the results (in terms of AUC) of the models built with combinations of the SVM-scores of the different data categories using a linear SVM. This table shows the p-values resulting from the Wilcoxon signed rank test in which the AUC values of the models are compared. Performances that are not significantly different at the 5% level from the top performance (ensemble model) are tabulated in bold face. Statistically significant underperformances at the 1% level as compared to the ensemble model are emphasized in italics.

	SD	LAL	FRI	BFFs	ensemble
SD	*0.002*	**0.9219**	*0.002*	*0.002*	-
LAL	-	**0.2324**	**0.2324**	**0.4316**	-
FRI	-	-	*0.002*	*0.002*	-
BFFs	-	-	-	*0.002*	-
ensemble	-	-	-	-	**1.000**

assess creditworthiness.

Using these models, the credit scoring process becomes an automated process. It can complement the manual screening that is traditionally applied in microfinance. It is nevertheless also important for the credit lender to understand the predictions of the model (Martens et al. 2007). In credit scoring one is likely to be interested in knowing why a particular applicant was predicted to be a potential defaulter. An instance-level explanation method, that was developed to explain document classification, could be used to explain the predicted class (Martens and Provost 2014). In this case an explanation would be defined as the minimal set of likes/interactions such that removing this set changes the class. A possible explanation could be: *If the user would NOT have liked ("Who cares about data science?" "Credit scoring is boring") then the class would change from default to non-default*. These explanations are useful to the model developer and can help to detect possible misclassifications by the model. Due to confidentiality reasons, we cannot publish the actual explanations of our predictions. For further information regarding the implementation of this method, we refer to (Martens and Provost 2014).

5 Conclusion

In this paper, we investigated the potential of Facebook data for microfinance credit scoring. The good predictive performance of the generated models allows to automate the credit scoring process for microfinance to massive settings, mainly thanks to the ability to include the difficult concept of character.

The splitup in different data categories shows that there is a significant difference in the predictive power of each, with interest-based data being the most valuable. It should be noted however that our methodology is limited to the setting where Facebook data is available, which is not always the case in microfinance lending. Also, the validity of our results is limited to this specific application on a dataset from the Philippines. It would be interesting to see to what extent these findings on BFFs and friends, as well as the superiority of interest-based data translate to other applications.

6 Acknowledgments

We would like to thank the data science team at Lenddo for providing us with the data and valuable feedback.

References

Banasik, J.; Crook, J.; and Thomas, L. 2003. Sample selection bias in credit scoring models. *Journal of the Operational Research Society* 54(8):822–832.

Copestake, J. 2007. Mainstreaming microfinance: social performance management or mission drift? *World Development* 35(10):1721–1738.

Domingos, P. 2005. Mining social networks for viral marketing. *IEEE Intelligent Systems* 20(1):80–82.

Emel, A. B.; Oral, M.; Reisman, A.; and Yolalan, R. 2003. A credit scoring approach for the commercial banking sector. *Socio-Economic Planning Sciences* 37(2):103–123.

Facebook Inc. 2014. Authorization and authentication based on an individual's social network. Patent.

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

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

Goel, S.; Hofman, J. M.; and Sirer, M. I. 2012. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*.

Hand, D. J.; Sohn, S. Y.; and Kim, Y. 2005. Optimal bipartite scorecards. *Expert Systems with Applications* 29(3):684–690.

Hermes, N., and Lensink, R. 2007. The empirics of microfinance: what do we know? *The Economic Journal* 117(517):F1–F10.

Hill, S.; Provost, F.; and Volinsky, C. 2006. Network-based marketing: Identifying likely adopters via consumer networks. *Statist. Sci.* 21(2):256–276.

Hu, J.; Zeng, H.-J.; Li, H.; Niu, C.; and Chen, Z. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, 151–160. ACM.

Junqué de Fortuny, E.; Martens, D.; and Provost, F. 2013. Wallenius Naive Bayes. Technical Report 2451/33545, New York University.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15):5802–5805.

Lu, Q., and Getoor, L. 2003. Link-based classification. In *ICML*, volume 3, 496–503.

Martens, D., and Provost, F. 2011. Pseudo-social network targeting from consumer transaction data. Technical Report CEDER-11-05, New York University.

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

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

Min, J. H., and Jeong, C. 2009. A binary classification method for bankruptcy prediction. *Expert Systems with Applications* 36(3):5256–5263.

Moeyersoms, J., and Martens, D. 2015. Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems* 72:72–81.

Morduch, J. 1999. The microfinance promise. *Journal of economic literature* 1569–1614.

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

Putnam, R. D. 1995. Bowling alone: America's declining social capital. *Journal of democracy* 6(1):65–78.

Raeder, T.; Stitelman, O.; Dalessandro, B.; Perlich, C.; and Provost, F. 2012. Design principles of massive, robust prediction systems. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1357–1365. ACM.

Schreiner, M. 2000. Credit scoring for microfinance: Can it work? *Journal of Microfinance/ESR Review* 2(2):105–118.

Schreiner, M. 2003. *Scoring: the next breakthrough in microcredit*. Consultative group to assist the poorest (CGAP).

Sharma, M., and Zeller, M. 1997. Repayment performance in group-based credit programs in bangladesh: An empirical analysis. *World development* 25(10):1731–1742.

Stankova, M.; Martens, D.; and Provost, F. 2015. Classification over bipartite graphs through projection. Technical Report 2015|001, Antwerp University.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Van Gestel, T.; Baesens, B.; and Martens, D. 2015. *Predictive Analytics: Techniques and Applications in Credit Risk Modelling*. Oxford University Press.

Van Gool, J.; Verbeke, W.; Sercu, P.; and Baesens, B. 2012. Credit scoring for microfinance: is it worth it? *International Journal of Finance & Economics* 17(2):103–123.

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

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

Wei, Y.; Yildirim, P.; Van den Bulte, C.; and Dellarocas, C. 2014. Credit scoring with social network data. *Available at SSRN 2475265*.

Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Philip, S. Y.; et al. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14(1):1–37.

Zeller, M. 1998. Determinants of repayment performance in credit groups: The role of program design, intragroup risk pooling, and social cohesion. *Economic development and cultural change* 46(3):599–620.