**Group 3A**

**Report of SAS Project - P000182DA**

**Data-Driven Home Loan Origination Scorecard Development**

**Team Members:**

Chelsea Wang s3886626

Jonathan Feng s3858705

Kar Yenc Natalie Mun s3991774

Livia Nathania Fireta s3980951

Phan Van Thai s3818387

Thach Duy Phong s3821853

**Academic Supervisors:**

Mali ABDOLLAHIAN

Yan Wang

Irene Hudson

**Industry Supervisors:**

Lucy Biasi

Bushra Khalid

Mahi Ramasubbu

Klara Vickov

**Students’ Contribution**

All six members have contributed equally to the project.

All members contributed to the data cleaning, pre-processing, model building, and scorecard development. The analysis and interpretation of the results were discussed and gathered by all members.

Contribution to this report for each member:

1. Chelsea Wang

Create the report.

Contributed to the literature review, the introduction part, and the scorecard development method and results.

Contributed to the final check.

1. Kar Yenc Natalie Mun

Contributed to the literature review, the disclaimer, acknowledgment, model evaluation method, discussion, and conclusion.

Contributed to the final check.

1. Livia Nathania Fireta

Contributed to the literature review, variables selection, model development methods, and results.

Contributed to the final check.

1. Jonathan Feng

Contributed to the executive summary and introduction.

1. Phan Van Thai

Contributed to the literature review, references, and data pre-processing.

1. Thach Duy Phong

Contributed to the references, and feature machine method and results.

**Disclaimer**

We declare the following to be our work unless otherwise referenced, as defined by the University’s policy on plagiarism. The work taken from other sources has been referenced using the referencing standards. We declare that we ensured our academic integrity.

**Acknowledgment**

We would like to express our gratitude and appreciation to the Royal Melbourne Institute of Technology (RMIT) University, for allowing us to gain industry experience by applying our knowledge and analytical skills before entering the real-world market. We would also like to extend our thanks to the following people for their continuous support and contributions to this project.

**RMIT Academic Supervisors**

We would like to thank our course coordinator, Dr Mali Abdollahian for organizing such an amazing opportunity to work with such an esteemed organization and for the continuous support. We would also like to extend our special thanks to our academic supervisors, Dr Irene Hudson and Dr Yan Wang for their continuous support, advice and guidance throughout this project during the bi-weekly meetings held.

**SAS Industry Supervisors**

We would like to thank SAS for providing such a valuable opportunity to gain real-world experience for us students. We would like to also extend our heartfelt gratitude to Lucy Biasi, Bushra Khalid, Mahi Ramasubbu, and Klara Vickov for providing their continuous guidance, comments, and suggestions for improvement leading to the completion of this project.

**Table of Contents**

[***Executive Summary 4***](#_heading=h.gjdgxs)

[***Introduction 6***](#_heading=h.30j0zll)

[***Literature Review 7***](#_heading=h.1fob9te)

[***Methods and Materials 11***](#_heading=h.3znysh7)

[*Data Pre-Processing 11*](#_heading=h.2et92p0)

[*Model Development 15*](#_heading=h.tyjcwt)

[*Model Evaluation 17*](#_heading=h.31qwk6jqjh7)

[*Segmentation Scorecard 17*](#_heading=h.3dy6vkm)

[***Results 19***](#_heading=h.1t3h5sf)

[*Feature Machine 19*](#_heading=h.4d34og8)

[*Variable Selection 19*](#_heading=h.2s8eyo1)

[*Decision Tree for Specific Variables 21*](#_heading=h.7yc9yotu1eu5)

[*Model Evaluation 24*](#_heading=h.17dp8vu)

[*Segmentation Scorecard 24*](#_heading=h.3rdcrjn)

[***Discussion 27***](#_heading=h.26in1rg)

[***Conclusion 29***](#_heading=h.lnxbz9)

[***Appendix 31***](#_heading=h.35nkun2)

[***References 38***](#_heading=h.1ksv4uv)

# Executive Summary

SAS (Statistical Analysis System), is a leading privately held software company, renowned for its innovative artificial intelligence (AI) and analytical products. SAS Viya, an open, cloud-native data, and AI platform enables organisations to gain valuable insights from data, facilitating intelligent decision-making and efficient implementation of necessary changes.

For our project, we aim to develop a data-driven home loan origination scorecard using advanced analytics and real banking data through SAS Viya. The goal is to enhance the accuracy and efficiency of the loan approval process within the Australian housing market.

The project commenced by setting the event-based sampling to the raw data, ensuring balanced data. Feature machine methods were applied to impute, transform, and variable selection to select relevant features for predictive analytics. Various advanced analytical models were then selected, fine-tuned, and evaluated via SAS Viya's pipelines, including Logistic Regression, Neural Networks, Support Vector Machines, and Random Forests. We used accuracy as our main evaluation metric, and Logistic Regression emerged as the most effective model, achieving the highest overall accuracy of 0.77.

The customer credit-risk scorecard was then developed leveraging the feature importance from the best performing model. This scorecard segmented customers into three risk levels: lower risk, medium risk, and higher risk based on the interquartile range (IQR) of the computed scores. The scorecard can streamline loan approval processes and reinforce decision-making and risk management for financial institutions.

Our project achieved the development of an improved predictive model and scorecard. Providing financial institutions with a practical tool to streamline credit risk assessment, reinforce decision-making, and enhance risk management through improved predictive accuracy. For future research, incorporating additional data variables and exploring alternative, updated modeling techniques can assist in further enhancing predictive power. Continuous monitoring of feature importance can also keep the model aligned with economic changes, ensuring its relevance and effectiveness over time. Overall, it is essential to perform regular updates and tuning of models according to the evolving market conditions and maintain competitive edges.

# Introduction

The Australian mortgage industry is a critical component of the national economy, driving growth and serving as an indicator of financial stability. However, the industry has become increasingly competitive in recent years due to rising interest rates, higher funding costs, and evolving regulations. One of the most significant challenges is the approaching fixed-rate cliff, further strained the market. Many financial institutions are engaging in aggressive lending practices to maintain market share, targeting riskier customers at thin margins. While this may provide short-term growth, it introduces long-term risks such as increased loan defaults and liquidity issues. These risks can have extensive consequences across the entire economy.

SAS, a global leader in advanced analytics, plays a crucial role in helping Australian banks navigate these complexities. SAS provides the tools and solutions needed to improve risk management and decision-making, particularly in the mortgage sector. However, traditional credit evaluation models may no longer be sufficient in this volatile market. Ineffective risk assessment mechanisms can lead to an increase in non-performing loans, endangering banks' financial health and subsiding customer satisfaction due to slow loan approval processes.

This research focuses on developing a data-driven home loan origination scorecard to improve risk management and enhance decision-making in the Australian mortgage market. Leveraging real banking data and SAS Viya, the project aims to develop predictive models that streamline loan approvals and assess customer credit risk with greater accuracy. The research addresses key challenges by refining data for better predictions, identifying the most accurate analytical models, and creating a scorecard to optimize loan approval processes.

The objectives of this project include performing data pre-processing to enhance the data quality and predictive quality of the models, evaluating multiple models to determine the best model, and building a scorecard that enhances both efficiency and risk management for financial institutions.

# Literature Review

**Data Imputation**

Sessa and Syed (2016) discusses many different techniques for data imputation including simple imputation methods such as mean, median, and mode imputation where the missing data is replaced with the mean, median, or mode of that feature for a given class depending on the dataset’s characteristics. These imputation methods are straightforward, yet they typically produce respectably good results when applied appropriately to the appropriate dataset. Their shortcomings include underestimating variance and overestimating sample size. Another imputation method that is recommended is the decision tree. According to Alam et al. (2023), the decision trees method is likely to offer the best accuracy for imputing the missing values, outperforming other techniques such as K-Nearest Neighbour, Neural Networks, and Support Vector Machine (SVM). This approach results in imputed values that demonstrate minimal deviation from the original, indicating that it preserves the underlying structure of the data. From these imputation methods, choosing the best imputation strategy is crucial to our project since it impacts the accuracy and reliability of the analysis.

**Outliers**

To improve model performance, Dinh and Thanh (2022) emphasised the significance of scaling numerical variables. Additionally, Nagajyothi (2020) emphasised the significance of resolving outliers before model creation, pointing out that their existence may result in forecasts that are not correct. Their research used logarithmic transformations to mitigate the impact of outliers. Their research used logarithmic transformations to mitigate the impact of outliers. Furthermore, Vimala and Sharmili (2018) suggested using binning techniques, particularly for categorical variables where the model's stability may be in danger because of an imbalance in value distribution.

**Feature Selection**

Feature selection is a crucial process used to extract important features in a dataset and prevents the model from being overfitted by reducing the model complexity and increasing the generalizability of the models (Sinap, 2024). The author also noted that the removal of unnecessary features from the models, reduces the computational time and resources and enhances the model’s interpretability. Furthermore, feature selection is important for several machine learning models. For example, Reddy and Kavitha (2010) noted that the process of feature selection is especially important if the neural network model is used for the loan default prediction since the addition of irrelevant variables could significantly affect the model performance of the neural network model. Therefore, feature selection is a necessary process as it not only improves the accuracy of the model but also optimizes the efficiency and interpretability of the model.

**Model Evaluation**

Kuhn and Johnson (2013) offer a practical guide to choosing the best models for predictive tasks, with a focus on techniques that boost accuracy and reliability. They start by highlighting resampling methods like cross-validation and bootstrapping, which provide a dependable way to evaluate a model’s performance and its ability to generalize to new data. Hyperparameter tuning is another key topic, where approaches like grid search and random search help fine-tune model parameters for the best results. To avoid overfitting, they recommend pairing tuning with cross-validation.

Choosing the right performance metric is crucial, as it depends on the specific goals of the project; for example, accuracy, RMSE, or AUC might each lead to different model choices. For binary classification problems, there are various types of evaluation metrics where the accuracy shows proportion for correct predictions, the F1-score balances the precision and recall, and the AUC score evaluates the model's ability to differentiate between positive and negative classes (Rainio et al., 2024). The authors also emphasize ensemble methods combining multiple models through techniques like bagging, boosting, and stacking to enhance accuracy, as ensembles typically outperform individual models. Additionally, feature selection plays an important role, especially when working with large datasets. They suggest methods like recursive feature elimination (RFE) and penalized models like LASSO to help remove irrelevant features that could cloud the model’s predictions.

**Model Algorithm**

Logistic Regression: According to Hosmer et al. (2013), logistic regression is a method for modeling binary outcomes, emphasising its interpretability and utility in practical applications. The authors thoroughly explain parameter estimation, variable selection, and diagnostic tools to assess model fit, making it especially useful in medical, social science, and risk assessment contexts. The book highlights logistic regression’s straightforward interpretation, which is beneficial for understanding relationships between predictor variables and outcomes.

Random Forests: Breiman (2001) introduces random forests as an ensemble learning technique that uses multiple decision trees to boost predictive accuracy. Breiman describes how random forests overcome overfitting issues common in single decision trees by averaging predictions, thus reducing variance. He also covers out-of-bag error estimates, which allow for built-in model validation, and variable importance measures that make random forests interpretable. This work establishes random forests as a robust method for handling high-dimensional, noisy data.

Neural Networks: Schmidhuber (2015) traces the evolution of neural networks and highlights innovations that have propelled modern deep learning. Schmidhuber explains architectures like convolutional and recurrent neural networks, which enable deep learning models to excel in complex tasks such as image recognition and natural language processing. He also discusses breakthroughs in training methods and hardware advances that have made deep learning scalable.

Support Vector Machines (SVMs): Cortes and Vapnik (1995) provide a comprehensive introduction to SVMs, focusing on maximising the margin between data classes for improved classification. The authors detail the "kernel trick," which allows SVMs to perform non-linear classification by mapping data into higher-dimensional spaces. This innovation makes SVMs powerful for complex classification tasks and sets the groundwork for their extensive application in text classification, bioinformatics, and image recognition.

**Scorecard Development**

Finlay (2012) lays out a practical, step-by-step guide to creating scorecards that are both effective and easy to interpret. He begins with data preparation, stressing the need for clean data by addressing missing values and outliers, and selecting the most predictive variables through statistical tests. For the model itself, Finlay recommends logistic regression because it’s straightforward and performs well for predicting outcomes like loan defaults. This model’s results can be easily turned into scores that are meaningful and actionable.

A big part of Finlay’s method involves Weight of Evidence (WOE) and Information Value (IV). WOE is used to convert variables into a format that works well with logistic regression, while IV helps pinpoint which variables add the most value, making the selection process clearer. He then explains scorecard scaling, where model outputs are adjusted to a standard range, making the scorecard easy to interpret as each score represents specific credit behaviors in a consistent format.

# Methods and Materials

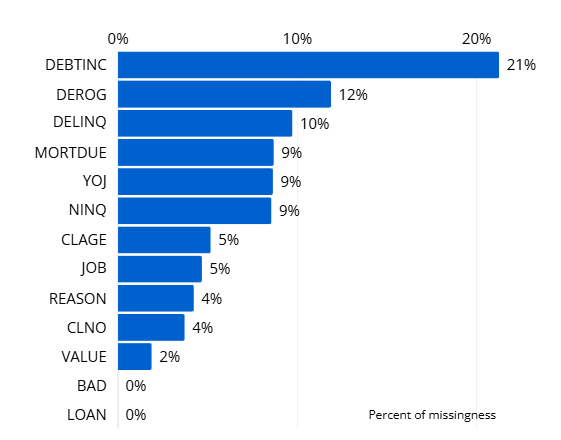
## Data Pre-Processing

The dataset comprises 5,960 observations across 12 variables, provided by SAS. The variables of the dataset are as below:

| **Variable Name** | **Type** | **Variable Details** |
| --- | --- | --- |
| **BAD** | Num | 1 = client defaulted on, loan 0 = loan repaid |
| **LOAN** | Num | Amount of the loan request |
| **MORTDUE** | Num | Amount due on the existing mortgage |
| **VALUE** | Num | Value of the current property |
| **REASON** | Char | DebtCon = debt consolidation, HomeImp = home improvement |
| **JOB** | Char | Type of occupation |
| **YOJ** | Num | Years at present job |
| **DEROG** | Num | Number of major derogatory reports |
| **DELINQ** | Num | Number of delinquent credit lines |
| **CLAGE** | Num | Age of oldest trade line in months |
| **NINQ** | Num | Number of recent credit lines |
| **CLNO** | Num | Number of credit lines |
| **DEBTINC** | Num | Debt-to-income ratio |

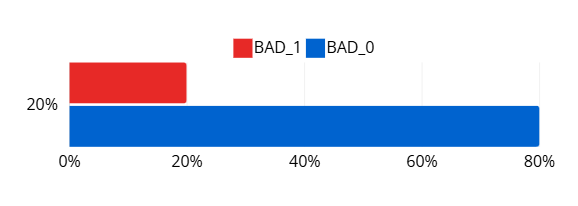
*Table 1: Description of each variable in this dataset*

Many variables in our dataset have substantial missing data, with five variables showing missingness rates between 2% and 5%, and six others between 9% and 21%.



*Figure 1: Missingness percentage of the dataset*

This presents a significant challenge that we addressed using imputation techniques to fill in missing values. Additionally, our binary dependent variable is imbalanced with an 80-20 distribution, requiring careful consideration in the modeling process to avoid biased predictions.



*Figure 2: Class proportion of target variable (BAD)*

To handle these challenges and improve model performance, we employed a few methods. The event-based sampling was employed to address the imbalance dataset. By utilising the event-based sampling, the data was sampled and resulted in data that has 50% of 0 and 50% of 1 as our target.

We implemented a data partitioning strategy to prepare our dataset for machine learning model training, validation, and testing. The partitioning was done by dividing the data into three distinct subsets: training (60%), validation (30%), and test (10%). The data was set to a higher proportion of validation due to developing complex models which requires more tuning to avoid overfitting. The higher validation data helped to better evaluate the model’s performance on the unseen data and make informed adjustments that improved the accuracy of the model. The partition method chosen is stratification, which maintains the distribution of key variables across the subsets, ensuring that each subset represents the overall dataset's characteristics.

Moreover, to handle the missing values in the data, several imputation methods were tested to find the best approach. Some methods that were implemented were the simple imputation like mean and median, and the decision tree imputation. However, these methods did not result in a satisfactory performance of the model. Consequently, the Feature Machine in SAS Viya was employed, which effectively provided a solution for the data challenges and helped to improve model performance. This method automatically imputes missing values with optimal methods and performs feature extraction and selection. This ensures we capture the most predictive aspects of each variable and enhances model accuracy.

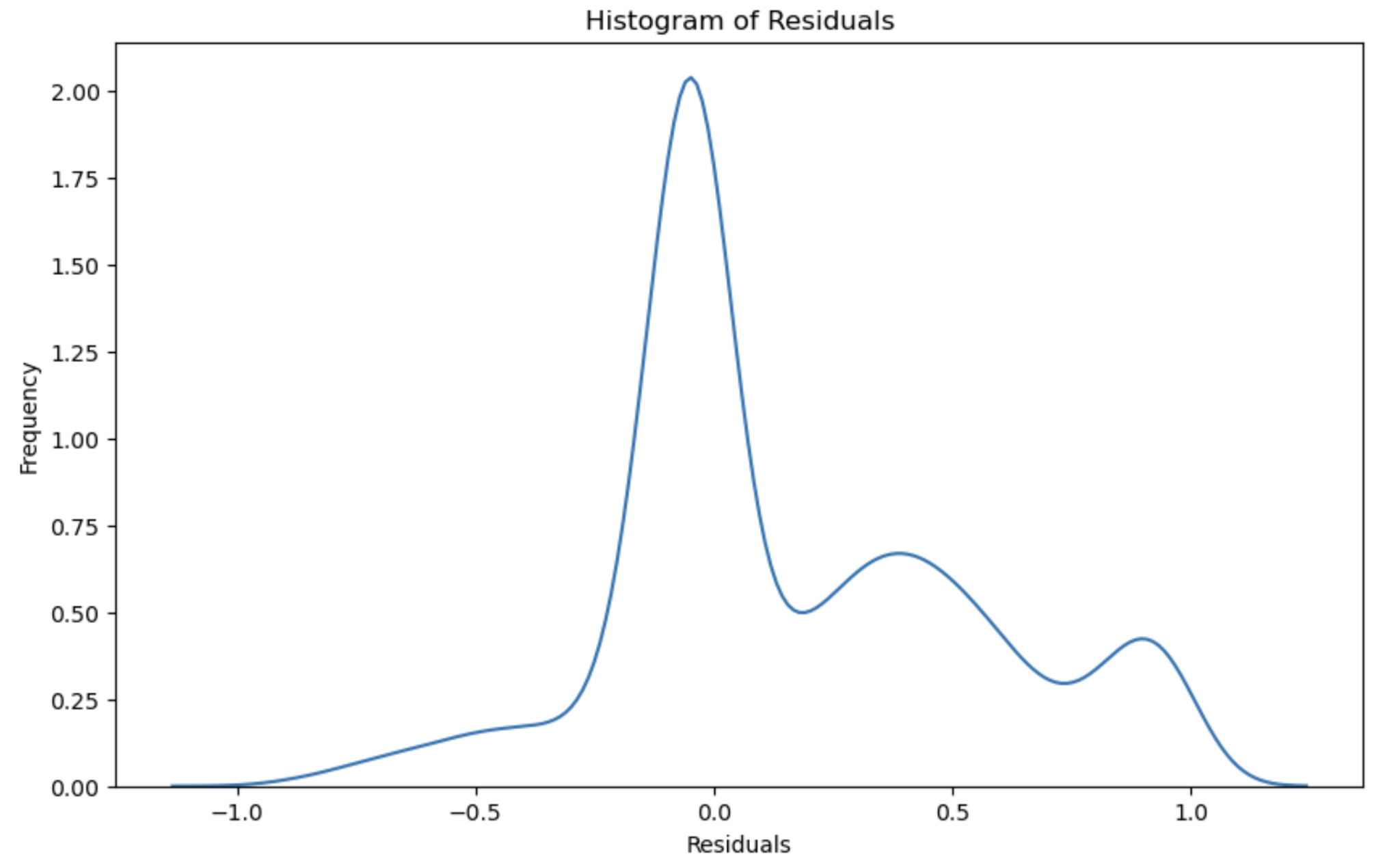
The Feature Machine generates new variables that address the identified transformation policy. The transformation policies set for our dataset were cardinality, missingness, outliers, and skewness. Addressing cardinality reduces dimensionality by consolidating high-cardinality categories, which simplifies the model without sacrificing critical information. Handling missing values ensures the dataset’s integrity, preventing biases that missing data could introduce. Outlier detection safeguards the model from being disproportionately influenced by extreme values, while skewness adjustments help normalize distributions, promoting model stability and interpretability.

In Input Variable Screening, settings like the coefficient of variation and grouping rare levels filter out variables with low variability and consolidate sparse categories, respectively, to improve both model robustness and computational efficiency. A leakage threshold of 90% minimizes the risk of information leaks from variables that could inadvertently reveal insights about the target variable, while the mutual information and redundancy thresholds help eliminate irrelevant and highly correlated features, ensuring that only the most informative variables are retained. Enabling feature selection and limiting to one feature per input streamlines the dataset by retaining the most impactful features, improving model interpretability, and reducing complexity. As a result, it identified several significant variables and transformations that improve the model’s predictive power.

## Model Development

Prior to model development, variable selection was conducted using the fast supervised selection method in SAS Viya where BIC was the stop criterion. The cumulative variance cut-off was set to 1, where the process of selection continuously proceeds until it takes into account all the variance in the data. The variable selection was applied only for Logistic Regression, Neural Networks, and SVM, but not for Random Forest due to its robustness toward irrelevant features.

For the model development, the model was built using the pipeline in SAS Viya. Four models were utilised based on the literature review. The first model was Logistic regression. This model was chosen as it was the most popular model to predict the loan default. The model was easy to implement and the relationships of the predictors and the target can be directly seen from the coefficients of the model.



*Figure 3: Histogram of the error distribution*

For the parameters of the model, the target binary link function that was chosen was Logit. These were the best parameters for the data since it was found that the error distribution was not normal (Figure 3). The selection method used was backward since it focused on the important predictors that gave the most relevance to the loan prediction. The second model was the Random Forest. Random Forest was known for its ability to model non-linear data and capture complex interactions between features, making this algorithm suitable for the data. After various trials and testing, the following parameters were chosen which resulted in the best model performance: maximum tree depth = 20, minimum leaf size = 5, class target criterion = information gain ratio, and maximum number of branches per split = 2. The third model was the Neural Network. Similar to the Random Forest, this algorithm is also a versatile tool that is able to capture the complex relationships and long-term dependencies through their hidden states. The parameters that obtained the best model performance were as follows: number of hidden layers = 1, activation function = tanh, and number of neurons per layer = 50. Lastly, we have the Support Vector Machine (SVM). SVM is based on a strong mathematical foundation, making the accuracy of this model reliable and the model to perform well. The model is also suitable for non-linear and high dimensional data which make it a suitable model for the loan default prediction. The parameters of the model were: kernel = radial basis function, penalty = 0.1, tolerance = 0.0001, and number of support vectors = 3,500.

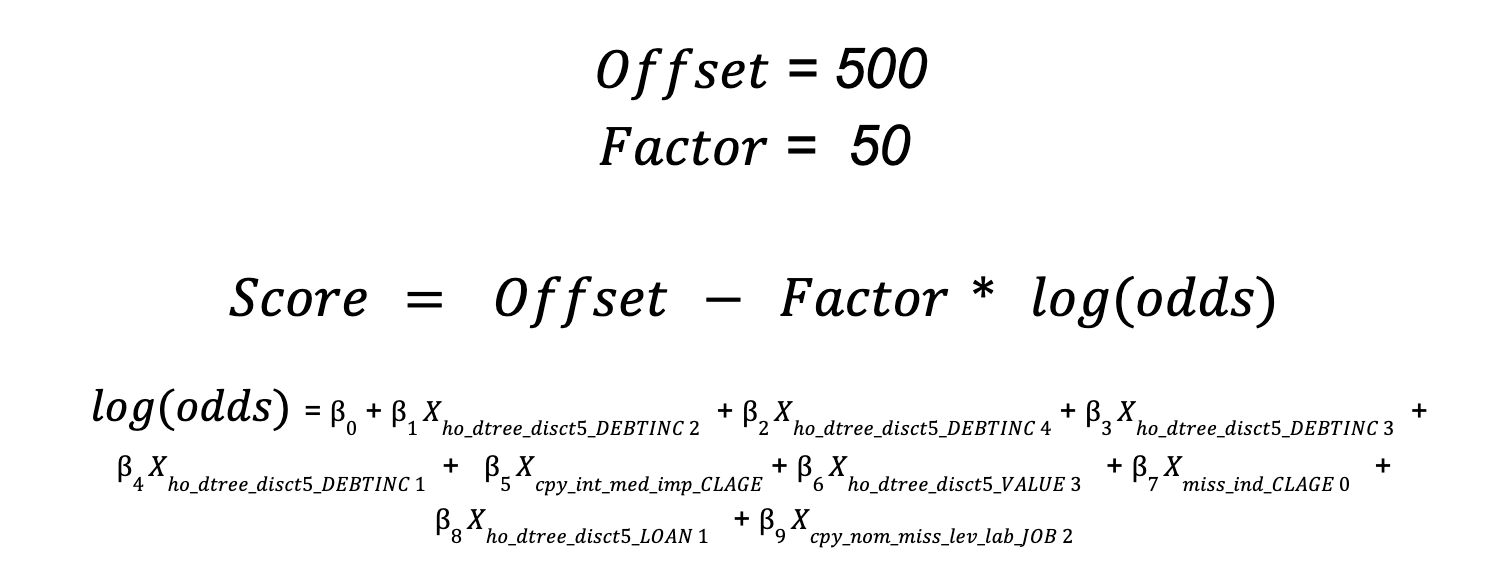
## Model Evaluation

To evaluate the models built in this project, the following evaluation metrics were used: accuracy, F1-score, and AUC score based on the previous literature since these metrics are often used for binary classification problems and are best suited for this dataset. Accuracy is the proportion of instances that a classification model predicts correctly. The F1-score is the harmonic mean of precision and recall. The area under the ROC curve (AUC) is obtained by plotting the sensitivity against the false positive rate (1 - specificity).

## Segmentation Scorecard

This scorecard enables the bank to classify borrowers into various risk categories based on the estimated probabilities delivered from selected variables in the model. For scorecard development, these probabilities are then assigned to a score using the G/B odds, and the study scales the points such that a total credit score of 500 points corresponds to G/B odds of 1 to 1, and that an increase in the credit score of 50 points corresponds to a doubling of G/B odds.

We will use the following equation to compute the scores in this case.



is the regression coefficient for each variable, is the intercept term from Logistic Regression.

The computed scores from the scoring model are linked to the predicted probability of good credit, but not exactly. The formula standardizes this relationship between variables from source data, allowing direct comparison of scores across different segments.

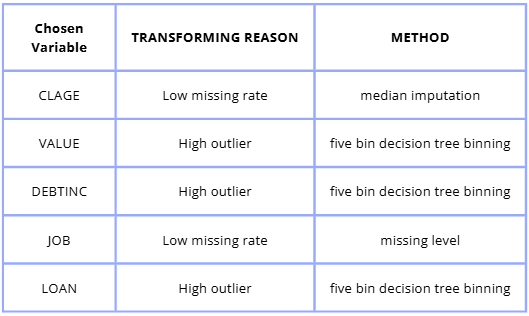
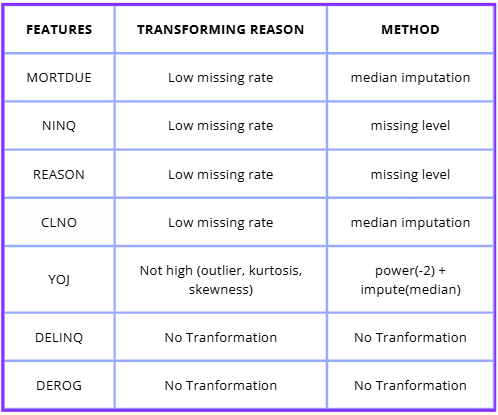
For safety risk management purposes, we categorized the profiles of mortgages into low-risk, medium-risk, and high-risk by using the IQR measure. Using bad rates to assess the probabilities of default, especially the BAD outcomes based on certain score ranges.

While a higher bad rate suggests that more bad accounts in that specific score range mean high risk for these profiles, a lower bad rate indicates a lower likelihood of default which means lower risk.

# Results

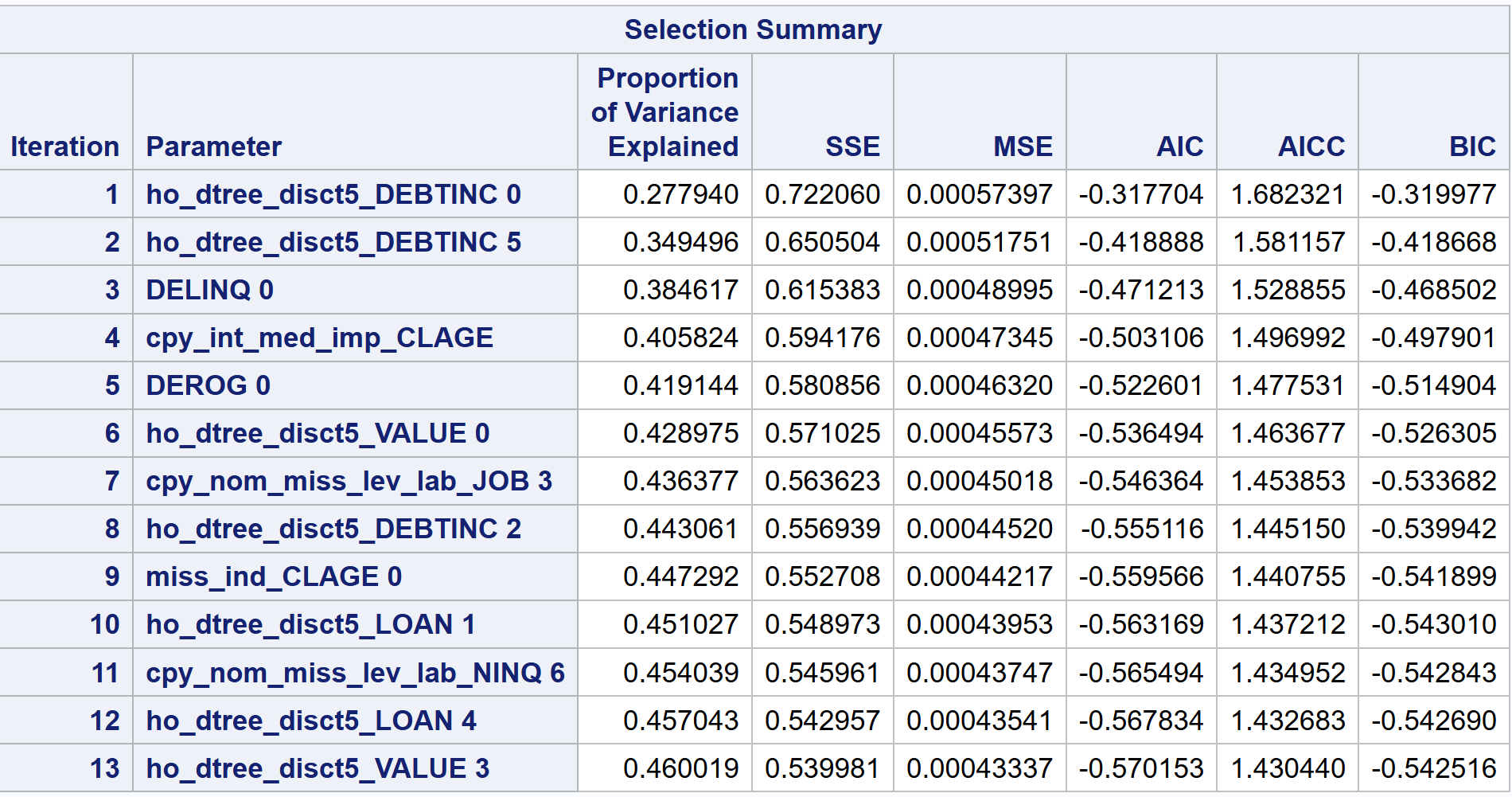
## Feature Machine

The Feature Machine generated a total of 41 new variables, with the best variables selected based on the ranking criteria it used. The reason for each transformation, along with the specific transformation method applied to each variable, is detailed in Table 2.



*Table 2: The variables transforming reason and method in the Feature Machine node*

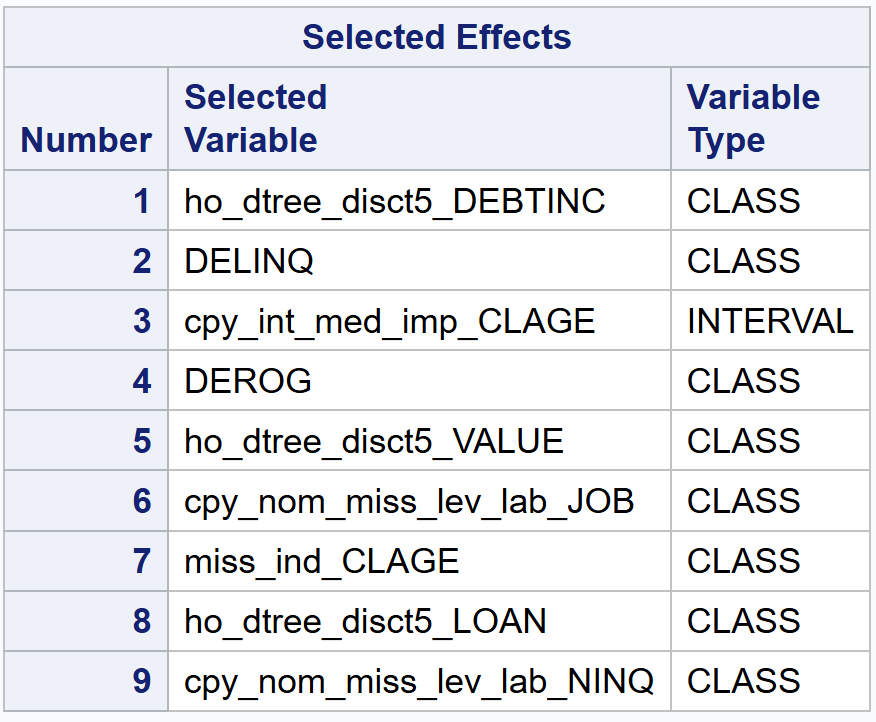
## Variable Selection



*Table 3: The selection summary of the Variable Selection node*

Table 3 presents the variables that were selected from the variable selection. These variables were the specific variables that were derived from the feature machine results.

For instance, ho\_dtree\_disct5\_DEBTINC 5 represents specific classes taken from the ho\_dtree\_disct5\_DEBTINC with class 5. Furthermore, the “proportion of variance explained” displays the proportion of total variance explained by each variable chosen, ordered in ascending order. The metrics, SSE, MSE, AIC, AICC, and BIC, in the table further help in evaluating the variables in terms of how well they account for the variance in the data. Table 3 was the expanded version of the variables in Table 2.

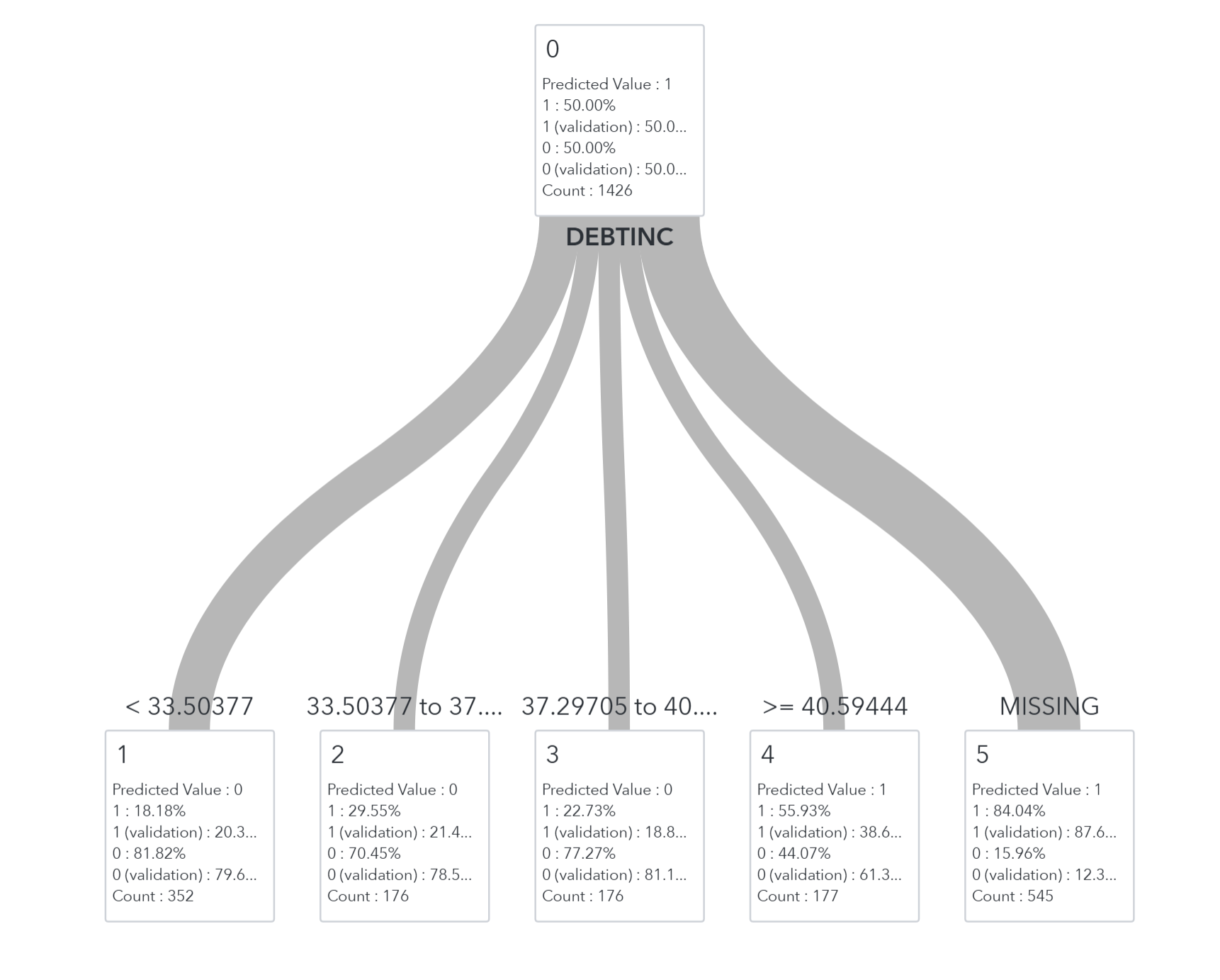


*Table 4: The selected variables from the Variable Selection node*

According to the variables that were shown in Table 4, Table 4 summarised the overall features (without their specific class) that were chosen from the variable selection process. These variables were used in the model development.

## Decision Tree for Specific Variables

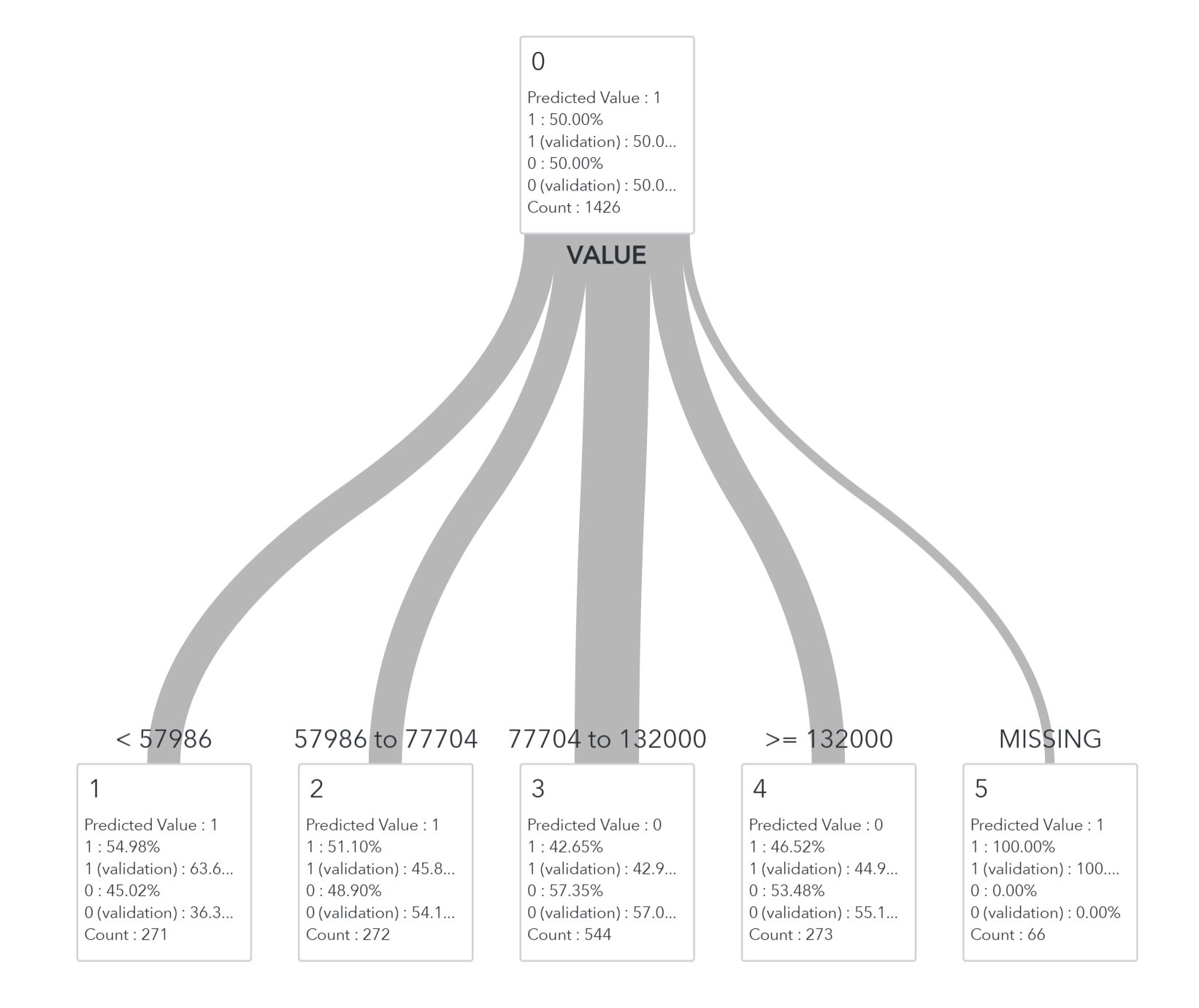
From the Feature Machine results, there were a few variables where decision trees were utilised to transform the variables. Specifically from the variable selection, the variables were ho\_dtree\_dsict5\_DEBTINC, ho\_dtree\_disct5\_LOAN, and ho\_dtree\_dsict5\_VALUE, where the decision tree split these variables into 5 classes. Hence, using the decision tree algorithm these variables’ transformations were replicated to further analyse how they were classified into specific classes.



*Figure 4: Decision Tree for DEBTINC*

The DEBTINC was split into 5 levels (Figure 4) by the decision tree as following, where each class represented a different range of DEBTINC:

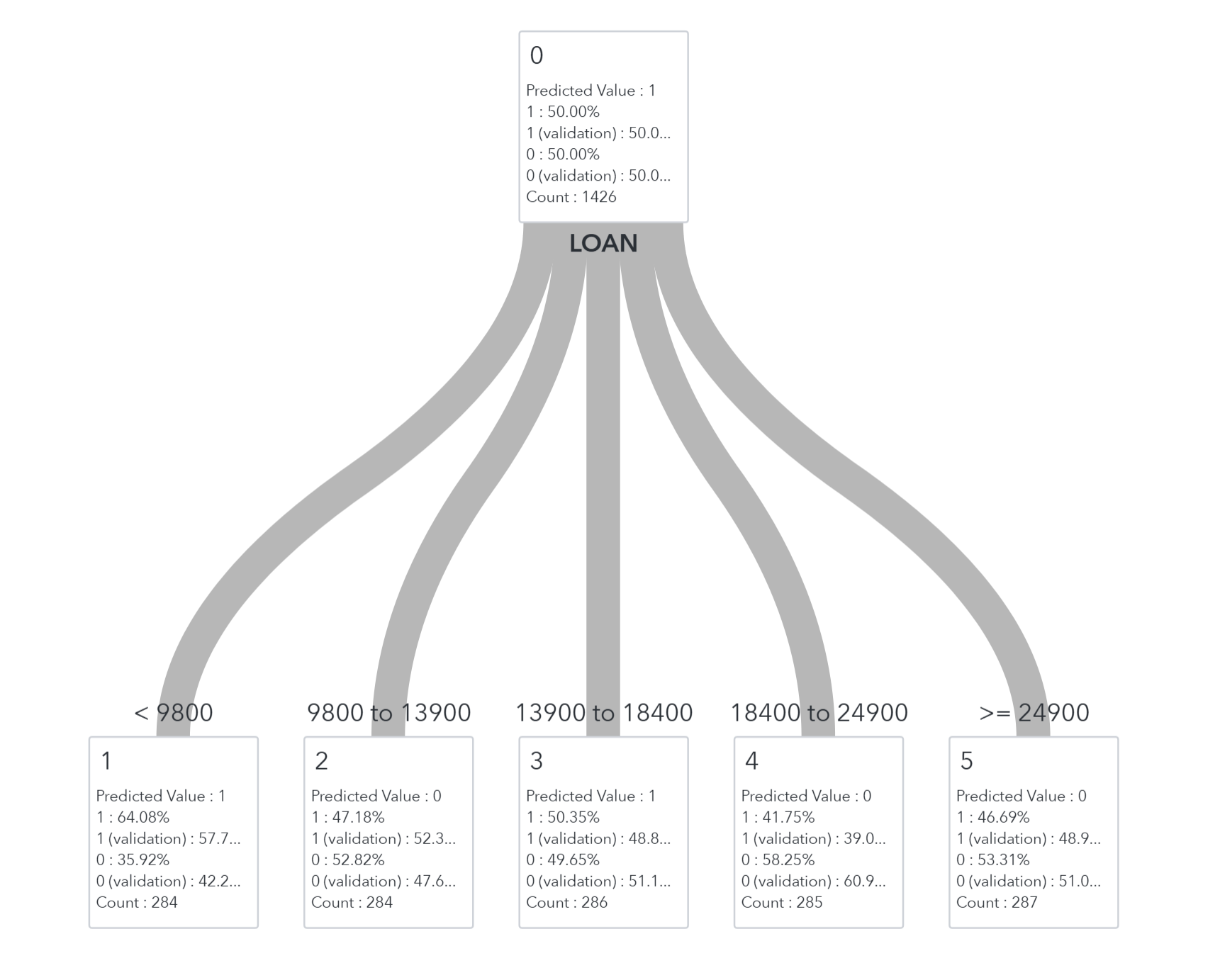
* ho\_dtree\_dsict5\_DEBTINC 1: DEBTINC ratio < 33.504
* ho\_dtree\_dsict5\_DEBTINC 2: DEBTINC ratio from 33.504 to 37.297
* ho\_dtree\_dsict5\_DEBTINC 3: DEBTINC ratio from 37.297 to 40.594
* ho\_dtree\_dsict5\_DEBTINC 4: DEBTINC ratio >=40.594
* ho\_dtree\_dsict5\_DEBTINC 5: DEBTINC ratio MISSING



*Figure 5: Decision Tree for VALUE*

The VALUE was splitted into 5 levels (Figure 5) by the decision tree as following, where each class represented different range of VALUE:

* ho\_dtree\_dsict5\_VALUE 1: VALUE is < 57986
* ho\_dtree\_dsict5\_VALUE 2: VALUE from 57986 to 77704
* ho\_dtree\_dsict5\_VALUE 3: VALUE from 77704 to 13200
* ho\_dtree\_dsict5\_VALUE 4: VALUE is >=13200
* ho\_dtree\_dsict5\_VALUE 5: VALUE is MISSING



*Figure 6: Decision Tree for LOAN*

The LOAN was splitted into 5 levels (Figure 6) by the decision tree as following,where each class represented different range of LOAN:

* ho\_dtree\_dsict5\_LOAN 1: LOAN is < 9800
* ho\_dtree\_dsict5\_LOAN 2: LOAN is from 9800 to 13900
* ho\_dtree\_dsict5\_LOAN 3: LOAN is from 13900 to 18400
* ho\_dtree\_dsict5\_LOAN 4: LOAN is from 18400 to 24900
* ho\_dtree\_dsict5\_LOAN 5: LOAN is >= 24900

## 

## 

## 

## Model Evaluation

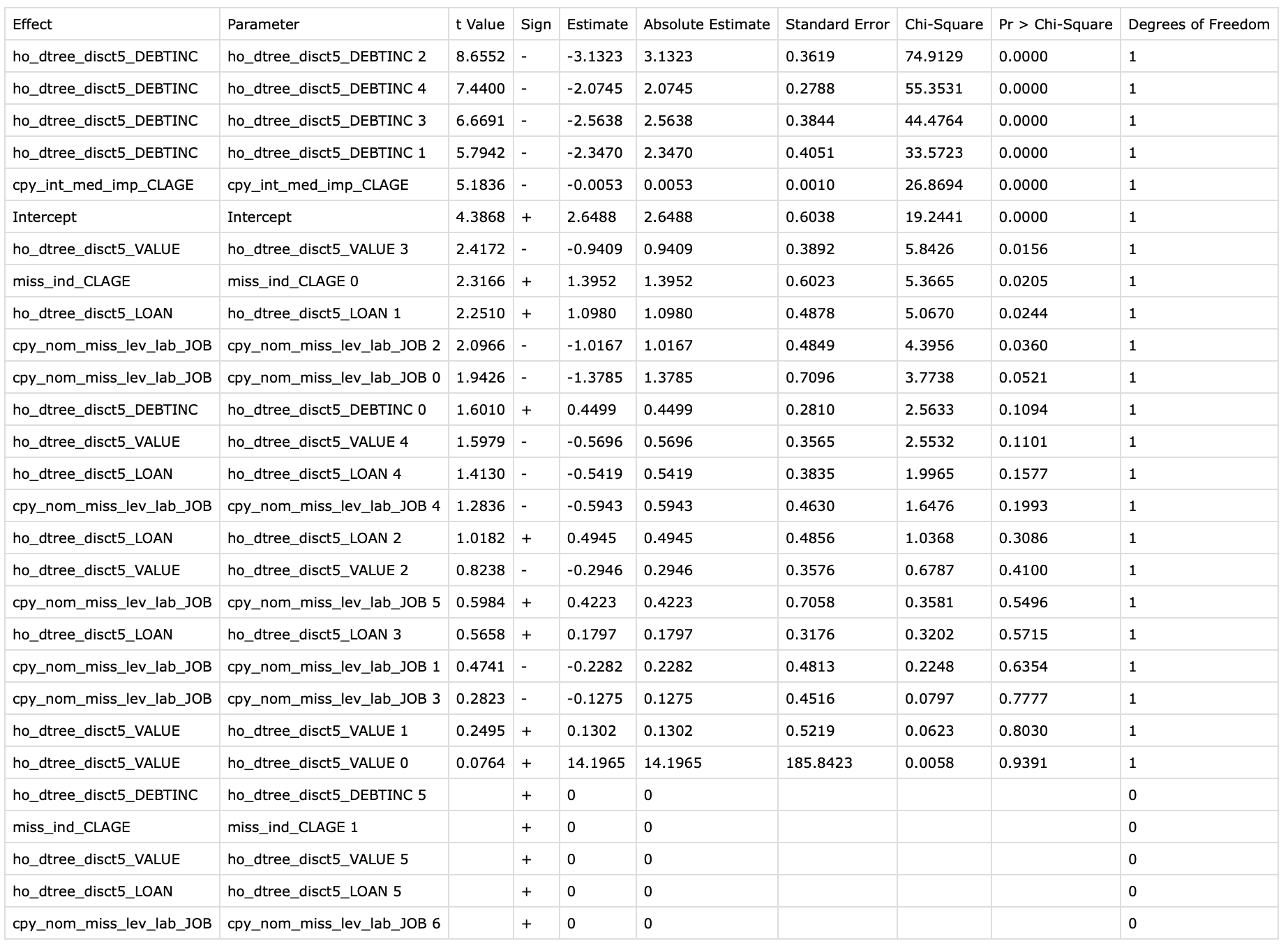
|  | Accuracy | F1-score | AUC score |
| --- | --- | --- | --- |
| Logistic Regression | 0.7731 | 0.7128 | 0.8973 |
| Random Forest | 0.7563 | 0.6848 | 0.9400 |
| Neural Network | 0.7311 | 0.6364 | 0.8987 |
| SVM | 0.6387 | 0.4416 | 0.9038 |

*Table 5: Evaluation metrics of the four model algorithms*

From the four models that have been built, the model performance was evaluated based on three metrics, including accuracy, F1-score, and AUC score. Table 2 shows the performance metrics of each model, where the rows depict the different model algorithms used and the columns indicate the metrics used to evaluate the model. The results were sorted in descending order based on the model with the highest accuracy.

## Segmentation Scorecard

From the estimates of variables computed by Logistic Regression as Figure 3 shows, as p-values are less than 0.05, there are 9 parameters significantly influencing the total credit scores, in contrast, 13 parameters are not significant.



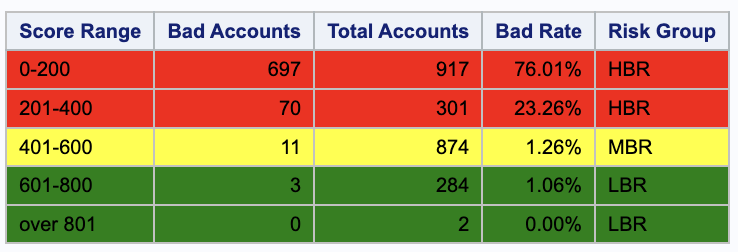
*Table 6: The estimates of parameters from Logistic Regression*

We construct the scoring model using the significant parameters, t[able](https://www.sciencedirect.com/science/article/pii/S0957417411012516#t0025) 3 indicates the 3 significant variables we used which are delinquent and non-delinquent credit related such as DEBTINC (debt to income ratio), VALUE(Value of the current property), LOAN(Amount of the loan request), the attributes, G/B odds, and attribute points for each variable.

| Variables | Attributions | G/B odds | Attributed Points |
| --- | --- | --- | --- |
| DEBTINC | < 33.5 | 1.3523 | 484.910 |
| 33.5 to 37.3 | 0.0269 | 680.790 |
| 37.3 to 40.6 | 0.0065 | 752.130 |
| >=40.6 | 0.0035 | 782.460 |
| VALUE | 77704 to 13200 | 95.2781 | 272.160 |
| LOAN | < 9800 | 0.8404 | 508.695 |

*Table 7: The Results of Scoring Model*

Using the scoring model to segment the mortgage, we classify 2378 accounts which are 50% of overall accounts as the event-based sampling applied in the analysis. Table 4 presents the 5 score ranges with the risk group. To better segment customer risk, these accounts are grouped into three risk categories based on their bad rates. The accounts with a bad rate above 23.26% are classified as HBR means high-risk which is below 400 points, between 1.26% and 23.26% MBR means medium risk which is between 401 and 600 points, and below 1.06% as LBR means low-risk which are above 601 points. The high-risk (red) group has an average bad rate of 49.64%, the medium-risk (yellow) group 1.26%, and the low-risk (green) group 0.53%. These clear risk distinctions are used in credit strategy applications based on the Logistic Regression-based scoring model.



*Table 8: The Segment results based on the Bad Rate*

# Discussion

To ensure the data is in a suitable format for analysis and model building, data exploration was carried out. The binary target variable was found to have a disproportionate number of observations where it has 4771 observations with the value 0 and 1189 observations with the value 1. Hence, the dataset was imbalanced with an 80-20 distribution. To address the imbalanced dataset, event-based sampling was performed so that the data contained 50% of the value 0 and 50% of the value 1.

Missing values were detected from all the independent variables within this dataset. To address these missing values, various imputation methods such as mean and median imputation methods were attempted based on the literature review. Furthermore, the feature machine method in SAS Viya was later employed to impute the missing values and transform each of the independent variables according to the traits of each of the variables. When comparing the results of these imputation methods (feature machine, mean, and median imputation), the feature machine method produced a much higher performance for all models compared to mean and median imputation. This output is not surprising considering that the feature machine deals with issues such as low or high missing rates, outliers, skewness, and kurtosis for each variable separately.

Variable selection was then carried out based on the proportion of variance explained by each of these variables to ensure only relevant and important variables were included in the model development process. As a result, these selected variables as ho\_dtree\_disct5\_DEBTINC, DELINQ, DEROG, miss\_ind\_CLAGE, cpy\_int\_med\_imp\_CLAGE, ho\_dtree\_disct5\_VALUE, cpy\_nom\_miss\_lev\_lab\_JOB, ho\_dtree\_disct5\_LOAN, and cpy\_nom\_miss\_lev\_lab\_NINQ were included in the model development process. The Decision tree was also employed for these variables: ho\_dtree\_dsict5\_DEBTIN, ho\_dtree\_disct5\_LOAN, and ho\_dtree\_dsict5\_VALUE to divide these variables into 5 separate classes so that these variables are in a suitable format to be used for model development.

From the literature review, these models: Logistic Regression, Random Forest, Neural Network, and Support Vector Machine were identified and used for the model development. Parameter tuning was performed to ensure optimal results were gained for each model while ensuring the models were not overfitted through trial and error. The final model pipeline was then built using the parameters set as shown in the Model Development subsection in the Methods section.

The literature review also showed that various types of evaluation metrics were used to quantify the quality of the prediction made by the models built. Most of the articles utilized accuracy, F1 score, and AUC score as evaluation metrics. For model evaluation purposes, these three evaluation metrics were used to compare the models built. From these evaluation metrics, accuracy was chosen as the main evaluation metric to choose the best model for the scorecard building. This is because accuracy is one of the most commonly used evaluation metrics and the dataset used does not have any class imbalance issue for the target variable. In this case, the logistic regression model had been chosen as the best model since it had the highest accuracy (0.7731). Among all the models built, this model also had the highest F1 score (0.7128) and an adequately high AUC score (0.8973).

The final scorecard was then built based on the G/B Odds metric and logistic regression model results. The log-odds part of the score function is the linear combination of the significant variables found in the logistic regression model. Based on previous literature review and experience, the offset value was set as 500 since it often sets the minimum score while the factor value was set as 50 so that the score increases by 50 points with each doubling of the odds allowing the scores to be adjusted relative to the risk. In addition, the resulting scores were then divided into three risk level categories: low, middle, and high using the interquartile range (IQR). By calculating the IQR, which is the range between the 25th percentile (Q1) and 75th percentile (Q3) of the bad rate, we can measure the variability and dispersion in the bad rate across different customer segments. Hence, we used the bad rate distribution revealed by the IQR, the scorecard can be fine-tuned to assign lower bad rates to low-risk segments (with under Q1 bad rates), moderate bad rates to medium-risk segments (with between Q1 and Q3 bad rates), and higher bad rates to high-risk segments (with above Q3 bad rates). As a result, the accounts were categorized into three risk groups based on the bad rate: high-risk (score below 400, bad rate above 23.26%), medium-risk (score between 401 and 600, bad rate between 1.26% and 23.26%) and low-risk (score above 601, bad rate below 1.06%). This provides a clear distinction between the three groups which allows banks to make efficient decisions during the loan approval process in the Australian housing market.

# Conclusion

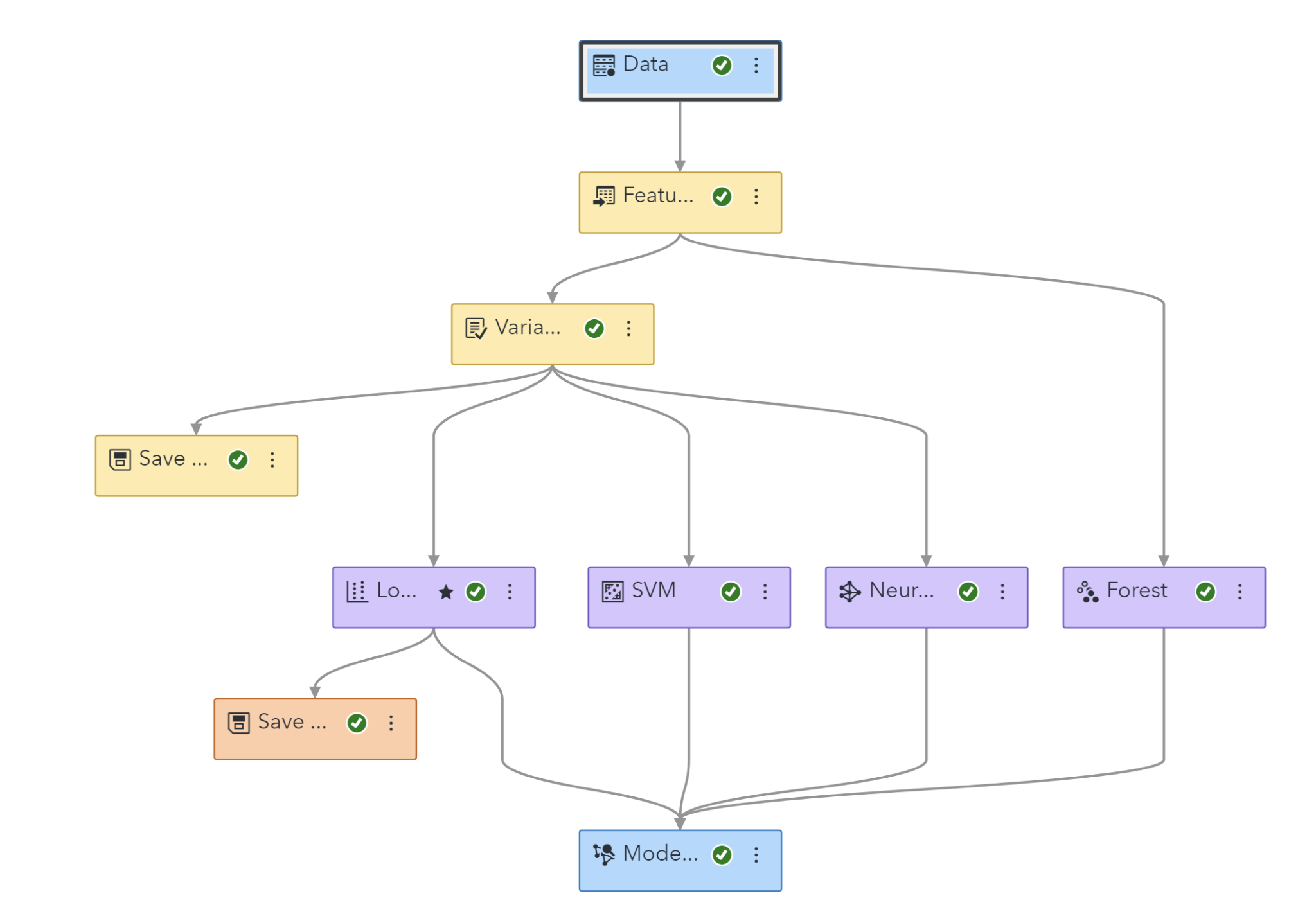
This project utilizes the significant coefficient estimates from the best model (logistic regression) to build the final scorecard which segments the accounts into three distinct risk level groups based on the scores calculated from the G/B Odds metric. By dividing the accounts into three distinct risk-level categories, it allows banks and financial institutions to plan and tailor their decision-making and risk-management strategies for these separate groups.

One of the limitations of our research is that after implementing the Feature Machine method, some variables that were classified into few classes became difficult to interpret as we were unable to trace the exact transformation applied. For instance, the JOB variable was transformed using the missing level method into cpy\_nom\_miss\_lev\_lab\_JOB, which was grouped into several classes. However, since the details of this method are unclear, we are unable to determine the specific jobs that were classified into cpy\_nom\_miss\_lev\_lab\_JOB 2. Due to this reason, there may be some limitations in the interpretability of our findings.

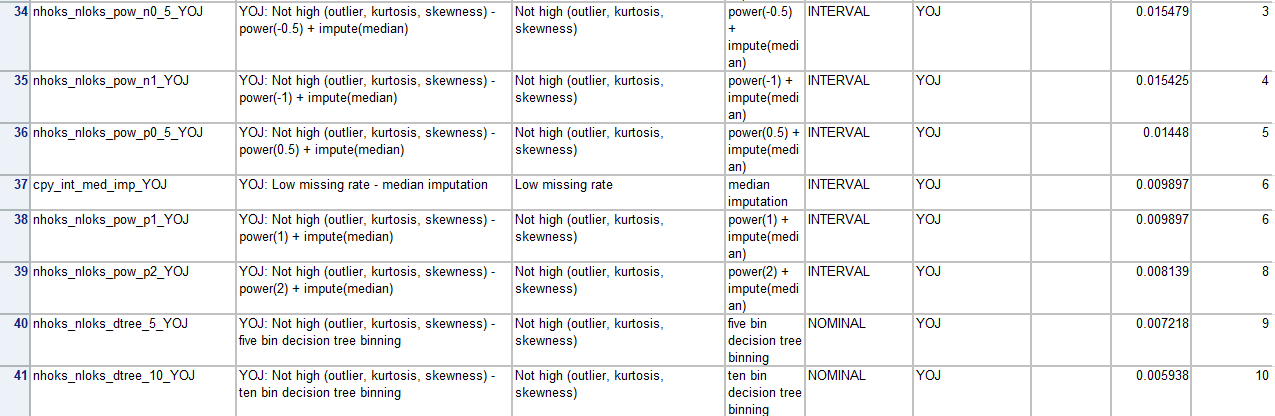
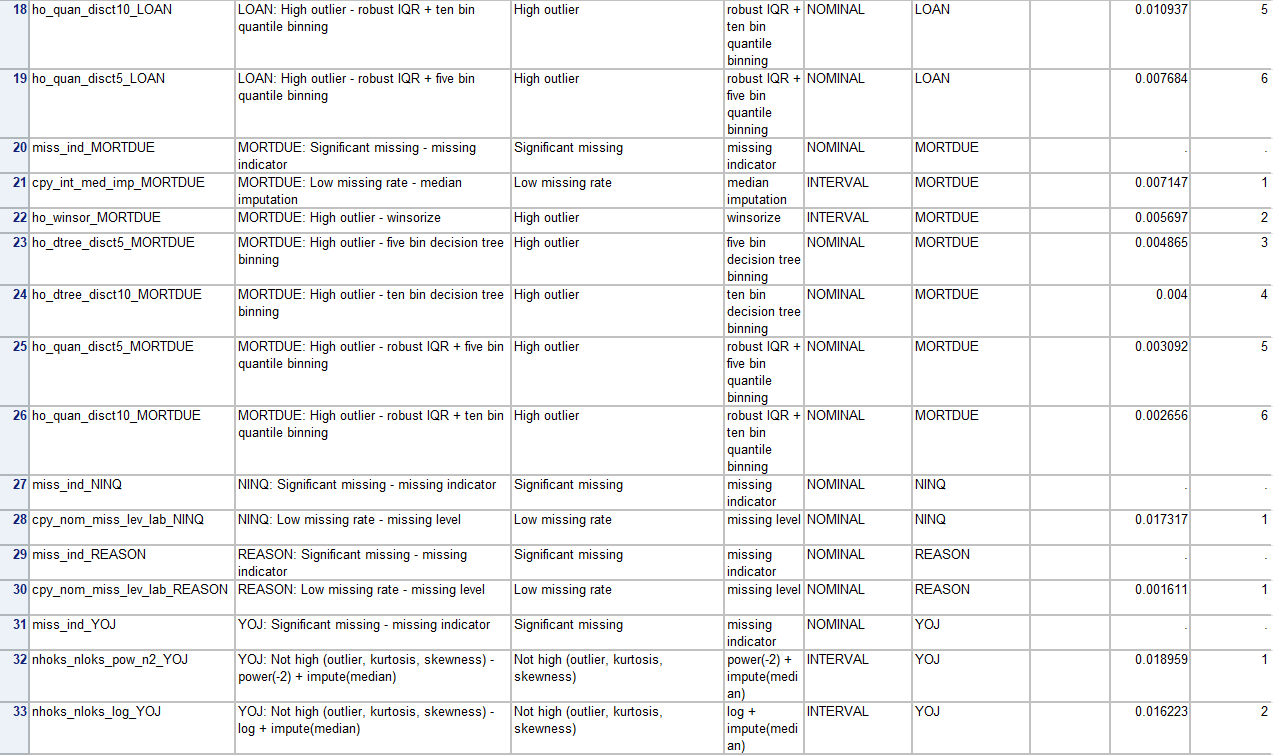
Future research could perform feature engineering or consider the addition of new variables, for instance, the borrower characteristics variables as education, and gender, the Collateral characteristics variables as Collateral area and Original loan-to-value ratio (Bo-Wen Chi and Chiun-Chieh Hsu, 2011). Feature engineering could enhance the model performance and efficiency by building a more robust model that includes only crucial client information (Shoaeinaeini, Shoaeinaeini, Harrison, and Jasemi, 2024). Although not included in this report, previous literature has also shown that various types of models had been utilized for loan prediction other than the ones utilized for this project. Therefore, future research could explore other models used for loan prediction available in SAS Viya as well and compare these models’ performances with the models’ performances in this project. As the home loan climate in Australia and the Australian mortgage market continue to evolve, it is important to update and tune the risk segmentation model according to these changes so that banks and financial institutions have a competitive edge against their competitors.

# Appendix

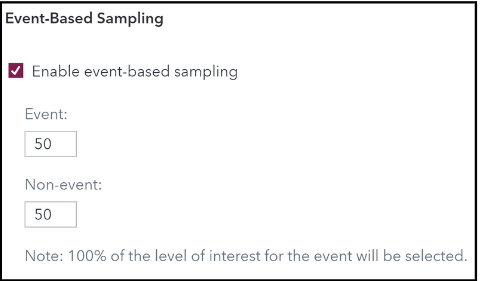
The final model Pipeline built-in SAS Viya



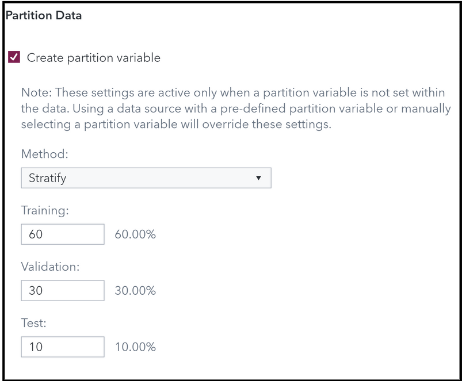
Full list of variables by Feature Machine



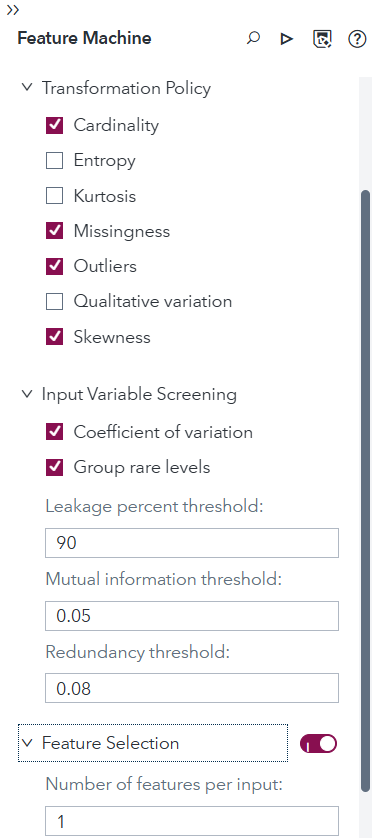
Event-Based Sampling setting



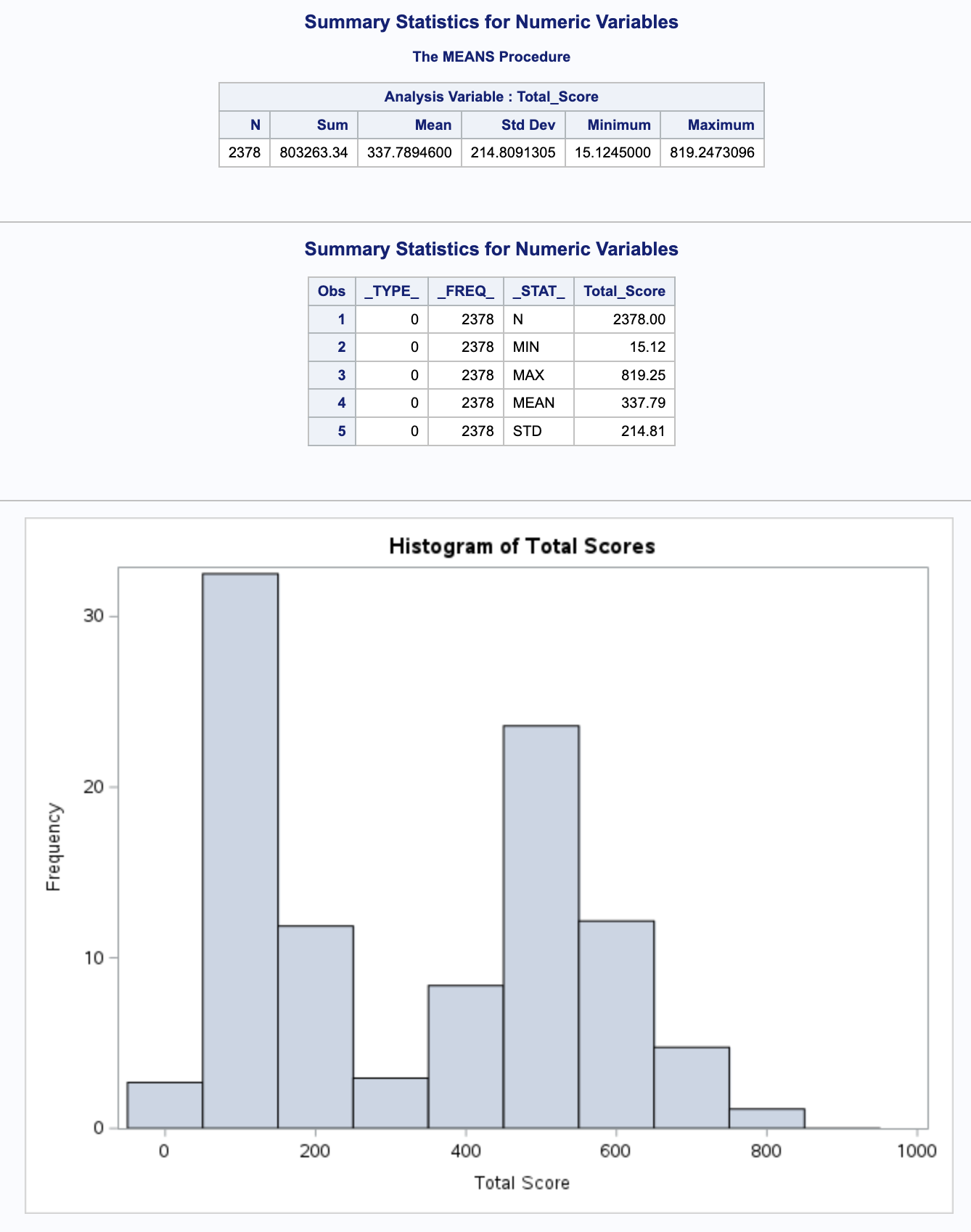
Data Partitioning setting

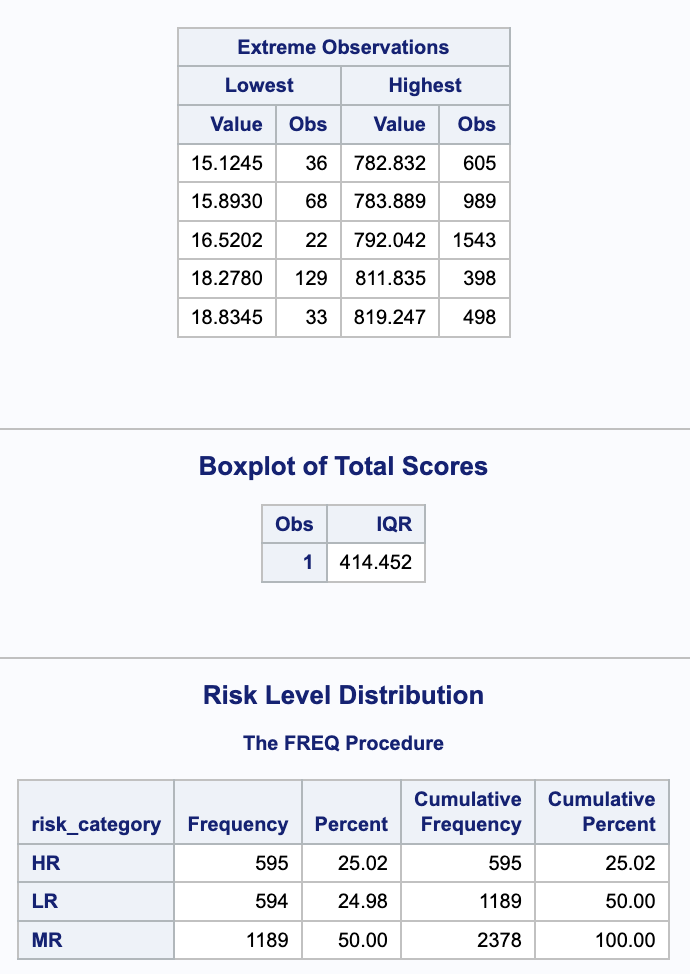
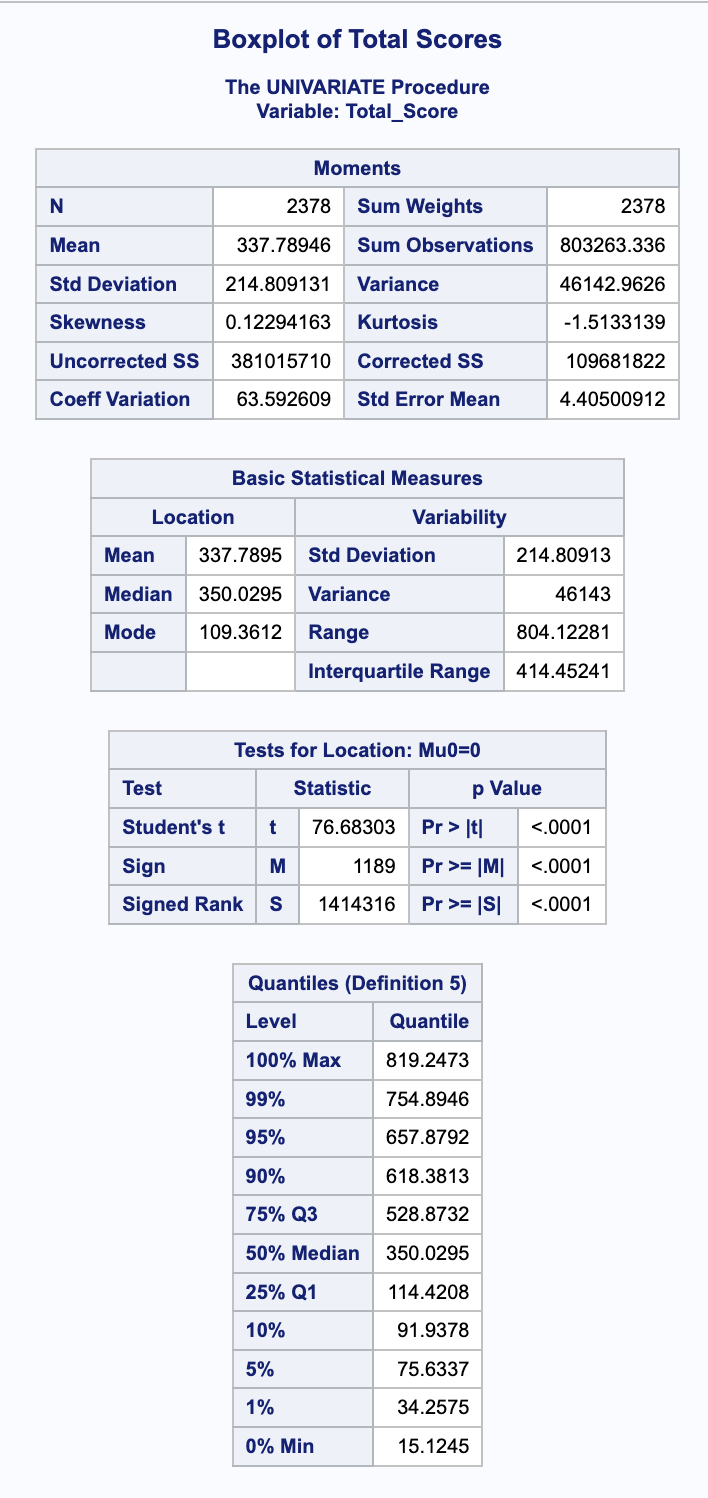


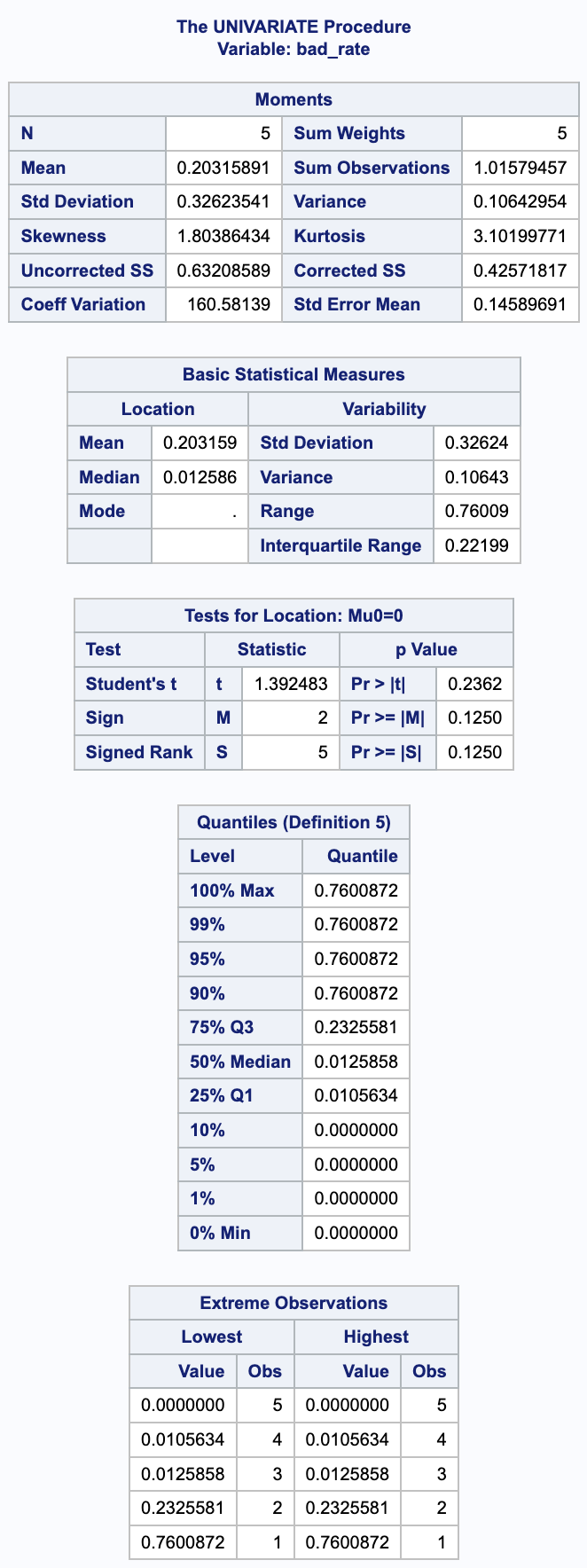
Feature Machine setting



SAS Codes Outputs for Scorecard Development







# References

Alam, S., Ayub, M.S., Arora S., & Khan, M. A. (2016). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, 9, 100341. https://doi.org/10.1016/j.dajour.2023.100341.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018.

Dinh, T. N. & Thanh, B. P. (2022). Loan Repayment Prediction Using Logistic Regression Ensemble Learning With Machine Learning Algorithms. *2022 9th International Conference on Soft Computing & Machine Intelligence*, 79-85.

Finlay, S. (2012). *Credit Scoring, Response Modeling, and Insurance Rating.* Palgrave Macmillan UK eBooks.

Hosmer, D., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression*. New John Wiley & Sons.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Nagajyothi, V. (2020). Loan approval prediction using KNN, decision Tree and Naive Bayes models. *International Journal of Engineering in Computer Science*, 2(1), 32-37. <https://doi.org/10.33545/26633582.2020.v2.i1a.30>

Rainio, O., Teuho, J., & Klén R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Reddy, M. V. J., & Kavitha, B. (2010). Neural networks for prediction of loan default using attribute relevance analysis. *2010 International Conference on Signal Acquisition and Processing*, 274-277. 10.1109/ICSAP.2010.10.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>.

Sessa, J. & Syed, D. (2016). Techniques to deal with missing data. *2016 5th international conference on electronic devices, systems and applications (ICEDSA)*, 1-4. https://doi.org/10.1109/ICEDSA.2016.7818486.

Sinap, V. (2024). A Comparative Study of Loan Approval Prediction Using Machine Learning Methods. *Journal of Science*, 12(2), 644-663. 10.29109/gujsc.1455978

Vimala, S., & Sharmili, K. C. (2018). Prediction of Loan Risk using NB and Support Vector Machine. *In International Conference on Advancements in Computing Technologies*, 4(2), 110-113.