Credit-worthiness Prediction in Microfinance using Mobile Data: A Spatio-Network Approach

Research-in-Progress

Tianhui Tan
Department of Information Systems,
National University of Singapore
13 Computing Drive, Singapore 117417
tianhui.tan@u.nus.edu

Prasanta Bhattacharya
Department of Information Systems,
National University of Singapore
13 Computing Drive, Singapore 117417
prasanta@comp.nus.edu.sg

Tuan Q. Phan
Department of Information Systems,
National University of Singapore
13 Computing Drive, Singapore 117417
disptq@nus.edu.sg

Abstract

Many communities in underdeveloped and developing economies of the world suffer from lack of access to personal credit via formal financial institutions, like banks. However, with the rapid increase in Internet and mobile phone penetration rates, firms are now trying to circumvent this problem using novel technology-enabled approaches. In this research, we leverage a real-world dataset obtained in collaboration with a microfinance firm to show that locational data from mobile phones, coupled with information about communication networks, can be effectively exploited to improve prediction of loan default rates. Specifically, we draw upon recent work in network cohesion based regression modeling to develop a model that uses locational predictors, but within a networked context. We contend that the results from our research can not only illuminate how locational data might be used in assessing creditworthiness, but also empower microfinance firms in resource-poor communities with novel methods for credit scoring.

Keywords: Microfinance, Credit scoring, Locational data, Network cohesion, Logistic regression

Introduction

Individuals in developing and underdeveloped countries often lack access to basic financial services. For instance, a recent report suggests that up to 2 billion people in the world still do not have a bank account, despite a 20% drop in this number (Financial Exclusion 2015). One key consequence of this exclusion is the difficulty in applying for personal loans or credit. Since payment institutions have no history of the applicant's creditworthiness, nor any personal information about their current social and financial status, most people in resource-poor communities have to look for alternate avenues (e.g. informal lending networks) to secure financial credit. There have been some recent advances at alleviating this problem through micro-loan institutions, like the Grameen Bank from Bangladesh (Yunus 2007), pro-social lending platforms like Kiva.org (Hartley 2010), and mobile-based financing systems like the M-Pesa (AllAfrica.com 2016). However, in the absence of any formal credit-scoring method for micro-financing institutions, the return from such ventures is often fraught with high risk and uncertainty (Chen et al. 2010; Assefa et al. 2013).An interesting observation, however, is that many of these countries have a high penetration of mobile devices and internet access (GSMA Intelligence 2015). As a result, some organizations in these countries are seeking to leverage emerging technologies, like smartphones, to offer access to basic banking facilities and credit (Green 2014; Dwoskin 2015; Groenfeldt 2015). Drawing on this emerging ecosystem, in our current research, we propose two important sources of information to improve assessment of creditworthiness in such communities – locational and social network data from mobile devices and telecommunication infrastructures. Further, we propose a model to predict creditworthiness based on such multivariate locational data, and perform a pilot study to validate our model on a real-world microloan dataset from an emerging economy in Southeast Asia.

Specifically, the research question we ask through this study is whether we can improve prediction of creditworthiness of an individual by incorporating information on his/her physical mobility. However, the empirical process of performing location-based prediction of user behavior is complicated by the network autocorrelation problem - that the observable behavior of two individuals tend to be more correlated if they are connected in some way, as opposed to when they are not (Leenders 2002; Doreian 1989). We address this empirical challenge by offering a method that leverages daily locational data about users, while also incorporating possible social ties among them.

To empirically validate the effectiveness of our approach, we obtained anonymized backend data from a smartphone-based financial services company offering microfinance services in a developing Southeast Asian country. As part of their loan-issuing process, the company requires their applicants to install their app and provide permissions to access their phone usage and social media usage data. In addition, the app also records the GPS-location data using a native Android service. We conjecture that the type of location most-visited by the borrowers can potentially act as a proxy for their consumption behavior as well as reflect their socio-economic status. Thus, we believe that information on user mobility might have significant power in predicting the default rate on loans issued. As evidenced in our empirical context, together with loan details, using phone data can predict loan default rates with an accuracy of over 70%. However, and as we mentioned earlier, this effect is confounded by the presence of an underlying social network connecting the different users i.e. locational clustering in data might be driven by individual preferences as well as by peer influence. We correct for this by offering a network-cohesion based regression model that generates consistent estimates after incorporating a measure of the underlying social network. Specifically, we find that on incorporating user network information from their phone-call, as well as the SMS (text message) network data, we obtained obtain around 4% increase in predictive accuracy over previous models, while also gaining improvements in explanatory power. Through these set of analyses, we show that accurate credit scoring and creditworthiness assessment can be achieved by leveraging locational data in a real-world context, even in the presence of a confounding social network.

The current study seeks to make a number of important contributions. To our knowledge, this is among the first studies to empirically assess creditworthiness based on dynamic locational data from a real world context. The findings from this study can potentially illuminate theoretical associations between the mobility behavior of a user and her creditworthiness. Further, we seek to make a methodical contribution by illustrating how locational analytics might be conducted in the presence of network-based confounds. Taken together, our research can benefit both traditional as well as microfinance based institutions by offering an understanding of how locational and network data can help predict creditworthiness of users. In resource-poor communities, where the credit access from traditional sources is problematic due to lack of a proper financial history, or collateral, such approaches would play an important role in assisting micro-financing organizations in developing credit scores in a cheap and quick, yet effective manner.

Background

A brief history of credit scoring

The idea of financial credit has existed for at least 4000 years, while organized banking has its roots in the middle ages (Marquez 2008; Lewis 1992). Despite this rich history, the amount of subjectivity involved in the assessment of potential debtors in terms of their repayment capabilities has remained much the same over the centuries. The use of credit ratings dates back to the early 20th century when organizations like Moody's investor services, and Fitch, devised letters-based rating schemes. A rating was considered to be an indicator of an investor's creditworthiness and was derived based on the applicants' current debt situation and the availability of securities. While this idea of credit rating-based creditworthiness gained popularity in the early 20th century, it was perhaps only after Durand's research on discriminating good and bad loans (Durand 1941), that this method was realized as a systematic approach to assist organizations in making credit decisions (Thomas et al. 2005). The widespread diffusion of credit scoring systems occurred in the early 1960s, with the rapid development of computer-based technology (Capon 1982). The domain of credit scoring grew as a collection of decision models that assessed the risk in lending to a particular borrower at the moment of underwriting (Mays 2001; Thomas et al. 2002).

Most of the early methods used to perform credit scoring were based on statistical modeling techniques (Thomas et al. 2002). For instance, studies using regression-based approaches were undertaken to compute credit scores and to compare results across multiple methods (e.g. Srinivasan and Kim 1987; Desai et al. 1997; Hand and Henley 1997; West 2000; Baesens et al. 2003). Among the regression models, binomial logistic regression was found to be a particularly simple, yet efficient, classification method for credit scoring, in terms of its predictive power. In recent years, more advanced credit scoring and decision-making methods such as neural network based models (Jensen 1992), nearest neighbor approaches (Chatterjee and Barcun 1970), support vector machine (SVM) based methods (Huang et al. 2007), genetic programming (Ong et al. 2005) and hybrid approaches (Hsieh 2005; Lee and Chen 2005) were developed.

The key to a reliable credit score is the availability of diagnostic information about the applicant. Historically, credit scores, particularly in the West, have been primarily derived from analyzing credit bureau data on the personal details and past repayment history of the applicant (Rosenberg and Gleit 1994; Thomas et al. 2005). The use of this information helps lower loan default rates, and improves borrowers' credit access (Klein 1992; McIntosh and Wydick 2005; Luoto et al. 2007). Additional data sources such as a borrower's capacity to repay (e.g. borrower's income) and availability of any collateral can additionally be exploited to help better predict loan outcomes (Mays 2001). However, this creates a challenge for financially underprivileged communities and small businesses who generally do not possess such structured financial history, and are hence likely to be underserved in this credit market. Thus, it has now become imperative to develop novel ways to compute and validate credit scores in such resource-poor communities, in the absence of structured applicant data. For instance, and in a recent study, Wei et al. (2016) presented an analytical framework to incorporate social network based measures into credit scoring. However, there still remains a lack of real-world application of credit-worthiness assessment using non-conventional user data. Our study addresses this gap by proposing a method for assessing creditworthiness using mobile data, in a networked context.

Location and mobility based predictive modeling

With the advent of new forms of information and communication technologies (ICT), like GPS-enabled smartphones, organizations now have access to fine-grained spatio-temporal data about the movement patterns of their customers. For instance, recent studies have shown how the mobility data of users over time has been used to predict and recommend user-level behavioral outcomes like traveling styles (Cho et al. 2011), next venue prediction (Noulas et al. 2012), friendship formation (Scellato et al. 2011), and bandwidth-based QoS prediction (Evensen et al. 2011). Most of the recent work in the area of location-based prediction leverages information on location "types", collected from data sources like mobile check-ins on apps like Foursquare and Gowalla (Noulas et al. 2011; Cho et al. 2011).

A more recent stream of literature looks at the geographical properties of social networks, and finds that the geographical distance between two individuals is often correlated to the probability of the two individuals being friends (Liben-Nowell et al. 2005; Scellato et al. 2010). Drawing on this insight, there have been recent work that looks at predicting or recommending social ties based on locational co-occurrence (Scellato et al. 2011). This idea of correlated geographical and network locations is also consistent with the concept of network autocorrelation in spatial statistics (Leenders 2002; Doreian 1989) or network cohesion in social networks (Gargiulo and Benassi 2000). Network autocorrelation suggests that the behaviors of two individuals who are connected are more likely to be correlated than that of two individuals who are randomly selected. Thus, geographical clustering in mobility patterns might be driven by, and in turn might drive, social connections. It, therefore, becomes an interesting empirical problem to be able to leverage data on geographical positions and mobility, in the presence of an underlying social network among the users.

A Spatio-network based Regression Model

When applying regression models for prediction, a key assumption is that observations are independently sampled from the population. However, in our context, individuals are potentially connected by an underlying social network. Within a network, connected individuals tend to behave similarly, a phenomenon termed as ‘network autocorrelation’ in the spatial statistics literature (Leenders 2002; Doreian 1989), or as ‘network cohesion’ in the social network literature (e.g. Gargiulo and Benassi 2000; Kossinets and Watts 2006; Lewis et al. 2008). Cohesive social ties can carry valuable information about social connections for predicting behavioral outcomes. By ignoring these latent connections among individuals, we would likely under- or over-estimate the model parameters and prediction performance. To resolve this issue in our study, we adapt a new method which takes networked observations into account by introducing a penalty term in the regression estimation (Li et al. 2016). The network cohesion model can be illustrated using the example of a linear regression with n observations and p covariates, as specified below:

$$\boldsymbol{Y} = \boldsymbol{\alpha} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

Where, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$. The network cohesion effect is fully captured using the individual node effects$\boldsymbol{\alpha}$. Next, we regularize $\boldsymbol{\alpha}$ by a network-based cohesion penalty, and the objective function becomes:

$$min \, \|\boldsymbol{Y} - \boldsymbol{\alpha} - \boldsymbol{X\beta}\|^2 + \lambda \sum (\alpha_u - \alpha_v)^2$$

where, u, v are connected observations, and $\lambda > 0$ is the tuning parameter. In other words, the difference for a connected pair is penalized by $(\alpha_u - \alpha_v)$. Note that the term $\sum(\alpha_u - \alpha_v)^2$ can be rewritten using graph-based notations. The network is denoted by a n-by-n binary adjacency matrix $\boldsymbol{A}$. The graph Laplacian can then be defined as: $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, where $\boldsymbol{D}$ is a diagonal matrix and d_v is the node degree. Hence, $\boldsymbol{\alpha}^T\boldsymbol{L\alpha} = \sum(\alpha_u - \alpha_v)^2$, and is termed as *cohesion penalty*. The optimization problem can then be explicitly solved to obtain the following analytical solution for the estimate:

$$(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda\boldsymbol{M})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{Y} \text{ where, } \widetilde{\boldsymbol{X}} = (\boldsymbol{I}_n, \boldsymbol{X}) \text{ and } \boldsymbol{M} = \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0}_{n\times p} \\ \boldsymbol{0}_{p\times n} & \boldsymbol{0}_{p\times p} \end{bmatrix}$$

As illustrated by the closed-form solution, this is a generalized framework that incorporates connections among individual observations, as opposed to a simple linear regression with individual fixed effects, where there are no connections between individuals and they have different intercepts. This method makes fewer assumptions about the underlying network, as compared to other studies attempting to address the similar issues (e.g. Lee 2007; Bramoullé et al., 2009). Furthermore, this method can easily be extended to various types of regressions models outside linear regressions, such as the logistic regression model, which we use in the present study. Assuming the binomial likelihood function is $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{\alpha} \in R^n$ captures the individual effects, the objective to maximize the penalized likelihood can be written as below:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{Y}) - \lambda\boldsymbol{\alpha}^T\boldsymbol{L\alpha}$$

Taking network cohesion into consideration in the logistic model implies that we simultaneously maximize the classic likelihood function which includes individual effects, and minimize the disagreement among the individual effects between two connected observations i.e. minimize the network cohesion penalty. Similar to iterative numerical-method solutions to the normal generalized linear model, such as the Newton-Raphson method, the optimization problem defined above can also be solved using similar numerical methods.

Empirical Context

We empirically validate our proposed method using anonymized backend data on loan issuance, phone usage and social media usage for a sample of users, from a microfinance firm located in a major Southeast Asian country, which provides micro-loans, aimed mainly at the emerging middle class. This Southeast Asian country has one of the largest populations in the world, but a per capita GDP of less than USD 3000, which categorizes it as a lower middle income economy. Since there exists no credit scoring bureaus in this country, it is increasingly difficult for individuals to get credit without providing a substantial amount of collateral. According to a recent report, over 460 million people in this part of the world have no bank accounts (Banning-Lover 2015). However, contrary to the state of banking, the adoption of Internet based technologies has been steadily increasing, with over 40% of the population having access to a smartphone (eMarketer, 2015). Leveraging this growing popularity of smartphones and social media, the micro-finance firm we study in this paper computes a credit score based on the borrowers' phone records and social media data. Thus, instead of offering collaterals and financial history, borrowers are encouraged to share their phone details and SNS (e.g. Facebook) accounts with the lender (i.e. the company itself) to obtain better credit scores. The process of issuing a loan on the app-based platform follows a simple 4-step process as follows:

Step 1: To start a loan application process, the borrower registers herself on the lender's website/app.
Step 2: The borrower connects her loan account on the app to her social network(s), and grants permission to the lender to access her phone and SNS details.
Step 3: The lender generates a credit score based on a proprietary algorithm and this score is provided to the borrower. A higher score guarantees a borrower a better credit outcome.
Step 4: Based on the credit score, the borrower is granted or denied a loan. In the event of a successful loan release, the lender informs the borrower of the latest return date.

From the data provided to us, we found that on average, the lender offered small loans, equivalent of USD 300 or less, for a duration of 3-6 months. The dataset we use consists of two main parts. The first includes data on loan details about a sample of micro-loans issued between 2011 and 2014, including with repayment/default information. The second part corresponds to anonymized Android phone-data for their registered users. This includes details about the number of contacts, details of phone calls and text messages sent, the various apps installed on the phone, and the different URLs visited on the phone browser over time. We also obtained information about mobility of the users in the form of geocoded location data recorded over time. This location data forms the basis for our focal predictors. To generate our focal variables about geographical location types we use the Google Maps API[1] to reverse map the user location geocodes to specific location types. The Google Maps API supports a comprehensive list of location types which classifies the address into a type of property e.g. neighborhood, hospital, airport, café etc. However, since this list of types is prohibitively large, we further categorize the location types into 6 major categories [2],[3] viz. Residential (R), Commercial (CO), Industrial (I), Agricultural (A), Public Service (P) and Entertainment and Recreation (E). The Residential type of locations includes neighborhoods and non-commercial properties privately owned by individual settlers. Commercial locations are addresses where financial transactions take place (e.g. supermarkets, bookstores etc.). Industrial locations include large factories and manufacturing plants (e.g. oil refineries, steel plants). Agricultural locations include farmlands and properties used for faming-related activities (e.g. livestock). Public Service locations are common utilities that are used for public consumption e.g. bus terminals, hospitals etc. Finally, Entertainment and Recreation are locations which the local communities visit at their leisure to relax and enjoy (e.g. movie theatres, amusement parks etc.). We also incorporate information about social ties among the borrowers in our sample by exploiting the call and text data from the borrowers' phone logs. Specifically, we construct an adjacency matrix A for the communication network, such that the value of $A(i,j) = 1$ if user i has contacted user j either via a phone call or using a text message. Conversely, $A(i,j) = 0$ if the user i has never contacted j. We incorporate this network information in the penalized regression method as described in the previous section. In the following section, we provide a detailed data description of our key variables.

---

[1] https://developers.google.com/maps/
[2] https://www.tax.ny.gov/research/property/assess/manuals/prclas.htm
[3] https://en.wikipedia.org/wiki/List_of_building_types

Data Description

As a pilot study to validate our proposed approach, we select a random sample of 140 borrowers with loans issued in 2014, and extract their locational and phone data from our dataset. We focused on the mobility information at a daily basis and obtained locational snapshots for these users over a 1 year period, from June 2014 till June 2015. Once the coding of location data was complete, we found that no address corresponded to the address types I and A. Thus, for our current research, we limit our location types to one of 4 major categories viz. Residential (R), Commercial (CO), Public Service (P) and Entertainment and Recreation (E). The summary statistics of the key variables are described in Table 1 below.

Table 1. Variable Descriptions and Summary Statistics

Variable	Description	Mean	St. Dev.	Min	Max
Defaulted	Binary variable to classify user as a loan defaulter or not	0.243	0.43	0	1
Loan Amount (Amount)	Total amount borrowed by user (in USD)	265.14	168.57	63.66	912.46
Loan Interest Rate (Interest)	Interest rate assigned by lender	6.445	1.742	4.93	8.75
Commercial (CO)	Proportion of yearly visits to commercial locations e.g. supermarkets	0.749	0.114	0.281	0.977
Residential (R)	Proportion of yearly visits to residential locations e.g. neighborhoods	0.034	0.035	0	0.189
Entertainment and Recreation (E)	Proportion of yearly visits to entertainment or recreational locations e.g. parks and museums	0.023	0.028	0	0.143
Public Service (P)	Proportion of yearly visits to public service locations e.g. schools and hospitals	0.194	0.115	0.023	0.667
Call Count (Call)	Average number of calls made to phone contacts per month	60.027	62.592	0	300.917
Message Count (SMS)	Average number of messages sent to phone contacts per month	377.584	469.778	0	3,456.75
Contact Count (Contact)	Total number of contacted connections on the phone	81.75	169.871	0	1,771
Calendar Events (Event)	Average number of events scheduled per month on phone calendar	11.357	30.432	0	198.083
App Count (App)	Total number of installed phone apps	195.643	142.877	0	748
URL Count (URL)	Average number of sites visited on the phone browser per month	55.242	137.137	0	1,364.92

Preliminary Results

We highlight some preliminary results in Tables 2 and 3 in the following page, where we implement both a logistic regression model (LR) as well as a logistic regression model that incorporates network cohesion among users (LR-NC), as described earlier. Our initial results show that LR-NC model outperforms the LR model in predictive accuracy, as reflected by a higher AUC and a lower mean absolute error (MAE) score. However, the increase in predictive accuracy is not as substantial, as perhaps the difference in the parameter estimates[4]. Interestingly, we see that in the LR model, none of the location types are statistically significant, while in the LR-NC model, the frequency of visit to Commercial(C) and Public Service (P) type locations are found to be statistically significant predictors of default rate, after controlling for all phone-usage related and loan related control variables.

Table 2. Prediction results (10-fold cross-validation)

	LR (without location)	LR (with location)	LR+NC (without location)	LR+NC (with location)
MAE	0.3219401	0.3240794	0.317933	**0.2986736**
RMSE	0.4134174	0.4188696	0.4174533	**0.4100722**
AUC	0.7355303	0.7363258	0.7376486	**0.7626894**

Table 3. Estimation results

	LR: no location	LR: with location	LR+NC without location	LR+NC with location
CO		-4.199		-4.098***
		(6.328)		(1.492)
P		-4.266		-4.271*
		(6.291)		(2.420)
E		4.433		6.780
		(10.702)		(9.458)
Amount	0.0001***	0.0001***	0.0001***	0.0001***
	(0.00003)	(0.00003)	(0.00004)	(0.00004)
Interest	0.109	0.150	-0.289 ***	0.172
	(0.135)	(0.142)	(0.008)	(0.175)
Contact	-0.003	-0.004	-0.008	-0.005
	(0.004)	(0.004)	(0.005)	(0.005)
App	-0.005**	-0.005**	-0.007***	-0.005**
	(0.002)	(0.002)	(0.002)	(0.002)
Event	-0.017	-0.017	-0.021	-0.019
	(0.012)	(0.012)	(0.014)	(0.015)
URL	0.006**	0.006**	0.009 ***	0.008 **
	(0.003)	(0.003)	(0.003)	(0.003)
Call	0.0001	0.0006	0.003	0.002
	(0.004)	(0.004)	(0.005)	(0.005)
SMS	-0.0001	-0.00008	-0.00003	-0.0001
	(0.001)	(0.0007)	(0.0009)	(0.001)

Note: *p<0.1; **p<0.05; ***p<0.01*

[4] *Note that in a networked context such as LR-NC, the process of computing penalized standard error remains an open problem and as such the comparison of prediction performance is principled. The current analysis uses un-regularized standard errors, which might potentially introduce some bias.*

Discussion and Roadmap

In the current research, we address the issue of generating accurate credit scores for individuals living in developing economies which generally have poor infrastructure for assessing creditworthiness of applicants using conventional approaches. Through this work-in-progress paper, we highlight our efforts at leveraging user mobility data obtained from the loan borrowers' smartphones to predict whether the borrower will default on her loan or not. However, since the correlation in geographical locations among users might be driven by an underlying social network, we offer a logistic regression based approach that incorporates these underlying social ties to generate consistent estimates and superior predictions. We leverage data on phone calls and text messages that were exchanged among users in our sample, to construct a communication network which we use as a proxy for their social network. However, since data on communication networks is neither required nor utilized by the company to make the decision on loan application, strategic gaming of social interaction by borrowers is unlikely to arise in our current context (i.e. users would not try to game the loan issuance process by forging fake connections). However, our results highlight that the addition of social information does indeed increase the predictive power of model, while also offering improved explanation over a baseline model that ignores the presence of any such network.

As next steps in this research, we hope to make three significant advances over what has been described in this paper. First, we plan to investigate whether or not users register any persistent change in most-visited locations *after* drawing a loan. This would help us understand the borrower's intentions behind issuing the loan, over and beyond what is self-reported by her. Second, we wish to expand our sample size to incorporate a much larger user base. This would help us replicate the findings from this pilot study on a more representative sample, while also allowing us to study heterogeneity among the applicants more effectively. Third and most importantly, we seek to understand the theoretical rationale behind the association between the different location types and the loan-default behavior. For example, our results show that individuals who frequent commercial institutions the most, have a lower tendency to default on loans as compared to individuals who mostly frequent residential properties. One reason for this could be the fact that such individuals might be richer than the average applicant. Alternately, one could also hypothesize that individuals who frequent commercial houses most often might be employed individuals while those who spent most of their time in residential areas have a higher probability of being unemployed. Thus, we need to do further analyses to truly understand the mechanisms that are driving these results, and appropriately offer recommendations to the microfinance institutions.

Conclusion

This research-in-progress paper describes our preliminary attempts at predicting loan default rates using mobile phone data about borrowers' mobility patterns and by leveraging a network-cohesion based logistic regression model. We show that by using a combination of technology-enabled data and computational approaches, microfinance companies in developing and underdeveloped countries can develop effective credit scores for their applicants. This would facilitate the issuance of micro-loans without the need for richer financial history or higher collaterals. While such methods are obviously not a substitute for traditional banking infrastructures and credit bureaus, they offer a timely and effective solution to firms operating in resource-poor environments.

References

Allafrica.com. 2016. "Africa: Vodafone Claims 25 Million M-Pesa Customers.,"AllAfrica (available at http://allafrica.com/stories/201605030063.html).

Assefa, E., Hermes, N., and Meesters, A. 2013. "Competition and the performance of microfinance institutions," *Applied Financial Economics* (23:9), Taylor & Francis, pp. 767–782.

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

Banning-Lover, R. 2015. "Robbed of opportunity: 460 million east Asians have no bank account," *The Guardian* (available at http://www.theguardian.com/global-development-professionals-network/2015/may/12/robbed-of-opportunity-460-million-people-across-east-asia-have-no-bank-account-financial-inclusion).

Bramoullé, Y., Djebbari, H., and Fortin, B. 2009. "Identification of peer effects through social networks," *Journal of Econometrics* (150:1), Elsevier, pp. 41–55.

Brewer, K. 2016. "How to help residents trapped by financial exclusion – livechat," *http://www.theguardian.com/* (available at http://www.theguardian.com/public-leaders-network/2016/apr/18/financial-exclusion-livechat-councils-charities-housing).

Capon, N. 1982. Credit scoring systems: A critical analysis. *The Journal of Marketing* (46:2) pp. 82-91.

Chatterjee, S., & Barcun, S. 1970. A nonparametric approach to credit screening. *Journal of the American Statistical Association* (65:329), pp.150-154.

Chen, G., Rasmussen, S., and Reille, X. 2010. "Growth and vulnerabilities in microfinance," *Focus Note* (61), pp. 1–21.

Cho, E., Myers, S. A., and Leskovec, J. 2011. "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082–1090.

Desai, V. S., Conway, D. G., Crook, J. N., and Overstreet, G. A. 1997. Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics 8(4)*, pp. 323-346.

Doreian, P. 1989. "Network Autocorrelation Models: Problems and Prospects," *Spatial Statistics: Past, Present, Future*, pp. 369–389.

Durand, D. 1941. Risk elements in consumer instalment financing. *NBER Books*.

Dwoskin, E. 2015. "Lending Startups Look at Borrowers' Phone Usage to Assess Creditworthiness," *The Wall Street Journal* (available at http://www.wsj.com/articles/lending-startups-look-at-borrowers-phone-usage-to-assess-creditworthiness-1448933308#:jEaK6s_N/xGMGA).

EMarketer. 2015. "Asia-Pacific Boasts More Than 1 Billion Smartphone Users," *www.emarketer.com* (available at http://www.emarketer.com/Article/Asia-Pacific-Boasts-More-Than-1-Billion-Smartphone-Users/1012984).

Evensen, K., Petlund, A., Riiser, H., Vigmostad, P., Kaspar, D., Griwodz, C., and Halvorsen, P. 2011. "Mobile video streaming using location-based network prediction and transparent handover," in *Proceedings of the 21st International Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 21–26.

"Financial Exclusion.," 2015. *The Economist* (available at http://www.economist.com/news/economic-and-financial-indicators/21648642-financial-exclusion).

Gargiulo, M., and Benassi, M. 2000. "Trapped in your own net? Network cohesion, structural holes, and the adaptation of social capital," *Organization Science* (11:2), INFORMS, pp. 183–196.

Green, H. H. 2014. "The Smartphone Does It Again: Building Credit Scores For Microentrepreneurs," *Forbes* (available at http://www.forbes.com/sites/ashoka/2014/06/10/the-smartphone-does-it-again-building-credit-scores-for-microentrepreneurs/#7a02507d4737).

Groenfeldt, T. 2015. "Lenddo Creates Credit Scores Using Social Media," *Forbes* (available at http://www.forbes.com/sites/tomgroenfeldt/2015/01/29/lenddo-creates-credit-scores-using-social-media/#14d86eda3f79).

Hand, D. J., & Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*. Series A (Statistics in Society), pp. 523-541.

Hsieh, N. C. 2005. Hybrid mining approach in the design of credit scoring models. Expert Systems with Applications (28:4), pp. 655-665.

Huang, C. L., Chen, M. C., & Wang, C. J. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* (33:4), pp. 847-856.

Intelligence, G. 2015. "The Mobile Economy," (available at http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf).

Jensen, H. L. 1992. Using neural networks for credit scoring. *Managerial Finance (18:6)*, pp. 15-26.

Klein, D. B. 1992. Promise keeping in the great society: A model of credit information sharing. *Economics & Politics* (4:2), pp. 117-136.

Kossinets, G., and Watts, D. J. 2006. "Empirical analysis of an evolving social network," *Science* (311:5757), American Association for the Advancement of Science, pp. 88–90.

Lee, L. 2007. "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects," *Journal of Econometrics* (140:2), Elsevier, pp. 333–374.

Lee, T. S., & Chen, I. F. 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications* (28:4), pp. 743-752.

Leenders, R. T. A. J. 2002. "Modeling social influence through network autocorrelation: constructing the weight matrix," *Social Networks* (24:1), Elsevier, pp. 21–47.

Lewis, E. M. 1992. An introduction to credit scoring. *Fair, Isaac and Company*.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. 2008. "Tastes, ties, and time: A new social network dataset using Facebook. com," *Social Networks* (30:4), Elsevier, pp. 330–342.

Li, T., Levina, E., and Zhu, J. 2016. "Regression with network cohesion," *arXiv preprint arXiv:1602.01192*.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. 2005. "Geographic routing in social networks," *Proceedings of the National Academy of Sciences of the United States of America* (102:33), National Acad Sciences, pp. 11623–11628.

Luoto, J., McIntosh, C., & Wydick, B. 2007. Credit information systems in less developed countries: A test with microfinance in Guatemala. *Economic Development and Cultural Change* (55:2), pp. 313-334.

Marquez, J. 2008. "An Introduction to Credit Scoring For Small and Medium Size Enterprises.," World Bank

Mays, E. 2001. Handbook of credit scoring. *Glenlake Publishing Company Ltd.*

McIntosh, C., & Wydick, B. (2005). Competition and microfinance. *Journal of Development Economics* (78:2), pp. 271-298.

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. 2012. "Mining user mobility features for next place prediction in location-based services," *2012 IEEE 12th International Conference on Data mining (ICDM)*, pp. 1038–1043.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. 2011. "An Empirical Study of Geographic User Activity Patterns in Foursquare.," *ICwSM* (11), pp. 70–573.

Ong, C.-S., Huang, J.-J., and Tzeng, G.-H. 2005. "Building credit scoring models using genetic programming," *Expert Systems with Applications* (29:1), Elsevier, pp. 41–47.

Rosenberg, E., & Gleit, A. 1994. Quantitative methods in credit management: a survey. *Operations Research* (42:4), pp. 589-613.

Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. 2010. "Distance Matters: Geo-social Metrics for Online Social Networks.," in *WOSN*.

Scellato, S., Noulas, A., and Mascolo, C. 2011. "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1046–1054.

Srinivasan, V., & Kim, Y. H. 1987. Credit granting: A comparative analysis of classification procedures. *Journal of Finance* (42:3), pp.665-681.

Thomas, L. C., Edelman, D. B., & Crook, J. N. 2002. Credit scoring and its applications. *Siam*.

Thomas, L. C., Oliver, R. W., & Hand, D. J. 2005. A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society*, (56:9), pp.1006-1015.

Wei, Y., Yildirim, P., den Bulte, C., and Dellarocas, C. 2015. “Credit scoring with social network data,” *Marketing Science* (35:2), INFORMS, pp. 234–258.

West, D. 2000. "Neural network credit scoring models." *Computers & Operations Research,* (27:11), pp. 1131-1152.

Yunus, M. 2007. Grameen Bank at a glance. *Grameen Bank*.