DEPARTMENT OF ENGINEERING MANAGEMENT

Who Cares About Your Facebook Friends?
Credit Scoring for Microfinance

Sofie De Cnudde, Julie Moeyersoms, Marija Stankova, Ellen Tobback, Vinayak Javaly & David Martens

UNIVERSITY OF ANTWERP
Faculty of Applied Economics
City Campus
Prinsstraat 13, B.226
B-2000 Antwerp
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
www.uantwerpen.be
Who Cares About Your Facebook Friends? Credit Scoring for Microfinance

Sofie De Cnudde, Julie Moeyersoms, Marija Stankova, Ellen Tobback, Vinayak Javaly & David Martens

RESEARCH PAPER 2015-018
OCTOBER 2015

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@uantwerpen.be

The research papers from the Faculty of Applied Economics are also available at www.repec.org
(Research Papers in Economics - RePEc)

D/2015/1169/018
ABSTRACT
Microfinance has known a large increase in popularity, yet the scoring of such credit still remains a difficult challenge. In general, retail credit scoring uses socio-demographic and credit data. We complement such data with social network data in an innovative manner i.e. with fine-grained interest and social network data from Facebook. Using a unique dataset of 4,985 microfinance loans from the Philippines, we show how the different data types can predict creditworthiness. A distinction is made between the relationships that the available data imply: (1) look-a-likes are persons who resemble one another in some manner, be it liking the same pages, having the same education, etc. (2) friends have a clearly articulated friendship relationship on Facebook, and finally (3) the “Best Friends Forever” (BFFs) are friends that interact with one another. Our analyses show two interesting conclusions for this emerging application. Firstly, applying relational learners on BFF data yields better results than considering only the friends data. Secondly, the interest-based data that defines look-a-likes, is more predictive than the friendship or BFF data. Moreover, the model built on interest data is not significantly worse than the model that uses all available data, including the friendship data. Hence begging the question: who cares about your Facebook friends when your interest data is available?

General Terms
Networks, Data mining, Default prediction, Microcredit

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 [1]. 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 [2, 3]. 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 [4]. 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, the world leader in social authentication and scoring technology [5]. 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 including 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

http://partners.lenddo.com
social networks, enabling Lenddo to provide unique insights about their creditworthiness. For purposes of this paper, only a small subset of Lenddo’s data was shared and analyzed. The analysis and methodology presented in this article are similar in concept to the approaches used by Lenddo, but they do not describe any of the algorithms and scoring solutions currently or previously used by Lenddo in its business. The data used is from Facebook and is categorized for this analysis 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 on 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). Look-a-likes (LAL) refer to people that are similar to one another. In this case this can be interpreted as persons that show similarities in certain socio-demographic characteristics, like the same pages on Facebook, have a Facebook-friend in common or are commenting on the same status for example. 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 each other at all. Take the example of a person that lives on the other side of the country, who likes the same music groups as you do and joined the same Facebook groups. Although it is clear that you have some similarities, it is very probable that you will never meet or be friends with this person. However, the information included in these similarities can be an important source of information to predict default behavior since similar behavior in one domain (e.g. preferences) might imply similarities in other domains (e.g. default) as well [3, 6, 7, 8]. 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. The contributions of this paper are three-fold, as illustrated in Fig. 1. 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 [3, 10] and balance sheets [11, 12], 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 [1]. In many cases the loan officer communicates with the local community of the client to get an idea about the client’s trustworthiness [13]. In the 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 [14]. Social capital - defined by Putman [15] 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 [16] and Sharma and Zeller [17] use group, community and borrower or program characteristics to describe credit risk of joint liability groups. Schreiner [1] remarks 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 [3] have investigated whether traditional credit scoring is applicable to microfinance lending. Using borrower, loan and lender characteristics they have built a credit scoring model for a Bosnian microlender. They find 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 [2] 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. The structured data include 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 [1] advises microlenders to search for personal character traits that are predictive of repayment behavior. Recently, Wei et al [8] 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.

2.2 Interest-based vs social network data

Different types of data are commonly used for predictive modeling in a retail setting [19]. Except for the conventional socio-demographic 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 [20], which uses the social relationships observed in calling behavior to predict product/service adoption. Other
studies have looked at call behavior as well to predict churn [21] and social network data for viral marketing [22]. 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. [23] and Junque de Fortuny et al. [24] looked at predicting different personality traits from a dataset of users liking Facebook pages. The studies of Goel et al. [25] and Hu et al. [26] predict demographic attributes and Raeder et al. [8] predict brand interest from people’s browsing history. Weber et al. [27] reveal the political views from history of videos watched on YouTube. For financial applications, Martens et al. [5] predict interest in financial products from transactional datasets of consumers making payments to merchants and Provost et al. [7] consider geo-location 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 customer characteristics. 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 that were made by 4,512 users. As stated previously and visualized in Fig. 1, we dispose of three data categories which we use to distinguish three levels of relations in the network in terms of look-a-likes, friends and BFFs. We use Fig. 2 to illustrate these. Table 1 shows a list of all the constructed data structures along with some relevant data characteristics. Note that any names or personally identifiable information shown in this paper are those of the authors, and not names or information of actual Lenddo members.
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 \((\rho)\) defined as \(\rho = N/(m \times n)\), with \(n\) the number of data instances.

<table>
<thead>
<tr>
<th>Name</th>
<th>Represented data</th>
<th>Category</th>
<th>(m)</th>
<th>(N)</th>
<th>(\rho)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAL_Sociodemo</td>
<td>Socio-demographic attributes of a person</td>
<td>SD-based LAL</td>
<td>29</td>
<td>111,989</td>
<td>83 %</td>
</tr>
<tr>
<td>LAL_Likes_Item</td>
<td>Persons liking a page on Facebook</td>
<td>Interest-based LAL</td>
<td>48,701</td>
<td>127,241</td>
<td>0.052%</td>
</tr>
<tr>
<td>LAL_LikesCat_Item</td>
<td>Persons liking a category of a page on Facebook</td>
<td>Interest-based LAL</td>
<td>238</td>
<td>53,441</td>
<td>4.504%</td>
</tr>
<tr>
<td>LAL_Groups_Item</td>
<td>Persons joined in a group on Facebook</td>
<td>Interest-based LAL</td>
<td>38,037</td>
<td>55,399</td>
<td>0.029%</td>
</tr>
<tr>
<td>LAL_Education_Item</td>
<td>Persons having a certain level of education</td>
<td>Interest-based LAL</td>
<td>4,620</td>
<td>11,015</td>
<td>0.048%</td>
</tr>
<tr>
<td>LAL_Employers_Item</td>
<td>Persons working for employers</td>
<td>Interest-based LAL</td>
<td>5,190</td>
<td>13,173</td>
<td>0.051%</td>
</tr>
<tr>
<td>LAL_Position_Item</td>
<td>Persons holding employment positions or business titles</td>
<td>Interest-based LAL</td>
<td>3,393</td>
<td>9,983</td>
<td>0.059%</td>
</tr>
<tr>
<td>LAL_Comments_All</td>
<td>Persons commenting on one of the statuses of the same person</td>
<td>Interest-based LAL</td>
<td>896,164</td>
<td>1,217,744</td>
<td>0.027%</td>
</tr>
<tr>
<td>LAL_Photos_All</td>
<td>Persons mentioned in one of the pictures of the same person</td>
<td>Interest-based LAL</td>
<td>731,574</td>
<td>235,645</td>
<td>0.047%</td>
</tr>
<tr>
<td>LAL_Links_All</td>
<td>Persons mentioned in one of the links of the same person</td>
<td>Interest-based LAL</td>
<td>262,761</td>
<td>105,170</td>
<td>0.068%</td>
</tr>
<tr>
<td>LAL_Status_All</td>
<td>Persons mentioned in one of the statuses of the same person</td>
<td>Interest-based LAL</td>
<td>630,749</td>
<td>490,942</td>
<td>0.016%</td>
</tr>
<tr>
<td>LAL_Videos_All</td>
<td>Persons mentioned in one of the videos of the same person</td>
<td>Interest-based LAL</td>
<td>46,078</td>
<td>30,899</td>
<td>0.014%</td>
</tr>
<tr>
<td>LAL_Likes_All</td>
<td>Persons liking one of the videos/statuses/photos/comments of the same person</td>
<td>Interest-based LAL</td>
<td>1,817,619</td>
<td>2,692,752</td>
<td>0.030%</td>
</tr>
<tr>
<td>LAL_Comments_Items</td>
<td>Persons commenting on the same status of the same person</td>
<td>Interest-based LAL</td>
<td>2,141,630</td>
<td>1,763,453</td>
<td>0.017%</td>
</tr>
<tr>
<td>LAL_Photos_Items</td>
<td>Persons mentioned in the same picture of the same person</td>
<td>Interest-based LAL</td>
<td>293,155</td>
<td>40,7358</td>
<td>0.028%</td>
</tr>
<tr>
<td>LAL_Links_Items</td>
<td>Persons mentioned in the same link of the same person</td>
<td>Interest-based LAL</td>
<td>267,298</td>
<td>80,411</td>
<td>0.024%</td>
</tr>
<tr>
<td>LAL_Status_Items</td>
<td>Persons mentioned in the same status of the same person</td>
<td>Interest-based LAL</td>
<td>27,442</td>
<td>33,602</td>
<td>0.024%</td>
</tr>
<tr>
<td>LAL_Videos_Items</td>
<td>Persons mentioned in the same video of the same person</td>
<td>Interest-based LAL</td>
<td>4,122,418</td>
<td>2,846,613</td>
<td>0.014%</td>
</tr>
<tr>
<td>LAL_Likes_Items</td>
<td>Persons liking the same video/status/photo/comment of the same person</td>
<td>Interest-based LAL</td>
<td>4,985</td>
<td>20,301</td>
<td>0.081%</td>
</tr>
<tr>
<td>LAL_Comments_Borrowers</td>
<td>Persons giving/receiving comments to/from each other</td>
<td>Relational LAL</td>
<td>4,985</td>
<td>30,347</td>
<td>0.122%</td>
</tr>
<tr>
<td>LAL_Photos_Borrowers</td>
<td>Persons mentioning one another in one of their photos</td>
<td>Relational LAL</td>
<td>4,985</td>
<td>9,199</td>
<td>0.037%</td>
</tr>
<tr>
<td>LAL_Links_Borrowers</td>
<td>Persons mentioning one another in one of their links</td>
<td>Relational LAL</td>
<td>4,985</td>
<td>14,318</td>
<td>0.057%</td>
</tr>
<tr>
<td>LAL_Status_Borrowers</td>
<td>Persons mentioning one another in one of their statuses</td>
<td>Relational LAL</td>
<td>4,985</td>
<td>9,949</td>
<td>0.040%</td>
</tr>
<tr>
<td>LAL_Videos_Borrowers</td>
<td>Persons mentioning one another in one of their videos</td>
<td>Relational LAL</td>
<td>4,985</td>
<td>1,496</td>
<td>0.006%</td>
</tr>
<tr>
<td>LAL_Likes_Borrowers</td>
<td>Persons liking each other’s video/status/photo/comment</td>
<td>Relational LAL</td>
<td>4,985</td>
<td>29,814</td>
<td>0.120%</td>
</tr>
<tr>
<td>FRL_FBFriends</td>
<td>Persons befriending one another</td>
<td>Friends</td>
<td>4,985</td>
<td>30,347</td>
<td>0.122%</td>
</tr>
<tr>
<td>BFF_Comments</td>
<td>Persons giving/receiving comments to/from one another AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>30,347</td>
<td>0.122%</td>
</tr>
<tr>
<td>BFF_Photos</td>
<td>Persons mentioning one another in one of their photos AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>8,609</td>
<td>0.035%</td>
</tr>
<tr>
<td>BFF_Links</td>
<td>Persons mentioning one another in one of their links AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>13,072</td>
<td>0.053%</td>
</tr>
<tr>
<td>BFF_Status</td>
<td>Persons mentioning one another in one of their statuses AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>9,469</td>
<td>0.038%</td>
</tr>
<tr>
<td>BFF_Videos</td>
<td>Persons mentioning one another in one of their videos AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>1,438</td>
<td>0.006%</td>
</tr>
<tr>
<td>BFF_Likes</td>
<td>Persons liking each other’s video/status/photo/comment AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>22,606</td>
<td>0.091%</td>
</tr>
<tr>
<td>BFF_All</td>
<td>Persons having any kind of interaction AND they are friends</td>
<td>BFF</td>
<td>4,985</td>
<td>24,243</td>
<td>0.098%</td>
</tr>
</tbody>
</table>
3.1 Socio-demographic data
The socio-demographic data is used to denote look-a-likes (LAL). People demonstrating similar socio-demographic characteristics may imply similar default behavior. The 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 LAL\_Sociodemo matrix. The number of missing values is approximately 16.65%. Note that a missing value might denote data intentionally left blank by users. We replaced missing values by the mean value and added missing value flags.

3.2 Interest data
In addition to traditionally available socio-demographic characteristics, we also dispose of fine-grained interest characteristics, which also let us determine look-a-likes.

First, there are interests which manifest themselves immediately. Liking a Facebook page or joining a Facebook group for example 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, Jane and Sofie both like the page of the University of Antwerp and therefore are look-a-likes. The constructed LAL\_\_Item matrices model persons with a common interest (page or category of that page), group, school, employer and employment position respectively. Fig. 3 shows the degree distributions for look-a-likes based on similar interests (pages and the categories of these pages) and groups. The distributions illustrate that a Facebook page, a category of a Facebook page and a group all have a low probability of having many likes or memberships respectively, i.e. a power law distribution is present as is common with most interest-based data [28]. Fig. 4 shows a network of users and four page categories. Two of them are discriminative for defaulters, the other two for non-defaulters. Already this shows the potential of using such data for default prediction. Secondly, interests can also become clear by looking at interactions between users. 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, David and Jane are look-a-likes as both of them comment on Ellen’s status. Also, Julie and Vinayak are look-a-likes based on Vinayak liking Julie’s status. Jeff and Julie are not friends, but both might be member of the WSDM group on Facebook which implies a common interest.

Three types of data matrices are constructed to model look-a-likes in the network. First, the LAL\_\_Borrowers matrix of size 4,985 x 4,985 represents borrowers directly interacting with one another through comments, photos, links, statuses, videos or likes. Since direct 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 and items. 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 Ellen become friends. This information is modeled in the FRL\_FBFriends matrix. Fig. 5 shows the friends connections between the borrowers. White nodes represent good borrowers, black nodes represent bad borrowers. The network is a large cluster in which no apparent pattern can immediately be distinguished. Two Facebook friends that actually interact with one another by e.g. liking one another’s statuses, makes them BFFs. When Jane comments on Ellen's status in the second interaction of Fig. 2, Ellen and Jane change from being just friends to being BFFs. Supposing Marija, Sofie, Julie and Ellen befriended one another in the past, Ellen tagging them in her status update, makes all of them BFFs. This data is modeled by combining the direct interactions in LAL\_\_Borrowers with the friends in FRL\_FBFriends. Fig. 6 shows a portion of a BFF network based on interactions with photos. Two users in this network are connected if they are friends and if one of them has shared a photo on the other person's wall or mentioned the other person in a photo. The entire BFF photos network consists of separate, smaller networks like the one presented in Fig. 6. The network clearly contains clusters of good and
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 interest-based datasets can be modelled as bipartite graphs (bigraphs) or matrices. The former are defined as

Figure 5: Network of friends. Black dots represent defaulters, white dots represent non-defaulters.

Figure 6: Network of BFFs interacting by photos (subset of network). Black dots represent defaulters, white dots represent non-defaulters.

Figure 3: Degree distributions for the pages, the categories of the pages, the groups, the friends and the BFFs.
graphs with two types of nodes where edges exist only between nodes of different type. In our case, one set of the nodes refers to the persons and the other set of nodes to the items of interest. We use the proposed three-step framework for node classification within bigraphs by Stankova et al. [29] to first create a weighted unigraph projection of the bigraph and then apply some of the existing relational learners for unigraphs. The projection is created by connecting two persons if they have at least one shared interest. Based on the empirical results from the study, we focus our attention to the following techniques: (i) the tangens hyperbolicum function for creating the top node weights, where we down-weight the very popular items as providing less information for the target variable, (ii) sum of the shared nodes as an aggregation function and (iii) the weighted-vote Relational Neighbor (wvRN) classifier [30] as an appropriate choice for problems that exhibit network assortativity and the network-only Link-Based classifier (nLB) [31] as a more powerful learner that can capture more complex patterns. Alternatively to this network based approach, we also look at this from a standard classification perspective, where we apply propositional learners on the matrix representation of the data. More specifically, we apply a linear SVM from the LibLinear package [32] to the sparse, high-dimensional feature data.

The social network data 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 use the standard relational classifiers, namely the wvRN and the nLB, directly on the unweighed unigraphs. In addition, we apply the previously discussed version of a linear SVM. For each of the loan applicants we also have a set of 29 socio-demographic variables, where we dummy encode the categorical variables.

Finally, we incorporate all the pieces of information into an ensemble model, where we combine the socio-demographic data with the scores from the different classification techniques applied over the interest-based and the social network data. As a classification technique for the ensemble we use a linear SVM, since we need to be able to understand the decisions made by the classifier. This is a very important issue in credit scoring and we further elaborate on it later. The experimental setting is as follows: we use 10 fold cross-validation where a subset of (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 et al. [6], it is very important 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
The results for all different data sources are given in Fig. 7 and 8 which give the performance for the SVM and nLB respectively since those methods gave the best performances. The Y-axis shows the AUC whereas the X-axis denotes the different data categories. The first thing to observe is that the look-a-likes data, especially Likes and Likes categories have the most predictive value as compared to other data sources. Surprisingly, for both methods the look-a-likes data is performing 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.

The socio-demographic data 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 BFFs data has a higher predictive value as compared to friends. This confirms the fact that real, active friendships are more predictive than just being connected on Facebook.

Lastly, the ensemble model, which includes all the data sources, seems to outperform the individual data sources. The latter result can also be seen from Table 2 that displays the p-values of the Wilcoxon signed rank test. For each model, a model being a combination of data types, we tested whether the 10-fold AUC values of the respective model differ significantly from the 10-fold AUC values of the top performing model, i.e. the ensemble model. The diagonal elements show the model where all data sources of the respective data type are included. The rest of the matrix indicate the results of the combinations of the corresponding data categories. 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), according to the Wilcoxon signed rank test, are tabulated in bold face. Statistically significant underperformances at the 1% level are emphasized in italics.

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>look-a-likes</th>
<th>friends</th>
<th>BFFs</th>
<th>ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>look-a-likes</td>
<td>0.002</td>
<td>0.232</td>
<td>0.002</td>
<td>0.002</td>
<td>-</td>
</tr>
<tr>
<td>friends</td>
<td>-</td>
<td>0.193</td>
<td>0.275</td>
<td>0.193</td>
<td>-</td>
</tr>
<tr>
<td>BFFs</td>
<td>-</td>
<td>-</td>
<td>0.002</td>
<td>0.002</td>
<td>-</td>
</tr>
<tr>
<td>ensemble</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.002</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 2: Results (in terms of p-value of Wilcoxon signed rank test) of the different models: The diagonal shows the results when only including the corresponding category of data. The rest of the matrix shows the results of the combinations of the corresponding data categories. Finally, the ensemble model includes all data. 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 are emphasized in italics.
Again, this confirms our previous finding that interest data gives more information than the social network data which evokes the question: who cares about your Facebook friends when there is interest data available? Moreover this implies that in this case, using one source of data (look-a-likes) is sufficient to build the predictive model.

Using these models, the credit scoring process becomes an automated process. It can substitute or 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. 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. 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?” “Sydney is boring”) then the class would change from default to non-default. Due to confidentiality

Figure 7: AUC results (anonymized) for the different data categories when using a linear SVM.

Figure 8: AUC results (anonymized) for the different data categories when using the network-only Link-Based classifier (nLB).
reasons, we cannot publish the actual explanations of our predictions. For further information regarding the implementation of this method, we refer to [34].

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 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
The authors would like to thank the Flemish Research Council (FWO) for financial support (Grants G.0827.12N, G.0319.14N and the fellowship for Ellen Tobback). Also, we would like to thank the data science team at Lenddo for providing us with the data and valuable feedback.

7. REFERENCES


