Logistic regression applied to a vehicle loan data set

Michael Meyer

Assignment presented in the partial fulfilment
of the requirement for the degree of
MCom (Financial Risk Management)
at the University of Stellenbosch

Degree of confidentiality: A

June 2021

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

This project aims to build a logistic regression model from a given vehicle loan data set in order to accurately predict whether or not an obligor will default. Various data manipulation and statistical techniques will be used to select input variables that are considered relevant to the response variable. A logistic regression model is built using these variables and back tested to see the validity of the model. The data set and final model will be evaluated to determine whether or not it can be used for prediction. The project will discuss some theoretical concepts necessary for the discussion that follows and the practical implementation is coded in SAS software.

The aim throughout this project is to build an accurate predictive model. Inference from the model is therefore a secondary objective and will not be discussed extensively. The final model will only be assessed on the predictive power. The main goal in predictive modeling is generalization. Generalization means the ability to accurately predict outcomes for new data. In predictive modeling, you aim the maximize predictive power as assessed by relevant metrics.

1.2 BASIC METHODOLOGY

The basic methodology for predictive modeling is summarized as below:

1. Prepare the inputs
    - Organize and clean the historical data.
    - Check for missing values, outliers or any irregularities.
    - Reduce variables by testing their significance level,relation to response variable, redundancy and variable screening.
2. Build the model
    - Decide on the types of models to use.
    - Estimate parameters of the models using these inputs.

3. Test flexibility of model

    - Asses the models to see how flexible they are to new data.
    - Decide on a final model to use.

Supervised classification is used when the outcome of the target variables is known and discrete. The method "supervises" the model by testing the input variables with the target variable outcomes. This will be used as we have a binary target variable. Too many variables can cause a lot of problems when trying to build the model so it is essential to go through a process of elimination to disregard any variables that are not of use when building the model.

1.3 ASSUMPTIONS

The data set given has no additional information except for the input description, and therefore certain assumptions were made and conclusion drawn from the data. It is assumed throughout that the model built is for online loan applications in India that can be used today. The process that follows therefore takes this into consideration when selecting inputs that are deemed relevant for this use. The currency of the amounts is assumed to be the Indian Rupee.

CHAPTER 2

DATA PROCESSING AND PREPARATION

2.1 BACKGROUND ON DATA

The data set given to build the model is a data set from India for car loans. The data consist of 233,154 observation; 39 input variables and 1 binary response variable. The data set was obtained from the internet with no additional descriptions or context of the data set. Certain assumption were made about the data in order to properly work with the data. The binary target variable is **loan_default**, which measures whether the loan went into default or not. The input variables and a short description of each variable is given in Appendix A.

2.2 TYPES OF DATA ELEMENTS

It is important to appropriately consider the different types of data elements at the start of each credit risk modeling exercise. The following types of data elements can be considered (Baesens, Roesch and Scheule, 2016).

- Continuous data: These are data elements that are defined on an interval. A distinction can be made between ratio data(includes a natural zero) and interval data(excludes a natural zero).
- Categorical data: These are data elements that are defined to different categories. The different types of categorical data is:
    1. Nominal. The categories have no meaningful ordering between them.
    2. Ordinal. The categories have a meaningful ordering between them.
    3. Binary. There are only two categories.

The variables from the vehicle loan data set are divided into the different types of data as follows:

- 23 Continuous Variables :
  disbursed_amount asset_cost ltv Age PERFORM_CNS_SCORE PRI_NO_OF_ACCTS PRI_ACTIVE_ACCTS PRI_OVERDUE_ACCTS PRI_CURRENT_BALANCE PRI_SANCTIONED_AMOUNT PRI_DISBURSED_AMOUNT SEC_NO_OF_ACCTS

SEC_ACTIVE_ACCTS SEC_OVERDUE_ACCTS SEC_CURRENT_BALANCE SEC_SANCTIONED_AMOUNT SEC_DISBURSED_AMOUNT PRIMARY_INSTAL_AMT SEC_INSTAL_AMT NEW_ACCTS_IN_LAST_SIX_MONTHS DELINQUENT_ACCTS_IN_LAST_SIX_MON Ave_Acc_Age Cre_His_Len NO_OF_INQUIRIES ;

- Categorical data:
  - 2 Nominal Variables: **Employment_Type State_ID**
  - 1 Ordinal Variable : **PERFORM_CNS_SCORE_DESCRIPTION**
  - 5 Binary Variables : **Aadhar_flag PAN_flag VoterID_flag Driving_flag Passport_flag MobileNo_Avl_Flag**

2.3 DATA PREPOSSESSING

2.3.1 Variable transformations

Some of the variables, given in the previous section, have been transformed from their original data to data that is more suitable for logistic regression purposes. The following changes have been made:

1. The variable **Age** was constructed by subtracting **Date_of Birth** from **Disbursal_Date** to obtain the age of the obligor when the loan was disbursed.
2. The variables **AVERAGE_ACCT_AGE** and **CREDIT_HISTORY_LENGTH** have been changed to the continuous variables **Ave_Acc_Age** and **Cre_His_Len**, repectively. These variables are now measured in months.
3. The variable **PERFORM_CNS_SCORE** ranges from 300 to 890 when a bureau score is given. However approximately 50% of the values do not have a score and then a score of zero is given. This will likely skew the estimates and there for a score is given of 500 when there is insufficent data to produce a score. The reason that this amount was chosen is that it still falls into the category of high risk.

.

2.3.2 Missing Values

Missing values can occur for various reasons; such as incorrect data entries or the obligor not disclosing certain information. The most popular schemes for dealing with missing values are (Baesens, Roesch and Scheule, 2016) :

- Replace(Impute). This implies replacing the missing values with a known value such as the mean or median.
- Delete. This the most straightforward and consists of deleting observation with lots of missing values.
- Keep. Missing values can be meaningful. For example, if a customer did not disclose his or her income because he or she is unemployed. This fact may have a relationship with default and needs to be considered as a separate category.

The only variable that has missing value for categorical data is **employment_type**. A new value will be assigned to the missing entries. The reason this method is used is to check whether or not additional information can be gained when loan applicants did not fill in their employment type, for example that they are unemployed and therefore leave it out. A lot of the continuous data had not necessarily missing data but had large proportions of zero values [1]. This can be interpreted as missing values though. A possible solution is to exclude cases where the value is zero and find coefficients based on the interval from 1 to the max amount and assign a binary variable whether the value was zero or not but this process is rather complicated and therefore the original values are used as is. From this we can already see that a lot of these variables will likely not be very useful in modelling as they do not generalize well to the whole population.

2.3.3 Judgemental Screening

The first step in reducing variables is to use a judgemental approach to remove any variables that could not be used to increase the prediction accuracy. Some of the variables are only meta data of the loan when it was processed and therefore are irrelevant for future prediction. These variables are :

[1] Most of these variables have 98%+ zero values

1. **UniqueID**.
2. **branch_id**.
3. **Current_pincode_ID**
4. **Employee_code_ID**
5. **DisbursalDate**

The disbursal date could possibly be used when modelling economic cycles but this process becomes too complicated for our purpose. Also the branch ID could be useful for inference about default rates at the different branches but not for prediction.

2.3.4 Construction of Training and Validation data

The entire data set is usually split into training and validation data sets. The training data set is used to build the model and the validation data set is used measure the classification error of the model (or any other metric for determining model accuracy) and to determine an honest assessment of the data. Simple random sampling involves randomly sampling from without any consideration for the proportion of the target variable; and stratified sampling involves sampling from the data so that the proportion of the target variable remains constant in the sampled data. Stratified sampling is used in the project. The split used is 75:25, which means that 75% of the observations was imported into the training data set and 25% into the validation data set.

CHAPTER 3

EXPLORATORY DATA ANALYSIS

The next step was check the descriptive statistics of the variables to see if any irregularities occur or any thing can be found to understand the variables a bit better. Significance levels are also tested for the variables and certain metrics for determining the association of the predictor variables to the response variable. **Proc Glm, Proc Univariate, Proc Sgscatter, and Proc BoxPlot** was used to evaluate the continuous variables by drawing histogram, boxplots and correlations .**Proc Freq** is used to evaluate the categorical variables to see how many level the variables contain and to preliminary check the predictive power of the variables. The different metrics used to asses the inputs:

1. χ^2 coefficient.
2. F statistic.
3. Cramer'V.
4. Pearson Correlation.
5. Spearman correlations.

3.1 TARGET VARIABLE

The variable **loan_default** is a binary variable that indicates whether the observations defaulted or not. The default percentage is 21.71% for the entire population. This is quite a high default rate so an accurate model to predict future defaults will be very helpful.

3.2 CATEGORICAL DATA

3.2.1 Binary data

The binary variable **MobileNo_Avl_Flag** has only values equal to one and therefore should be excluded from the model as it does not contain any useful information. The variables **Driving_flag** and **Passport_flag** have 2.32% and 0.21% when the indicator function is 1, so that these variables are very skewed in their distribution. An F-test was done to see if the default means between the

categories differ. The test however showed that the means are statistical different. An important factor to remember is that when working with such a large sample size, even a slight difference in the mean will be significant. The default rates across the two groups were therefore manually inspected and very little difference was found in default means and seeing how one category accounts for only +- % 2 of values these variables were disregarded. The variables **PAN_flag** was found be not statistically significant using an F-test and was subsequently disregarded from the analysis.

3.2.2 Nominal and Ordinal data

The **Employment_Type** has three levels and therefore dummy variables are constructed in order for them to be modeled. **State_ID** has 22 levels, and the assumption that the model is for online loan applications implies that this input could be potentially useful; otherwise the state of the potential obligor would have been irrelevant, if the model was used at a branch. **PERFORM_CNSDESCRIPTION** has 20 levels with some of the levels containing almost no observations.**Manufacturer_id** has 10 levels and **supplier_id** has 2906 levels, therefore **supplier_id** will also be excluded as when modelling a categorical with so many levels and few observations with each category it is likely to overfit the data on the training set,[1].Both the binary and categorical data association to the response variable was tested using Cramer's V. As a rule of thumb, inputs with Cramer's V values less 0.1 can be disregarded. However all the input variables had values less than 0.1, so it was decided to include all of them in the analysis.

3.2.3 Smoothed Weight-of-Evidence coding

A categorical variable with too many levels can cause problems with modelling. A technique for working with categorical inputs is to replace the values with a single column that represents the event rate for each category. The weight-of-evidence method (WOE) replaces the categorical values with the log odds of the event. The WOE method is however, typically overfits the training data. Smoothed weight-of-evidence coding (SWOE) is a more general approach that avoids overfitting by taking into account the sampling variability. SWOE uses adjusted log odds ratios. The categorical variable is converted to a continuous variable. SWOE coding is applied to the following variables; **manufacturer_id**,**State_ID** and **PERFORM_CNS_SCORE_DESCRIPTION**.

[1]High dimensionality and quasi-complete separation could also lead to problems

3.3 CONTINUOUS VARIABLES

The continous variables were first tested using Pearson Correlation to see if they have any relation to the target variable. All the variables have correlations less then 0,1. This is an indication that these variables have at most weak linear associations. When drawing a histogram of the ages of the obligors we see that most of the obligors are 20-30. This could indicate that a lot of variables such as past credit history, secondary number of accounts etc. will have missing values.

Outliers are extreme observations that are very dissimilar to the rest of the population. Two types of outliers are considerded namely: valid observations and invalid observations. The maximum for **PRI_NO_OF_ACCTS** is 453 while the mean is 2.44. The observation for the maximum is therefore invalid since as I fail to see how a person could have 453 loans for personal use. If it is possible, then other consideration could be taken into account when determining the loan, but for our regression purposes it will be excluded [2]. Histograms and boxplots were constructed to see how the variables are distributed. From the various graphs we can see that, for example, that there are very few observations with **PRI_NO_OF_ACCTS** more than 25,and the mean is close to zero, we therefore exclude these observations from our analysis. A similar approach was followed for all the other variables.

The variables **SEC_NO_OF_ACCTS SEC_ACTIVE_ACCTS SEC_OVERDUE_ACCTS SEC_NO_OF_ACCTS SEC_ACTIVE_ACCTS SEC_OVERDUE_ACCTS SEC_CURRENT_BALANCE SEC_SANCTIONED_AMOUNT SEC_DISBURSED_AMOUNT** have values eqaul to zero in at least 98.5% of observations. This makes modeling these variables very difficult as everything is almost zero and then the other 1.5% is scaled up on a continuous basis. These variables will therefore also be excluded from the model.

There is also different methods of scaling continuous variables so that the regression coeffients in the final logistic regression model is not to small. One could calculate z-scores, based on the normal distribution, standardize values or divide by factors of 10, which is also called decimal scaling. None of the variables follows a normal distribution and therefore decimal scaling.

[2] A fraud detection could be put in place to immediately exclude the loan

CHAPTER 4

REDUCING INPUTS

4.1 REDUCING REDUNDANT VARIABLES

Having too many predictor variables could cause problems in model development,interpretation and scoring. By reducing redundant variables that explains similar variation in the response variable fewer inputs can be used to assist in more accurate model development. Redundant variables can cause the following problems :

- Destabilize the parameter estimates.
- Increase the risk of overfitting.
- Increase computation time.
- Increase the cost of data collection and augmentation.

Therefore it is always a good idea to reduce the number of input variables. There are different methods for reducing variables. We will look at variable clustering, which involves clustering variables into groups that are highly correlated with each other and not highly correlated with variables in other groups. The procedure **Proc Varclus** in SAS will be used reduce redundant variables.

The underlying algorithm that Proc VarClus uses is called divisive clustering. At each stage of the algorithm, divisive clustering splits a given subset of variables into two groups. All variables start in one cluster. Then, a principal components analysis is done on the variables in the cluster to determine whether the cluster should be split into two subsets of variables. The algorithm uses eigenvalues, which are weighted sums of the principal components.Principal components measure the unique variation explained by a variable. A cutoff value is assigned to the second eigenvalue. If the second eigenvalue is lower than the cutoff value then algorithm stops dividing the cluster. The reason the second eigenvalue is chosen is that it explains the additional variation that can be reduced it the cluster is divided further. The following output is given using this procedure.

A represented input variable should then be chosen from each cluster. We ideally want a vari-

able that is highly correlated within its own cluster and has a weak correlation with its own cluster. A metric to determine this the $1 - R^2$ ratio for which the equation is

$$1 - R^2 ration = \frac{1 - R^2_{\text{Own Cluster}}}{1 - R^2_{\text{Next Closest}}}$$

The numerator increases if the correlation within its own cluster is low and the denominator decreases when the correlation with the next closets cluster is high; therefore a low $1 - R^2$ ratio is preferred . This ratio in combination with subject matter knowledge and the inputs relation to the target variable should be used to choose the a representative input variable from each cluster. The output for the variable clusters is given Appendix B.1 The following variables are chosen using the criterion : **PRI_ACTIVE_ACCTS Per_CSD VoterID_flag PRI_DISBURSED_AMOUNT Employ_Sal Ave_Acc_Age asset_cost PRI_OVERDUE_ACCTS disbursed_amount NO_OF_INQUIRIES manufacturer_id SEC_INSTAL_AMT PRIMARY_INSTAL_AMT Age State_ID PRI_CURRENT_BALANCE ltv**

4.2 VARIABLE SCREENING

4.2.1 Detecting non-linear relationships

Logistic regression is constrained to testing linear associations between input and response variables. However there could non-linear associations between the input and response variables that can not be modelled using logistic regression. In simple regression, scatter plots can be constructed to asses non-linear association. However with a binary response variable this is not possible. We can however plot the logits by the input variable to test for non-linear associations. This can be time consuming and an easier method is to use the Spearman correlation in combination with Hoeffding's D statistic to test for non-monotonic relationships.

Hoeffding's D is a general and robust similarity measure that detects a wide variety of associations between two variables, not only monotonic associations. The value of Hoeffding's D does not give indication of how strong the non-linear association is but we can rank the variables according to this metric. By ranking the variables according to Spearman correlation and Hoeffding's D we can test for non-linear relationships. When the Spearman rank is low and the Hoeffding's D rank is high, the association is non-monotonic. The output for this procedure is given in Appendix

B.2 From the results it is apparent that there is no variable with a relatively high Hoeffding's D rank and a low Spearman rank. We can see that the variables**PRIMARY_INSTAL_AMT SEC_INSTAL_AMT asset_cost** have very low correlations and therefore are excluded from the modeling inputs.

CHAPTER 5

FITTING THE MODEL

5.1 THE LOGISTIC MODEL

When modeling a binary target variable D using linear regression we get the following equation :

$$D = B_0 + B_1 x_1 + ... + B_n X_n$$

When estimating this using ordinary least squares (OLS) or maximum likelihood estimation there are two key problems that arise:

1. The targets are not normally distributed but follows a Bernoulli distribution with only two outcomes.
2. There is no guarantee that the target is between 0 and 1; it would be handy if it were, because then it could be interpreted as a probability.

We can use following transformation which bounds the outcome between zero and one[1].

$$f(z) = \frac{1}{e^{-z}}$$

By combining linear regression and the bounding function the general formulation of the logistic regression model becomes:

$$P(D = 1|X_1, .., X_n) = \frac{1}{1 + e^{B_0 + B_1 X_1 + ... + B_n X_n}}$$

We can rewrite this in terms of the odds and take natural log to get the equation:

$$ln(\frac{P(D = 1|X_1, .., X_n)}{P(D = 0|X_1, .., X_n)} = B_0 + B_1 x_1 + ... + B_n X_n$$

.

[1] This is also known as a sigmoidal transformation

The **Proc Logistic** procedure in SAS uses this equation and MLE to get the ML estimates for the regression coefficients. Different types of models could be specified such as the logit(the default option in SAS) model, probit model or the cloglog model to use in the **Proc Logistic** procedure . Little difference was found using these different models so the more simple logit model was used.

5.2 TYPES OF VARIABLE SELECTION

5.2.1 Forward Selection

Forward stepwise selection is a selection method that starts with an empty model. The method then adds inputs incrementally until no more inputs meet the entry criterion for statistical significance. The usual criteria used to add inputs from the model are based on the p-value from a test of the null hypothesis that the parameter estimate for a variable is zero. Forward Stepwise selection was devised to give a computationally efficient alternative to examining all subsets. It is not guaranteed to find the best subset and it can be shown to perform badly in many situations (Harrell Jr, 1997).

5.2.2 Backward Selection

Backward elimination is a method of subset selection that starts with a full model. At each step, the test statistics are computed, and the variable with the largest p-value that exceeds the criterion is removed from the model. Backward elimination is less inclined to exclude important inputs or include spurious inputs than forward stepwise selection.

5.2.3 Best Subset Selection

The best subset selection method does the most thorough search of the input variables. This method considers all possible models and rank-orders them based on the score chi-square. The model can be specified to include only the best model for each number of input variables included in the model.

5.2.4 Interactions

When you are selecting the most predictive inputs for a predictive model, it is important to look for interactions between inputs. Interaction occurs when the relationship between an input variable

and the target differs by the level of another input variable. Interaction can found by using the forward selection method and specifying the factor of interaction to include. For example if two factor interaction terms are specified only interactions between two variables are included.

5.3 MODEL FIT STATISTICS

For predictive modeling a significance level that is based on a model's fit statistics is used in order to asses and compare models. We will use the Bayesian information criterion (BIC), which is a fit statistic measure that penalizes for model complexity, .i.e to many input variables. Selecting the model with the smallest BIC favors a tight fit to the training data (that is, a large likelihood) and a small number of parameters. The formula for the BIC is negative 2 times the log likelihood plus the number of parameters times the log of the sample size.

5.4 SELECTION PROCEDURE FOLLOWED

A significance level based on the BIC is calculated for entry and exit of predictors variables when using forward and backwards selection. Once the significance level was specified the forwards selection approach was used to detect important two-factor interactions among the screened variable. The following interaction were found : **PRI_ACTIVE_ACCTS*Ave_Acc_Age**
PRI_ACTIVE_ACCTS*disbursed_amount
PRI_DISBURSED_AMOUNT*PRI_OVERDUE_ACCTS PRI_CURRENT_BALANCE*ltv
Ave_Acc_Age*PRI_OVERDUE_ACCTS Ave_Acc_Age*disbursed_amount

Backward selection was also used but forward selection was found to produce a better model. The final steps was to use best subset selection to find the best model for each number of input variables. The variables were then ranked by their BIC value to find the best model. The output for this in given in Appendix B.3.

5.5 FINAL MODEL

The final model selected based on the AUC, C-statistic,BIC value, and the number of terms in the equation. The best model under best subset selection had 20 terms, however in the performance measures there was little difference between the model with 20 terms and 10 terms so the model with 10 inputs was chosen to reduce complexity. A model with fewer parameters is likely to be more flexible. The output for this model is given in Appendix B.4. The final variables selected are:

- X_1 = Per_CSD
- X_2 = Employ_Sal
- X_3 =PRI_OVERDUE_ACCTS
- X_4 =disbursed_amount
- X_5 =NO_OF_INQUIRIES
- X_6 =manufacturer_id
- X_7 =Age
- X_8 =State_ID
- X_9 =ltv
- X_{10} =PRI_ACTIVE_ACCTS

With regression coefficients:

- β_0 = -0.6065
- β_1= 0.7913
- β_2= -0.1908
- β_3= 0.1422
- β_4= 0.9921
- β_5= 0.1616
- β_6= -0.00401

- β_7= -0.00973
- β_8= 1.0713
- β_9= 0.0247
- β_{10}= -0.1537

And the model is the given as:

$$ln(\frac{P(D=1|X_1,..,X_n)}{P(D=0|X_1,..,X_n)} = B_0+B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5+B_6X_6+B_7X_7+B_8X_8+B_9X_9+B_{10}X_4*X_{10}$$

.

The model has only one interaction term, however the interaction term uses a input that is not in the main inputs of the model. This does luckily not matter as with predictive modelling we are not interested in the hierarchy principal of inputs. The value of the c-statistic (which we will discuss later) on training data is 0.635, which is not very high. Although the aim of this paper is not inference, we will draw some inference from the regression coefficients to check if the model makes sense. The probability of default is positively related to the regression coefficients. One of the inputs in the model is the age of the obligor. From previous credit risk modelling we know it is usually the case that age is inversely related to the probability of the default. We see that the regression coefficients is negative for age and therefore if age increases the probability of default decreases, which is what we would expect. The standardized estimates of coefficients also shows that Loan to Value is the most important predictor variable when determining the probability of default of an obligor.

CHAPTER 6

MODEL VALIDATION

6.1 GENERAL MODEL PERFORMANCE METRICS

Now that we have a model we need to test the model on a different data set to check whether it has any predictive power. The validation data set will be used to test the model and the same preparation should be applied as for the training set. The measurements used to validate the model is the c-statistic, ROC curve, lift curve, K-S statistic and the confusion matrix.

6.2 C STATISTIC

The c statistic is measurement that that calculates the percentages of concordant pairs. Concordance measures are frequently used for assessing the discriminative ability of risk prediction models. The c statistic is measurement that calculates the percentages of concordant pairs. A value of 1 indicates the model has perfect discrimination, while a value of 0.5 indicates the model discriminates no better than random chance. The c-statistic for the model is 0.632 while for the training data it was 0.635. This shows that the model only the validation data is only slightly worse than the training data and the model generalizes very well. Values over 0.70 usually indicates a good model, which shows that the model is not very good at discriminating between between defaults and non default but better than random chance.

6.3 ROC CURVE

A more comprehensive measure to model performance across all of the cutoff probabilities is by using the receiver-operating characteristic curve, also called the ROC curve. The ROC curve displays the sensitivity on the y-axis and 1 minus specificity on the x-axis for the entire range of cutoffs. Sensitivity (true positive rate) measures the proportion of positives that are correctly identified. Specificity (true negative rate) measures the proportion of negatives that are correctly identified. The baseline displaying a 45-degree angle from (0,0) to (1,1) represents a random model for which the area under the curve is 0.50, which is represented by the c statistic. This baseline represents the accuracy of a model that predicts classes no better than random chance. The more the ROC

Figure 6.1: ROC curve for validation

curve bulges from the 45-degree line, the more accurately the model predicts compared to a random model. A model that predicts events perfectly would reach the (0,1) point on the ROC curve. The more the ROC Curve bulges. The ROC curve for the final model is given shown in Figure 6.1. We can see that there is some bulge but that it does not deviate from the model and therefore is not a very good model using this metric.

6.4 LIFT CURVE

The lift chart displays the lift on the y-axis and the depth on the x-axis Depth equals the total percentage of cases that are allocated to the default class while the the lift equals the positive predicted value divided by the proportion of defaults. The positive predictive value is the probability that a observation that is classified as a default is a default. As the cutoff probability increases, the depth decreases. The lift basically displays how much more likely we are to correctly classify someone as a default at a certain level of cutoff probability (or depth) in comparison with a random

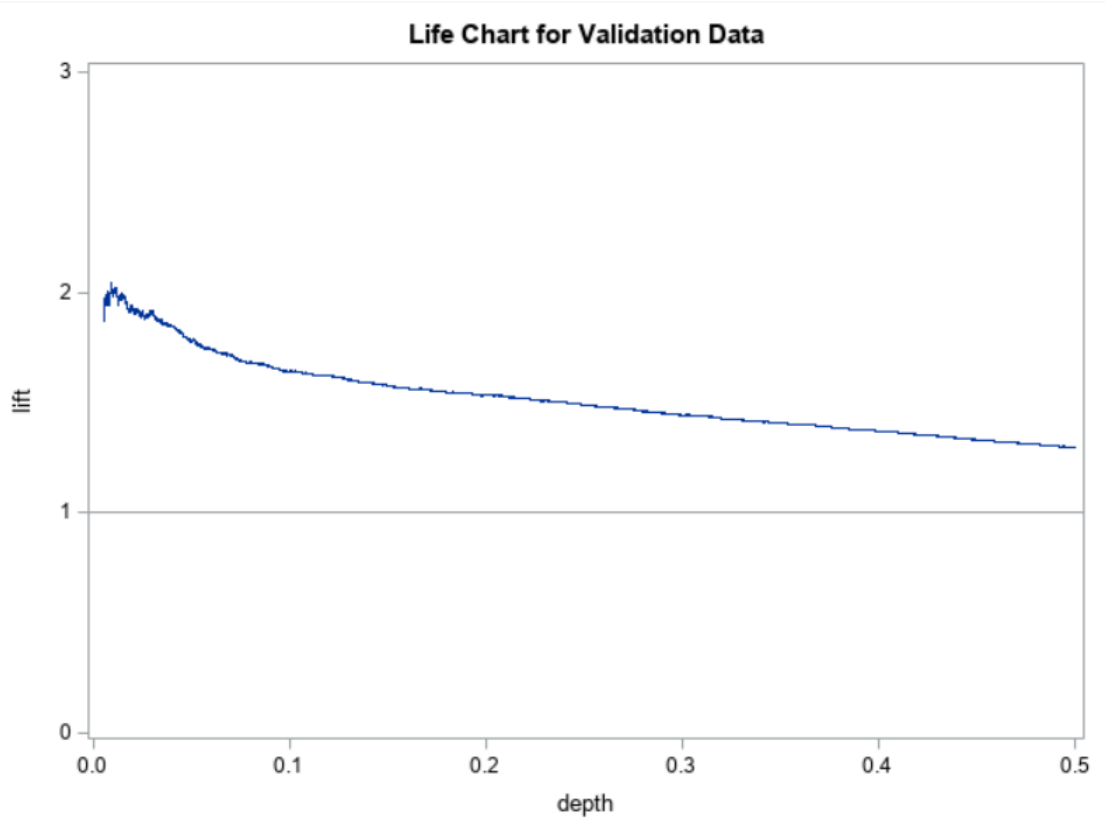

Figure 6.2: Lift Chart for Validation Data

sample at the same level. The lift chart also shows the baseline where the value of a random sample is 1. This means we are equally likely to predict whether someone will default or not based on random chance. Figure 6.2 displays the lift chart from our results. From the graph we can deduce that when we have a very high cutoff probability we are 2 times more likely to classify someone who will default. When the depth is 0.5 the lift is approximately 1.3. This shows that the model is not a lot more likely to correctly classify defaults than random chance.

6.5 K-S STATISTIC

The Kolmogorov-Smirnov two-sample test is commonly used to assess how well a model distinguishes between events and non-events. It is based on the distance between the empirical cumulative distribution functions (Conover and Iman, 1980). The Kolmogorov-Smirnov test produces a value called the K-S statistic. Use of the K-S statistic for comparing predictive models is popular

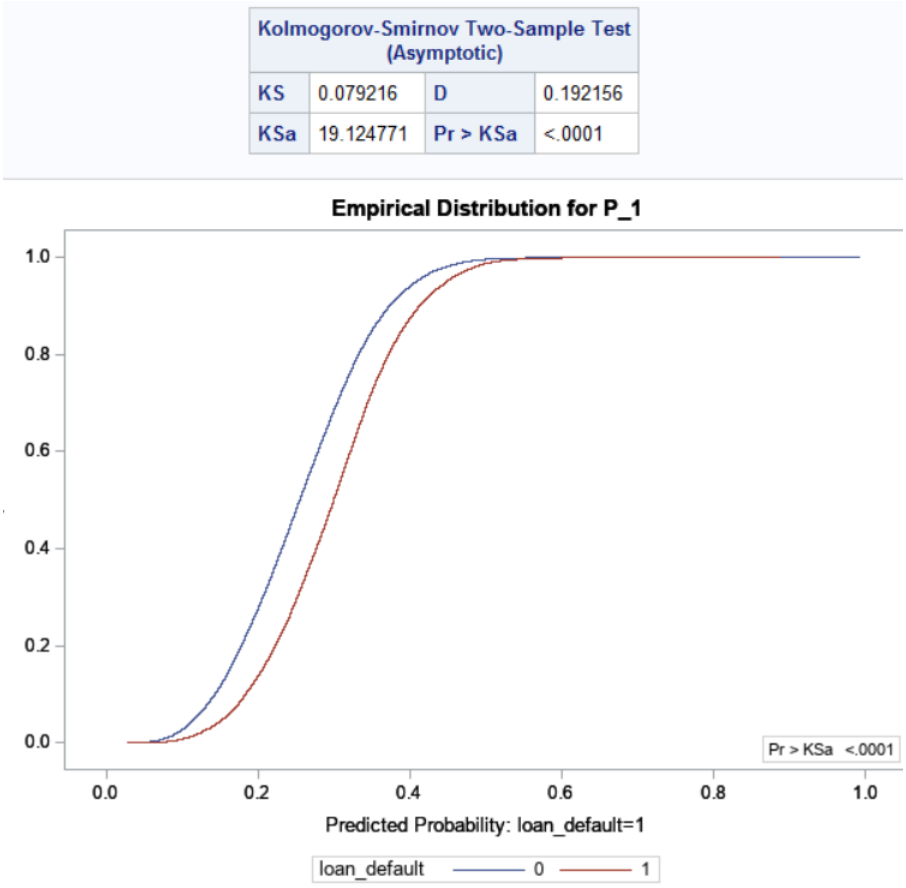

Figure 6.3: K-S Statistic

in credit risk modeling. On the left is the class separation graph for a model with a smoothed probability density function on the Y axis. The same classification model is plotted in the empirical cumulative distributions graph on the right, where the Y axis represents the proportion of observations that have predicted probabilities below a given posterior probability. The D statistic equals the maximum vertical difference between the cumulative distributions for class 0 and class 1. Figure **??** displays our results for this test. The D-statistic is 0.192 which is not very high and we can see from the graph that the distance between the cumulative distribution of the defaults and non-defaults is close to each so the model does not distinguish very well between defaults and non-defaults. When can see there are few observations with a probability higher than 0.5 for the defaults. When using a cutoff of 0.5, the model will thus not predict a lot of defaults.

Table of Actual by Predicted			
	Predicted		
Actual	0	1	Total
0	45413	222	45635
1	12486	166	12652
Total	57899	388	58287

Figure 6.4: Confusion Matrix

6.6 CONFUSION MATRIX

The fundamental assessment tool for model performance is the confusion matrix. The confusion matrix is simply a cross tabulation of predicted classes and actual classes. It quantifies the confusion of the classifier. The confusion matrix from our model is given in Figure 6.4. The rows display the predicted number of non-defaults and defaults while the columns display the number actual defaults and non-default. There are 166 true positive, 45413 true negatives, 222 false positives and 12486 false negatives. The accuracy ratio is the correctly classified cases divided by the total cases and the error rate is the incorrectly classified cases divided by the total cases. The accuracy for the model is 78.2 % and the error rate is 21.8 %. These value don't look to bad but on closer inspection the model falls apart for the purpose it was intended for. The model predicts a meager 388 defaults where only 166 are correct out of 12652 defaulters. The sensitivity of the model is 0,01312. With this low predicted defaults the model is basically saying "just give everyone loans" and is therefore not useful model.

A more conservative approach could be used by lowering the cutoff value of the predictive defaults so that more predictive defaults are obtained and therefore the depth increase and less loans will be disbursed disbursed. The default cutoff value is 0.5 when using **Proc logistic**. However when looking at the cdf of the defaults and non-defaults at 0.5 the empirical distributions are identical. We can see that that the maximum distance between the cdfs occurs at approximately 0.35. So we lower the cutoff value to 0.35 and to get the confusion matrix given in Figure 6.1 :

Table of loan_default by Pred			
	Pred(Predictive)		
loan_default(Actual)	0	1	Total
0	38736	6899	45635
1	9151	3501	12652
Total	47887	10400	58287

Figure 6.5: Confusion Matrix with cutoff = 0.35

We can clearly see that there are now more predicted defaults. There are now 3501 true positives. The sensitivity now improves to 0.276 and the specified will decrease. The accuracy of the model slightly decreases to 0.725. The lift is now 1.55, which means that we are 1.55 times more likely to predict a default than with random chance. The model can therefore be used predict defaults to a limited capacity. The cutoff value is usually determined by a profit matrix to optimize the profit. The financial institution using this model should take into consideration the profit generated from the interest rates on the loans and losses incurred when obligors default. The results are summarized in Table 6.1

	Performance Measure				
Cutoff probability	Specificity	Sensitivity	Accuracy	Positive predicted value	Negative PV
0.5	0.013	0.995	0.782	0.423	0.784
0.35	0.277	0.850	0.725	0.337	0.809

Table 6.1: Comparing accuracy ratios with different cutoffs

CHAPTER 7

CONCLUSION

A thorough process was followed in order to try and established a model that could predict whether or not a potential obligor will default on their vehicle loan. Various data manipulation techniques and statistical methods have been used to find a model that best fits the given data for the purpose intended. The data given had a lost of missing values which made using these predictor variables less useful and more difficult to work with. The predictor variables that were entirely filled in like Age and Loan to Value showed the most predictive power although predictive power was limited. A more complete data set will aid a lot further in the model development. Throughout the process various models were built and tested that were not included in the report. The variation between the models was not a lot and I decided to work on one relatively simple model to show the methodology for building a model accurately.

The aim of this project was achieved to build a predictive model. The accuracy of the model was limited in its capacity to classify defaults and non-defaults. The model can still be used as it gives better odds than randomly discarding potential obligors. The financial institution using this model should choose a cutoff probability of the predicted defaults that maximizes their profit, or limit their risk by taking a more conservative approach when disbursing loans.

REFERENCES

Baesens, B., Roesch, D. and Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons.

Conover, W. and Iman, R.L. (1980). The rank transformation as a method of discrimination with some examples. *Communications in Statistics-Theory and Methods*, vol. 9, no. 5, pp. 465–487.

Harrell Jr, F.E. (1997). Regression modeling and validation strategies. *Website: http://www. healthsystem. virginia. edu/internet/hes/biostat/docs/teaching/clinicians/model. pdf Retrieved*, vol. 8, no. 13, p. 04.

APPENDIX A

VARIABLE DESCRIPTION

Variable Name	Description
UniqueID	Identifier for customers
loan_default	Payment default in the first EMI on due date
disbursed_amount	Amount of Loan disbursed
asset_cost	Cost of the Asset
ltv	Loan to Value of the asset
branch_id	Branch where the loan was disbursed
supplier_id	Vehicle Dealer where the loan was disbursed
manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)
Current_pincode	Current pincode of the customer
Date_of_Birth	Date of birth of the customer
Employment_Type	Employment Type of the customer (Salaried/Self Employed)
DisbursalDate	Date of disbursement
State_ID	State of disbursement
Employee_code_ID	Employee of the organization who logged the disbursement
MobileNo_Avl_Flag	If Mobile no. was shared by the customer then flagged as 1
Aadhar_flag	If aadhar was shared by the customer then flagged as 1
PAN_flag	If pan was shared by the customer then flagged as 1

VoterID_flag	If voter was shared by the customer then flagged as 1
Driving_flag	If DL was shared by the customer then flagged as 1
Passport_flag	If passport was shared by the customer then flagged as 1
PERFORM_CNS_SCORE	Bureau Score
PERFORM_CNS_SCORE_DESCRIPTION	Bureau score description
PRI_NO_OF_ACCTS	Count of total loans taken by the customer at the time of disbursement,Primary accounts are those which the customer has taken for his personal use
PRI_ACTIVE_ACCTS	Count of active loans taken by the customer at the time of disbursement
PRI_OVERDUE_ACCTS	Count of default accounts at the time of disbursement
PRI_CURRENT_BALANCE	Total Principal outstanding amount of the active loans at the time of disbursement
PRI_SANCTIONED_AMOUNT	Total amount that was sanctioned for all the loans at the time of disbursement
PRI_DISBURSED_AMOUNT	Total amount that was disbursed for all the loans at the time of disbursement
SEC_NO_OF_ACCTS	Count of total loans taken by the customer at the time of disbursement, Secondary accounts are those which the customer act as a co-applicant or gaurantor
SEC_ACTIVE_ACCTS	Count of active loans taken by the customer at the time of disbursement,
SEC_OVERDUE_ACCTS	Count of default accounts at the time of disbursement

SEC_CURRENT_BALANCE	Total Principal outstanding amount of the active loans at the time of disbursement
SEC_SANCTIONED_AMOUNT	Total amount that was sanctioned for all the loans at the time of disbursement
SEC_DISBURSED_AMOUNT	Total amount that was disbursed for all the loans at the time of disbursement
PRIMARY_INSTAL_AMT	EMI Amount of the primary loan
SEC_INSTAL_AMT	EMI Amount of the secondary loan
NEW_ACCTS_IN_LAST_SIX_MONTHS	New loans taken by the customer in last 6 months before the disbursment.
DELINQUENT_ACCTS_IN_LAST_SIX_MON	Loans defaulted in the last 6 months
AVERAGE_ACCT_AGE	Average loan tenure
CREDIT_HISTORY_LENGTH	Time since first loan
NO_OF_INQUIRIES	Enquries done by the customer for loans

APPENDIX B

SAS OUTPUT

B.1 VARIABLE CLUSTERING

Cluster	Variable	1-RSqaure Ratio
Cluster 1	PRI_NO_OF_ACCTS	0.2914
	PRI_ACTIVE_ACCTS	0.1592
	NEW_ACCTS_IN_LAST_SIX_MONTHS	0.3156
Cluster 2	PERFORM_CNS_SCORE	0.2124
	Cns_Score	0.0265
	Per_CSD	0.2487
Cluster 3	Aadhar_flag	0.0765
	VoterID_flag	0.0749
Cluster 4	PRI_SANCTIONED_AMOUNT	0.0008
	PRI_DISBURSED_AMOUNT	0.0008
Cluster 5	Employ_Sal	0.0323
	Employ_Self	0.0324
Cluster 6	Ave_Acc_Age	0.1223
	Cre_His_Len	0.1329
Cluster 7	asset_cost	0.0000
Cluster 8	PRI_OVERDUE_ACCTS	0.3023
	DELINQUENT_ACCTS_IN_LAST_SIX_MON	0.2960
Cluster 9	disbursed_amount	0.0000
Cluster 10	NO_OF_INQUIRIES	0.0000
Cluster 11	manufacturer_id	0.0000
Cluster 12	SEC_INSTAL_AMT	0.0000
Cluster 13	PRIMARY_INSTAL_AMT	0.0000
Cluster 14	PAN_flag	0.0000
Cluster 15	Age	0.0000
Cluster 16	State_ID	0.0000
Cluster 17	PRI_CURRENT_BALANCE	0.0000
Cluster 18	ltv	0.0000

Figure B.1: Variable Clustering

B.2 SPEARMAN AND HOEFFDING RANK

Spearman and Hoeffding rank

Obs	Variable	Spearman rank of variables	Hoeffding rank of variables	Sperman Correlation	Spearman p-value	Hoeffdin Correlation	Hoeffding p-value
1	ltv	1	1	0.09782	<.0001	0.00122	<.0001
2	Per_CSD	2	3	0.09409	<.0001	0.00092	<.0001
3	disbursed_amount	3	2	0.09311	<.0001	0.00113	<.0001
4	State_ID	4	4	0.07177	<.0001	0.00062	<.0001
5	PRI_OVERDUE_ACCTS	5	12	0.04725	<.0001	0.00006	<.0001
6	VoterID_flag	6	11	0.04487	<.0001	0.00007	<.0001
7	NO_OF_INQUIRIES	7	13	0.04028	<.0001	0.00005	<.0001
8	PRI_DISBURSED_AMOUNT	8	6	-0.03993	<.0001	0.00015	<.0001
9	Age	9	5	-0.03840	<.0001	0.00018	<.0001
10	PRI_ACTIVE_ACCTS	10	8	-0.03753	<.0001	0.00013	<.0001
11	Ave_Acc_Age	11	7	-0.03718	<.0001	0.00014	<.0001
12	PRI_CURRENT_BALANCE	12	9	-0.03567	<.0001	0.00012	<.0001
13	manufacturer_id	13	10	-0.02790	<.0001	0.00009	<.0001
14	Employ_Sal	14	14	-0.02632	<.0001	0.00005	<.0001
15	asset_cost	15	15	0.01753	<.0001	0.00004	<.0001
16	PRIMARY_INSTAL_AMT	16	18	-0.01193	<.0001	0.00001	0.0237
17	SEC_INSTAL_AMT	17	16	-0.00742	0.0019	-0.00001	1.0000
18	PAN_flag	18	17	0.00232	0.3322	-0.00001	1.0000

Figure B.2: Spearman and Hoeffding rank

B.3 FIT STATISTICS USING BEST SUBSET SELECTION

Fit Statistics

Obs	model	AUC	AIC	BIC	MisClass	AdjRSquare	BrierScore
1	18	0.636869	176186.8	176378.2	0.2173	0.05943	0.16345
2	19	0.636878	176186.3	176387.8	0.2173	0.059451	0.163448
3	17	0.636716	176214.3	176395.6	0.2173	0.059181	0.163468
4	20	0.636892	176185.6	176397.1	0.2173	0.059475	0.163445
5	16	0.636609	176230.1	176401.3	0.2173	0.059029	0.163479
6	13	0.63638	176269.2	176410.2	0.2173	0.058647	0.163508
7	14	0.636532	176259.2	176410.2	0.2173	0.058749	0.163494
8	15	0.636457	176255	176416.2	0.2173	0.058801	0.163497
9	12	0.635661	176331.6	176462.6	0.2173	0.058101	0.163571
10	11	0.635338	176361.5	176482.4	0.2173	0.05783	0.16361
11	10	0.634953	176413.4	176524.1	0.2173	0.057374	0.163679
12	9	0.633341	176539.3	176640	0.2172	0.056288	0.163787
13	8	0.631157	176749	176839.6	0.2172	0.054489	0.163994
14	7	0.629314	176971.1	177051.7	0.2173	0.052582	0.164199
15	6	0.627241	177189.8	177260.3	0.2172	0.050703	0.164394
16	5	0.622887	177531.5	177591.9	0.2172	0.047771	0.164715
17	4	0.617535	177945.4	177995.8	0.2171	0.044215	0.165114
18	3	0.614182	178293.3	178333.6	0.2171	0.041217	0.165441
19	2	0.598327	179544.3	179574.6	0.2171	0.03043	0.166683
20	1	0.568494	181285	181305.2	0.2171	0.0153	0.16836

Figure B.3: Fit statistics using best subset selection

B.4 REGRESSION COEFFICIENTS OF SELECTED MODEL

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.7554	0.0750	101.5549	<.0001	
Per_CSD	1	0.7772	0.0292	707.0439	<.0001	0.1051
VoterID_flag	1	0.0986	0.0177	31.1807	<.0001	0.0191
Employ_Sal	1	-0.1860	0.0123	229.2086	<.0001	-0.0506
PRI_OVERDUE_ACCTS	1	0.2673	0.0201	176.1512	<.0001	0.0805
disbursed_amount	1	0.9717	0.0509	363.9946	<.0001	0.0699
NO_OF_INQUIRIES	1	0.1549	0.00792	382.8864	<.0001	0.0599
manufacturer_id	1	-0.00394	0.000274	206.7495	<.0001	-0.0482
Age	1	-0.00939	0.000633	220.0997	<.0001	-0.0508
State_ID	1	1.0032	0.0323	963.1501	<.0001	0.1051
ltv	1	0.0250	0.000625	1600.7242	<.0001	0.1579
disbursed*PRI_ACTIVE	1	-0.1352	0.00762	314.8991	<.0001	-0.0847
ltv*PRI_CURRENT_BALA	1	-0.00108	0.000152	50.8308	<.0001	-0.0447
PRI_OVERD*Ave_Acc_Ag	1	-0.00407	0.000581	49.0222	<.0001	-0.0405

Figure B.4: Regression coefficients of selected model