MODELLING OF TENANT'S BEHAVIOR INFORMATION CHARACTERIZATIONS FOR CREDIT SCORING USING LOGISTIC REGRESSION IN MALAYSIA

LING KIM SIA

A thesis submitted in
fulfillment of the requirement for the award of the
Degree of Master of Science

Faculty of Applied Sciences and Technology
Universiti Tun Hussein Onn Malaysia

JUNE 2023

I dedicate this thesis to my beloved family members and my friends whose support me.

ACKNOWLEDGEMENT

First of all, I would like to express my appreciation to my main supervisor, Assoc. Prof. Ts. Dr. Siti Suhana binti Jamaian and co-supervisor, Dr. Syahira binti Mansur for their comments, remarks and encouragement throughout this study. Thanks for their time and efforts in giving me her valuable knowledge to complete this study. Besides, I would like to thank to my parents and friends for their endless supports. They direct or indirectly helping me to overcome the problems faced along the journey of completing this study.

ABSTRACT

In Malaysia, the mortgage applications of the low income group are usually rejected by banks and financial institutions as they have poor credit scores due to too little or even no credit history. Therefore, they normally rent a property, but their rental payment history does not accountable in the mortgage applications. This study aims to reduce the credit unscorable of low income group with limited credit history. In this study, the logistic regression is applied to compute the credit score of tenants based on their characteristics, without relying on the tenant's credit history. The penalized maximum likelihood estimation with ridge regression is utilized to find the parameters of the logistic regression model as the existing separation in training data. The initial 9 factors considered affecting tenants' credit score were gender, age, marital status, monthly income, household income, expense-to-income ratio, number of dependents, previous monthly rent and number of months late rental payments. The marital status factor was then removed from the logistic regression model as it is insignificant to the model. Meanwhile, k-fold cross-validation with Grid Search was applied to determine the appropriate regularization strength value for maximum likelihood estimation. The main factors of the tenant's credit score are the number of months late payment, gender, expense-to-income ratio, previous monthly rent and age. Besides, there is no underfitting or overfitting in the proposed credit scoring model. Meanwhile, the accuracy of the proposed tenant's credit scoring model on testing data is 0.90. Lastly, a graphical user interface was developed for tenant's credit scoring.

ABSTRAK

Di Malaysia, permohonan pinjaman gadai janji golongan berpendapatan rendah lazimnya ditolak oleh bank dan institusi kewangan kerana mereka mempunyai skor kredit yang rendah disebabkan ejarah kredit yang terhad atau tiada sejarah kredit. Oleh itu, mereka biasanya menyewa hartanah, tetapi sejarah pembayaran sewa mereka tidak dipertimbangkan dalam permohonan pinjaman perumahan. Kajian ini menggunakan regresi logistik untuk mengira skor kredit penyewa berdasarkan ciri-ciri mereka, tanpa bergantung kepada sejarah kredit penyewa. Penalti anggaran kebolehjadian maksimum dengan regresi ridge digunakan untuk mencari parameter model regresi logistik kerana pengasingan dalam *training data*. Faktor awal yang dipertimbangkan mempengaruhi skor kredit penyewa ialah jantina, umur, status perkahwinan, pendapatan bulanan, pendapatan isi rumah, nisbah perbelanjaan kepada pendapatan, bilangan tanggungan, sewa bulanan sebelumnya dan bilangan bulan lewat pembayaran sewa. Faktor status perkahwinan kemudiannya dikeluarkan daripada model regresi logistik kerana ia tidak penting kepada model. Sementara itu, *k-fold cross-validation* dengan *Grid Search* telah digunakan untuk menentukan nilai *regularization strength* yang sesuai untuk anggaran kebolehjadian maksimum. Faktor utama skor kredit penyewa ialah bilangan bulan lewat pembayaran, jantina, nisbah perbelanjaan kepada pendapatan, sewa bulanan sebelumnya dan umur. Selain itu, tiada *underfitting* atau *overfitting* dalam model pemarkahan kredit yang dicadangkan. Ketepatan model pemarkahan kredit penyewa yang dicadangkan pada *testing data* ialah 0.90. Akhir sekali, satu *graphical user interface* dibangunkan untuk pemarkahan kredit penyewa.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

d	-	Search Direction
g	-	Jacobian Matrix
$g(x)$	-	Logit of the Logistic Regression
H	-	Approximation to the Inverse of the Hessian Matrix
i	-	Number of Subintervals
$I(\beta)$	-	Fisher Information Matrix
k	-	Number of Iterations
L_0	-	Likelihood Function of the Simpler Model
L_1	-	Likelihood Function of the Model that Considered All Factors
$L(\beta)$	-	Log-likelihood Function
m	-	Number of Corrections Stored
n	-	Total Number of Independent Observations
p	-	Number of Independent Variable Considered
$R(\beta)$	-	Ridge Regression
$S(\beta)$	-	Regularizer
x	-	Independent Variable or Factor
y	-	Binary Outcome Dependent Variable
$\hat{m}$	-	Minimum Value of k and $m-1$
α	-	Step Size
β	-	Logistic Regression Parameter
$\ell(\beta)$	-	Likelihood Function for Logistic Regression
λ	-	Regularization Strength
$\pi(x)$	-	Probability of Default
AUC	-	Area under the Receiver Operating Characteristic Curve
BFGS	-	Broyden-Fletcher-Goldfarb-Shanno

CCRIS	-	Central Credit Reference Information System
CTOS	-	Credit Tip-Off Service
FICO	-	Fair Isaac Corporation
FN	-	Number of False Negatives
FP	-	Number of False Positives
FPR	-	False Positive Rate
LASSO	-	Least Absolute Shrinkage and Selection Operator
RAMCI	-	RAM Credit Information Sdn. Bhd.
ROC	-	Receiver Operating Characteristic
TN	-	Number of True Negatives
TP	-	Number of True Positives
TPR	-	True Positive Rate

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of research

Affordable housing has always been a hot topic in numerous countries around the world, including Malaysia. The National Housing Department Malaysia (2019) stated that the median multiple methodology is implemented as the key indicator to measure housing affordability in Malaysia. According to the median multiple methodology, a house is deemed affordable if its price is not over three times the annual household income. The median monthly Malaysian household gross income decreased from RM5,873 in 2019 to RM5,209 in 2020 (Department of Statistics Malaysia, 2021). Therefore, affordable housing for Malaysians with median household income is priced at RM187,524 and below. Based on the National Property Information Centre (2022), the median house price in 2021 is RM310,000, which is 1.65 times the price of affordable housing.

The household income classification in Malaysia is divided into three categories: B40, M40, and T20. The B40 represents the bottom 40% of the Malaysian household group whose household income is below RM4,850 per month. Meanwhile, M40 is the middle 40% of the household group whose the household income is between RM4,850 to RM10,959 per month. And, the T20 represents the top 20% class with a household income of at least RM10,960 per month (Department of Statistics Malaysia, 2020). The housing issue has become more severe as 20%, or about 600,000 households in the M40 group,

have slipped into the B40 group due to the Covid-19 crisis in 2020 (The Star, 2021).

Therefore, some housing schemes are introduced by the government in Malaysia to assist the M40 and B40 groups in owning a house, such as Perumahan Rakyat 1 Malaysia (PR1MA), Program Perumahan Rakyat (PPR), and the Rent-to-Own scheme (Liu & Ong, 2021). However, not all low household income groups will benefit from the housing schemes due to limited units. Thaker (2020) stated 48% demand for affordable homes while the supply is only 28%. According to the Central Bank of Malaysia (2018), the key reasons for mortgage rejection include insufficient income to support debt repayment, adverse credit history, and inadequate financial documentation. Additionally, due to borrowers' age or poor credit scores, banks and financial institutions reject 60% of mortgage applications of people looking to purchase affordable housing (The Sun Daily, 2021).

In Malaysia, the Central Credit Reference Information System (CCRIS) is a system created by the Central Bank of Malaysia to synthesize the credit information of borrowers without credit scoring and is available to every financial institution. The CCRIS report shows the outstanding loans, special attention accounts, and the number of approved or rejected loan or credit facility applications made in the past 12 months (Ebekozien *et al*., 2019). Moreover, Malaysians can obtain their credit reports with credit scores through private credit reporting agencies in Malaysia, such as Credit Tip-Off Service (CTOS) and RAM Credit Information Sdn. Bhd. (RAMCI).

In the past, credit bureaus such as Fair Isaac Corporation (FICO) and Experian only set credit history as the credit score factor. The credit scoring model that depends only on credit history cannot be used to gain credit scores for those individuals with little or no credit history. As a result, some credit bureaus have generated credit scoring models using additional non-financial data, i.e., the use of rental payment records by Experian and the use of utility data, evictions, and other variables by FICO (Djeundje *et al.*, 2021). Besides, the research papers that use non-financial data such as rental payment records, utility data, criminal history, and delinquency also are reviewed by Njuguna &

Sowon (2021). Some researchers utilized other non-financial data such as individual characteristics, loan characteristics, and behavioural variables to compute the probability of default or credit score (Lin, Li & Zheng, 2017; Chamboko & Bravo, 2019; Adzis *et al.*, 2020; Saha, Lim & Siew, 2021).

The chance of individuals who are lack credit histories to get a mortgage will increase if financial institutions use non-financial data to develop the credit scoring model. In this research, we focus on computing tenants' credit scores based on their characteristics, monthly rent, and rental payment behaviour, particularly the credit score of the B40 group who rents a house.

1.2 Problem statement

Individuals with too little credit history or thin files are referred to as 'credit unscored' and those without any credit history are referred to as 'credit invisible' (Njuguna & Sowon, 2021). According to the Central Bank of Malaysia (2022), the minimum income eligibility for new credit card holders is set at RM24,000 per annum. The B40 group with at least RM 2,000 monthly income in rural areas can apply for a credit card, but some challenges are faced such as internet connection problems and lack of financial education (Sharizan, Redzuan & Rosman, 2021). The B40 category in rural areas is usually 'credit unscored' or 'credit invisible', where they have no credit records or poor credit scores due to insufficient credit history to support their mortgage application (Turner & Walker, 2019; Djeundje *et al.*, 2021). Hence, they normally rent a property since they are not affordable to own. However, their rental payment records are not accountable in mortgage applications.

Credit score has extended from banks to areas such as rental property, car and home insurance (Njuguna & Sowon, 2021). For example, TransUnion introduced "ResidentScore" which utilizes rental data to predict the likelihood of an eviction (TransUnion, 2021). In addition, Turner & Walker (2019) showed that adding rental payment data as a factor in FICO and VantageScore credit scoring models tends to reduce credit unscorable dramatically. Therefore, in this research, a credit profile for the tenants based on their rental payment behaviour such as late payment is proposed to measure the creditworthiness of tenants.

This credit profile is aimed to increase the confidence of banks, future property investors and developers to select the 'credit unscored' or 'credit invisible' B40 group as their potential customers. Besides, this credit profile has potential to support the tenant's mortgage application. Additionally, this credit scoring will assist government agencies in offering the designated B40 group appropriate solutions and incentives.

1.3 Objectives of research

The objectives of this research are

(i) to predict the probability of tenants defaulting based on their rental information records by using the logistic regression model,

(ii) to analyze the effect of the factors considered on the tenant's credit score according to the values of logistic coefficients,

(iii) to evaluate the performance of the proposed tenant's credit scoring model, i.e., accuracy, precision, recall and area under the receiver operating characteristic curve,

(iv) to develop a graphical user interface for tenant's credit scoring using HTML and Flask library in Python.

1.4 Scopes of research

In this study, multivariable logistic regression is implemented to develop a credit scoring model based on tenants' characteristics, monthly rent, and rental payment behaviour. The 9 factors that are considered to affect the credit score of tenants are their gender, age, marital status, monthly income, household income, expense-to-income ratio, number of dependents, previous monthly rent and number of months late payment. The landlord company, Homiee in Malaysia is where the rental payment records are obtained. The parameters of the multivariable logistic regression were also estimated using the maximum likelihood method and thus, the probability of tenant's default is generated. Meanwhile, the tenant's credit scoring model is proposed and the effect of the factors considered on the tenant's credit score is analyzed. Lastly, the

performance of the proposed tenant credit scoring model is evaluated and a graphical user interface is developed for the proposed model.

1.5 Significance of research

This study proposes a credit scoring model that is independent of credit history. This study aims to increase the credit scorable of low income group with limited credit history and hence might increase the approval rate of their mortgage application.

1.6 Framework of research

This thesis consists of six chapters. The first chapter discusses the background of research, problem statement, objectives of the research, scope of research, significance of the research and framework of research.

The literature review of this research is discussed in Chapter 2. This chapter introduces the existing credit scoring models in commerce, the factors affecting credit score and the machine learning methods used to generate credit scoring models in previous studies. Besides, the methods to evaluate the performance of machine learning classier are presented. The machine learning classier applied in this study is logistic regression. Therefore, the theory for maximum likelihood estimation is reviewed to determine the parameters of the logistic regression. The collected data may under separation or overlap. Thus, penalized maximum likelihood estimation for solving separation is discussed in this chapter. Meanwhile, the significance tests of the coefficient in logistic regression are also mentioned to decide the significant factors of credit score.

Chapter 3 focuses on the methodology of this research. In this research, the types of data considered as factors of credit score, linear programming for separation detection and maximum likelihood estimation method for logistic regression are investigated to achieve the objectives of this research. Furthermore, the algorithm of the proposed credit scoring model, factor reduction in model and the performance of logistic regression for classification are also presented in this chapter.

Besides, the proposed tenant credit scoring model is discussed in Chapter 4. The parameters of the logistic regression model are obtained using the maximum likelihood estimation method and hence the probability of the tenant defaulting can be computed. The effect of the factors considered on the tenant's credit score and the performance of the proposed model is also discussed in this chapter.

In Chapter 5, a graphical user interface for the proposed tenant's credit scoring is developed. Lastly, the conclusion of this research and recommendations for future study are included in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Credit scoring model

A credit score is a creditworthiness indicator used by banks and financial institutions to determine their potential borrowers' likelihood of defaulting on a loan. The higher the loan applicant's credit score, the higher the chance of the loan application being approved. FICO score created by Fair Isaac Corporation (FICO) and VantageScore introduced by United States national consumer reporting agencies i.e., Experian, Equifax and TransUnion are common credit scores used in the United States (Albanese, 2021). In Malaysia, the Credit Tip-Off Service (CTOS) score is the most common credit score applied. The three credit scores utilize similar factors, i.e., payment history, credit amounts owed, length of credit history, credit mix and new credit but with different proportions (Fair Isaac Corporation, 2021; VantageScore Solutions, 2021; Credit Tip-Off Service, 2021). The mathematical calculations behind these three credit scores are confidential, therefore, not public. Many research papers applied machine learning, such as neural networks, support vector machine, decision trees, logistic regression, fuzzy logic and genetic programming for developing credit scoring models (Abdou & Pointon, 2011; Louzada & Fernandes, 2016). Furthermore, Munkhdalai, Lee & Ryu (2020) proposed a hybrid credit scoring model using neural networks and logistic regression, while Kumar, Shanthi & Bhattacharya (2021) proposed a hybrid credit scoring model using neural networks and k-means algorithm.

The credit scoring model that depends only on credit history cannot be used to gain credit scores for those individuals with little or no credit history. Therefore, some credit bureaus have generated credit scores using additional non-financial data, i.e., the use of rental payment records by Experian and utility data, evictions and other variables by FICO (Djeundje *et al.*, 2021). Besides, the research papers that use non-financial data such as rental payment records, utility data, criminal history, and delinquency also are reviewed by Njuguna & Sowon (2021).

Some researchers utilized other non-financial data such as individual characteristics, loan characteristics and behavioural variables to compute the probability of default or credit score (Lin *et al.*, 2017; Chamboko & Bravo, 2019). These papers involved loan characteristics such as the loan amount, loan term, installment size, and loan interest rate. Meanwhile, the behavioural variables included are the number of missed payments and the average length of delinquency spells. Furthermore, Adzis *et al.* (2020) and Saha *et al.* (2021) investigated the factors contributing to home mortgage loan default, i.e., individual characteristics and loan characteristics, by utilizing borrowers' default data in Malaysia. The individual characteristics covered in the papers mentioned above include gender, ethnicity, age, marital status, place of residence, the status of children, level of education, occupation, income, debt-to-income ratio, payment-to-income ratio, and others. The individual characteristics that significantly affect the default rate according to these papers are presented in Table 2.1.

Table 2.1: Individual Characteristics that Significantly Affect Default Rate

Reference	Individual Characteristics				
	Gender	Age	Marital Status	Income	Debt-to-income ratio or payment-to-income ratio
Lin *et al.*, 2017	/	/	/		/
Chamboko & Bravo, 2019	/	/		/	/
Adzis *et al.*, 2020	/				
Saha *et al.*, 2021	/	/			/

In addition, some papers generated credit scoring models without using credit history. Berg *et al.* (2020) proposed a credit scoring model using only digital footprint variables such as device type, operating system, and email host. In order to create a different model for comparison, the paper used credit bureau scores and digital footprint variables. The paper concluded that digital footprint variables complement rather than a substitute for credit bureau information. Additionally, Djeundje *et al.* (2021) used email usage variables such as the fraction of emails sent in certain periods, the fraction of emails sent or received from non-top financial product providers, and the number of contacts sent to build a credit scoring model. The paper also found that a model that incorporates email usage and psychometric variables performs better than a model that incorporates only individual characteristics. Shema (2019) generated a credit scoring model only based on mobile airtime recharge or top-up history.

2.2 Tenant screening report

Most landlords will not rent to tenants with criminal records, eviction records, or poor credit scores. Tenant screening services provide tenant screening reports with the tenant's information, such as criminal records, eviction records, and credit score databases. Landlords can use tenant screening services to decide who to rent their property to (So, 2022). The landlords can screen potential

tenants to make more informed decisions to reduce the risk of non-payment and avoid eviction hassle due to problematic tenants (Credit Tip-Off Service, 2023).

In United States, Experian is the first consumer reporting agency incorporating rental history and credit history into a tenant credit check report (Experian, 2022). Besides, TransUnion introduced "ResidentScore" which specifically utilizes rental records data to determine the likelihood of an eviction (TransUnion, 2021). For Malaysia, CTOS tenant screening report provides identity verification, financial checks, litigation, bankruptcy checks, and "know-your-customer" screening, but the credit score of the tenant is not provided (Credit Tip-Off Service, 2022).

2.3 Performance of machine learning classifier

In machine learning, the collected data is split into training and testing data to generate a model for a classification task. The training data is used to train the classifier model and then the performance of the model is tested with testing data (Xie *et al.*, 2011). Random forest, decision tree, support vector machines and logistic regression are examples of machine learning classifier models. When the model correctly classifies the default class, the prediction is called a true positive. When the model incorrectly classifies the non-default class as the default class, the prediction is a false positive. Similarly, a false negative is an incorrect prediction that the default class is not the default class, and a true negative is a correct prediction of the non-default class. These four outcomes can calculate the classification performance, i.e., accuracy, precision, and recall (Hackeling, 2017).

Based on Hackeling (2017), there are two fundamental causes of the machine learning classifier's prediction error, i.e., the model's bias and its variance. A model with high variance overfits the training data, while a model with high bias underfits the training data. Overfitting occurs if the accuracy of training data is significantly higher than testing data, while underfitting occurs if the accuracy of both training and testing data is low (Gu *et al.*, 2016).

Besides, the receiver operating characteristic (ROC) curve is commonly used to visualize a classifier's performance. ROC curve plots the

REFERENCES

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, *18*(2-3), 59–88.

Adzis, A. A., Lim, H. E., Yeok, S. G., & Saha, A. (2020). Malaysian residential mortgage loan default: a micro-level analysis. *Review of Behavioral Finance*.

Agarwal, A., & Saxena, A. (2018). Malignant tumor detection using machine learning through scikit-learn. *International Journal of Pure and Applied Mathematics*, *119*(15), 2863–2874.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Albanese, J. (2021). *BIG DATA & BIG ERRORS*. Washington: Student Borrower Protection Center.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1–10.

Andruski-Guimaraes, I. (2016). Detecting separation in logistic regression via linear programming. *Proceedings of the XVIII Latin-Iberoamerican Conference on Operations Research, CLAIO 2016*, 38-44.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, *12*(4), 387–415.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, *33*(7), 2845–2897.

Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression

model with emphasis on medical research. *Journal of data analysis and information processing*, *7*(4), 190–207.

Bolton, C. (2010). *Logistic regression and its application in credit scoring* (Unpublished doctoral dissertation). University of Pretoria.

Botes, M. (2013). *Comparing logistic regression methods for completely separated and quasi-separated data* (Unpublished doctoral dissertation). University of Pretoria.

Bujang, M. A., Sa'at, N., Bakar, T. M. I. T. A., Joo, L. C., et al. (2018). Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *The Malaysian journal of medical sciences: MJMS*, *25*(4), 122.

Central Bank of Malaysia. (2018). *Risk Developments and Assessment of Financial Stability in 2017.* Retrieved November 30, 2021, from https://www.bnm.gov.my/documents/20124/856368/cp01.pdf/8976bcd6-dd90-84bc-34ea-4087d8311318?t=1585713818627

Central Bank of Malaysia. (2022). *Credit Card.* Retrieved December 23, 2022, from https://www.bnm.gov.my/PD-CreditCardl

Chamboko, R., & Bravo, J. M. (2019). Frailty correlated default on retail consumer loans in zimbabwe. *International Journal of Applied Decision Sciences*, *12*(3), 257–270.

Chin, X. Y., Lau, H. Y., Chong, Z. X., Chow, M. P., & Salam, Z. A. A. (2021). Personality prediction using machine learning classifiers. *Journal of Applied Technology and Innovation (e-ISSN: 2600-7304)*, *5*(1), 1.

Credit Tip-Off Service. (2021). *5 Factors That Can Impact Your Credit Score.* Retrieved November 30, 2021, from https://ctoscredit.com.my/personal/5-factors-can-impact-credit-score/

Credit Tip-Off Service. (2022). *CTOS Tenant Screening Report Sample.* Retrieved September 15, 2022, from https://uat-tenantscreening.ctoscredit.com.my/tenancy/pdf/CTOSTenantScreeningReportSample.pdf

Credit Tip-Off Service. (2023). *CTOS Tenant Screening.* Retrieved January 28, 2023, from https://ctoscredit.com.my/business/evaluate-new-customer/tenant-screening/

Cule, E., & De Iorio, M. (2012). A semi-automatic method to guide the choice of ridge parameter in ridge regression. *arXiv preprint arXiv:1205.0686*.

Department of Statistics Malaysia. (2020). *Household Income and Basic Amenities Survey 2019 Report*. Putrajaya: Department of Statistics Malaysia.

Department of Statistics Malaysia. (2021). *Household Income Estimates and Incidence of Poverty Report, Malaysia, 2020*. Putrajaya: Department of Statistics Malaysia.

Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, *163*, 113766.

Duffy, D. E., & Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics-Theory and Methods*, *18*(3), 959–980.

Ebekozien, A., Abdul-Aziz, A.-R., & Jaafar, M. (2019). Housing finance inaccessibility for low-income earners in malaysia: factors and solutions. *Habitat International*, *87*, 27–35.

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In *Informatics* (Vol. 8, p. 79).

Experian. (2022). *Tenant credit report.* Retrieved September 15, 2022, from https://connect.experian.com/credit-report/tenant-credit-report.html

Fair Isaac Corporation. (2021). *What's in my FICO® Scores?* Retrieved November 30, 2021, from https://www.myfico.com/credit-education/whats-in-your-credit-score

Febrianti, R., Widyaningsih, Y., & Soemartojo, S. (2021). The parameter estimation of logistic regression with maximum likelihood method and score function modification. In *Journal of physics: Conference series*

(Vol. 1725, p. 012014).

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Gu, Y., Wylie, B. K., Boyte, S. P., Picotte, J., Howard, D. M., Smith, K., & Nelson, K. J. (2016). An optimal sample data usage strategy to minimize overfitting and underfitting effects in regression tree models based on remotely-sensed data. *Remote sensing*, *8*(11), 943.

Gupta, A., Sharma, A., & Goel, A. (2017). Review of regression analysis models. *Int. J. Eng. Res.*, *6*(08), 58–61.

Hackeling, G. (2017). *Mastering machine learning with scikit-learn*. Packt Publishing Ltd.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, *21*(16), 2409–2419.

Jason, B. (2022). *Tune Hyperparameters for Classification Machine Learning Algorithms.* Retrieved December 23, 2022, from `https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/l`

Konis, K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models* (Unpublished doctoral dissertation). University of Oxford.

Konis, K., & Fokianos, K. (2009). Safe density ratio modeling. *Statistics & probability letters*, *79*(18), 1915–1920.

Kumar, A., Shanthi, D., & Bhattacharya, P. (2021). Credit score prediction system using deep learning and k-means algorithms. *Journal of Physics: Conference Series*, *1998*(1), 012027.

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in china. *Applied Economics*, *49*(35), 3538–3545.

Liu, D. C., & Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, *45*(1), 503–528.

Liu, J., & Ong, H. Y. (2021). Can malaysia's national affordable housing policy guarantee housing affordability of low-income households?

Sustainability, *13*(16), 8841.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117–134.

Mansournia, M. A., Geroldinger, A., Greenland, S., & Heinze, G. (2018). Separation in logistic regression: causes, consequences, and control. *American journal of epidemiology*, *187*(4), 864–870.

Marime, N., Magweva, R., & Dzapasi, F. D. (2020). Demographic determinants of financial literacy in the masvingo province of zimbabwe. *PM World J*, *9*(Iv), 1–19.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC medical research methodology*, *7*(1), 1–10.

Munkhdalai, L., Lee, J. Y., & Ryu, K. H. (2020). A hybrid credit scoring model using neural networks and logistic regression. In *Advances in intelligent information hiding and multimedia signal processing* (pp. 251–258). Springer.

National Housing Department Malaysia . (2019). *National Affordable Housing Policy*. Putrajaya: National Housing Department Malaysia.

National Property Information Centre. (2022). *Residential Prices Yearly Update 2021*. Retrieved August 12, 2022, from `https://napic.jpph.gov.my/portal/web/guest/main-page?p_p_id=ViewStatistics_WAR_ViewStatisticsportlet&p_p_lifecycle=2&p_p_state=normal&p_p_mode=view&p_p_resource_id=fileDownload&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_count=1&fileURI=21960`

Njuguna, R., & Sowon, K. (2021). Poster: A scoping review of alternative credit scoring literature. In *Acm sigcas conference on computing and sustainable societies* (pp. 437–444).

Nocedal, J. (1980). Updating quasi-newton matrices with limited storage. *Mathematics of computation*, *35*(151), 773–782.

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., … Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, 56–69.

Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, *63*(7), 07TR01.

Pereira, J. M., Basto, M., & da Silva, A. F. (2016). The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, *39*, 634–641.

Rayner, J. (1997). The asymptotically optimal tests. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(3), 337–345.

Saha, A., Lim, H.-E., & Siew, G.-Y. (2021). Housing loan repayment behaviour in malaysia: An analytical insight. *International Journal of Business and Economics*, *20*(2), 141–159.

Salim, M., & Ahmed, A. (2018). A family of quasi-newton methods for unconstrained optimization problems. *Optimization*, *67*(10), 1717–1727.

Saputro, D. R. S., & Widyaningsih, P. (2017). Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). In *Aip conference proceedings* (Vol. 1868, p. 040009).

Sharizan, S., Redzuan, N. H., & Rosman, R. (2021). Issues and challenges of financial inclusion among low-income earners in rural areas of malaysia. *Turkish Journal of Economics*, *8*, 277–299.

Shema, A. (2019). Effective credit scoring using limited mobile phone data. In *Proceedings of the tenth international conference on information and communication technologies and development* (pp. 1–11).

Šinkovec, H., Heinze, G., Blagus, R., & Geroldinger, A. (2021). To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC medical research methodology*, *21*(1), 1–15.

So, W. (2022). Which information matters? measuring landlord assessment of tenant screening reports. *Housing Policy Debate*, 1–27.

Starkweather, J., & Moske, A. K. (2011). *Multinomial logistic regression.* Retrieved February 27, 2022, from https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf

Sur, P., & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, *116*(29), 14516–14525.

Thaker, H. M. T. (2020). Housing prices and affordability in malaysia: A look into the supply-side drivers of housing prices. *Spotlight on Research*, *5*.

The Star. (2021). *Roughly 600,000 families went from M40 to B40 due to pandemic, says Tok Pa.* Retrieved October 29, 2021, from https://www.thestar.com.my/news/nation/2021/10/25/roughly-600000-families-went-from-m40-to-b40-due-to-pandemic-says-tok-pa

The Sun Daily. (2021). *BMF Offers Recommendations to Put Affordable Homes Within Reach of B40 Families.* Retrieved November 30, 2021, from https://www.thesundaily.my/local/bmf-offers-recommendations-to-put-affordable-homes-within-reach-of-b40-families-AI7794763

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

TransUnion. (2021). *ResidentScore.* Retrieved November 30, 2021, from https://www.transunion.com/product/resident-score-reseller?fbclid=IwAR1BnF7_nuNpwUhuriHlzht_Bg0oAGAL-YI64K6arRFVxZQv0WDoL0PYt8Q

Turner, M., & Walker, P. (2019). Potential impacts of credit reporting public housing rental payment data. *Available at SSRN 3615881*.

Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Statistical methods in medical research*, *29*(11), 3166–3178.

VantageScore Solutions. (2021). *What Data Does VantageScore Use?*

Retrieved November 30, 2021, from https://vantagescore.com/lenders/why-vantagescore/how-it-works

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, *710*, 120–127.

Wang, Y.-P., Tao, S.-L., & Chen, Q. (2015). Retrieving the variable coefficient for a nonlinear convection–diffusion problem with spectral conjugate gradient method. *Inverse Problems in Science and Engineering*, *23*(8), 1342–1365.

Xie, X., Ho, J. W., Murphy, C., Kaiser, G., Xu, B., & Chen, T. Y. (2011). Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, *84*(4), 544–558.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274–13283.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, *23*(4), 550–560.

VITA

The author was born on December 30, 1997, in Perak, Malaysia. She went to SMK Nan Hwa, Sitiawan, Perak, Malaysia for her secondary school and pre-university. She received a degree of Bachelor of Mathematics Technology in 2021 and is currently pursuing her Master of Science degree at Universiti of Tun Hussein Onn Malaysia.