International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Machine Learning Model for Loan Program Subscription from Private Lending Institutions

Ronil S. Goloran[1], Gilbert Alerta[2], Fernando T. Herrera[3]

[1]Saint Joseph Institute of Technology – Graduate School
[2]Department of Science and Technology, Philippine Science High School-Caraga Region Campus in Butuan City
[3]Ph.D., Caraga State University, Ampayon, Butuan City

ABSTRACT

The primary interest of all financial and lending institutions is to determine their market. Determining the probability of loan subscription for every qualified individual remains a challenge. This study develops a machine learning model using logistic regression that classifies loan subscribers from non-subscribers. A total of 300 (150 subscribers and 150 non-subscribers) qualified respondents were surveyed. Subscribing to the machine learning algorithm, 80% (240) and 20% (60) of the respondents were used for training and testing sets, respectively. Input features include: (1) number of existing loans, (2) monthly salary, (3) number of credit cards, (4) awareness to loan program of the company, and (5) required time for loan approval by the company. The data showed that the model was able to correctly classify 87.69% of the training set and 87.50% of the testing set. This result provides empirical evidence that the developed model can be used by lending or financing institutions in future market planning and decision-making.

Keywords: Loan program subscriptions, Private Lending Institutions, Machine Learning Model

1.0 Introduction

Lending or financing institutions depend heavily on the number of customers. In recent years, loan programs have been designed to attract and address the changing needs of the customers. Continued subscription of customers to lending programs plays a critical role in the sustainability of lending companies. The ability to predict the likelihood of customers' subscription to lending programs is a very essential step to ensure the growth and stability of companies. Predictions will aid companies in developing proactive decisions and anticipating customer needs in advance (Gu & Wu, 2019).

Loan programs are the lifeblood of lending companies and financing institutions. Increasing the likelihood of customers to avail or subscribe to loan programs is an essential task to fatten its market and increase profit (Mong & Muraya, 2019). Known strategies to improve loan subscription among customers vary from one company to another. In addition, customers also have different views and preferences on loan subscriptions. Hence, finding the best mechanism to attract customers to subscribe in loan programs is a major challenge.

In the recent literature, modeling the likelihood of customers subscribing to a loan program is given less attention. Most recent studies are focused on the factors related to loan eligibility to safeguard companies from the risk of failure to pay (Antwi et al. 2012). This indicates that there is a scarcity of research related to market preferences in the context of services and benefits provided by the lending institutions. In light of the dearth of papers regarding the likelihood of customers to avail or subscribe to a loan program, this study aims to develop a logistic regression model that will be used to predict the likelihood or chance of loan subscription among target customers (Mong & Muraya, 2016).

At the local scale, the number of lending institutions in the Caraga region has increased. According to the report from the Philippine Statistics Authority (PSA), the number of lending and financing companies in the region has increased in 2023. This means that competition in the market is getting rough, and companies should strategize ways to improve their services and understand the likelihood of customers availing of loan programs. Besides the challenges induced by the rising number of loan institutions, the prior interviews with several marketing specialists in the city can speak of the need to predict the chances of loan subscription among the target market.

The use of machine learning models is already acknowledged as the most scientific method in prediction and classification problems. For example, Rajagopal (2020) used a machine learning algorithm to predict the chances of being admitted to a particular university or college. In another study by Sahragard & Chahouki (2015), a logistic regression model was used to classify species of plants given several features or independent variables. Definitely, predictive analytics has the power to revolutionize almost any industry. It's especially useful for financial institutions like banks and lending firms because they help gain access to the subconscious behavior of customers (Choudhary, 2022; Wang, 2020). Anchored on these sample applications, this study will employ a logistic regression model to predict the likelihood of the target customers to avail or subscribe to a loan program.

The findings of this paper can be used by lending institutions for better decision-making. In addition, the depth of analysis and the advanced knowledge that will be produced from the study will open opportunities for management courses to use data science models to understand market demands. In addition, the study will be very useful for companies as an aid in understanding the needs of their clients.

2.0 Methodology

Research Design

This study employed a predictive research design. Correlation analysis is an essential step before the development of prediction models. Under predictive analytics, a set of predictors or input features must be highly correlated to the dependent variable. Correlation analysis ensures better predictive ability of the model. Hence, the mentioned research design is appropriate for the objectives of the study. In the context of the study, the predictive research design was deemed highly appropriate because of the main goal of predicting the loan subscription behavior of the target market.

Research Participants

The target participants of this study were the qualified government employees who are qualified to subscribe to or avail of a particular loan program of a known company. In compliance with data privacy law, the name of the company and its salary loan program were not disclosed in detail. Satisfying the sample size requirement for the machine learning algorithm and specifically employing the logistic regression model, a total of 300 respondents were considered. These 300 respondents were divided equally. One hundred fifty (150) respondents are those who availed the loan program, while there 150 did not. This was designed to optimize the estimation accuracy of the logistic model. Furthermore, eighty (80%) of the whole data set to be collected were used for training the model while the other twenty (20%) were used for testing. This practice of partitioning is highly suggested and used in the literature (Hu, 2023).

Research Locale

This study was conducted in Butuan City. Butuan is the commercial, industrial, and administrative center of the Caraga region. It is a strategic trading hub in Northern Mindanao with major roads connecting it to other main cities on the island such as Davao, Cagayan de Oro, Malaybalay, Surigao, and soon, Tandag. It hosts one of the busiest airports in the country, the Bancasi Airport, serving around 525,000 passengers in 2012. Cebu Pacific and Philippine Airlines serve the airport. Meanwhile, the nearby Nasipit International Port and in-city Masao Port are providing for its shipping and cargo needs.

The total number of businesses registered in 2013 was 9,619, reflecting a growth of 9.86% and almost 3 times that of the next major Caraga city. New businesses registered numbered 2,032 with a combined capitalization of P504,598,667, an expansion of 75.63 from 2012. As further proof of its dynamic economy, Butuan's local income reached P330,510,000 in 2013, besting other major cities in the country. By 2014, its local income is expected to reach P513,870,000.00 or register a growth of 55%; and total income (including IRA) will be P1,515,970,000. The city was ranked 4th and 16th Most Competitive City for the years 2012 and 2014 by the National Competitiveness Council of the Philippines.

Data Gathering

This study complied with the existing protocols and procedures to ensure ethically sound research. The following steps were followed. As a market specialist in the Caraga region, the proponent already has prior knowledge and linkages in reaching the respondents. The researcher determined the target respondents through the following steps.

The researcher sent a formal invitation to any government employees who are qualified for a particular loan program. A qualified respondent is defined as a government employee who holds the plantilla for at least one year. In the form of a letter request, the need to conduct the study and its implied benefits were also discussed.

After all the necessary documents are complied with, the proponent conducts an orientation to the target customers regarding the content and interest of the ongoing research. It was assured that all respondents were informed of the fact that their participation is voluntary in nature. The actual gathering of data was done face-to-face. All gathered data were tabulated in Excel in preparation for data analysis. With the guidance from the statistician, the modelling was implemented using R software.

Machine Learning Model

Binary Logistic Regression was fitted to the collected data to predict the likelihood of loan subscription given several input features. The modelling employed the machine learning algorithm, where 80% of the dataset was used for training the model while the 20% was used for testing the model. The model was implemented in R software.

3.0 Results and Discussions

This chapter presents the descriptive statistics of the input features that were considered for the logistic regression model. Table 1 shows the frequency and percentage distribution of the variables. It highlights the variation of the characteristics of the different respondents with respect to the number of existing loans, number of credit cards, monthly salary, awareness to loan program, and time requirement for loan approval. In terms of the number of existing loans, the majority of the respondents are paying 1-3 loans, as evidenced by the percentage share of 75% (226). When it comes to monthly salary, the majority of the respondents, or 66% (197), are receiving a monthly salary that ranges from Php 29,000.00 to Php 38, 999.00. A percentage share of 22% (65) earns Php 49,000.00 to Php 58,999.00 a month. The same table also posits that 7% (21) and 6% (17) are respectively enjoying a monthly salary of Php 69,000.00 to 78,999.00 and Php 39,000.00 to 48,999.00. According to Antwi et al. (2012) monthly salary is a significant indicator of the success of loan approval, though few studies have focused on its effect on the decision of a person to subscribe to a particular loan program.

Table 1.
Distribution of respondents across input features

Input Features	**Frequency**	**Percentage**
Number of Existing Loans		
None	74	25%
1-3	226	75%
Total	**300**	**100%**
Number of Credit Cards		
None	149	50%
Holder of 1	151	50%
Total	**300**	**100%**
Monthly Salary		
29000-38999	197	66%
39000-48999	17	6%
49000-58999	65	22%
69000-78999	21	7%
Total	**300**	**100%**
Awareness to Loan Program		
Very Low	0	0%
Low	61	20%
Average	87	29%
High	87	29%
Very High	65	22%
Total	**300**	**100%**
Time required for loan approval		
0.5-2.5	150	50%
2.5-4.5	56	19%
4.5-6.5	86	29%
6.5-8.5	8	3%
Total	**300**	**100%**

Pertaining to the respondents' awareness of the loan program, 22% (65), 29% (87), 29% (87), and 20% (61) disposed a very high, high, average, and low levels, respectively. Choudhary (2022) also noted that the presence of credit cards and ownership of valuable properties are predictors of being lone eligibility. Moreover, when the respondents are surveyed about their preferred time required for loan approval, 50% (150) say that they prefer an approval and release of 0.50 to 2.50 hours. This observation reflects the reality about the majority of the customers in the lending institutions. Many customers subscribed to loan programs that have fast approval time (Choudhary, 2022). In a similar table, 29% (86) preferred to have an approval time of 4.50 to 6.50 hours. While it would be very difficult to maintain the quality control of many financing institutions, reducing the time requirements for the approval and loan release increases the subscription rate to loan programs (Hamayel et. al., 2021).

Furthermore, as part of the research design, half of the respondents are known subscribers to a particular loan while the other half are not. The profile of the participants as input features or predictors to loan subscription depicts evident variability, which is potential for correlation against the dependent variable of interest. In the paper of Li et. al. (2022), it was mentioned that common predictors of a successful loan subscriber include number of existing loans, awareness of the program, interest rates, and timeliness of loan approval and release. In another article by Mong & Muraya (2019), it was particularly noted that qualified loan subscribers are those who have fewer existing loans, with valuable assets or properties, and a net monthly salary. Definitely, the differences in the profile variables of the respondents could suggest the appropriateness of the logistic regression model as one of the most useful classification machine learning models.

Following the data distribution in Table 1, a logistic regression model is fitted to the training set, where the variables are tested for statistical significance as shown in Table 2, before the Machine Learning Algorithm was performed.

Table 2.
Logistic regression model using the five significant input features

Input Features	Estimate	Z-value	P-value	Remarks
Existing Loans	-1.707	-3.975	0.001	**Significant**
Monthly Salary	0.813	2.623	0.008	**Significant**
Credit Cards	0.806	2.059	0.040	**Significant**
Awareness to Loan Program	0.819	2.892	0.004	**Significant**
Time required for loan approval	-1.298	-5.677	0.001	**Significant**

With emphasis on the results from Table 2, significant input features or predictors include the number of existing loans, monthly salary, number of credit cards, awareness of the loan programs, and the time required for loan approval. When it comes to variable importance, the negative z-value of -5.677 indicates that timeliness of loan release and approval emerges as the most influential predictor of the likelihood of loan subscription. The negative coefficient of -1.298 indicates that the majority of those who decided to subscribe to a loan package desired a lesser time of approval. This is conformant to the experiences of the proponent as a marketing specialist from a financial institution or company. Also, Hamayel et al. (2021) noted that a large percentage of the target market prefers loan programs with a very fast time of approval. The number of existing loans was exposed as the second contributory predictor to loan subscription, as evidenced by the z-value of -3.975. The negative coefficient of -1.707 indicates that those who have 2-3 loans are more likely to avail another loan package. Further, Monthly salary and number of credit cards are also shown as significant predictors with positive coefficient estimates and z-values. These empirically imply that those respondents with higher income and credit card holders displayed a higher likelihood of subscribing to a loan program or package. This particular result is similarly observed in the study of Li et al. (2022), which asserted that monthly income and credit cards are essential determinants of loan eligibility, along with the person's loan power.

The findings presented in the model only speak of the reality regarding the characteristics of loan subscribers. Significant findings not only provide additional knowledge in the literature but also yield relevant inputs and innovation to the marketing strategy of the financial institutions in the country. While it is known that machine learning algorithms have led to developments in business innovations, the results of this paper also open the opportunity for business institutions in the development of technology or devices that will be used to predict the likelihood of loan subscription among the target market.
Machine Learning Algorithm to fit the logistic regression model with 10-fold validation was then performed, where a confusion matrix was produced for both the training and testing sets. These are shown in Tables 3 and 4 below.

Table 3.
Confusion matrix of the model using the training data

Predicted	**Actual**	
	Subscriber	***Non-subscriber***
Subscriber	108 (45%)	17 (6.92%)
Non-subscriber	13 (5.38%)	102 (42.69%)
Correct Classification Rate	87.69%	
Misclassification Rate	12.31%	

For the training data, Table 3 shows that the logistic regression model correctly classifies both the subscribers and the non-subscribers, as evidenced by the correct classification rate of 87.69%. By rule of thumb, this classification rate is already highly acceptable (Li et al., 2019; Witten et al., 2017; Mong & Muraya, 2019). However, by underpinning the principle of machine learning algorithms, this confusion matrix only provides a preliminary fit of the data and does not necessarily measure the predictive or classification ability of the model (Li et al., 2019). This is because it is only tested on the data that is utilized for training and hence very prone to biases and overfitting (Hosmer et al., 2013). Table 4 presents the confusion matrix using the testing data. Meaning, the model is now tested on a new set of data that was not utilized in the training of the logistic regression model.

Table 4
Confusion matrix of the model using the testing (20%) data

Predicted	**Actual**	
	Subscriber	***Non-subscriber***
Subscriber	26 (42.50%)	4 (7.50%)
Non-subscriber	3 (5.00%)	27 (45.00%)
Correct Classification Rate	87.50%	
Misclassification Rate	12.50%	

Remarkably, a classification rate of 87.50% consistently substantiates the robustness or ability of the logistic regression model to classify a subscriber from a non-subscriber considering the input features presented. Li et al. (2019) noted that a correct classification rate of at least 70% is already useful for decision-making. Hence, the model output of this paper is highly useful for better marketing strategy among financing institutions in the country.

4.0 Conclusions

The application of the Machine Learning algorithm through the logistic regression model for business organizations is established in this study. By deploying the model developed in this study, lending institutions will be assisted on its target market with higher chances of subscription to their loan programs. Moreover, the results of the study supplement the literature about the predictive power of input features, namely: (1) number of existing loans, (2) monthly salary, (3) number of credit cards, (4) awareness to loan program of the company, and (5) required time for loan approval by the company, to predict loan subscription behavior of a qualified individual.

5.0 References

1. Antwi, S., Mills, E., Mills, G., & Zhao, X. (2012). Risk Factors of Loan Default Payment in Ghana: A case study of Akuapem Rural Bank. *International Journal of Academic Research in Accounting, Finance and Management Sciences. Volume 2, Issue 4 (2012). ISSN: 2225-8329*
2. Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. *Expert Syst. Appl.* 2019, *134*, 209–224. [Google Scholar] [CrossRef] [Green Version]
3. Byanjankar, A., Heikkila, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. In Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 7–10 December 2015; pp. 719–725. [Google Scholar] [CrossRef]
4. Bishop CM (2010) Pattern recognition and machine learning. Springer, New York
5. Choudhary, A. (2022). Predictive Analytics In Banking: How It Helps Lenders Succeed. https://atrium.ai/resources/predictive-analytics-in-banking-how-it-helps-lenders-succeed/
6. Chopra, A., & Bhilare, P. (2018). Application of ensemble models in credit scoring models. Business Perspectives and Research, 6(2), 129-141.
7. Evermann J, Rehse J-R, Fettke P (2016) A deep learning approach for predicting process behaviour at runtime. In: Dumas M, Fantinato M (eds) PARISE 2016 proceedings, Rio de Janeiro
8. Fishbein, M. (2002). Belief, attitude, intention, and behavior: An introduction to theory and research. Reading: Addison-Wesley.
9. George, J.F., 2004. The Theory of Planned Behaviour and Internet Purchasing. Internet Research, 14(3), pp.198-212.
10. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
11. Gu, S., & Wu, Y. (2019). Using the Theory of Planned Behaviour to Explain Customers' Online Purchase Intention. World Scientific Research Journal. Volume 5 Issue 9, 2019. ISSN: 2472-3703. DOI: 10.6911/WSRJ.201909_5(9).0026
12. Hamayel, M. J., Abu Mohsen. M. A., & Moreb, M. (2021). "Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine," 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 2021, pp. 33-37, doi: 10.1109/ICIT52682.2021.9491636.
13. Haykin SS (2009) Neural networks and learning machines: A comprehensive foundation. Pearson, New York
14. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. Wiley, Hoboken
15. Käfer, B. (2017). Peer-to-Peer Lending—A (Financial Stability) Risk Perspective. *Rev. Econ.* 2017, *69*, 1–25. [Google Scholar] [CrossRef] [Green Version]

16. Li, Pan, S., Huang, L., & Zhu, X. “A machine learning based method for customer behavior prediction.” (Tehniˇcki Vjesnik, 26(6), 2019) pp. 1670–1676.
17. Li, Ying and Wu, Zengyuan, "Prediction of Customers’ Subscription to Time Deposits Based on SMOTEENN-XGBoost Model" (2022). WHICEB 2022 Proceedings. 61. https://aisel.aisnet.org/whiceb2022/61
18. Ma, H.-Z. & Wang, X.-R. (2016). Influencing factor analysis of credit risk in P2P lending based on interpretative structural modeling. *J. Discret. Math. Sci. Cryptogr.* 2016, *19*, 777–786. [Google Scholar] [CrossRef]
19. Mild, A., Waitz, M., & Wöckl, J. (2015) How low can you go?—Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *J. Bus. Res.* 2015, *68*, 1291–1305. [Google Scholar] [CrossRef]
20. Kang B, Kim D, Kang S-H (2012) Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction. Expert Syst Appl 39(5):6061–6068
21. Mong, D. & Muraya, M. (2019). Analysis of Individual Loan Defaults Using Logit under Supervised Machine Learning Approach. Asian Journal of Probability and Statistics. 3. 1-12. 10.9734/ajpas/2019/v3i430100.
22. Pentzlin, D. (2011). Why to integrate sustainability criteria in financial regulation. Friends of the Earth. URL: www. foeeurope. org.
23. Rajagopal, S.K.P. (2020). Predicting Student University Admission Using Logistic Regression. European Journal of Computer Science and Information Technology. Vol.8, No.3, pp.46-56, June 2020. Published by *ECRTD- UK.* Print ISSN: 2054-0957 (Print), Online ISSN: 2054-0965 (Online)
24. Sahragard, P., & Chahouki, M. (2015). An evaluation of predictive habitat models performance of plant species in Hoze soltan rangelands of Qom province. Ecological Modelling. 309. 10.1016/j.ecolmodel.2015.04.005.
25. Salem, A. S., & Mahamad, S. (2020). PITEH: Providing Financial Identities to Those without Credit Score. 2020 International Conference onComputational Intelligence (ICCI). Doi:10.1109/icci51257. 2020.9247 779
26. Sentosa, I. and Mat, N.K.N., 2012. Examining a Theory of Planner Behaviour (TPB) and Technology Acceptance Model (TAM) in Internet Purchasing Using Structural Equation Modeling. International Refereed Research Journal, Apr, Vol. III, Is. 2(2), pp. 62-77.
27. Teinemaa I, Dumas M, La Rosa M, Maggi FM (2019) Outcome-oriented predictive process monitoring: review and benchmark. ACM Trans Knowl Discov Data 13(2):17
28. Trafimow, D., Sheeran, P., Conner, M. and Finlay, K.A., 2002. Evidence that Perceived Behavioural Control is a Multidimensional Construct: Perceived Control and Perceived Difficulty. British Journal of Social Psychology, Vol. 41, pp. 101-121.
29. Wang, Q., Xiong, X., & Zheng, Z. (2020). Platform Characteristics and Online Peer-to-Peer Lending: Evidence from China. *Financ. Res. Lett.* 2020, *38*, 101511. [Google Scholar] [CrossRef]
30. Witten IH, Frank E, Hall MA, Pal CJ (2017) Data mining: Practical machine learning tools and techniques. Morgan Kaufmann/ Elsevier, Amsterdam
31. Ziemba, P., Zalas, A. R., & Becker, J. (2020). Client evaluation decision model in the credit scoring tasks. Procedia Computer Science, 176, 3301-3309.