National College of Ireland

Prediction of Loan Defaulters in Microfinance Using Social Network Data

MSc Research Project
Data Analytics

David Murphy
x16124227

School of Computing
National College of Ireland

Supervisor: Paul Laird

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	David Murphy
Student ID:	x16124227
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Paul Laird
Submission Due Date:	23/04/2018
Project Title:	Prediction of Loan Defaulters in Microfinance Using Social Network Data
Word Count:	XXX

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	20th April 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Loan Defaulters in Microfinance Using Social Network Data

David Murphy
x16124227
MSc Research Project in Data Analytics

20th April 2018

Abstract

Faced with growing competition in the microfinancing market and higher operational risk, it is ever more important for an MFI to be able to leverage less conventional customer data to improve the efficiency of their lending models. Most MFIs are active in developing countries where financial history is generally non-existent on their user base which increases the difficulty in assessing the credit worthiness of individuals. Instead, an alternative source of data such as mobile phone call and SMS logs can be utilised to assist with this problem. In this study, call and SMS logs from the borrowers of a MFI operating in the Kenyan marketplace are featurised and used to train various classification models. The results show how such data is a valuable commodity in predicting the default class, particularly when relationship tie-strength features are introduced. The influence of an existing borrower's loan outcome on a new loan applicant within their social network is also modelled using the spreading activation method as an alternative approach to traditional classification, but results indicate that they are not effective.

1 Introduction

Many microfinancing institutions (MFI) are operating in developing countries where micro-loans are a valuable credit source for segments of the population that are in many ways restricted from traditional banking credit lines. This market participation adds to the overall operational risk of a MFI however, as the majority of borrowers are those with little to no credit history. This lack of any credit trail invariably opens up the information asymmetry gap between the loan recipient and the MFI which subsequently increases the challenge in distinguishing between good and bad loans.

Existing research indicates classification models, both parametric and non-parametric, can help improve the prediction of loan outcomes for MFIs (Blanco et al., 2013; Cubiles-De-La-Vega et al., 2013), but there is still high associated misclassification costs due to the presence of information asymmetry (Baklouti and Bouri, 2013; da Kammoun and Triki, 2016). Much of the research in this domain has also been applied to data sets composed of demographical attributes that are not always readily available in a digital app driven business. In P2P lending, various forms of social network data have been introduced to mitigate the information asymmetry gap and the integration of such data has been

shown to improve default rate prediction (Zhang et al., 2016; Ge et al., 2017). Therefore, to reduce this information asymmetry in the absence of conventional data a different approach is applied in this research by introducing new features to the classification techniques, based entirely on the call and SMS log data of the loan applicants.

The overarching objective of this research is to therefore exploit the network data existent in the call and SMS logs of a loan applicant to measure how effectively such data can predict the outcome of micro-loans. Two methodologies are considered to harness this information: the first will transform the call and SMS logs into various feature sets ranging from general usage features such as total call duration to network related features such as tie-strengths to neighbouring nodes and these will then be used to train classification models. As we generally expect people to behave similarly to those they are closest to from the concept of homophily (McPherson et al., 2001), the addition of tie-strengths are interesting features to consider and their impact to the classification model performance will also be assessed.

The second method borrows from promising research in the prediction of churn subscribers in the telco domain where network graphs using the call detail records of customers are exploited and relational learner models are used to diffuse influence across the nodes (Dasgupta et al., 2008; Kim et al., 2014; Backiel et al., 2015). The underlying hypothesis is that churned customers can influence those they are closest to into churning and the relational learners provide a conduit to spread such influence around the network. Again the underpinning of this application has roots in the concept of homophily as the strength of the relationship is the factor that determines the influence a person propagates to a neighbouring node in their network. In applying this to a loan default problem, the hypothesis is that an outcome can be predicted for a new loan applicant by measuring the amount of influence passed to them through the topology of their social network. The spreading activation method (SPA) is adopted for this purpose and the results are assessed against the traditional classification methods.

Based on the above objectives there are three questions that are analysed:

Q1. Does the spreading activation method perform better in predicting the loan outcome compared with traditional classification methods with featurised network variables?

Q2. Does the featurisation of a loan applicant's communication logs enable classification models to effectively predict defaulters?

Q3. Does the introduction of tie-strength features improve the performance of a classification algorithm?

There are three main contributions to the field from this study. Firstly, we provide answers to the three novel research questions by empirically evaluating the performance of several SPA and classification models using telecommunication data in the domain of microfinance. Thus, it is determined whether relational learners offer a beneficial alternative over traditional classifiers. Secondly, it is demonstrated how call and SMS log data can be harvested into features to efficiently predict default using classification models. Finally, it is found that the introduction of tie-strength features has a very positive impact on many classification models.

The remainder of the paper is structured as follows. Chapter 2 presents various related works. Section 3 details the methodology followed for integrating the spreading activation method and featurising the communication logs for the classification algorithms. Section 4 outlines the experimental setup used to assess the model performances along with the models adopted for the analysis. In Section 5, the experimental results from the analysis analysis are discussed and Section 6 gives concluding remarks on the overall findings.

2 Related Work

No prior research was found that focused on the integration of telecommunications data for default prediction in microfinance, so this review of past literature includes work from the domains of microfinance, peer-to-peer lending (P2P) and churn prediction.

2.1 Default Prediction in Microfinance

The existing research conducted in the microfinance domain has focused on the comparative performance of various classification algorithms in predicting defaulters with positive results with both parametric and non-parametric models featuring prominently in the studies. Blanco et al. (2013), Baklouti and Bouri (2013), and Baidoo and Arku (2015) all apply logistic regression (LR) to MFIs operating in Peru, Tunisia and Ghana respectively and observe good AUC results but misclassification is somewhat worrying. In Baklouti and Bouri (2013) for instance, there is a misclassification of 20% which can lead to significant operational costs. Given a parametric model is dependent on numerous underlying assumptions, research has also focused on non-parametric models ranging from decision trees to black box algorithms such as neural networks (Blanco et al., 2013; Cubiles-De-La-Vega et al., 2013). Both studies present a comprehensive comparison of classification models with neural networks providing the highest AUC and lowest misclassification cost. Cubiles-De-La-Vega et al. (2013) also tested the performance of decision trees with and without bagging and although standard decision trees were actually the worst performing model in their analysis, results improved dramatically when random forests were used. Features used in this past research are based mainly on demographical data that are not available for this research so it is unclear how the performance achieved in the aforementioned research will associate using communication log data.

2.2 Social Network Impact in P2P Lending

In the P2P lending domain there has been a wide focus on introducing social network data into classification models to improve prediction accuracy. There are two strands of social network data that were used in these studies: using internal and external data. In separate work by Lu et al. (2012), Freedman and Jin (2017) and Lin et al. (2013), social data that is internal to the P2P lending platform is leveraged with mixed results. Contrary to the concept of homophily, it was determined that social connections made online were more likely to lead to default than a real world friend (Lu et al., 2012; Freedman and Jin, 2017). These online connections would generally be considered weak friendship ties, similar to connections on social media sites like Facebook so such an influence is a strange result. Lin et al. (2013) also finds that a borrower's online connections can signal loan outcome but in their analysis an extra dimension is added which indicates the type of friendship (verified or unverified friend) which effectively strips out the noise and it is found that verified friends (closer ties) provide the strongest signal. A drawback of using internal platform data is that it is difficult to accurately determine the real relationship between two connections and it is open to social network fabrication, which people will likely attempt to appear more creditworthy (Wei et al., 2016). This limitation is addressed by Zhang et al. (2016) and Ge et al. (2017) that use alternative sources of social data. Of particular interest is the result of Ge et al. (2017) as they incorporated data from Weibo which resulted in a strong improvement to the prediction of default.

2.3 Classifying Churners Using Customer Call Records

Whereas traditional churn models generally use standard customer data to assign a churn score to each customer, the integration of social network features to models has enhanced performance (Zhang et al., 2012; Kusuma et al., 2013; Backiel et al., 2016).In using network features the predictive performance of various classification models improve than if they are excluded (Backiel et al., 2016). These network features are extracted from call records on customers so this literature provides a good foundation for how call and SMS logs could be employed for the focus of this paper. In addition, the positive results in this domain provide some gravity to their performance potential in microfinance. In terms of network features, Zhang et al. (2012) exploit call records to add features such as neighbour composition to the classification models as well as tie-strengths, where a neighbour is labelled as a churner or non-churner. In the microfinance setting, such labels would become default or non-default. Óskarsdóttir et al. (2017) introduces another interesting feature transformed from the call records that takes into account the time of day and week the communication occurs under the assumption that closer ties are more likely to contact each other during the evening rather than during work. Finally, there are various features introduced to take account of tie-strengths within the network. In most cases, only call data is available so tie-strength is measured as the total or average call duration to a connection. In Kusuma et al. (2013) they have access to call and SMS logs and integrate them by converting the value of one SMS into the duration of a call.

2.4 Relational Learners in Churn Prediction

Relational learners applied on telecommunications data have found them to offer improved performance over traditional classification methods in predicting churners (Dasgupta et al., 2008; Phadke et al., 2013; Abd-Allah et al., 2014; Verbeke et al., 2014; Backiel et al., 2015; Óskarsdóttir et al., 2017). There are numerous relational learners that can be used such as the network-only link based and class-distribution relational neighbour classifier which one of four considered by Verbeke et al. (2014). The most commonly used relational learner however, is the spreading activation model (SPA) first introduced to churn prediction by Dasgupta et al. (2008). In their analysis they compared its performance against decision trees with featurised network variables with the SPA proving far superior.

3 Methodology

The aim of this research is to assess how social network telecommunications data can be used to improve the prediction of defaulters in a Kenyan microfinancing company. More specifically there are two subsets to consider in the analysis: new loan applicants that are present in an existing loan user's social network, and more general loan applicants. In this first subset the assumption is that since a user has social contact to an existing defaulter/non-defaulter, this relationship may influence the outcome of the new applicant's loan. To analyse this effect, a relational learner called the Spreading Activation Model (SPA) is introduced in Section 3.3. For the second subset Section 3.4 outlines how the social network data is featurised for application in classification models, providing a generalised method to predicting default.

Call log		SMS log	
number	phone number of connection	number	phone number of connection
call type	incoming/outgoing/missed	direction	incoming or outgoing
duration	length of time in seconds	date	time in milliseconds since the epoch
date	time in milliseconds since the epoch		

Table 1: Call and SMS JSON keys

3.1 Data Extraction and Preparation

Access was given to the call and SMS logs of the customers of a Kenyan MFI, as well as to user and loan database tables that store some basic user information and the terms of conditions of each approved loan respectively. While this data is structurally stored in a relational database, the SMS and call log data is stored in JSON objects. For efficient harvesting of this data, the primary requirement was to create a relational structure of the data contained within each object. Table 1 shows the keys from each JSON object that were extracted and locally stored for the various processing stages outlined below.

Only communication to mobile phone numbers are fed through to the analysis. Communication to service numbers are not considered as these are assumed to be transactional events that do not form a user's social network. This is particularly relevant since a measurement of tie-strength is of interest in the analysis, as the inclusion of transactional log data can lead to a dilution of tie-strength between two nodes in relative terms.

Close to 40,000 loans have been approved by the company to date but not all of these were suitable for this analysis. The company only began requesting call and SMS logs from customers as of June 2017 so a high proportion of loans given prior to this point in time are removed as there is no communication log data with which to use. As the business operations going forward will function by having such data as a prerequisite for future applicants, the removal of these loans naturally make sense as their inclusion to any prediction model based on social data is redundant. It should also be noted that due to batch processing issues when the company first rolled this out, in some cases the SMS logs are blank for a user, while the call logs are available. Although not having complete sources for each user is a drawback, the call and SMS log data are instead combined into single features to ensure completeness. With all of this considered 16,861 loans remain available for the analysis with a default rate of 13.8%.

3.2 Social Network Graph

Using the communication logs for each user, a network graph can be generated. The nodes within each graph relate to each individual person that had interactions and the edges are established between the nodes that had observed communication such as a call or SMS. In a directed network, the edge represents the communication that originates at the source node and ends at the destination node. In an undirected network, two-way communication is included (incoming and outgoing). For this analysis the edges will take the value of the relative tie-strength which will be composed of both call and SMS communication and this calculation will be described in Section 3.3.1. Figure 1 shows an example of a network flow using the total directed tie-strength while Figure 2 shows the relative tie-strengths.

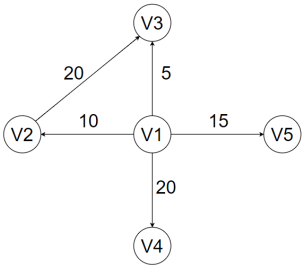

Figure 1: Graph using total directed tie-strength

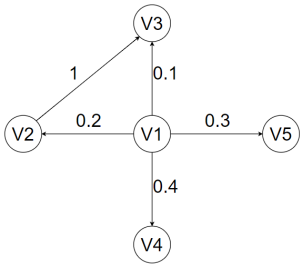

Figure 2: Graph using relative directed tie-strength

3.3 Spreading Activation Model

Given the positive results achieved with the application of the spreading activation model (SPA) in churn prediction (Dasgupta et al., 2008; Kim et al., 2014; Backiel et al., 2015) the algorithm will be harnessed for this research to assess if such a relational technique is more beneficial than standard classification methods.

In the churning prediction problem, the objective is to determine which customers currently availing of a telco provider's service are potentially going to churn in the future. The method is applied at a point in time and gathers the customers that have churned up to that point as seed nodes, sets their energy level to 1, and diffuses this energy across the network to determine what other nodes (customers) attain an energy level indicative of a churner (based on an energy threshold value).

For the default prediction, such methodology cannot be readily applied on such a broad network scale. At a given point in time, while the defaulters up to that point can be easily identified, any active user node has already obtained a loan so labelling them as a defaulter/non-defaulter at this point becomes redundant. Instead, the methodology will be applied to uniquely defined networks for each new loan applicant with the loan applicant effectively acting as the central hub.

3.3.1 Tie-Strength Measurement

Research by Wiese et al. (2015) indicated that SMS and call logs are useful (but not perfect) for measuring tie-strength and instead should be combined with multiple touch points to enhance their utility. The availability of additional messaging apps such as WhatsApp or Facebook messenger would indeed be a beneficial resource but in reality, access to such a range of communication apps is very unlikely. Much of the research in churn prediction has focused on using only the duration of calls to measure tie-strength, but analysis carried out by Baras et al. (2014) demonstrates that significant improvement in prediction is possible by combining call and SMS communication. We therefore align with an approach similar to that followed by Kusuma et al. (2013) where tie strength is calculated using the duration of a call made by a user to a connection and the number of SMS messages also sent to this connection.

Tie strength will be measured in seconds so the value of an SMS message in seconds is needed. Only one telecom provider is currently supported by the company. This telco charges 10 Kenyan shilling (KES) for a bundle of 200 SMS messages while the cost of a phone call per minute costs 1KES. This equates 3 seconds of a phone call per SMS so the weight for each SMS count will be factored by 3. So tie-strength is measured as:

$$w_{x,y} = \sum_{i=1}^{N} F_i * w_{x,y,i} \tag{1}$$

Where F (1 for call, 3 for SMS) is the factor to apply to the outgoing communication i of total outgoing communications N from node X to node Y.

This equation of tie-strength covers a directed graph where communication sent from X to Y is only considered. To make this undirected, the communication sent from Y to X can also be included so that:

$$w_{x,y} = \sum_{i=1}^{N} F_i * w_{x,y,i} + \sum_{j=1}^{M} F_j * w_{y,x,j} \tag{2}$$

3.3.2 Default Influence Diffusion

Taking X to be a defaulter, this node is initiated with a value of 1 for the first iteration of the algorithm. To propagate this energy to a connecting node Y, a transfer function using linear edge weight normalisation is applied with the value of tie-strength between X and Y relative to total tie-strength value of all neighbouring nodes N connected to X:

$$T_{x,y} = \frac{w_{x,y}}{\sum_{i}^{C} w_{x,i}} \tag{3}$$

Given this fraction of the influence from X to be passed to Y, the total amount of influence received by Y at step i will be:

$$I_{x,y} = d * T_{x,y} * I_{x(i)} \tag{4}$$

Where $I_{x(i)}$ is the default influence energy on node X on iteration i and d is the spreading factor to be chosen. The spreading factor d (between 0 and 1) represents the amount of energy that a node will spread to their neighbours while retaining (1-d) amount of energy. By initialising a high value of d, it is possible for a node to influence other nodes that are further away in the network, while a low value restricts the influence around their more immediate connections.

3.3.3 SPA Algorithm Stages

The process of the SPA relies on two main stages: initialisation and spreading.
Initialisation:

- Identify defaulters in network
- Initiate the default energy value on these nodes to 1
- Set the value of the spreading factor to be used
- Define a minimum default energy threshold a node must have to be active

Spreading steps:

1. Begin algorithm by finding active node seeds (nodes with default energy > threshold)
2. Diffuse the default energy to each neighbouring node using the transfer function and spreading factor
3. Identify the new nodes that have sufficient default energy to be used as activation nodes in next iteration
4. Repeat from step 2 until max iteration or there is no more activated nodes

Calculating new default energy levels for each loan applicant, ROC curves can be generated to compare the performance of the technique versus alternative classification methods.

3.4 Traditional Classification

To provide a model applicable to any new loan application and as an alternative approach to the spreading activation method, traditional classification methods are introduced. As there is an absence of demographic variables for each subscriber, the models will rely on communication features of an individual's call and SMS log. To incorporate further features, the network structure within each of these logs are also transformed through feature engineering. In the context of a loan applicant's social network, this featurisation enables the log records to be split between the types of connections within the network which is especially useful when we consider that a connection can be an existing in-app user that has either defaulted or not defaulted on their loan. In this way, not only will basic features of communication such as the total number of incoming communication events be generated, but also the total number of incoming communication events from a defaulter or a non-defaulter.

3.5 Feature Engineering

To overcome the issue of missing SMS logs that affect some users, the method used to calculate tie-strength in Section 3.3.1 is applied where the SMS is equated to the price of a call in seconds. Instead of having a separate feature for a call and for a SMS, the events are combined to ensure that features passed to the classification models are complete. For example, the count of outgoing SMS and calls are added together to create a count of all outgoing communication events.

Within each loan applicant's communication logs, the node on each log entry is determined to be an external source or an in-app user based on whether the phone number in the log entry exists within the user table in the database. In addition, a loan outcome label is imputed onto the user-friend instance to enable more features to be defined by whether the communication was between a defaulter or a non-defaulter in the network.

It is important to note that the definition of a defaulter in this situation is different from the operational definition that labels a loan as defaulted if the loan remains unpaid after 120 days from the loan start date. For this purpose, a loan is labelled as defaulted if it remains unpaid after the initial loan period of 28 days expires. Such a reduced time frame may seem overly harsh but this is done to provide more relevancy to the model as waiting 120 days before labelling the loan outcome is subject to understate the default status of the loan book on each day that a loan application is received.

Finally, the start time of a communication log is used to group each log entry into three categories: weekday work hours (Mon-Fri: 08:00-17:30), weekday evening hours (Mon-Fri: 17:30-08:00), and weekend days. The time of day or week that communication takes place can be indicative of strong ties and Motahari et al. (2012) show that friends and family are more likely to contact each other during the weekend or in the evening. Since it is assumed that influence is more likely to exist between stronger ties, featurising this information can add valuable information.

With this additional grouping information now part of a loan applicant's communication logs, a number of local and network features can be generated, a subset of which are shown in Table 3.

The generation of tie-strength features and the addition of the default absorption features again borrow from the methodology of the spreading activation model. In this application however, the tie-strength (both total and relative) are also calculated with the feature splits between the type of connection and time of day/week.

Usage Features	Breakdown of Usage Features	Normalised Usage Features
Number of communications	Number of communications during weekday work time	Daily communications
Number of incoming/outgoing communications	Number of communications on weekday evening	Daily incoming/outgoing communications
Duration of communications	Duration of incoming/outgoing communication on weekend	Daily duration of incoming/outgoing communications
Duration of incoming/outgoing communications		Daily communications during weekday work time
Number of log days available		
Network Features		
Number of communications to a defaulter	Number of communications during weekday work time to in-app user	
Number of incoming/outgoing communications to defaulter	Number of communications on weekday evening to in-app user	
Duration of communications to non-defaulter	Duration of incoming/outgoing communication on weekend to defaulter	
Tie-strength Features		
Total tie strength directed	Total tie strength directed to defaulter	Relative tie strength directed to defaulter
Total tie strength undirected	Total tie strength undirected to defaulter	Relative tie strength undirected to defaulter
Total tie strength undirected on weekend	Total tie strength undirected to defaulter on weekend	Relative tie strength undirected to defaulter on weekend
Default Energy Absorption		
Default absorption directed	Default absorption undirected	

Table 2: Local and Network Features

Feature Set	Group	Content	Number
1	Basic	Usage features only	38
2	Connection Types	Usage features and network features	149
3	Tie-strength	Usage features, network features and tie-strength	213
4	Default Absorption	Usage features, network features and default absorption	151
5	Every Feature	Every available feature	215

Table 3: Feature set composition

Two extra features were added to the tabular classification data: directed and undirected default absorption from neighbour. Using the real time probability index from the curves in Figure 3, an absorbed default value is calculated based on the tie-strength and probability of default for each friend in a users communication log (provided a loan exists for that user at the application date). This is a similar concept to the SPA except the loan applicant is absorbing the default energy rather than being sent it. The default energy absorbed by a loan applicant from their connection N, at application time t is:

$$\sum_{i=1}^{N} w(x,i) * D(i,t) \tag{5}$$

where w(x,i) represents the directed or undirected relative weight to connection i and D(i,t) is the real time default value of the connection at the time of application.

With the complete set of features defined, five feature sets were compiled in order of easiest to most difficult to set up and each then used in training the classification models in order to test the benefit in adding both complexity and information (see Table 3).

3.6 Probability of Default Curves

Historical loans were also used to generate default probabilities based on whether any payment was received from the loan recipient by a certain day since the loan was drawn

		Days from beginning of loan to first payment						
		>= 0	>=7	>=14	>=21	>=28	>=35	>=42
Loan Outcome	Default	6981	6927	6828	6777	6741	6676	6621
	Repaid	31320	22718	15071	11058	7943	4712	3322
Total observed outcomes		38301	29645	21899	17835	14684	11388	9943
P(D|first payment >X days)		18.2%	23.4%	31.2%	38.0%	45.9%	58.6%	66.6%

Table 4: Probability of default given the number of days to first loan repayment

down. Table 4 shows an example of this calculation across different intervals of days to first loan repayment received. What is interesting to note is that when a borrower has paid nothing by day 14, they carry a risk of default of 31.2% and this risk only increases as day from the start of the loan increases.

In order to make this information accessible to the analysis, daily probability curves are generated using only the information available in the system up to that point in time. This ensures that the probability of default passed to any model is dependent on the knowledge known at that moment so loans that have an observed outcome up to that date are the only consideration (so effectively a loan must have existed for > 120 days by that date). Figure 3 illustrates how the probability curves evolve as time move forward and more loan outcomes become available. Shown are four dates since the company first approved a loan. Date 1 relates to a date on which no observed loan outcome was available, hence the probability is 0%, whereas Date 4 is associated to a more recent day for which the full sample of loan outcomes are available.

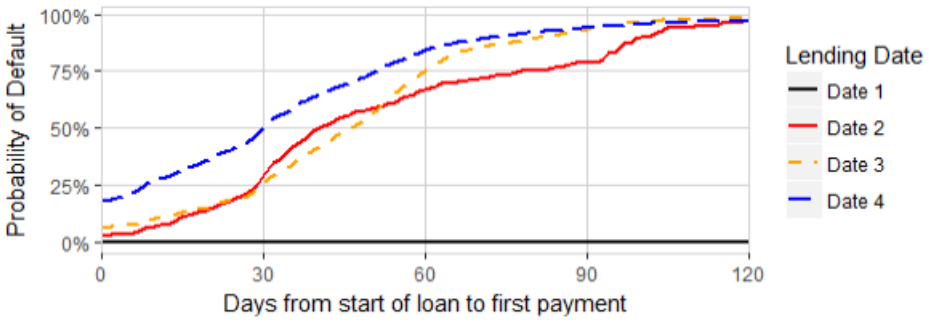

Figure 3: Probability of default given no payment received by day X

4 Implementation

In this section, the experimental setup is described along with the various models that were implemented to test each experiment.

4.1 Experimental Design

Aside from the overall objective of improving default prediction, another concern of the microfinancing company is that users who default on their loans are an influencing factor for defaults observed on future loan users. The model proposed to predict such an effect is the SPA method. For a loan instance to be considered using this model however, there is a prerequisite that must be passed which is that:

- At the time a user applies for a loan, they are present within an existing loan user's communication logs. Since the SPA method spreads influence from node X to Y, node Y can only receive influence if they exist in node X's network.

	Loan Count			Default Rate		
	Training	Test	Total	Training	Test	Total
Influencer Loans	2165	927	3092	17.1%	17.2%	17.1%
Standard Loans	9639	4130	13769	13.1%	13.1%	13.1%
All Loans	11804	5057	16861	13.8%	13.8%	13.8%

Table 5: Loan size and default rates for sample sets

Of the 16,861 loans available for analysis, only 3,092 loans satisfy this condition. Three samples of loans are therefore considered, sample set A which are composed of 3,092 loan which will be labelled as "Influencer Loans", sample set B containing 13,769 which will be labelled as "Standard Loans" and sample set C which will be labelled as "All Loans" and contains all 16,861 loans.

For each sample, 70% was used as the training set and 30% used as the test set. The split was applied using stratified sampling. Stratified sampling is used since there is a moderate imbalance of the target default class to ensure that the proportion of defaults between each set remain the same. The train and test sets for the "All Loans" sample set are created from the train and test splits that are made for "Influencer Loans" and "Standard Loans" to ensure consistency. To train the classification models 10-fold cross validation with stratified sampling for the folds on the 70% training set was used. The size and default characteristics of each of these sample sets is shown in Table 5. To label the classes for train and testing, the operational definition of default is used so that a loan is defaulted if it is not fully paid by 120 days from the draw down date.

4.1.1 Analysis of SPA versus Classification Algorithms

To test whether the SPA model provides a more effective method to predict a loan default based on influence spread from social network connection, the "Influencer Loans" sample set is used. The default energy output from the SPA models for the training set loans are used to tune an optimal threshold value to class the loan as default or non-default. This threshold value is then applied to the test set to measure the models performance. For the Classification algorithms, 10-fold cross validation is applied on the training set and the optimal models are then tested to evaluate their performance. The performance results of the SPA models are then compared to the classification algorithms.

4.1.2 Analysis of a Single Model or Combined Models

To build and evaluate a general model using a loan applicant's communication logs, the analysis is designed to test whether a classification model trained on the "All Loans" data performs better than two separate models, independently trained on the "Influencer Loans" and "Standard Loans" subsets. This is to determine if there are potential characteristics within each subset that may be better served with an independent model.

4.1.3 Analysis of the Benefit from Network Features

As a follow on to the general classification model evaluation, all of the classification models are trained on each of the feature sets from Table 3 to determine how the performance of models benefit from adding extra features with emphasis on network features such as tie-strengths between individuals from a users social graph.

4.2 SPA Models

There are many variations of model setup that can be explored with the SPA method. The relative tie-strengths can be directed or undirected, the spreading factor d can take any value between 0 and 1, any value of max iterations can be set and for this application three methods of node activation energy can be used to impute values upon initialisation of the model (outlined in Section 4.2.2). For the spreading factor d, values incremented at 10% intervals are applied and models with the max iterations parameter set as 1, 2, and 3 are also used. Using all parameter value combinations a total of 162 SPA models were built and applied to the communication networks of each loan applicant in the sample.

4.2.1 Communication Networks

The communication networks of a loan applicant are created to incorporate a community depth of 2. This means influence can be spread to the loan applicant from an immediate friend, and from a friend of an immediate friend. The max iteration parameter values were selected to support this network depth where in the case of an iteration value of 3, influence can reach the loan applicant by spreading through three edges in the network.

4.2.2 Activation Energy

In the initialisation phase in section 3.3.3 it was stated that the energy value for identified defaulters gets set to 1. In doing so, this inputs the default information into the network to be propagated. For this analysis, three states of default are considered for this activation: operational, fixed-value and real-time. The rational in doing so is to analyse how the introduction of different default definitions affect the model performance.

The operational default state is simply the definition used by the company to label a defaulter which occurs if the loan remains unpaid beyond 120 days. The fixed-value default state is similar to the operational except that it labels the loan as defaulted if it is unpaid beyond 28 days which is the original loan term length. The real-time default state imputes a probability of default on each node based on the status of their loan on the loan creation date for the new applicant being assessed (see Figure 3). Figure 4, 5 and 6 illustrate these three states and it can be seen that in the real-time state, all nodes contain some degree of default probability.

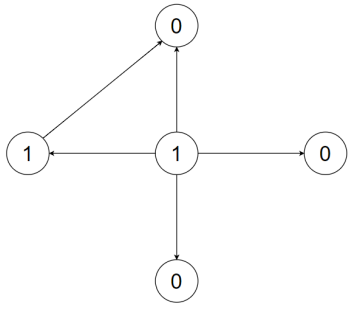

Figure 4: Activation using operational default

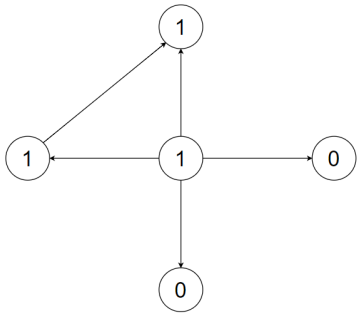

Figure 5: Activation using fixed-value default

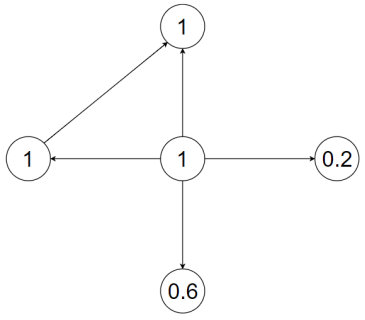

Figure 6: Activation using fixed-value default

4.3 Classification models used

A number of supervised machine learning algorithms were trained and evaluated. These included logistic regression, logistic regression with lasso regularisation, decision trees,

random forests, neural networks and support vector machines (with radial based function (RBF), polynomial and linear kernel although only the RBF provides reasonable results). The caret package in R was used to tune the parameters of each algorithm.

Although the black box algorithms of NNs, SVMs and random forests suffer from opaqueness issues and restrict the understanding about how an instance is classified, they can be top performers (Cubiles-De-La-Vega et al., 2013; Blanco et al., 2013).

4.4 Performance Measurement

To evaluate the performance of each model, two metrics are considered. Since there is moderate imbalance in the target variable (17% default rate), classification accuracy is not suitable as the performance of the model on the non-default level is likely to overwhelm the overall performance. Instead the true positive rate (TPR) and the false positive rate (FPR) are used to generate ROC plots for each model with the area under the curve (AUC) providing a comparative metric for how well each model can correctly predict true instances of default without misclassifying good loans.

The second metric of model performance is the potential profitability of the model predictions where both the cumulative profit and the return on investment (ROI) is measured. This is calculated by associating the loans predicted as non-defaulters by the model (true negatives and false negatives) to the profit/loss realised when a loan is approved to a non-defaulter and defaulter. For example, the profit of each good loan is equal to the interest rate charged which is 15% while the loss incurred is equivalent to the value of the loan itself which is 100%. Therefore the total profit of the model is given by:

$$\text{Profit} = 0.15\% * TN - 1 * FN \tag{6}$$

and the return on investment given by:

$$\text{ROI} = \frac{\text{Profit}}{TN + FN} \tag{7}$$

5 Evaluation

In this section, the results of the implementation scenarios presented in Section 4 are evaluated. As indicated, the performance of the models are analysed with respect to the AUC and also from a business operational point of view, using profit and ROI.

5.1 Evaluation of SPA versus Classification Algorithms

The first question of interest was to determine if the spreading activation model provided a more effective solution than traditional classification methods in predicting defaulters by exploiting the underlying communication network that a loan applicant exists in.

The top 5 best performing models in order of overall cumulative profit is shown in Table 6 along with the two best SPA models. Between the relational model and classification algorithms trained and tested, the traditional methods of feeding tabular data into classification models result in better performing models. The random forest trained on the communication features along with the added tie-strength features provides the highest AUC of 0.75 along with maximum profit equating to the monetary value of 38.9 loans out of 497 approved (loan approval rate of 54%). From a business operational point of view, depending on capital resource/availability, the random forest classifier

Feature Set	Algorithm	AUC	Max Profit	Loans Approved	% Loans Approved	ROI
3	RF	0.7503	38.9	497	53.6%	7.8%
5	RF	0.7382	37.25	555	59.9%	6.7%
4	RF	0.7314	35.65	345	37.2%	10.3%
1	RF	0.7238	34.9	432	46.6%	8.1%
2	RF	0.7257	33.2	482	52.0%	6.9%
	SPA	0.5544	0	0	0.0%	0.0%
	SPA	0.5535	0	0	0.0%	0.0%

Table 6: Top classification algorithms for 3092 subset of loans meeting criteria for SPA

trained on the feature set that includes the default absorption predictor variables is also worthy of consideration given a higher return on investment (ROI) of 10.3% on a lower loan approval count (less capital required).

The spreading activation models do not perform well. Of the SPA models tested, the maximum AUC is 0.554 which is barely better than a toss of a coin. The best performing model used the fixed-value default activation energy, with two transfer iterations and undirected tie-strength. The spreading factor for this model also proved insignificant as any value on this model parameter provided almost the same outcome. In terms of profit, there is nothing to be gained with a 0% return only possible by effectively taking a "do nothing" approach and classifying all loans as default.

Using the real-time default value as the initial node activation also performs quite similarly with an AUC of 0.553. The model used just one transfer iteration and directed tie-strength with a spreading factor of 20% or 30%. While the result is only marginally lower than the best SPA model, it is somewhat surprising that the real-time default probability did not have a more positive impact than the fixed-value given it was providing default energy to all nodes in the relational model.

A possible reason for this is the fact that the probability values imputed on the nodes are dependent on loans that were drawn down 120 days previously, so it is quite reasonable to assume that the information did not reflect the current state. This is hinted at in Figure 3 which shows how the shape of the default probability curve steepens as time moves on. What this effectively means is that the imputed probabilities applied on the network of a loan application at time T in the SPA model are lower than the actual effective value at that time T. The fixed-value on the other hand labels an existing loan as default if it is still unpaid at the end of the original loan term of 28 days so it potentially adds information more reflective of the actual state.

Looking at Figure 7 it is interesting to note the profit value at a probability threshold of 100%. This effectively indicates the outcome of the current regime where every loan is indiscriminately approved - so any profit value above this point is representative of improvement. Although the SPA model does indeed provide some benefit in this regard (and also relative to the random forest model at some points), it sill leads to negative profit so it is not a model that should be recommended within this analysis.

Sample=5057		Single Model					Combination of Models				
Feature Set	Model	AUC	Max Profit	Loans Approved	% Loans Approved	ROI	AUC	Max Profit	Loans Approved	% Loans Approved	ROI
3	RF	**0.7548**	**282.60**	**3402**	**67.3%**	**8.3%**	0.7503	273.40	3126	61.82%	8.7%
5	RF	0.7551	280.00	3446	68.1%	8.1%	0.7438	268.75	3003	59.38%	8.9%
2	RF	0.7558	273.35	3463	68.5%	7.9%	0.7337	264.30	3004	59.40%	8.8%
1	RF	0.7492	263.35	3289	65.0%	8.0%	0.7373	268.20	3007	59.46%	8.9%
4	RF	0.7487	268.10	2968	58.7%	9.0%	**0.7371**	**259.15**	**2686**	**53.11%**	**9.6%**
1	Nnet	0.7113	259.50	2811	55.6%	9.2%	0.5960	155.55	2210	43.70%	7.0%
4	Nnet	0.7038	252.65	2727	53.9%	9.3%	0.5915	192.10	2377	47.00%	8.1%
3	GLM net	0.7207	246.15	2768	54.7%	8.9%	0.7209	239.65	2671	52.82%	9.0%
2	Nnet	0.7158	244.65	2666	52.7%	9.2%	0.6089	193.95	2420	47.85%	8.0%
2	GLM net	0.7290	243.25	2764	54.7%	8.8%	0.6950	237.85	2498	49.40%	9.5%

Table 7: Top results comparison between single and combination models

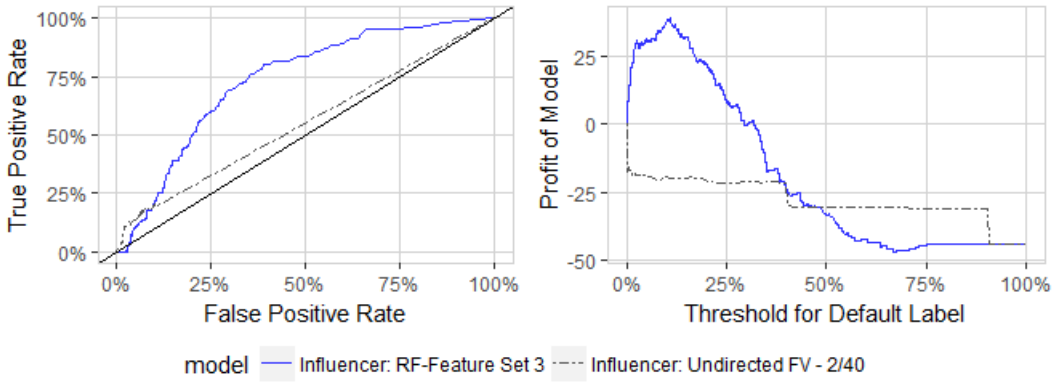

Figure 7: ROC Plots and Profit: Best classification model versus best SPA model

5.2 Evaluation of a Single Model or Combined Models

Through the applied methodology the data set was split into three loan sample sets: "Influencer Loans", "Standard Loans" and "All Loans". As the SPA models in section 5.1 underperform the traditional algorithms and provide no operational benefit, they are discounted from this stage of the analysis. Although it might be preferable to consider a single model, we evaluate the performance of a classification model trained on all of the loans and compare this against the outcome of two independent classifiers trained on the "Influencer" and "Standard" loan samples. It is possible that the characteristics of these samples has some degree of uniqueness and may be better served by individual models.

Only the top 10 model comparisons are presented in Table 7 and although the random forests dominate the results, note that neural nets and logistic regression with lasso regularisation do perform admirably and generally offer higher ROI than many of the random forest alternatives. Overall, the single model implementation outperforms the combination of models. From an academic point of view the preferred model is more commonly measured using AUC which would indicate the random forest trained on feature set 2 (includes usage features and network features) as the marginally better model with an AUC of 0.7558, but from a business perspective the random forest that includes tie-strength features results in the highest profit with a monetary value of 282.6 loans through acceptance of 67.3% of the sample test loans.

Considering that capital resource may be an important consideration for the business, the ROI may be a more central measurement of model performance in which case the combination of models could offer more flexible benefit. A random forest model training individually on each loan sample set including the default absorption features has the potential to achieve a ROI of 9.6%. What this is effectively telling us is that for every

Prediction Group	Trained Model	AUC	Max Profit	Loans Approved	% Loans Approved	ROI
Influencer	Single	0.7275	34.05	549	59.2%	6.2%
	Influencer	0.7351	38.90	497	53.6%	7.8%
Standard	Single	0.7584	248.55	2853	69.1%	8.7%
	Standard	0.7503	234.50	2629	63.7%	8.9%

Table 8: Overall comparison between single model and subset model predicting specific subset

100 loans approved and disbursed by the company, the value of 109.6 loans are returned.

Considering the top profit random forest model, Figures 8, 9 and Table 8 show ROC and profit plots comparing the performance between the single model and each model trained on the "Influencer" and "Standard" subsets respectively. In both cases the predicted results of the single model are only related to the relevant test subsets so we get a better understanding as to where the single model outperforms the individual builds.

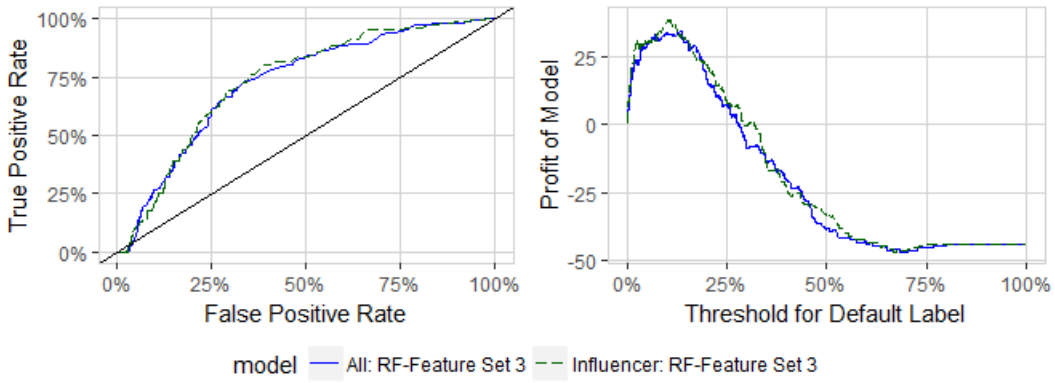

Figure 8: ROC Plots and Profit: Single model versus model trained on "Influencer" subset

In Figure 8 it is actually the model trained solely on the "Influencer" subset that delivers better performance than the single model. Although the difference is only marginally beneficial, it does suggests that this special cohort of loan applicants who know existing borrowers have different characteristics than the overall sample and are better classified using an independently trained model. The main victory is against the "Standard" subset model which likely indicates that the single model was able to use patterns from the "Influencer" subset that allowed it improve its generalisation. With this result, an optimal strategy for the company to consider deploying would be to utilise both approaches, where two models are trained, one on the "Influencer" subsets and one on the complete subset (to attain better generalisation). The "Influencer" model would be used to predict the smaller subset of loans while the main model would classify all others.

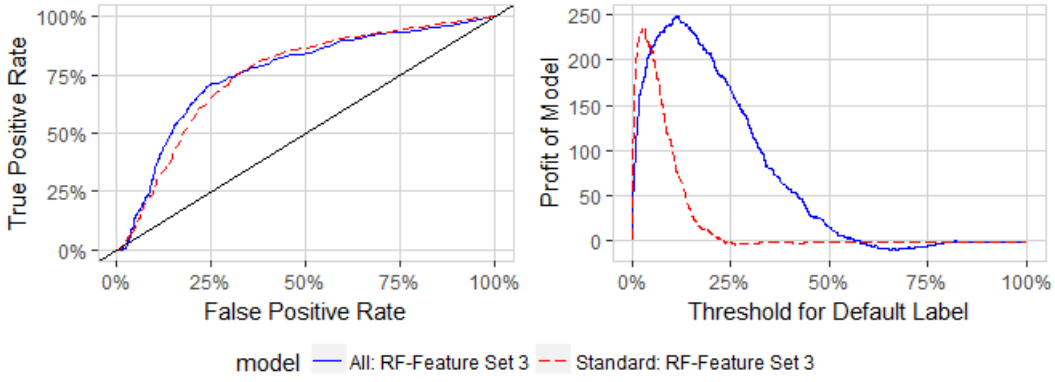

Figure 9: ROC Plots and Profit: Single model versus model trained on "Standard" subset

5.3 Evaluation of the Benefit from Network Features

As an aside to the overall model development, different feature sets of varying properties were also used to train the model to assess the impact of their inclusion. The result of this analysis is based on the single model approach.

Considering the AUC in Figure 10, the performance of the random forest, GLM net, and SVM radial models are relatively stable with different feature set properties. The GLM logistc regression model on the other hand shows a significant drop in performance when tie-strength features are added. This is very likely caused by multicollinearity between tie-strength features and usage features given they are a dependent function.

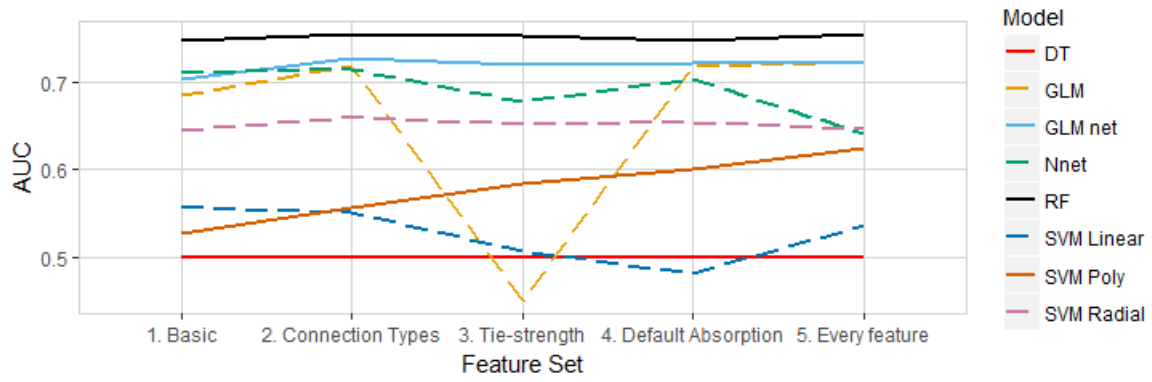

Figure 10: Comparison of AUC on single model using each feature set

Looking closer at Figure 11 the AUC is improved in the random forest model by adding both network features and tie-strength features. Taking the model profit into account the addition of the tie-strength features increases profit by 7% from a monetary value of 263.35 loans using the basic features to 282.6. Seeing how much the network features improve performance is a welcome result and ties in with existing research in churn prediction. With respect to the research question, adding the tie-strength features does improve performance but it is model dependent. Logistic regression was negatively affected but training a model on these features only many prove more reliable and is something that can be looked at going forward.

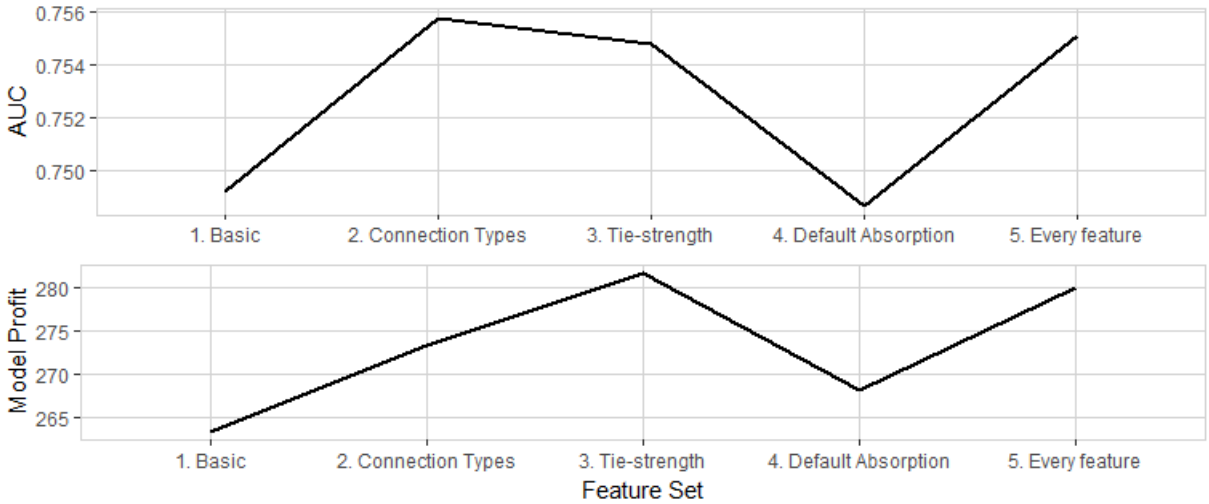

Figure 11: Comparison of AUC and profit on random forest using each feature set

6 Conclusion and Future Work

The objectives of this research had three primary questions to answer. First was to introduce a relational learning method and determine if it was better in default prediction over traditional classification methods. While results indicate that the predictions from the spreading activation model are generally inferior to the classification algorithms and do not offer meaningful operational benefit to the company, only one of many relational learner methods was tested. Given that network sparsity may also have a part to play in the poor results, future work could introduce other relational learners such as a network-only link based classifier when a denser sample set of loans are available. Additionally, researching various transfer functions may yield more promising findings. Only a single function was used in this analysis based on linear edge weight transfer. Introducing non-linear edge weights would effectively penalise weak connections and buffer the influence of strong ties. Another element that could be considered going forward is how featurising the score from the SPA method and adding it to a classification model improves performance.

The second central question was to understand whether classification models dependent on communication data could effectively predict defaulters. To provide a comprehensive answer to this, three different data subsets were used to train the models to understand if particular cohorts of loan applicants were better classified using independent models. The findings were interesting in this regard. While a single random forest model does provide the best AUC and profit when compared with the combination of a model trained separately on the "Influencer" and "General" loan subsets, it is shown that the model from the "Influencer" subset should be considered as a stand alone classifier in the operational process.

Finally, the various feature sets generated from communication logs were used to train several classifiers to analyse their benefit. Of most interest was tie-strength features and it was found that their addition mostly had a positive impact on performance, particularly on the random forest algorithm which realised a 7% increase in overall profit compared with a model dependent only on basic usage features. This result provides evidence that social network information is a very valuable asset. It supports findings in the P2P lending where such data improved P2P loan prediction but in this presented research, as the social data is from a user's call and SMS logs, the measure of tie-strength is more objective and less prone to user fabrication, thus adds more legitimacy to the results.

Based on these findings, further work could look at better refinement of the tie-strength measurement. A linear combination of call duration and SMS counts were used to calculate tie-strength but introducing extra data sources to this measurement may improve results further. For example, adding context to the relationship by using similarity of interests was initially considered for this research by comparing the overlap of downloaded apps between two users. However, this data proved too sparse at the time of research but in the future it is certainly something that should be considered.

Acknowledgement

I would like to extend my gratitude to Paul Laird for his time and helpful supervision throughout the course of this research.

References

Abd-Allah, M. N., Salah, A., and El-Beltagy, S. R. (2014). Enhanced Customer Churn Prediction using Social Network Analysis. *Proceedings of the 3rd Workshop on Data-Driven User Behavioral Modeling and Mining from Social Media*, pages 11–12.

Backiel, A., Baesens, B., and Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, 67(9):1135–1145.

Backiel, A., Verbinnen, Y., Baesens, B., and Claeskens, G. (2015). 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 - ASONAM '15*, pages 651–658.

Baidoo, I. K. and Arku, D. (2015). *Ghanaian journal of economics.*, volume 3. The African Finance & Economics Consult.

Baklouti, I. and Bouri, A. (2013). Can credit-scoring models effectively predict micro-loans default? Statistical evidence from the Tunisian microfinance bank. *Global Credit Review*, 3:57–69.

Baras, D., Ronen, A., and Yom-Tov, E. (2014). The effect of social affinity and predictive horizon on churn prediction using diffusion modeling. *Social Network Analysis and Mining*, 4(1):1–12.

Blanco, A., Pino-Mejías, R., Lara, J., and Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1):356–364.

Cubiles-De-La-Vega, M.-D., Blanco-Oliver, A., Pino-Mejías, R., and Lara-Rubio, J. (2013). Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques. *Expert Systems with Applications*, 40(17):6910–6917.

da Kammoun, A. and Triki, I. (2016). Credit Scoring Models for a Tunisian Microfinance Institution: Comparison between Artificial Neural Network and Logistic Regression. *Review of Economics & Finance*, 6(1):61–78.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. (2008). Social Ties and their Relevance to Churn in Mobile Telecom Networks. *Edbt*, pages 1–10.

Freedman, S. and Jin, G. Z. (2017). The information value of online social networks: Lessons from peer-to-peer lending. *International Journal of Industrial Organization*, 51:185–222.

Ge, R., Feng, J., Gu, B., and Zhang, P. (2017). Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. *Journal of Management Information Systems*, 34(2):401–424.

Kim, K., Jun, C. H., and Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications*, 41(15):6575–6584.

Kusuma, P. D., Radosavljevik, D., Takes, F. W., and van den Putten, P. (2013). Combining Customer Attribute and Social Network Mining for Prepaid Mobile Churn Prediction. *BENELEARN 2013: Proceedings of the 22nd Belgian-Dutch Conference on Machine Learning*, pages 50–58.

Lin, M., Prabhala, N. R., and Viswanathan, S. (2013). Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *Management Science*, 59(1):17–35.

Lu, Y., Gu, B., Ye, Q., and Sheng, Z. (2012). Social influence and defaults in peer-to-peer lending networks. *Thirty Third International Conference on Information Systems*, pages 1–17.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444.

Motahari, S., Mengshoel, O., Reuther, P., Appala, S., Zoia, L., and Shah, J. (2012). The Impact of Social Affinity on Phone Calling Patterns : Categorizing Social Ties from Call Data Records. *Proc. of the Sixth Workshop on Social Network Mining and Analysis.*, 12:1–9.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., and Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, 85:204–220.

Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D., and Doran, D. (2013). Prediction of subscriber churn using social network analysis. *Bell Labs Technical Journal*, 17(4):63–75.

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

Wei, Y., Yildirim, P., den Bulte, C., and Dellarocas, C. (2016). Credit Scoring with Social Network Data. *Markeing Science*, 35(March):234–258.

Wiese, J., Min, J.-K., Hong, J. I., and Zimmerman, J. (2015). "You Never Call, You Never Write": Call and SMS Logs Do Not Always Indicate Tie Strength. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 765–774.

Zhang, X., Zhu, J., Xu, S., and Wan, Y. (2012). Predicting customer churn through interpersonal influence. *Knowledge-Based Systems*, 28:97–104.

Zhang, Y., Jia, H., Diao, Y., Hai, M., and Li, H. (2016). Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending. In *Procedia Computer Science*, volume 91, pages 168–174.