Feature Selection of Credit Score Factor Based on Smartphone Usage using MCFS

Lathifah Alfat, Mia Rizkinia, Riri Fitri Sari
Department of Electrical Engineering
University of Indonesia
Depok, Indonesia
lathifah.alfat@ui.ac.id, mia@ui.ac.id, riri@ui.ac.id

Daniela M. Romano
Department of Information Studies
University College London
London, United Kingdom
D.Romano@ucl.ac.uk

Abstract—**Credit as a part of our consumptive life has helped a lot of people. As a financial product, it is used widely along with the growth of economic and financial services. Therefore, credit is very risky so that motivating the financial institution to use a system called credit scoring to make a decision about acceptance. However, the conventional credit scores is calculated from the user's financial history in financial institution. This method causes people without any financial history or account, such as students, being unnoticed by the system, then their credit proposal is declined. This phenomenon insists the researchers thinking about a new credit scoring system that facilitates people from different economic background. Knowing that people nowadays will spend any money and time on their smartphone, we make a hypothesis about how smartphone usage behavior can be the answer. Then, the survey is conducted to 90 respondents from low to high economic background to model their credit limit. This paper shows that smartphone usage has some insights that can be computed through Multi-Cluster Feature Selection (MCFS). The selected features are Brand of phone, Frequency of changing the phone, Occupation, Data usage for game, Cost for phone, Data usage for social media, Reason of changing the phone, Money for data, Protect personal interest, Age, Spend last money, and Pay for you.**

Keywords—credit, score, phone, financial, MCFS

I. Introduction

In the world of capitalism, credit has been a part of life. Houses, vehicles, smartphones, even fancy clothes and bags can be afforded by credit. It is still being utilized widely as long as people become more consumptive toward their needs of comfort. The growth of middle economic class in developing countries also contributes to this. Although the spreading is far, there are still a group of people who cannot access the credit because of their lack in financial history. The conventional banks use financial history to decide the acceptance of customers' credit. Hence, researchers aim to looking a better solution to facilitate people needs of credit.

Financial institutions conduct credit scores as a method to maintain the risk to customers. Credit scores represent customers' ability and inability to pay their debt before the due date. From the information form, some variable are chosen to make financial user models. Financial history is an important factor to compute and generate the credit scores. Therefore, the financial institutions will likely to accept a credit proposal from a customer who has a long financial history that is accessible [1]. This method has a drawback that people who are reliable to pay their debt but do not have any financial history will be declined when they propose a credit.

Researchers have been motivated to find another model to conduct a new credit scoring that facilitates people from lower economic background and people who responsible but do not ever have any financial account. As the source of information, demographic aspects and socioeconomic status is studied thoroughly. The result shows that both factors are related to the reliability of financial behavior and lead to some act, such as impulsiveness, education level or marital status [2, 3]. The era of smartphone opens a new paradigm of people's consumption. Smartphone is used almost in all aspect. People call, text, chat, play the game, browse the information from their smartphone. There is a hypothesis that mobile phone data usage may correlate to people's spending and financial risk [1]. This is also supported by a study about mobile phone usage data to model the users, such as inferring personality traits or socioeconomic status [4, 5, 6]. The study shows that the financial behavior has relationship to the mobile phone usage data as well [7].

After gathering a survey data about the smartphone behavior, there is a need to decide the method to classify which feature is the most important factor to decide the credit limit. Hence, a feature selection method named Multi-Cluster Feature Selection (MCFS) is used. Feature selection is different from feature extraction. Feature selection choose a subset of existing features without doing any transformation. While, feature extraction needs to transform the existing features into lower dimensional space [8]. The MCFS method select features with consideration of correlation that be produced among different features. Hence, it can optimize the feature subset, also the multi-cluster structure inside the data can be preserved well [9].

The contributions of this paper are composing a new method of credit limit and finding which variable is impactful. The questionnaire survey is filled by 90 respondents from different background, such as students, civil servants, housewives, and employees. This study benchmarks whether different background of person may affect their decision to use their money that correlates to their credit trustworthiness. This paper is conducted in six part. The second part is about literature review related to credit score, mobile phone usage, and MCFS. The third part explains related works that conduct the research about phone usage as trustworthiness parameter in the past. The fourth part tells about the survey and implementation of MCFS. The fifth part contains the result and analysis and the last is the conclusion of the research.

II. Literature Review

A. Credit Score

Before credit scoring exist, banking use underwriter experience to evaluate customers' application of credit. Customers' information is gained from relationship between customer and credit institution staff. This makes a movement border between customer and credit giver [10]. Credit, as evaluation process, where bank manager takes part based on 5C criteria:

- Character – is customer or their family member is acknowledged by the organisation?
- Capital – how much from the deposit is customers' demand and how much the amount of credit they ask?
- Collateral – what security is demanded by the customers?
- Capacity – what about the customers' ability to repay?
- Condition – how about the current economic condition?

That kind of process has some drawbacks, especially related to the consistency and reliability to include them in word that represent credit decision. Hand [11] has noted the key drawbacks, such as (i) the credit decisions are influenced by daily mood of bank manager; (ii) the credit decisions cannot be repeated since different manager may give different decisions; (iii) there is still no decision formalization to make decision resulting in teaching difficulty; (iv) evaluation based on people's perception may be applied on some certain condition that can cause loss.

The imperfection of credit evaluation can predict the risk solely based on the historical notes, not the future condition or potential. This causes system failure to identify customer who leave their responsibility to pay credit in the future [12]. This condition may correlate to some unpredictable reasons, such as: (i) fraud; (ii) divorce; (iii) lack of financial intelligence; and (iv) income loss caused by credit. Hence, there is a high demand of new credit system evaluation that can differentiate customers who is able and unable to repay in the future based on the behaviour [12]. Both researchers and practitioners agree that even a small change in credit risk scoring will give a significant impact to society in term of savings [13].

B. Mobile Phone Usage

Having the fourth biggest population in the world, Indonesia is a huge market in mobile technology business. The high economic growth potential can be seen in in which approximately 87 percent of Indonesian has a mobile phone. Then, in 2017, the number is growing from 32.6 to 43.2. The prediction says that in 2021, Indonesian smartphone users could be 96.2 million people [14].

Indonesia offer a huge market for internet market in mobile platform. The growing of smartphone users also triggers the using of internet in mobile. One significant feature is the usage of social media. It is proven by 48% of Indonesian who has Facebook account and 38% has Twitter account from data in the end quarter of 2016. Other fact revealed in 2012. When Jakarta was mentioned as the first Twitter city, people make Tweets from there more than anywhere in the world. However, social media as a product of mobile internet in Indonesia play significant factor. Not only as branding, but also to form public opinions about something.

The idea to correlate mobile phone usage data to people behaviour toward financial has been a topic among researchers. Personality traits or socioeconomic status may have crucial role [4, 5, 6]. In 2015, MobiScore was developed as new way to predict people credit scoring from mobile data usage [1]. While, in 2018, Daniel Björkegren and Darrell Grissen was working on the research to reveal people behaviour to predict loan repayment from their mobile data usage.

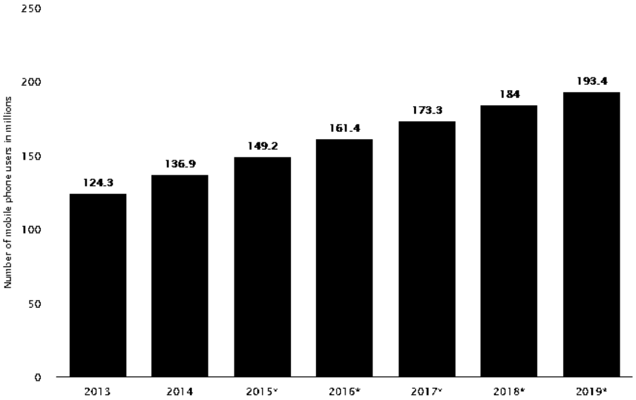


Fig. 1. Number of mobile phone users in Indonesia from 2013 to 2019 (in millions)

C. Multi-Cluster Feature Selection (MCFS)

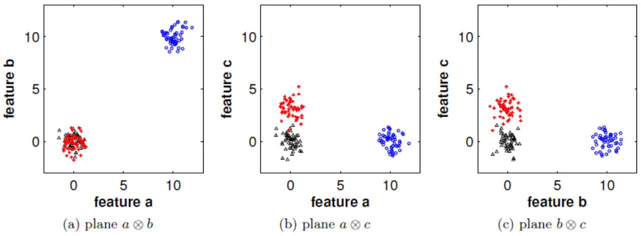
With the rapid growing of data, the dimensionality becoming highly increases. It needs more time and space for fulfilling the basic requirements so that the data processing can run well. Furthermore, other specific process, such as classification and clustering, may grow huge in spaces resulting in dimensions of hundred or thousand. Although, the analytical and computational stages can be managed best in low dimensional spaces [15]. Feature selection methods are offered to solve this problem. It is a work by reducing the dimension of data by defining which feature subset that relevant. After getting the relevant features in small quantity, analysis of data can be done.

The purpose of feature selection is choose the feature subset that are most relevant according to several characteristics. It is a solution to solve the computation complexity that can be expensive. Usually, feature selection methods tend to choose the highest ranking features according to independent scores to solve this problem. With the help of the scores, a feature can state their differentiate power in a classes or clusters. Although this strategy is good on binary classes or clusters, but it is considered as fail in multi classes or clusters. Figure 2 describe how three Gaussians in a three dimensional space work. Some prominent unsupervised feature selection methods will arrange the highest features to a > b > c without adding the label information. When two features are asked, methods will generate the features as a and b. This is actually called as sub-optimal. In the case of multi classes or multi clusters data, different features may not have same powers on different classes or clusters diversification (e.g., cluster 1 vs. cluster 2 and cluster 1 vs. cluster 3). On the other side, some research about supervised feature selection [17] have idea to manage this problem. The application of unsupervised feature selection methods seems unclear if the similar ideas is given without label information.

Usually, data that occurs naturally have structure that consists of multiple clusters. Hence, to make a good feature selection, algorithm should have these aspects:

- The selected features able to maintain the data's cluster structure optimally. The research about unsupervised feature selection [18, 19, 20] commonly apply Gaussian shape on the clusters. On the other side, recent research have finding that data from human may be fragmented from a sub complex that related to Euclidean space [21, 22, 23]. In the process of calculating the integrity of the clusters, there should be a consideration about intrinsic manifold structure [24].

The selected features able to represent all the possible clusters inside the data. There are different power on differentiate the cluster among different features. Hence, there are undesired condition that selected feature will differentiate cluster 1 and cluster 2 rightly, but differentiate cluster 1 and cluster 3 falsely.



(a) plane $a \otimes b$ (b) plane $a \otimes c$ (c) plane $b \otimes c$

Fig. 2. A failed example for binary clusters or classes feature selection methods.

Algorithm 1 Algorithm for MCFS

```
Construct p-nearest neighbour graph W
Solve generalized Eigen problem to get K eigenvectors corresponding to the smallest eigenvalues
Solve K L1-regulairzed regression to get K sparse coefficient vectors
Compute the MCFS score for each feature
Select d features according to MCFS score
```

III. Related Work

There are some research about phone usage to determine people' credit limit. The research is listed and explained below.

A. *MobiScore: Towards Universal Credit Scoring from Mobile Phone Data*

Pedro et al. make the paper [1] titled "MobiScore: Towards Universal Credit Scoring from Mobile Phone Data". The concept of credit and its risk are explained in the paper. The problem exists when the credit score uses past financial history resulting people who do not have any is neglected. Then, the authors offer MobiScore to replace the traditional credit score. MobiScore uses mobile phone usage data to model customer' profile. It also provides a credit access to people without financial history and additional information for traditional credit score system. The research confirms the data from a telecommunications operator and a financial institution in a country in Latin America. The result proves that the model is accurate compared by the traditional credit system.

B. *Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment*

Daniel Björkegren and Darrell Grissen write the paper [16] titled "Behaviour Revealed in Mobile Phone Usage Predicts Loan Repayment". The paper sees that the problem in developing countries is many households lack formal financial histories. This causes difficulty for financial institutions to approve and extend credit. Looking from another side, the households have mobile phones that save meaningful data about behaviours. Hence, this research predicts credit amount from mobile phone usage behaviours. Call records are the information resource, while the research was applied to a South American country. In conclusions, the research shows that people in the highest quintile are 2.8 times more default than in the lowest quintile. The method profiles the customers, with almost no financial history, using credit bureau information on a certain time period. The research successfully forms a basic method to reach people from unbanked sector.

IV. Survey and Method Implementation

To collect the data, a questionnaire that depicts the respondents' profiles and their act toward smartphone has been arranged. The respondents participate to give their data and opinion. Then, the data are resumed and compared. The process can be explained in detail as follows.

A. *Questionnaire*

The questionnaire contains 31 questions that classified into three different categories: background (5 questions, e.g. age, gender), smartphone usage parameter (17 questions, e.g. frequency of changing smartphone, reason of changing smartphone), trustworthiness (9 questions, e.g. whether many of their bills are past due). A scale of 1 to 5 is used on some quantitative question in mobile phone usage parameter categories to rate amount of data they invest on the kind of smartphone services (where 1 = a few, and 5 = many) and trustworthiness categories to rate how respondent react to the sentence depicting their condition (where 1 = strongly disagree and 5 = strongly agree). Other questions are answered in multiple choices.

B. *Respondents*

The 90 respondents are from different background of job, i.e. private employee, civil servant and state owned cooperation employee, while the rest are teacher, lecturer, also student. The rest background job of the respondent are divided into housewives, entrepreneur, and other jobs. The respondent' average of age was 31.5 years old, with the youngest was 17 years old, and the oldest was 65 years old. The respondents was given the link of online questionnaire. Then, they answer the questionnaire on voluntary basis based on their usage of smartphone.

C. *MCFS Implementation*

After collecting data from the survey, the pre-processing begins with data cleansing. Some data with almost the same input will be corrected based on their similarity. Then, data normalization take place in the next process. Data will be categorized in numbers based on its different classification. Hence, the data is ready to be implemented by MCFS.

The implementation of MCFS begins with normalizing the original data based on the minimum and the maximum value of the data. Then, removing the mean variable-wise (rowwise). The next stage is calculating the value of y. Because the process is unsupervised feature selection, the parameter k should be tuned. k default value itself is 5. The other important value is the Eigen functions that will be used. After that, the 1st and 2nd input that derived from data matrix and the number of features, respectively. Then, the algorithm will generate the result in the form of ranking of the features selected. The complete process can be seen in Figure 4 below.

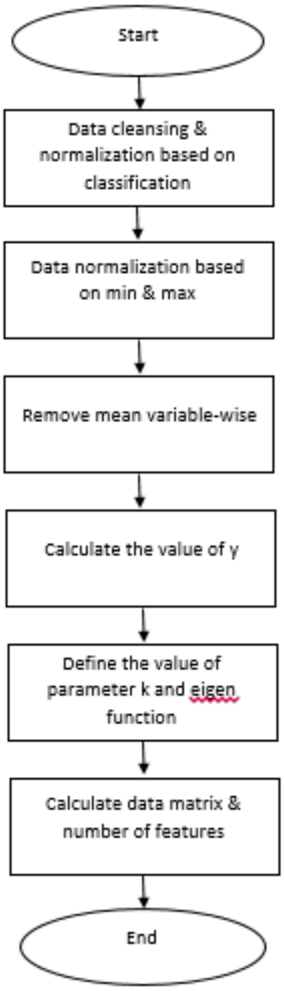

Start
Data cleansing & normalization based on classification
Data normalization based on min & max
Remove mean variable-wise
Calculate the value of y
Define the value of parameter k and eigen function
Calculate data matrix & number of features
End

Fig. 3. Flowchart of MCFS

This implementation is applied on 31 variables which are qualitative variables and quantitative variables. The division of the variables is depicted in Table I and Table II below:

- Qualitative Variables:

TABLE I. QUALITATIVE VARIABLES OF SURVEY DATA

Parameter	Variables
Background	Gender
	Occupation
	Education
Smartphone Usage	Operator
	Kind of Card
	Pay for you
	Pay for others
	Data usage for call
	Data usage for internet
	Data usage for email
	Data usage for game
	Data usage for social media
	Data usage for other app
	Brand of phone
	Frequency of changing the phone
	Reason of changing the phone
Trustworthiness	Protect personal interest
	Something to be proud
	Solve problems
	Spend last money
	Money for future
	Spend to feel better
	Tense of bills
	Finance cause panic
	Bills are past due

- Quantitative Variables:

TABLE II. QUANTITATIVE VARIABLES OF SURVEY DATA

Parameter	Variables
Background	Age
	Income
Smartphone Usage	Cost for phone
	Minute to call
	Money for data
	Amount of pay for others

V. RESULT AND ANALYSIS

After processing the data with MCFS, the ranking is produced from the highest to the lowest. The most impactful variables are chosen from the top 12 of the rank, with the assumption that 12 are considered as major of 31 variables. The result variables vary among the three parameters with the domination of smartphone usage parameter. There are 2 variables from background parameter, 8 variables from smartphone usage parameter, and 2 variables from trustworthiness parameter. The selected features are Brand of phone, Frequency of changing the phone, Occupation, Data usage for game, Cost for phone, Data usage for social media, Reason of changing the phone, Money for data, Protect personal interest, Age, Spend last money, and Pay for you. Looking from this result, it can be said that smartphone usage are proven impactful in people decision nowadays. The detailed findings can be summarized in the Table III below.

VI. LIMITATIONS

The respondents of the questionnaire are people around the writer. This result might not represent whole Indonesian population and their opinion. The survey was conducted in June 2019. The answer of the questions were based on the technology and situation at that time. The answer might change depends on the communication and technological evolution.

VII. Conclusions

The conclusions of this work can be summarized as mentioned. Calculation of feature selection of credit score factor based on smartphone usage has been done using MCFS method. The survey is carried with 90 respondents answering 31 questions about background, smartphone usage, and trustworthiness. With the MCFS calculation, our research concludes that 12 factors that impactful, which is dominated by smartphone usage parameter. There are Brand of phone, Frequency of changing the phone, Occupation, Data usage for game, Cost for phone, Data usage for social media, Reason of changing the phone, Money for data, Protect personal interest, Age, Spend last money, and Pay for you.

TABLE III. Summary of MCFS Ranking Result

Ranking	Variable Number	Variable Name	Parameter
1	20	Brand of phone	Smartphone Usage
2	21	Frequency of changing the phone	Smartphone Usage
3	3	Occupation	Background
4	17	Data usage for game	Smartphone Usage
5	8	Cost for phone	Smartphone Usage
6	18	Data usage for social media	Smartphone Usage
7	22	Reason of changing the phone	Smartphone Usage
8	10	Money for data	Smartphone Usage
9	23	Protect personal interest	Trustworthiness
10	1	Age	Background
11	26	Spend last money	Trustworthiness
12	11	Pay for you	Smartphone Usage

Acknowledgment

We thank University of Indonesia for financial support for this research under Smart City Grant and Resolv Program.

References

[1] J. S. Pedro, D. Proserpio, and N. Oliver, “MobiScore: Towards Universal Credit Scoring from Mobile Phone Data,” ResearchGate. DOI: 10.1007/978-3-319-20267-9_16, June 2015.

[2] J. E. Grable and S.-H. Joo. Environmental and biophysical factors associated with financial risk tolerance. Journal of Financial Counseling and Planning, 15(1), 2004.

[3] J. M. Henegar, K. Archuleta, J. Grable, S. Britt, N. Anderson, and A. Dale. Credit card behavior as a function of impulsivity and mothers socialization factors. Journal of Financial Counseling and Planning, 24(2):37-49, 2013.

[4] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In Social Computing, Behavioral-Cultural Modeling and Prediction, pages 48-55. Springer, 2013.

[5] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In Proceedings of International Conference on User Modeling, Adaptation and Personalization, UMAP'11. Springer, 2011.

[6] R. de Oliveira, A. Karatzoglou, P. Concejero Cerezo, A. Armenta Lopez de Vicuna, and N. Oliver. Towards a psychographic user model from mobile phone usage. In CHI '11 Extended Abstracts on Human Factors in Computing Systems, CHI EA '11, pages 2191{2196, New York, NY, USA, 2011. ACM.

[7] J. Gathergood. Self-control, financial literacy and consumer over-indebtedness. Journal of Economic Psychology, 33(3):590 { 602, 2012.

[8] P. Schrater. Feature Selection/Extraction Dimensionality Reduction. The Vision Research Laboratories. Psychology Department. University of Minnesota.

[9] Deng Cai, Chiyuan Zhang, Xiaofei He. Unsupervised Feature Selection for Multi-Cluster Data. KDD 2010. ACM 978-1-4503-0055-1/10/07. Washington, DC, USA. 2010.

[10] Anderson, R. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press, USA. 2007.

[11] Hand, D. Modelling consumer credit risk. IMA Journal of Management Mathematics. 2001.

[12] Finlay, S. Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research. 2011.

[13] Hand, D. & Henley, W. Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1997.

[14] Statista. Indonesia: Smartphone User Penetration. https://www.statista.com/statistics/257046/smartphone-user-penetration-in-indonesia/. Accessed on Aug 1, 2019.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2000.

[16] D. Björkegren & D. Grissen, “Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment”, Brown University, Department of Economics, February 2018.

[17] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research, 3:1229–1243, 2003.

[18] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. Journal of Machine Learning Research, 5:845–889, 2004.

[19] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9):1154–1166, 2004.

[20] V. Roth and T. Lange. Feature selection in clustering problems. In Advances in Neural Information Processing Systems 16. 2003.

[21] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in Neural Information Processing Systems 14, pages 585–591. 2001.

[22] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326, 2000.

[23] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323, 2000.