UNIVERSITI TEKNOLOGI MARA

MATHEMATICS AND STATISTICS

UNDERGRADUATE RESEARCH PROCEEDINGS 2025

UiTM CAWANGAN NEGERI SEMBILAN

Prediction Model of Home Loan Eligibility Status

Danisya Adlina Ruzi[1], Siti Aisyah Mohamed Kher[1], Isnewati Ab Malek[1,], Haslinda Ab Malek[1*,]
[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Negeri Sembilan Branch, Seremban Campus, 70300 Seremban, Negeri Sembilan, Malaysia*
[*]Corresponding: haslinda8311@uitm.edu.my

Abstract

A home loan is known as a mortgage which is a type of loan used to purchase a home or real estate. Home loans are essential for many individuals who are looking to buy a home without paying the entire price upfront. Confronting a prevalent challenge involving a notable of 60% rejection rate among home loan applicants, this issue was identified during the pre-pandemic year. Therefore, this study aims to develop an optimized home loan eligibility status prediction model. The study systematically evaluated Decision Tree approaches, incorporating variable selection for enhanced precision in model comparison. The findings of this study found that loan amount, marital status, property area, and co-applicant income as pivotal factors influencing home loan eligibility status. Notably, the Decision Tree Entropy model emerges as the optimal model, achieving a remarkable average squared error of 19.71%. This model exhibits superior performance, evidenced by a high sensitivity of 84.16% and accuracy of 59.78% according to the confusion matrix. The results are hoped to provide valuable insights for improving house loan eligibility prediction models, which could improve the banking sector decision-making procedures and in turn, improve consumer satisfaction.

Keywords*:* Housing loan, Loan Approval, Decision Tree.

Introduction

A home loan is a loan that helps a borrower purchase a home. The amount borrowed from a home is based on the appraised value of the home and the amount of money paid as a down payment. The home will be put up as collateral and loan applicants will make monthly mortgage payments for a specified amount of time (typically 15 or 30 years) until the loan is paid off. According to a study done by [1], in Malaysia, home loans can be divided into three main categories: Basic Term Loans, Semi-Flexi Loans, and Full-Flexi Loans.

In this study, the home loan eligibility status is predicted. The term 'home loan eligibility status' defines an applicant's suitability for a home loan based on the requirements set by the lender. It indicates if the applicants meet the requirements for a home loan based on variables such as income and loan amount. Home loan eligibility status is an important factor in the lending industry, and financial institutions must predict the likelihood of loan approval accurately. However, according to the finding by [2], an average of 60% of the home loan applicants failed to get a home loan during the pre-pandemic year. This has been a problem and a significant issue for the home loan applicants.

Therefore, this study was conducted to construct the home loan eligibility status model in determining the type of model that is best to use. The model can help to analyse applicant data more rapidly and improve the overall customer experience by offering faster and more accurate predictions. The application of these models can assist banks and financial institutions to make more informed home loan approval choices. Moreover, these models can reduce the risk of loan default and improve the overall efficiency of the home loan approval process.

As the demand for home loans in banks grows by the day, [3] performed a Modern Approach to Home Loan Sanctioning in Banks Using Machine Learning. The goal is to apply machine learning tools that use classification algorithms to predict worthy applicants for home loan approval as an appropriate option to reduce human efforts and make successful decisions in the home loan approval process. A system to develop a model is built in their research by training the system with records and approval results of previously applied home loan applicants

Moreover, study done by [4] employed two machine learning classification models, namely Decision Tree and Random Forest, for loan prediction. The three findings indicate that the highest level of accuracy achieved is

62.12%, accompanied with a confidence factor of 0.15. Other than that, [5] employed decision trees to construct an accurate and efficient predictive model for forecasting home loan approval. Experiments were conducted utilising several tree methodologies, spanning from the simplest and most comprehensible decision tree to the most intricate random forests.

Hence, [6] did a comprehensive study to make a detailed comparison between the Random Forest and Decision Trees algorithms. Both methodologies were employed on the identical dataset, and the outcomes revealed that the Random Forest algorithm exhibited much superior performance in terms of accuracy compared to the Decision Tree strategy, achieving a score of 80% as opposed to 73%. Moreover, [7] conducted a study to forecast the eligibility of consumers for home loan approval. Three machine learning models employed were the Logistic Regression Model, the Decision Tree, and the Random Forest. The study revealed that based on the comparison of the three models, Logistic Regression exhibits the highest accuracy rate of 89.7059%, followed by Decision Tree with an accuracy rate of 85.4054% and Random Forest with an accuracy rate of 77.4566%.

Based on the feature analysis run by [8], it showed that those who have a higher credit history have a better probability of getting a loan compared to those who do not. Furthermore, [9] conducted a study stated that the home loan approval status is determined to be independent of the applicant's income, gender, and loan amount. However, it appears that the credit history has the greatest impact on loan acceptance, suggesting that an applicant's credit history is used as an initial filter for loan applications.

The findings of this study will provide banking institutions with the best model for determining eligibility for housing loans. The greater demand for property loans justifies the need for a more effective, speedy, and prompt approach for bankers in predicting potential candidates. For the researchers, the study will help them uncover the best model for predicting the loan approval process that many researchers were not able to explore. Moreover, this study not only helps the applicant but also helps the bank institution by minimizing the risk and reducing the number of defaulters.

Methodology

Source of data

This study employed secondary data collected from GitHub. The dataset originally sourced from Analytics Vidhya's Loan Prediction Challenge, comprised 443 observations and 11 variables, with Loan Status serving as the dependent variable of interest [10]. The variables used in the analysis are summarized in Table 1.

Table 1: Description of the Variables

No	Variable	Type of Variables	Scale	Description
1	Gender	Categorical	Nominal	Male, Female
2	Marital Status	Categorical	Nominal	Yes, No
3	Number of Dependents	Continuous	Interval	0, 1, 2, 3+
4	Education Status	Categorical	Nominal	Graduate, Not Graduate
5	Self-Employment	Categorical	Nominal	Yes, No
6	Applicant's Income	Continuous	Interval	$0 - $22000
7	Co-Applicant's Income	Continuous	Interval	$0 - $11000
8	Loan Amount	Continuous	Interval	$15000 - $375000
9	Loan Term	Continuous	Interval	126.68 - 480 (months)
10	Property Area	Categorical	Nominal	Urban, Semiurban, Rural
11	Loan Status	Categorical	Nominal	Yes, No

Decision Tree

The purpose of employing a Decision Tree in this study is to develop a model that can be used to predict the class or value of the home loan eligibility status by learning simple decision rules inferred from prior data. This study employed three of the most common types of Decision Trees, namely Gini, CHAID, and Entropy. These types are distinguished from each other by the mathematical model used to select splitting property when extracting Decision Tree rules [11]

Model Assessment

Accuracy measures can be used to evaluate the ability of the estimated model to predict the target. The Receiver Operating Characteristic (ROC) Chart, Confusion Matrix, Misclassification Rate, Average Squared Error (ASE), and Receiver Operating Characteristic (ROC) Index are the tools from which different accuracy measures are derived.

Misclassification Rate

The misclassification rate is the frequency with which a classification model guesses a sample's class or category in a dataset incorrectly. It serves as an indicator of how well the model classifies the data. When assessing a classification model, the misclassification rate is determined by comparing the sample true class labels with the predicted class labels. The misclassification rate, for example, would be 10% if a model properly classified 90 out of 100 samples, meaning that 10% of the data were misclassified.

Average Squared Error

The average squared error (ASE) measures how close a regression line comes to a set of data points. The model with a smaller value of ASE indicates a better fit. Hence, it is better suited to be used in predicting the target. Equation 1, as shown, can be used to calculate the average squared error

$$\mathrm{ASE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{1}$$

Receiver Operating Characteristic (ROC) Index

The Receiver Operating Characteristic (ROC) index is the area under the value (AUC) in the ROC chart. The AUC is widely used to measure the accuracy of diagnostic tests. The higher the value of the ROC index, the greater the sensitivity and accuracy of the model.

Confusion Matrix

Accuracy is measured regarding the error rate percentage of records classified correctly or incorrectly. A confusion matrix can be used to determine this measurement. The number of accurate and inaccurate predictions made by the classification model concerning the actual target variable in the data is displayed in a confusion matrix.

Results and Discussion

Model comparisons between Decision Tree (DT) Gini, Decision Tree (DT) Entropy and Decision Tree (DT) CHAID were computed to compare the training sample and validation sample based on the values of the misclassification rate, average squared error, and Receiver Operating Characteristic (ROC) index.

Table 2: Model Comparison Based on Misclassification Rate, Average Squared Error and ROC Index

Model	Misclassification Rate			Average Squared Error			ROC Index		
	Valid	Train	Gap	Valid	Train	Gap	Valid	Train	Gap
DT Gini	0.430	0.307	0.123	0.278	0.197	0.081	0.494	0.735	-0.241
DT Entropy	0.413	0.303	0.110	0.291	0.197	0.094	0.570	0.722	-0.152
DT CHAID	0.458	0.386	0.072	0.247	0.232	0.015	0.559	0.621	-0.062

Table 2 shows the values of valid, train, and gap for the misclassification rate, average squared error, and ROC index for each model. The valid and gap values are used to find the best model. The gap values are calculated by finding the difference between valid and train values. The negative sign is not taken into consideration for the gap value. From Table 2, it is found that the DT Gini has the highest gap value for the misclassification rate (0.123) and ROC index (0.241). Therefore, DT Gini is found to be the overfit model.

Additionally, the valid value is evaluated to determine the best fit model. The best model has the lowest valid values for the misclassification rate and average squared error. In contrast, the model must have the highest valid value of the ROC index. DT Entropy has the lowest valid misclassification rate (0.413) and the highest ROC index valid value (0.570). Hence, DT Entropy is concluded as the best model since it meets most of the conditions.

Table 3 summarise that there is no underfit model since all model performances have the best validation results and no negative valid value. However, there is one overfit model, which belongs to the Decision Tree Gini. An overfit model can be determined when it has the largest gap value between valid and train value. Thus, the Decision Tree Gini is an overfitted model because it has the largest gap values for the misclassification rate and the ROC index. In choosing the best models, the valid values for the misclassification rate and average squared error need to be the lowest, while the value for the ROC index needs to be the highest. The Decision Tree Entropy meets most of the requirements, hence, it is chosen as the best model.

Table 3: Summary of Model Comparison

Assesment of Generalization	**Model Description**
Underfit	None
Overfit	Decision Tree Gini
Best Fit	Decision Tree Entropy

Best Model

The confusion matrix is observed to further confirm that the Decision Tree Entropy is the best model. A confusion matrix is a table that describes the performance of a classification model on a set of test data. The factors in the confusion matrix can be examined to yield several performance measures such as sensitivity, specificity, and accuracy, which provide insights into the model's performance.

Table 4: Result Analysis of Confusion Matrix

Model	**Sensitivity**	**Specificity**	**Accuracy**
Decision Tree Gini	0.6931	0.4231	0.5754
Decision Tree Entropy	0.8416	0.2821	0.5978
Decision Tree CHAID	0.4752	0.6281	0.5419

Based on Table 4, the Decision Tree CHAID has the highest specificity values at 62.81%, showing that the model is the best model for classifying the 'No' group of home loan eligibility status. Meanwhile, the Decision Tree Entropy

has the highest sensitivity and accuracy values at 84.16% and 59.78%, respectively. This confirmed that Decision Tree Entropy is the best model for classifying the 'Yes' group of home loan eligibility status and the best in predicting the overall performance of the model.

Conclusion

The specific requirements for obtaining a home can be ambiguous, causing applicants to question the qualifying conditions and the procedures involved in applying. Thus, this study has been conducted using secondary data obtained which includes ten variables that are believed to have an impact on the approval of house loans. This dataset has 443 observations with a binary target variable. The Decision Tree Gini is found to be overfit while none of the models are underfit. It was discovered that implementing Decision Tree Entropy produces the best model for predicting a house loan applicant's eligibility status with an average squared error of only 19.71% and highest accuracy values of 59.78%. Consequently, the Decision Tree Entropy model was deemed to be the most reliable in predicting home loan eligibility status. It is hoped that this research provides valuable insights for home-loan applicants, banking institutions and other researchers for their own interest. This study is in line with a study done [12], it indicates the entropy reduction decision tree technique outperforms the other seven methods in terms of overall classification accuracy rates at the 0.5 cut-off.

For future studies, it is recommended explore other models apart from DT Gini, Decision Tree Entropy and Decision Tree CHAID. One example of an alternative decision tree model is the Random Forest algorithm. Random Forest combines multiple decision trees to make predictions, offering improved accuracy and reducing the risk of overfitting. By utilising a variety of models, researchers can evaluate their performance and compare the results to identify the most effective model for predicting home loan eligibility status.

References

[1] Hassan, M. M., Ahmad, N., & Hashim, A. H. (2022). Housing Property Investment Opportunity in Malaysia: Things that new investor should know. International Journal of Academic Research in Business and Social Sciences, 12(1), 924–941. https://doi.org/10.6007/IJARBSS/v12-i1/12003.

[2] Sooi, C. C. (2021, Sep). Mitigating malaysia's worrying trend of high home loan rejection focus malaysia. Retrieved from https:// focusmalaysia.my/mitigating-malaysias-worrying-trend-of-high-home-loan-rejection.

[3] Rath, G. B., Das, D., & Acharya, B. (2021). Modern approach for loan sanctioning in banks using machine learning. In Advances in machine learning and computational intelligence: Proceedings of icmlci 2019 (pp. 179–188).

[4] Gautam, et al. (2020). Loan prediction using decision tree and random forest. International Research Journal of Engineering and Technology (IRJET), 7(08), 853–856.

[5] Alaradi, M., & Hilal, S. (2020). Tree-based methods for loan approval. In 2020 international conference on data analytics for business and industry: Way towards a sustainable economy (icdabi) (pp. 1–6).

[6] Madaan, et al. (2021). Loan default prediction using decision trees and random forest: A comparative study. In Iop conference series: Materials science and engineering (Vol. 1022, p. 012042).

[7] Dutta, P. (2021). A study on machine learning algorithm for enhancement of loan prediction. International Research Journal of Modernization in Engineering Technology and Science, 3.

[8] Sunitha, T., Chandravallika, M., Ranganayak, M., Suma Sri, G., Jagadeesh, T. V. S., & Tejaswi, A. (2021). Predicting the Loan Status using Logistic Regression and Binary Tree. Paper presented at ICICNIS 2020. http://dx.doi.org/10.2139/ssrn.3769854.

[9] Alaradi, M., & Hilal, S. (2020). Tree-based methods for loan approval. In 2020 international conference on data analytics for business and industry: Way towards a sustainable economy (icdabi) (pp. 1–6).

[10] Temburwar, S. (2019). Loan Prediction Dataset [Data set]. GitHub. https://github.com/shrikant-temburwar/Loan-Prediction-Dataset.

[11] Frempong, N. K., Nicholas, N., & Boateng, M. (2017). Decision tree as a predictive modeling tool for auto insurance claims. International Journal of Statistics and Applications, 7(2), 117–120

[12] Zurada, J. (2010). Could decision trees improve the classification accuracy and interpretability of loan granting decisions? In 2010 43rd hawaii international conference on system sciences (p. 1-9). doi: 10.1109/HICSS.20