BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Bank Negara Malaysia Working Papers
WP11/2017

Logistic Regression Models for Malaysian Housing Loan Default Prediction

By Foo Lee Kien, Chua Sook Ling, Daniel Chin and Muhamad Kamal Firdaus

October 2017

Logistic Regression Models for Malaysian Housing Loan Default Prediction

Foo Lee Kien
Multimedia University

Chua Sook Ling
Multimedia University

Daniel Chin
Bank Negara Malaysia

Muhamad Kamal Firdaus
Bank Negara Malaysia

Abstract

The aim of this paper is to develop a credit scoring model of housing loan defaults in Malaysia. We applied the logistic regression model on data obtained from the Malaysian Central Credit Reference Information Systems (CCRIS). The data used for model fitting is based on behavioural scoring, which assesses the risk of existing borrowers based on their recent loan servicing patterns. Variations of logistic regression models were developed to evaluate the model's predictive accuracy in credit scoring. The performance of the models was also evaluated using an in-time out-of-universe dataset.

JEL Classification: G21, D86, D832

Keywords: credit score, logistic models, loan defaults

1 Introduction

The 2008 US and European Financial Crisis highlighted the need to include a more macro perspective of housing loan performance in the arsenal of prudential tools available to regulatory bodies to monitor the stability of the financial system. Although Malaysia emerged from that crisis relatively unscathed, regulators remain vigilant especially given the rising level of household debt. The household debt stood at 88% of GDP in 2016, of which nearly half comprise housing loans to individual borrowers. The need for better early warning tools to monitor the default risks of the housing loan system has never been more pressing. While individual banks may employ credit assessment models to monitor their portfolios, no system-wide tool exists for Malaysian regulators to easily monitor the health of the entire population of loans.

There are two broad fields of research typically described in literature for developing credit scoring models. The first can broadly be termed as market-based models, which applies option pricing theory to default prediction. This is the basis of the famous Merton or Black Scholes Merton (BSM) models. There exists considerable research in this area. For instance, Order (2007) describes the underlying notions of this approach: default on a housing loan can be viewed as exercising a put option, and that the place to look in modelling default is the extent to which the option is in the money (the extent to which the borrower has negative equity in the property). In other words, the borrower is incentivised to default if the house value is less than the current outstanding loan. This approach applies to some degree in countries where the housing loans are non-recourse and the borrower can just "hand over the keys to the lender and walk away". Some US states for example, have non-recourse mortgages so option based models works reasonably well there. However, there is no such restriction in Malaysia, and the lender typically has recourse to the borrower for any shortfall after liquidation of the collateral. This suggests that market-based models may not be a strong predictor of housing loan defaults in Malaysia.

The second field of research is in operations research and statistics, and is the focus of this paper. This field is partly motivated by the need to develop robust credit scoring models for the purpose of regulatory capital. Various classification techniques have been adopted for developing credit scoring models. These include traditional statistical techniques such as Linear Discriminant Analysis (LDA) (Gurný and Gurný, 2013) and Generalised Linear Model (GLM) logistic regression (Wang, Xu and Zhou, 2015). More recently, non-parametric

statistical models have been explored such as Support Vector Machine (SVM) (Harris, 2015) and Neural Network (NN) (Zhao et al., 2014).

There are also works that attempt to use ensemble approach to improve the performance of classification (Dahiya et al., 2015; Xiao et al., 2016). In these works, multiple learning models (such as SVM, NN, Decision Trees, etc.) are trained to construct a set of classifiers. Classification is then performed by taking the weighted vote of their predictions.

The field of retail credit scoring can be sub-divided into two general types. The first type is application scoring. Such models rely mainly on borrower demographic factors such as age, occupation and salary, since little else is known about the borrower at the point of loan origination. The second type is behavioural scoring, which is typically used for existing borrowers during the course of servicing the loans. These models rely primarily on explanatory variables linked to borrower behaviour patterns. For example, behavioural type variables may include a borrower's delinquency status, credit card utilisation and limit breaches. This paper focuses on behavioural scoring, specifically borrowers who have been servicing their housing loans for at least 12 months.

The main objective of this paper is to develop and compare the default prediction ability of a range of different model classification techniques applied to the system of housing loans in Malaysia. The focus is on model performance, given a set of explanatory variables, rather than on the explanatory variables themselves. Therefore the choice of explanatory variables is not discussed in detail, although an attempt is made to discuss the impact of pre-regression variable selection and transformation.

This paper adds to previous research on housing loan defaults in two main ways. Firstly, it leverages data from the Malaysian Central Credit Reference Information Systems (CCRIS), which is a repository of account-level borrower information compiled from participating financial institutions. To our knowledge, this is the first published attempt to use CCRIS data to develop a system-wide credit scoring model for housing loan borrowers. Second, to-date, very little has been published on the modelling of housing loan defaults in Malaysia. A deeper understanding of the predictive ability of different classification techniques relevant to the Malaysian market may be useful for macro-level monitoring of housing loans.

The remainder of the paper is structured as follows: Section 2 describes the theory behind the classification techniques used in this paper. Section 3 elaborates on the data used for the

analysis, the sample selection procedure, and list of explanatory variables. Section 4 then describes the framework used for developing the various variations of models as well as the validation framework employed to evaluate the effectiveness of each variation. Section 5 provides a discussion of the results, while Section 6 presents the conclusions and suggests for future work.

2 Classification techniques

This section describes various classification techniques that are typically used to develop credit scoring models. Generalised Linear Models are first discussed, of which the logistic regression model used in this paper is a particular case of the GLMs. The logistic regression model is one of the main approaches most commonly used by banking institutions to develop credit scoring models and thus multiple model variations based on the logistic regression model are compared in this paper.

2.1 Generalised Linear Models

GLMs can be viewed as a generalisation of the standard logit models. GLMs and their assumptions are outlined below following similar notation used by Fitzpatrick (2013) and Veneables et al. (2002).

The first assumption for GLMs is that the response variable is observed at independent fixed values of the predictor variables X_i. The predictor variables affect the response through a linear predictor function $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$

Second, the distribution of *y* has a density which follows:

$$f(y_i; \theta_{i;} \Phi_i) = exp\left[A_i \frac{y_i \theta_i - \gamma(\theta_i)}{\emptyset} + \left(y_i, \frac{\Phi}{A_i}\right)\right] \quad (1)$$

In this density φ is known as the scale parameter, A_i is a known prior weight, and θ_i depends on the linear predictor. The density belongs to the exponential family of distributions. The third assumption relates to the link function which specifies the mean value of the response and its relationship to the predictors. The function $l(\mu)$ in equation (2) is known as the link function and its mean value, μ, is a function of the linear predictor.

$$\mu = m(\eta),$$

$$\eta = m^{-1}(\mu) = l(\mu) \quad (2)$$

These set of assumptions provide a general framework for several models where the response follows any distribution from the exponential family (for example, normal, Poisson, binomial etc.) In a standard regression a dependent variable Y is related to the linear combination of predictors $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$

In the application considered here, the dependent variable is a binary outcome variable. Therefore a new function *g* can be used to link the outcome:

$$g\big(E(Y|X)\big) = g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m \quad (3)$$

The relationship between η and the modelled mean μ is now through the link function

$$\mu = p = prob(Y = 1|X)$$

$$g(\mu) = \log\left[\frac{\mu}{(1-\mu)}\right] = Logit(\mu) \quad (4)$$

Here, μ is the mean of the dependent variable. *g()* is the monotonic differentiable link function. This general formulation, a GLM, nests the logistic regression model.

$$Logit(p) = \log\left[\frac{p}{(1-p)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m \quad (5)$$

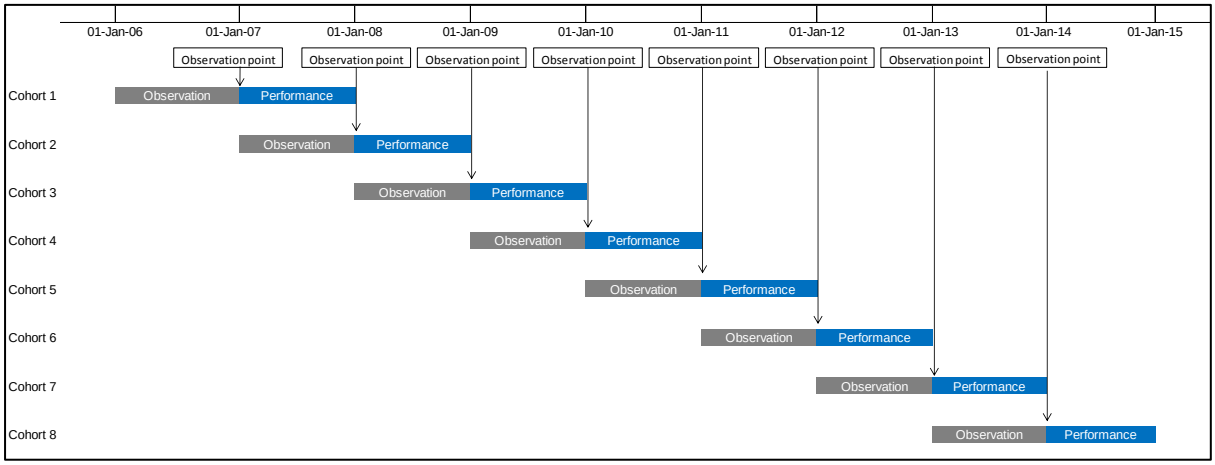
3 Dataset

The data used in this paper was extracted from the Malaysian Central Credit Reference Information System (CCRIS). CCRIS contains account level credit performance data such as outstanding balance and arrears information for each housing loan originated by all participating financial institutions in the country and is updated on a monthly basis. This includes all banking institutions and most non-banking financial institutions.

The credit scoring models in this paper are intended to predict the probability that a borrower will default on its housing loan within a 12 month time horizon. This means that the data needed to build such models must reflect information and knowledge that a lender has one year before the default event. Data from 2007 to 2014 is used. For each year, a cohort is specified, defined as all active borrowers who were not in default as of 1 January of that cohort year, and have been servicing their housing loan for at least 12 months. The rationale for only including borrowers with 12 months on book is to ensure sufficient data history to compute certain

behavioural factors. For each borrower in a cohort, their performance is tracked over the next 12 months (performance period) to observe whether a default occurs. Multiple defaults within the 12 month period were counted as a single default event. Behavioural factors were calculated based on data taken from the 12 months prior to the observation point. In this manner, eight distinct cohorts were defined, one for each calendar year from 2007 to 2014. This is shown in Figure 1.

Figure 1: Dataset selection



3.1 Default definition

The default definition employed for this analysis is based on the Basel II default definition. A default is considered to have occurred when the loan is past due for more than 90 days.

In our dataset, we classify defaulters based on the Months in Arrears (MIA) flag in CCRIS. An account is classified as default when the MIA exceeds three months. The MIA flag is chosen to best approximate the 90 days past due criterion, since the more granular days past due data is not readily available in CCRIS. Non-defaulters are classified as those that have not met the MIA condition throughout the performance period of the respective cohorts.

3.2 Sample selection

In theory, it would be possible to develop the credit scoring models without sampling, using instead the entire universe of housing loan borrowers available in CCRIS. However, computing resource constraints meant that using a smaller sample would be significantly faster and more manageable.

Two different samples were selected. One was used to develop the models and another to test the performance of the models after it has been finalised. The latter sample was not used in any way during the development process.

Given the relatively small number of defaults in the population of about 3% to 4% each year, there exists a large degree of class imbalances. The decision was made to under-sample the majority class of non-defaulters in the ratio of 1 defaulter to 3 non-defaulters to reduce the impact of class imbalance in the sample used to develop the models. A sample of 15,000 defaulters was selected randomly from the eight yearly cohorts followed by 45,000 non-defaulters in a similar manner. The random sampling across cohort years was to ensure both defaulted and non-defaulted borrowers were proportionately reflective of the population across each time period.

The second sample comprises another 60,000 housing loan accounts drawn randomly from the same population of eight yearly cohorts. No under-sampling of the majority class was carried out for this test set. As such, this set represents the actual proportion of defaulters to non-defaulters inherent in the population. The table below summarises the size of the population, the samples size of both the model developing and model testing datasets.

Table 1: Datasets used for model development

Dataset	Number of defaults	Number of non-defaults
Population	489,590	12,383,483
Training set	15,000	45,000
Population test set	2,323	57,677

3.3 Explanatory variables

The selection of explanatory variables is not the main focus of this paper. The explanatory variables considered here are a collection of borrower demographic and behavioural indicators that are typically used by banking institutions in the design of their retail credit scoring models. A total of 21 variables were tested. Table 2 shows the list of variables considered in this paper.

4 Methodology

This section describes the methodology used for model development, model validation and the performance metrics used for comparison of model performance.

4.1 Model Development Framework

Before preparing the data for model development, it is necessary to ensure that the source data used for model development as described in Section 4.1, are cleaned. Incomplete data are often unavoidable and prevalent in population-based data, such as those collected about the borrower information in CCRIS. Training a classifier on incomplete data ("garbage in") will generally produce incorrect predictions and misleading decisions ("garbage out"). From the sample data, there are a total of 6303 records with missing values for some variables. Since the sample size is large and there is no obvious pattern found in the missing data of any of the variables, we resort to handle missing values by removing them instead of imputation. The final sample consists of 53697 records with 40519 non-defaulters and 13178 defaulters.

Although care has been taken in the sampling stage to under-sample the non-defaulters, the ratio of non-defaulter to defaulter in this dataset after removing records with missing values is still 3:1. Training a logistic regression model on this imbalanced class dataset where the data are bias towards the majority class of defaulters may not only affect the classifier performance but also increase the number of false negatives. In order to investigate the impact of imbalanced class in the training set, two training sets are considered in this research: one with ratio of 3 non-defaulters to 1 defaulter, and another with ratio 1 non-defaulter to 1 defaulter. We called the former as the imbalanced class dataset and the latter as the balanced class dataset to reflect the ratios of non-defaulter to defaulter in the training sets. Random subsampling without replacement was applied when preparing these training sets. Two different variations of logistic regression models are developed for each of these training sets.

Table 2: Long list of variables

No.	Variable	Definition
1	GENDER	Gender
2	MARSTS	Marriage status
3	RESSTS	Residency
4	SINGLE_JOINT	Flag whether an account is account for a single entity or is joint account for two or more entity

No.	Variable	Definition
5	AGE_YEAR	Age in years
6	ASSET_VAL	Asset value of property
7	CURR_LTV	Loan to value ratio calculated as at the end of observation period divided with asset value
8	BIZ_TYPE	A housing loan account that is based on Islamic banking concepts is of Islamic Business Type and otherwise, it is of Conventional Business Type
9	ORI_MAT_MONTHS	Tenure of loans which is the agreed duration in months
10	TYP_PRC	Type of interest rate pricing applicable to the housing loan account e.g. Floating interest rate, Fixed interest rate
11	H_CURDEL	Current delinquency which is the number of months in arrears reported for housing account as at the end of observation period
12	H_MTHARRL12	The number of months since the housing account was last reported as delinquent in the last 12 months to the end of observation period
13	H_VOLDELL12	The number of times the housing account reported was reported as delinquent in the last 12 months to the end of observation period
14	H_VOLDELL6	The number of times the housing account reported was reported as delinquent in the last 6 months to the end of observation period
15	H_MTHDELL6	The number of months the housing account was reported as delinquent in the last 6 months to the end of observation period
16	H_WORSTDELL12	The highest number of months in arrears reported for the housing account in the last 12 months to the end of observation period
17	H_WORSTDELL6	The highest number of months in arrears reported for the housing account in the last 6months to the end of observation period
18	H_MOB	The number of months since the month and year the housing account was reported in CCRIS to the end of observation period
19	INDIC_CARDS	Number of credit card accounts held by the entity
20	INDIC_OTHERS	Number of other facilities held by the entity
21	EXPOSURE_TYPE	Combinations of types of accounts held by the entity i.e housing only accounts, housing and credit card accounts, housing and accounts in other facilities, housing and credit card and accounts in other facilities.

4.1.1 Imbalanced Class Dataset

The training sets were prepared by randomly selecting 70% of non-defaulter records and 70% of defaulter from the sample. There are a total of 28363 non-defaulters and 9225 defaulters in this training set. The remaining 30% of records were used as test set. This process is repeated 10 times, which results in 10 different training-test splits. The performance of the ten models was then aggregated. No bagging or boosting procedures were employed.

The first variation of model development is to fit a logistic regression on this training set with all the explanatory variables summarised in Table 2, without any pre-filtering for significance or correlation with other variables. We called this model as 'Logit UNBAL NOVARSEL'. Since no variables were pre-filtered, it is likely that highly correlated variables are included in the final model. The inclusions of highly correlated variables may cause interpretation problems due to multi-collinearity effects.

We have also investigated whether variable selection for model construction can improve the predictive accuracy of the classifier. A process of pre-regression variable selection and

transformation was applied. The variable selection is carried out based on Accuracy Ratio (Sobehart, 2000), which is described further in Section 5.2. The variable selection process used is a univariate selection procedure where each factor was examined independently of other variables and evaluated across a range of selection attributes. The selection process was ultimately judgmental but guided by the following criteria: firstly variables were ranked by their ability to predict default. The Accuracy Ratio (AR) is used as the statistical performance metric. Generally, only variables with AR greater than 30% were selected, although some variables below 30% AR were also selected if they exhibit low correlation with any other selected variables. Low AR variables may still contribute to the overall predictive power of a multi-variable model provided they have low correlation to other stronger predictive variables. Each variable was then examined for an intuitive and preferably monotonic relationship with default and discarded if necessary. Finally for remaining groups of highly correlated variables, only a single variable with the highest AR was selected.

In this manner, 7 of the original 21 variables were selected, which include ASSET_VAL, CURR_LTV, ORI_MAT_MONTHS, H_CURDEL, H_MTHARRL12, H_WORSTDELL12, INDIC_CARDS. These variables are transformed to ensure that they were monotonic in relation to the dependent variable (default) and to reduce the impact of noise caused by extreme values in the regression. As it turns out, only continuous variables were selected. These variables were 'transformed' by fitting equations of the following form where X is the raw variable value and Y is the sample default rate for each given value of X:

$$Y = a + \frac{b}{1+e^{cX+d}} \tag{6}$$

The sample default rate Y is then converted into the logit space by applying the following equation and standardising to a mean of 0 and standard deviation of 1.

$$Z = -\log\left(\frac{Y}{1-Y}\right) \tag{7}$$

Finally a logistic regression procedure was applied where variables with positive coefficients were manually removed and the regression rerun until only variables with negative coefficients remained. The pre-regression variable transformation process had already aligned the relationship to default across all variables. Therefore any variable with a positive coefficient in the regression was interpreted as unintuitive and excluded. We called this model as 'Logit UNBAL VARSEL'.

4.1.2 Balanced Class Dataset

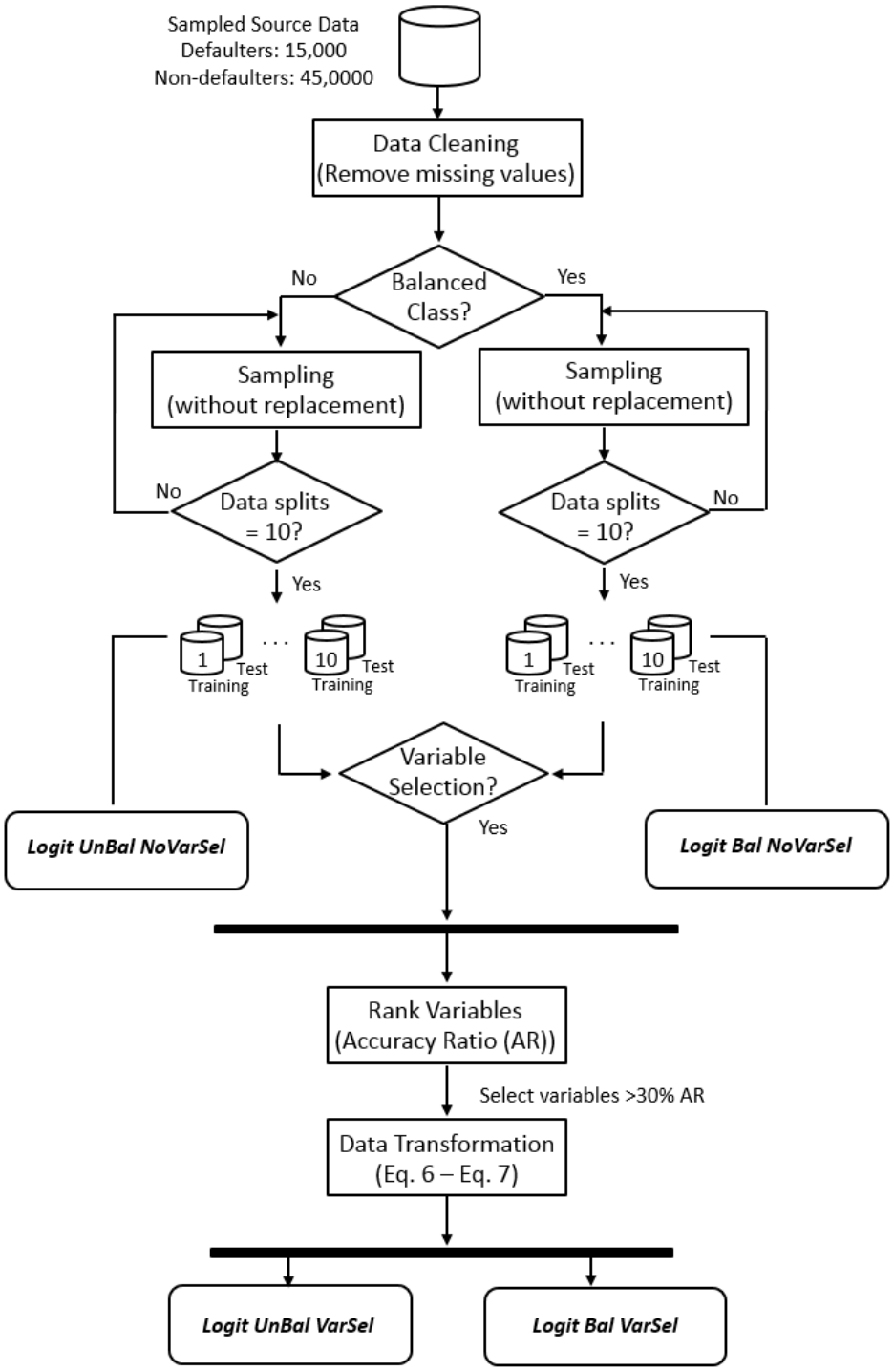
Since class imbalanced may produce classifiers, which are bias towards the majority class, we attempt to balance the class of defaulters and non-defaulters, and investigate how far this can improve the prediction performance of loan default. To obtain a balanced class training set, we used the exact equal number of instances in each class, i.e., the ratio of defaulter to non-defaulter instances is 1:1.

We randomly select 70% of instances from the default class and select the same number of instances randomly from the non-default class. This results in an equal number of defaulter and non-defaulter in the training set (i.e., 9225 each). All the remaining data are used as test set, which consists of a higher number of non-defaulters than defaulters. We repeated the sampling (without replacement) process for 10 times, resulting in 10 training-test splits.

Two variations of logistic models were trained on these training sets – the first trained on all the 21 variables and we called this model as 'Logit BAL NOVARSEL', and the second trained on only the 7 selected and transformed variables (as described in section 5.1.1), and we named this model as 'Logit BAL VARSEL'.

Figure 2 summarises the procedures of model development in this paper.

Figure 2: Summary of Model Development



Note: Processes within the synchronization bar (depicted by thick horizontal line) show the parallel flows for both imbalanced and balanced classes. Rounded rectangles represent the models developed in this study.

4.2 Model Validation Framework

The performance of credit scoring models can be highly sensitive to the data sample used for validation. To avoid embedding unwanted sample dependency, models should ideally be validated on observations not included in the sample used to build the model. Sobehart et al. (2000) describes four ways to do this that are differentiated along time and population dimensions. Firstly, an in-time out-of-sample dataset could be used comprising borrowers drawn randomly from the full training dataset that were not used in model development. Second, an out-of-time out-of-sample dataset could also be used, comprising borrowers drawn from a subsequent time period from the training set. Third, an in-time out-of-universe dataset would comprise borrowers whose distribution may differ from the population used to build the model. Finally, an out-of-time, out-of-universe dataset represents the most stringent of the approaches, as this comprise borrowers drawn from both a different time period and population from the training set.

This paper relies on the third approach. A panel data set corresponding to the same time period as the sample used for model development and containing another 60,000 housing loans randomly selected from the universe of housing loans in the system was used as the test set. This approach is more stringent than merely using an in-time out-of-sample dataset and has the advantage of assessing model performance against a sample representative of the actual population of housing loans.

For each evaluation, we calculate the confusion matrix and measure the accuracy performance. The confusion matrix for a 2-class problem is defined as follows:

		Predicted Class	
		+ve	-ve
Actual Class	+ve	True Positive (TP)	False Negative (FN)
	-ve	False Positive (FP)	True Negative (TN)

To compare models, we rely on four metrics. The first is the Percentage Correctly Classified *(PCC)*. The PCC measures the proportion of correctly classified cases on a sample of data. The PCC is calculated as follows:

$$PCC = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

The next two metrics are the corresponding the sensitivity and specificity of the model predictions. The sensitivity measures the proportion of positive examples that are predicted to be positive.

$$Sens = \frac{TP}{TP+FN} \tag{9}$$

The specificity measures the proportion of negative examples that are predicted to be negative.

$$Spec = \frac{TN}{FP+TN} \tag{10}$$

It is often useful to have a single measure that summarises the predictive accuracy of each risk measure for both Type I and Type II errors into a single statistic. For this purpose, the fourth metric used is the Accuracy Ratio (AR) as proposed by Sobehart (2000). This metric is also the most commonly used metric by banking institutions in the country for assessing credit scorecard performance. The graphical two-dimensional illustration of the AR is the corresponding Cumulative Accuracy Profile (CAP). To plot accumulative accuracy profiles, borrowers are first ordered by model score, from riskiest to lowest risk. For a given fraction x% of the total borrowers, a CAP curve is constructed by calculating the percentage y(x) of the defaulters whose risk score is equal to or lower than the one for fraction x. The AR can be obtained by comparing the CAP of a model with that of the perfect model. The closer the model CAP is to the perfect CAP, the better the model performance. The maximum area that can be enclosed above the Random CAP is identified as the ideal CAP. The AR is then the ratio of the area between a model's CAP and the random CAP to the area between the ideal CAP and the random CAP. The AR is a fraction between 0 and 100%. Models with ARs close to 0% display little advantage over a random assignment of model scores while those with ARs near 100% display almost perfect predictive power.

5 Results and Discussion

Table 3 reports the test set PCC, sensitivity, specificity and AR across all four model variations. We have used a notational convention whereby the best variation for each metric is underlined and denoted in bold face. Overall, there is no clear winner across all four performance metrics. LOGIT UNBAL NOVARSEL performs best when measured on a PCC and Specificity basis. LOGIT UNBAL VARSEL has the highest AR, although the result was only marginally higher than the rest of the model variations. Conversely LOGIT BAL VARSEL has the highest Sensitivity and second highest AR.

Table 3: Test set classification accuracy assuming cut-off of 0.5

Technique	PCC (%)	Sens (%)	Spec (%)	AR (%)
LOGIT UNBAL NOVARSEL	**91.7**	67.1	**92.7**	81.4
LOGIT BAL NOVARSEL	84.0	84.6	84.0	81.5
LOGIT UNBAL VARSEL	91.2	70.9	92.0	**81.6**
LOGIT BAL VARSEL	84.7	**84.6**	84.7	81.6

Given the mixed results, perhaps the best model variation would largely depend on the intended application of the model. From a regulatory capital computation perspective, the Accuracy Ratio is preferred as the Malaysian banking regulator has specified that banks on the Internal Ratings Based (IRB) approach use at a minimum the Accuracy Ratio. On that basis, all four model variations yield very similar results given that the difference between the best and worst variation is only 0.2 percentage points.

Interestingly, even though all four variations have approximately the same AR, there appears to be a significant difference in PCC, Sensitivity and Specificity, particularly depending on whether a balanced dataset (1 defaulter to 1 non-defaulter) or unbalanced dataset (1 defaulter to 3 non-defaulters) is used. At first glance, the two imbalanced datasets appear to perform better than the balanced datasets. The two imbalanced dataset variations have the highest PCC at 91.7% and 91.2% respectively. Conversely the balanced datasets have PCC's of only 84.0% and 84.7% respectively.

However, closer scrutiny of the sensitivity and specificity may suggest a different conclusion. The imbalanced datasets have lower sensitivity but higher specificity than their balanced counterparts. In fact, it appears that the high PCCs of the imbalanced datasets are largely driven by the high specificity rather than sensitivity. This implies that the model variations based on imbalanced datasets are better at correctly predicting non-defaulters, but relatively worse at predicting defaulters.

From a loan portfolio monitoring perspective, banking institutions would likely be more interested in models that are better at predicting defaulters (i.e. high sensitivity) rather than predicting non-defaulters (i.e. high specificity). This is because the cost of keeping a loan that subsequently defaults is greater than the opportunity cost of wrongly exiting a loan that does not subsequently default. From this perspective, the results suggest that using the balanced dataset variations performs better, with the LOGIT BAL VARSEL marginally better than LOGIT BAL NOVARSEL.

The model variations with variable selection perform marginally better when measured on AR and Sensitivity metrics than the variations without variable selection. However, the more significant benefit of variable selection is model parsimony. Only 7 variables were needed to produce the same or better performance compared to using all 21 variables. This may have significant practical considerations for banking institutions. Fewer variables translate to lower cost of storing, extracting and analysing superfluous data.

Table 4: Number of variables in each model variation

Technique	No of variables in model
LOGIT UNBAL NOVARSEL	21
LOGIT BAL NOVARSEL	21
LOGIT UNBAL VARSEL	7
LOGIT BAL VARSEL	7

Note that the PCC, Sensitivity and Specificity metrics in Table 3 were calculated assuming a cut-off value of 0.5 on the classifier's output. The outcome may vary depending on the cut-off threshold used. Hence it is interesting to examine the impact of using other cut-off values. Table 5 reports the results with a cut-off value of 0.25. The AR does not depend on cut-off value and therefore not reported again.

Table 5: Classification accuracy on test sets assuming cut-off of 0.25

Technique	PCC (%)	Sens (%)	Spec (%)
LOGIT UNBAL NOVARSEL	**83.9**	84.4	83.9
LOGIT BAL NOVARSEL	76.7	89.7	76.2
LOGIT UNBAL VARSEL	76.0	**90.5**	75.4
LOGIT BAL VARSEL	84.6	85.0	**84.6**

Using the alternative cut-off, it was observed that the LOGIT UNBAL NOVARSEL retains the highest PCC. However, the model variation with the best Sensitivity shifts from LOGIT BAL VARSEL to LOGIT UNBAL VARSEL. Similarly, the model variation with the best Specificity shifts from LOGIT UNBAL NOVARSEL to LOGIT BAL VARSEL. The results suggest that the PCC, Sensitivity and Specificity metrics are sensitive to the cut-off value used and different conclusions may be drawn depending on the exact cut-off. From this perspective, the Accuracy Ratio is perhaps the more reliable metric.

Table 6 and Table 7 reports the accuracy performance on the test sets but split by yearly cohorts. Results suggest that the performance across all variations is relatively stable across time. No model variations perform exceptionally poorly in any single year.

Table 6: Classification accuracy on test sets by yearly cohorts assuming cut-off of 0.5

Technique	2007			2008			2009			2010		
	PCC	Sens	Spec	PCC	Sens	Spec	PCC	Sens	Spec	PCC	Sens	Spec
LOGIT UNBAL NOVARSEL	89.3	68.3	90.4	90.3	69.9	91.4	91.4	65.8	92.7	91.8	64.6	93.0
LOGIT BAL NOVARSEL	80.2	86.4	79.8	81.1	84.9	80.9	83.0	85.4	82.9	83.5	81.0	83.6
LOGIT UNBAL VARSEL	88.7	71.9	89.6	88.9	76.6	89.5	90.6	68.2	91.8	91.1	69.1	92.0
LOGIT BAL VARSEL	81.3	86.1	81.0	81.1	85.2	80.8	83.9	84.6	83.8	84.5	81.9	84.6

Technique	2011			2012			2013			2014		
	Sens	Spec	PCC	Sens	Sens	Spec	PCC	Sens	Sens	Spec	PCC	Spec
LOGIT UNBAL NOVARSEL	69.4	93.0	92.3	64.6	69.4	93.0	92.3	64.6	69.4	93.0	92.3	93.0
LOGIT BAL NOVARSEL	84.8	84.6	85.1	87.0	84.8	84.6	85.1	87.0	84.8	84.6	85.1	83.6
LOGIT UNBAL VARSEL	72.4	92.3	92.2	68.3	72.4	92.3	92.2	68.3	72.4	92.3	92.2	92.0
LOGIT BAL VARSEL	81.3	85.4	85.9	86.6	81.3	85.4	85.9	86.6	81.3	85.4	85.9	84.6

Table 7: Accuracy ratio on test sets by yearly cohorts

Technique	2007	2008	2009	2010	2011	2012	2013	2014	Std. Dev
LOGIT UNBAL NOVARSEL	79.3	80.0	81.5	79.4	82.0	83.7	84.8	77.8	2.4
LOGIT BAL NOVARSEL	79.3	80.2	81.6	79.5	82.1	83.7	84.8	78.0	2.3
LOGIT UNBAL VARSEL	79.4	79.5	80.5	80.5	82.6	83.8	86.1	78.4	2.6
LOGIT BAL VARSEL	79.3	79.4	80.5	80.5	82.5	83.7	86.0	78.4	1.6

6 Conclusions

In this paper, the performance of various variations of the logistic regression technique was explored. Variations include the use of balanced vs imbalanced class in the training set, with and without variable selection and transformation. The performance of each variation was

assessed using the percentage correctly classified cases (PCC), sensitivity, specificity and Accuracy Ratio (AR).

It was found that all four model variations perform well on an AR basis, with the best being the unbalanced dataset with variable selection. However, the variation based on the balanced dataset with variable selection yielded the highest model sensitivity (assuming a cut-off value of 0.5), which arguably is the most practical metric for banking institutions and regulatory bodies using the model for portfolio monitoring. Results appear to be sensitive to the cut-off value used. A lower cut-off value of 0.25 would change the conclusion of best sensitivity model to the unbalanced dataset with variable selection.

The development of an industry-wide housing loan credit scoring model such as that discussed in this paper has many potential benefits for banking regulators. Firstly, it improves macro-surveillance efforts by complementing traditional backward-looking metrics such as non-performing loan (NPL) ratios. The more forward-looking nature of a behavioural credit scoring model can serve as an early-warning indicator of the overall health of the housing loan market. As the credit score of individual loans deteriorate, this shifts the risk profile of the entire industry faster than an NPL-based metric, allowing regulators more time to react. Secondly, it provides regulators with a micro-surveillance tool for monitoring the loan health of individual financial institutions. As each housing loan is assigned a credit score, regulators would be able to quickly identify any deterioration in credit quality and isolate the cause to say, a particular institution, loan vintage or customer demographic. This in turn could help inform appropriate policy or supervisory measures to be taken. Thirdly, if the credit scoring model is calibrated to produce a Probability of Default (PD) measure, it could also be useful as a sanity check by regulators on banks' using the Internal Ratings Based approach for calculating capital requirements for the housing loan portfolio. A comparison of individual bank's PD estimate against that produced by this credit scoring model may help identify in-accuracies in model estimates and help mitigate model risk.

This study raises several interesting topics for future research. One interesting topic would be to explore the use of machine learning techniques in comparison with the logistic regression used here. Another area is to investigate in greater detail the nature of the most predictive variables for identifying housing loan defaults in Malaysia. This study is limited to the information provided by the CCRIS database. Future studies could also explore if "Big Data" analytics could further improve the default prediction power of such models, e.g. by exploring

the use of unstructured data such as social media postings of the borrowers as predictive variables.

References

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring, *The Journal of the Operational Research Society*, 54(6), pp. 627-635.

Berg, Daniel (2005). Bankruptcy Prediction by Generalized Additive Models, *Statistical Research Report No. 1*, University of Oslo.

Dahiya S., Handa S.S. and Singh N.P (2015). Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set, *Industrija*, 43(4), pp. 163-174.

Fitzpatrick, T. (2013) Classification methods for mortgage distress-preliminary draft, *Credit Scoring and Credit Control XIII Conference*, Credit Research Centre, University of Edinburgh.

Gurný P. and Gurný M. (2013). Comparison of credit scoring models on probability of default estimation for US banks, *Prague economic papers*, 22(2), pp. 163-181.

Harris T. (2015). Credit scoring using clustered support vector machine, *Expert systems with applications*, 42(2), pp. 741-750.

Order, R. V. (2007). Modeling and evaluating the credit risk of mortgage loans: a primer, Ross School of Business Working Paper Series, Working Paper No. 1086, University of Michigan.

Sobehart, J.R., Keenan, S. and Stein, R. (2000). Rating methodology – benchmarking quantitative default risk models: A validation methodology, Technical report, Moody's Investors Service.

Veneables, W. and Ripley, B. (2002). Modern Applied Statistics with S, Chapman-Hall 4 ed.

Wang H., Xu Q. and Zhou L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLoS One*, 10(2), pp. 1-20.

Xiao J., Xie L., Liu D., Xiao Y. and Hu Y (2016). A Clustering and Selection Based Transfer Ensemble Model for Customer Credit Scoring, *Filomat,* 30(15), pp. 4015-4026.

Zhao, Z. and Xu, S. and Kang, B. H. and Kabir, M. M. J. and Liu, Y. (2014). Investigation of multilayer perceptron and class imbalance problems for credit rating, *International Journal of Computer and Information Technology*, 3, (4) pp. 805