Credit Risk Analysis using Lasso Logistic Regression and Random Forests

Kobamo Nelton Nnoka *

November 2022

Introduction

The pupose of this article is to interpret and discuss the credit risk analysis performed in the R files found in my Credit Risk Analysis repository on GitHub. The data was sourced from `https://www.kaggle.com/surekharamireddy/credit-data`. It starts with an exploratory data analysis and goes on to discuss the models (LASSO and Random Forests) and their validation metrics.

Exploratory Data Analysis

The information value (IV) of the variables home ownership, number of dependents, minor derogatory reports, major derogatory reports and employment type are presented in figure 1. From the figure, it can be seen that the variables home ownership and number of dependents have an information value between 0.02 and 0.1 implying that the variables have the ability to predict the default status though very weakly. However their predictive strength is relatively stronger than the predictors minor derogatory reports, major derogatory reports and employment type which have information values below 0.02. According to the information value criterion, variables with an IV below 0.02 are generally insignificant predictors.

With respect to home ownership status, analysis has shown that individuals who rent their homes have a higher default rate when compared to individuals who own thier homes. This can be seen in figure 2. From figure 2, it can then be postulated that home ownership is

*BA (Economics and Statistics) - University of Botswana(2022) - kobamonnoka@gmail.com

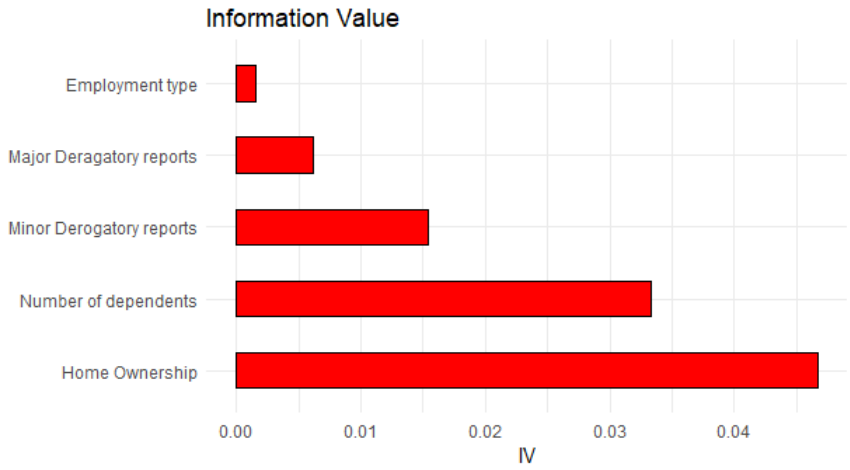

Figure 1: Information Value of variables

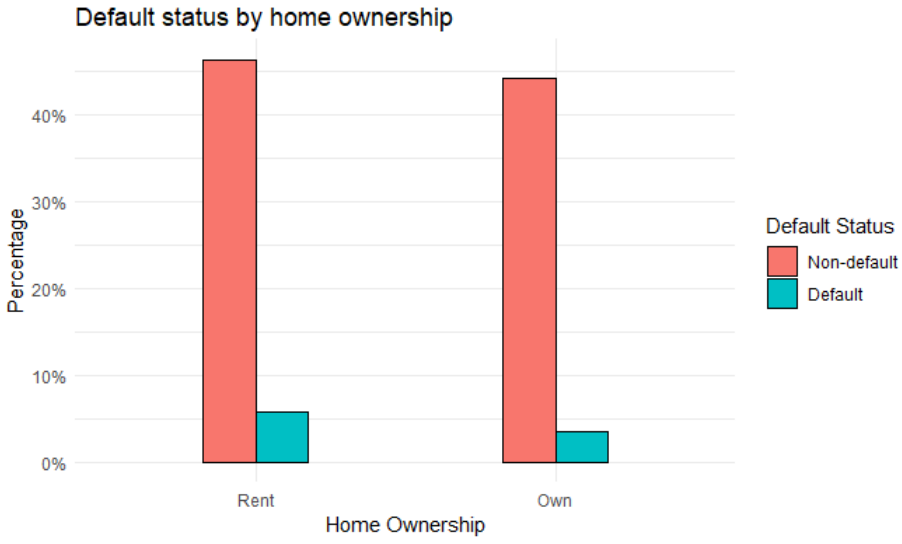

Figure 2: Proportion of default by home ownership

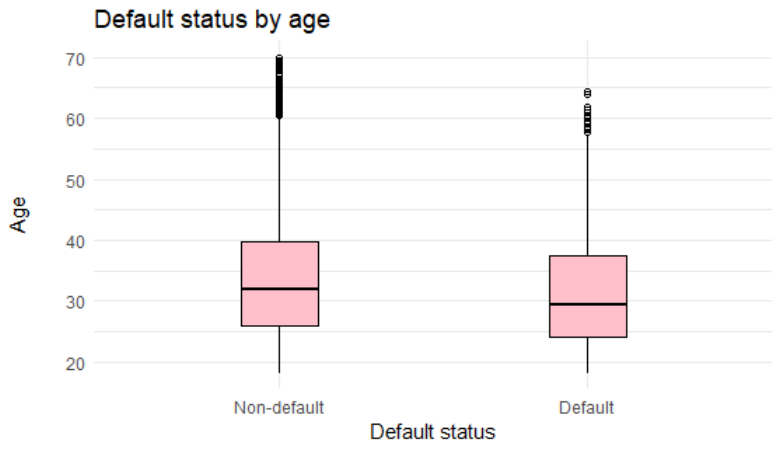

Figure 3: Default status by age

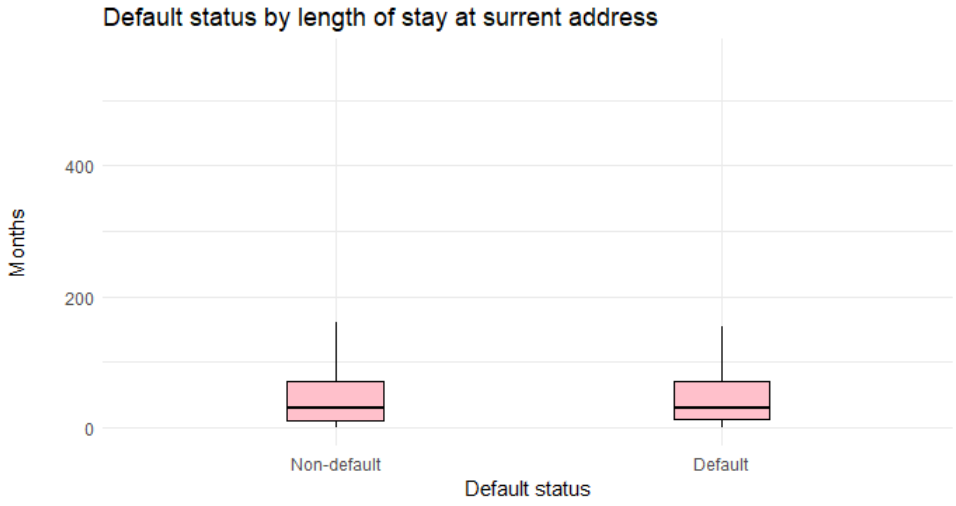

Figure 4: Default status by length of stay at current address

a significant variable in predicting default status. These results are consistent with those of the IV criterion though this criterion classifies home ownership as a weak predictor.

We can also speculate that relatively younger customers are more likely to default as compared to their older counterparts as shown on figure 3. Customers who default have a median age of 29 years whilst non-defaulting customers have a slightly higher median age of about 32 years, suggesting that older customers have a lower default risk.

The variable of lenth of stay at current address seems to have no value in differentiating between defaulters and non defaulters as seen on figure 4. From the data the mean length of stay at current address in months for defaulters and non-defaulters is 53.63 and 55.85 respectively. The median length of stay in months is 31 and 30 months for defaulters and non-defaulters respectively. The mean and median length of stay in months for defaulters

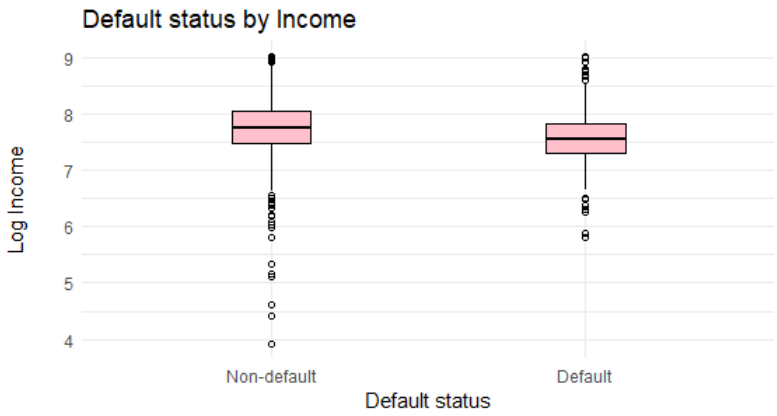

Figure 5: Default status by income

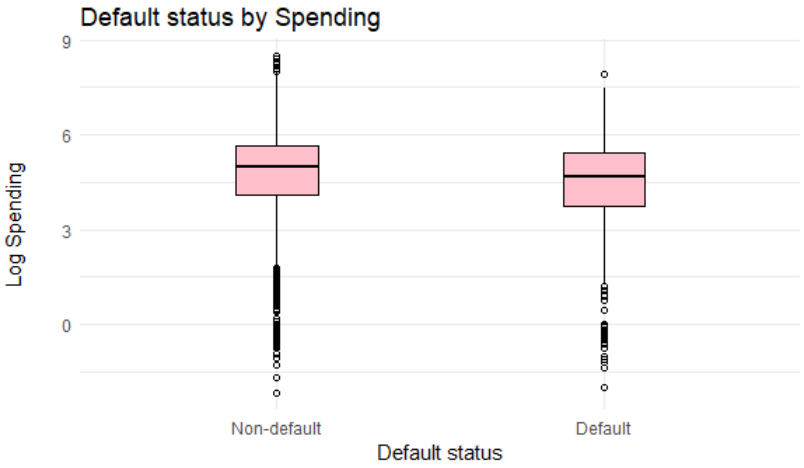

Figure 6: Default status by spending

and non-defaulters in months is not too different which may hint that length of stay is not a significant variable

From figure 5, individuals with higher incomes fair better at paying back their credit card debt timely. It is expected that individuals with slightly higher incomes will be less likely to default since they generate enough money to cover the payments as compared to their counterparts in lower tax brackets. The default group has a mean income of 2156.243 while the non-default group has a mean income of 2654.696. This then suggests that income is a useful variable in the classification of a customer as a defaulter or non defaulter.

In addition, non-defaulters have higher spending as shown in figure 6 which is probably influenced by their relatively higher incomes. The differences in spending between the two groups can therefore suggest that this variable is useful in the analysis of default status of customers.

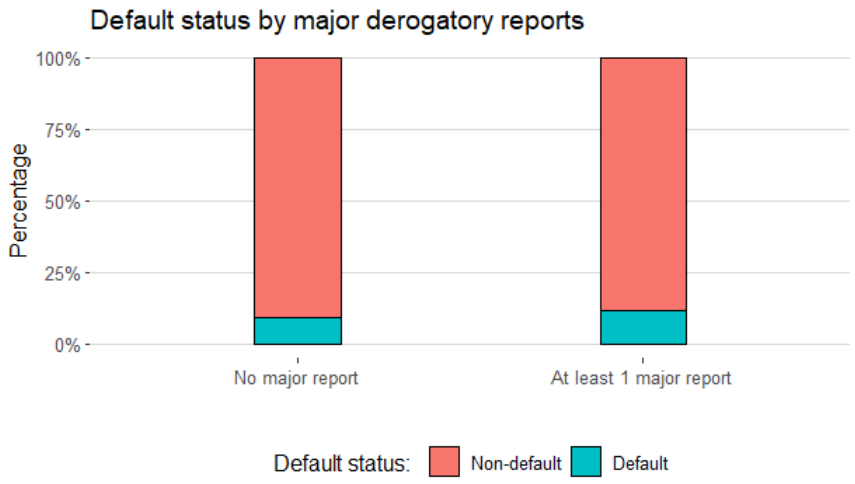

Figure 7: Default status by major derogatory reports

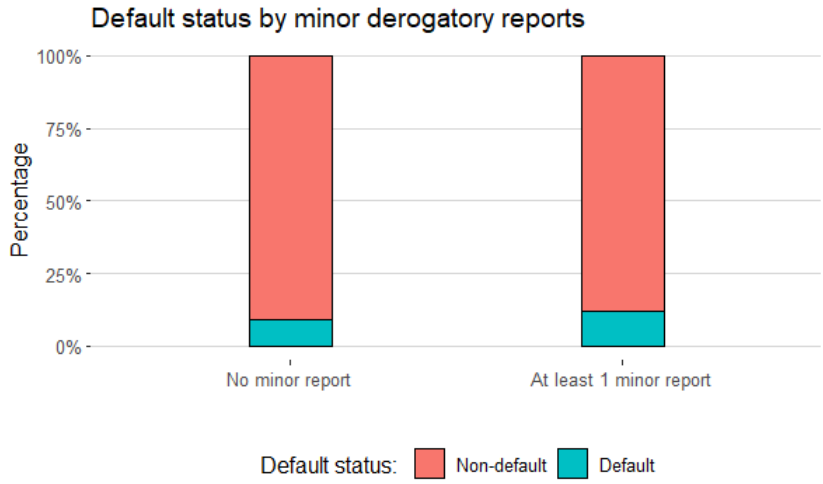

Figure 8: Default status by minor derogatory reports

Figures 7 and 8 show the default status by major derogatory reports and minor derogatory reports respectively. It can be seen that individuals with at least 1 major or minor derogatory have a slightly higher default rate as compared to those without a single report. However, in the figures the differences between the 2 are very minimal and can even be considered negligible, suggesting that derogatory reports are very weak predictors as predicted by the information value criterion.

Lasso Logistic Regression

Parameter Estimation

80% of the observations were taken as the training set and a lasso logistic model was fit using the glmet package in R. The value of λ was obtained through cross validation using the function cv.glmnet(). Thereafter the value of λ chosen is one that yields the most regularized model is within 1 standard error of the minimum value of λ that minimises the cross validation error. The value of λ was 0.01253751. The dashed line on the right in

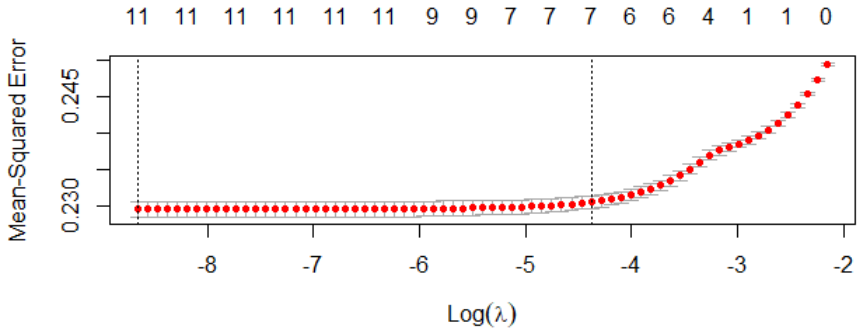

Figure 9: Optimal λ

figure 9 represents the selected value of λ. The numbers at the top represent the number of predictors in the model and it can be seen that 1 standard error value of λ will produce a model with approximately 7 predictors.

Running the lasso logistic regression resulted in 4 predictors out of a possible 11 being dropped from the model, leaving 7 predictors. The lasso algorithm can therefore be seen as an automation of the variable selection process. The coefficients of the model are as given in the table 1.

Variables with positive coefficients positively affect or actually increase the log odds in favour of default. Using the same reasoning, variables with negative coefficients decrease the odds in favour of default. For instance, a unit increase in the variable Log Income will decrease the log odds in favour of default by 0.902. This is to say the the more income an individual has, the less their chances of defaulting. In the context of binary variables (coded 0 or 1 in this article), the analysis is made on the level assigned the value 1. As an example, the variable home ownership has 0 representing renting and 1 owning a house. Therefore if an individual is renting their home their log odds in favour of default will be -0.284 implying that owning a house decreases the chances an individual will default.

Table 1: Coefficients of Lasso logistic regression

Variable	Coefficient
Intercept	8.648
Age	−0.0001
Months of stay at current address	0
Number of dependents	0
Major derogatory reports	0.297
Minor derogatory reports	0.343
Home Ownership	−0.284
Employment type	0
Expenditure to Income	0
Log Spending	−0.092
Log Income	−0.902
Log Income per dependent	−0.174

Variables with the coefficient 0 will be dropped from the model as they are insignificant.

Model Perfomance

It must be noted that the lasso predicts the probabilities of default for individuals given the values of the independent variables associated with them. A cut-off point is then necessary to classify a customer as a defaulter or non-defaulter. The optimal cut-off value will be the one which minimises the misclassification error that is where sensitivity and specificity are simultaneously high. The cut off point was obtained with the assistance of R using the function roc() and was 0.464. Table 2 shows the confusion matrix for this cut-off value. From the confusion matrix, the lasso model has a 63.05% accuracy rate.

Table 2: Actual values versus predicted values

		Actual Values	
		Defaulter	Non-Defaulter
Predicted Values	Defaulter	137	694
	Non-Defaulter	76	1177

This is to say that the model will correctly classify about 63 out of 100 applicants for a credit card. The sensitivity or true positive rate is 64.3%. A sensitivity of 64.3% means

the model will perform just above average in classifying defaulters. However the model has a specificity or true negative rate of 62.9% which implies the model have an above average correct prediction of non-defaulters. The model also has an AUC of 0.6754.

Random Forests

Variable Importance

The random forests algorithm generally lacks interpretability. However we can use variable importance to order the variables in terms of their importance in classification. An important variable in the classification procedure will be one with a high mean decrease in Gini and the higher the mean decrease in Gini the more important the variable. Figure 10 shows the variable importance of the independent variables used to determine default status. From the figure 10 it can be seen that the random forests procedure considers age to

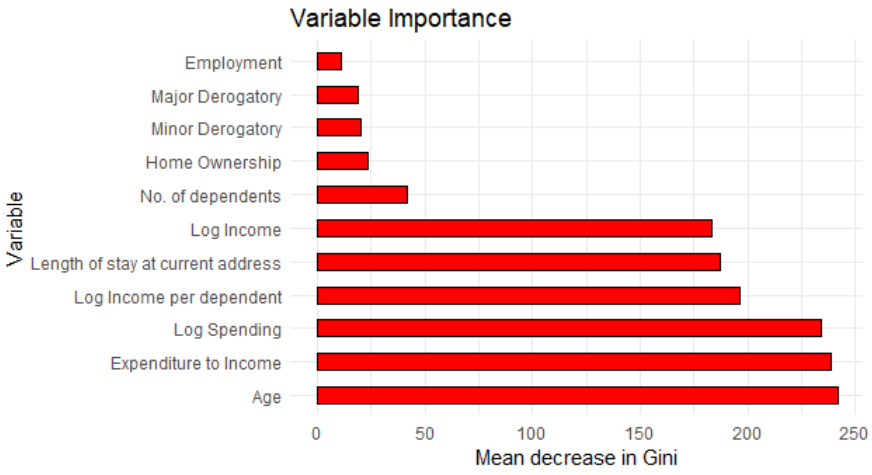

Figure 10: Importance of predictors

be the most important variable for classification and Employment as the least important variable.

Model performance

From the confusion matrix in table 3, the random forests algorithm had an accuracy of 89.88%, with a 1.88% and 99.9% sensitivity and specificity rate respectively. These reveal that the model is weak at correctly classifying defaulters but fairs really well in correctly classifying non-defaulters. In addition, the AUC for the random forests model is 0.5089. It can be noted that the random forests have a very low AUC despite their high accuracy

Table 3: Actual values versus predicted values

		Actual Values	
		Defaulter	Non-Defaulter
Predicted Values	Defaulter	4	2
	Non-Defaulter	209	1869

rate. This happens because the AUC metric is more concerned with correct classification into the two distinct groups (default and non-default) whilst the accuracy metric just concerns itself with correct classification in any of the groups. Since random forests were not able to classify a good number of of defaulters regardless of its great correct prediction of non-defaulters it yielded a low AUC.