# ­Loan Default project Documentation

**Project Background:**

As an analytics consultant, you are tasked with the following individual project:

In spite of non-interest incomes on the rise, over half the money made by banks still comes from net interest earnings. A bank’s success heavily relies upon how many loans it can give out while maintaining low default rates, where default means the inability of the borrowers to pay back the loan in time. Provide Mammoth Bank with a strategy to predict customers who default or not. Use the loan\_default dataset for this assignment.

Clearly presenting results to clients or stakeholders is a critical part of your job. Therefore, please write in complete sentences for our client and clearly label and introduce your figures and tables. You will lose points not only for getting the answer incorrect but for incomplete sentences and not explaining figures. Provides answers below each question.

**Project Objective and introduction (using question and answer format):**

1. Introduction:
   1. The purpose of this analysis is to provide Mammoth Bank with a strategy to predict which customers would default on a loan and which would not.
      1. What is the business problem?  
         Mammoth Bank is seeking a way to predict customers who would default on their loan versus customers who would not default on their loan. This strategy would enable Mammoth Bank give out more loans confidently and to the right kind of customers, who have a very low probability of defaulting on their loan. The purpose of this analysis is to provide Mammoth Bank with the right strategy so they can loan out more money while maintaining a low default rate.
   2. What did you do?
      1. After assessing the dataset, I began my predictive analysis process with Exploratory data analysis and getting a summary statistic of my data. I access that the response variable in my dataset was categorical (filled with 1s and 0s), 1 for loan customers who defaulted and 0s for loan customers who did not default. After the initial exploratory data analysis and summary statistics on my data, I partitioned my data and prepared it for model training and testing. I partitioned my data into two halves 80/20, 80 for my training dataset and 20 for my testing dataset. I used a couple of modeling techniques on my training dataset and then used the models I had trained to make predictions on my test dataset. The modeling techniques I used for training my model were, subset selection’s backward elimination and forward selection and LASSO model.   
           
         After this, I evaluated my models and compared which was the best model to use in predicting whether a loan customer would default on their loan or not using the AUC values I got for each model.

**Exploratory Data Analysis:**

1. Exploratory data analysis
   1. Data Contents:
      1. What is the response variable?  
         The ‘Default’ Variable is the response variable.
      2. What are the number of observations and predictor variables?

There are 1000 observations and 15 predictor variables.

* 1. Explain Exploratory Data Analysis findings.
     1. Summary Statistics:
        1. Are there missing values or outliers? (List variables with either)  
           There are no missing values in our dataset. However, there are outliers in the following variables of our dataset.
* Checking\_amount
* Term
* Credit\_score
* Amount
* Saving\_amount
* Age
* No\_of\_credit\_acc
  + - 1. Provide the summary from the skim function in R.

A screenshot of a computer

Description automatically generated

* + - 1. How many categorical variables and numerical variables are there?   
         There are 8 categorical variables and 8 numerical variables in our dataset. The response variable, ‘Default’ is a categorical variable as well, and is included in the count of 8 categorical variables.
    1. How many 1s and 0s are there in the response variable? (make a table for this one and also report how many 1s and 0s are in the response variable)  
        A number with numbers in black

       Description automatically generated with medium confidence  
       The response variable ‘default’ is a categorical variable with two entries. O for those who did not default on their loan and 1 for those who defaulted on their loan. Based off of the table above there are 700 customers who did not default on their loan from Mammoth Bank and 300 customers who defaulted on their loan. Meaning out of 1000 customers who were loaned money by Mammoth Bank 30% of them defaulted on their loans and 70% did not.



* + 1. Provide two boxplots of a continuous variable grouped by the response variable and describe any insights.

A graph of a box diagram

Description automatically generated with medium confidence



The chart above shows the distribution of the checking account balances for all loan customers of Mammoth Bank by loan defaults. It shows the distribution of the checking account balances for those who defaulted on their loans represented by 1 and those who did not default on their loan represented by 0. From the chart we can see that the normal bound of the checking account balance for customer who did not default on their loan (0) is generally higher than that of those who defaulted on their loan (1).

A graph of a credit score

Description automatically generated

The box plot above shows the distribution of the credit scores for all the loan customers of Mammoth Bank in our dataset, those who defaulted on their loan denoted by 1 and those who did not default on their loan denoted by 0. From the chart we can observe that the normal bounds of the credit score of those who did not default on their loan (0) is higher than that of those who defaulted on their loans.

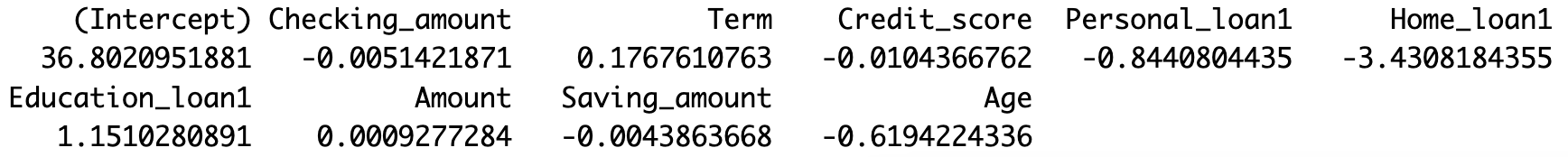
**Modeling and Predictive analysis:**

1. Modeling and results
   1. Generalization Approach
      1. Use 80% of the data as training data with 5-fold cross validation, and 20% of the data as testing data. Describe why training data with cross validation and testing data are used as part of your modeling process?
   2. Model
      1. Subset Selection Methods
         1. Fit forward and backward selection.
         2. What is the best model for each algorithm?

A screenshot of a computer screen

Description automatically generated

The model above is the final model with the lowest AIC from the subset selection’s backward elimination approach. This final model provides a good balance between model fit and complexity, as measured by AIC.

  
The above is a table showing the coefficients of the variables of the final model from the backward elimination approach.

A screenshot of a computer screen

Description automatically generated

The final forward selection model above is the model with the lowest AIC.

* + 1. LASSO
       1. Fit a LASSO logistic regression model to the data and tune the model with alpha =1 and lambda = seq(0.0001, 1, length = 20).

A screenshot of a computer code

Description automatically generated

* + - 1. What is the best lambda value and describe why lasso prevents overfitting compared to least squares.

**The best lambda value:**

The best lambda value is the one that provides the optimal trade-off between model complexity (sparsity) and goodness of fit. In the case of our data analysis the output shows that the best lambda value, selected based on accuracy, is ‘0.0001’.

**Why LASSO prevents overfitting compared to least squares:**

LASSO (L1 regularization) adds a penalty term to the linear regression objective function, which includes the absolute values of the coefficients. This penalty encourages some of the coefficients to be exactly zero, leading to a sparse model. By shrinking some coefficients to zero, LASSO effectively performs feature selection, which helps prevent overfitting. In contrast, traditional least squares regression does not have this feature selection property and may include all available predictors, potentially leading to overfitting when there are many predictors and a limited sample size.

* 1. Interpret results (5 sentences or more)

The model was trained using the ‘glmnet’ package, which implements elastic net regularization. The result shows the performance metrics (Accuracy and Kappa) across different values of the tuning parameter lambda. The tuning parameter ‘alpha’ was held constant at 1, indicating pure LASSO regularization.

The cross-validation accuracy and kappa values are provided for each lambda value. The optimal model is selected based on the largest accuracy.

The output suggests that as lambda increase, the model becomes more regularized, and accuracy decreases. The best lambda, 0.0001, is chosen as it maximizes accuracy while preventing overfitting.

The final model, with alpha=1 and lambda=0.0001, is the one selected based on the cross-validated performance metrics.

In summary, the LASSO logistic regression model with the chosen lambda value is expected to provide a good balance between accuracy and simplicity, preventing overfitting by penalizing the absolute values of the coefficients and encouraging sparsity in the model. The final model can be used for predictions, and the selected lambda represents the strength of the regularization applied.

* + 1. Explain an ROC curve, and why it is used in classification problems.

An ROC (Receiver Operating Characteristic) curve is a graphical representation that illustrates the performance of a binary classification model across different discrimination thresholds. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for various threshold values. The curve visually shows the trade-off between correctly classifying positive instances (true positives) as the decision threshold of the model varies. An ROC is good for a classification model because it provides a comprehensive view of a classification model’s performance by visualizing its ability to discriminate between classes at various decision thresholds. It is a powerful tool for model evaluation, selection and customization based on specific application requirements. Also, an ROC is used for classification problems because it provides a comprehensive and visually intuitive way to assess and understand the performance of a binary classification model.

* + 1. Provide the ROC curve for the testing set as shown below (only fill out the chart below). – the ROC curve for LASSO, forward & backward

A graph with numbers and lines

Description automatically generated

**Forward selection ROC curve**

A graph of a positive rate

Description automatically generated

**Backward elimination ROC curve**

A graph with a white background

Description automatically generated with medium confidence

**LASSO Logistic Regression**

|  |  |
| --- | --- |
| Method | Testing set AUC |
| LASSO Logistic regression | 0.982381 |
| Forward Selection | 0.9827381 |
| Backward Elimination | 0.982619 |

**Recommendation and Closing:**

1. What are your recommendations to the client?

Based on the output from our analysis above we can clearly see that the following predictor variables have the most impact on predicting the outcome of whether a customer would default on their loan or not:

* Age
* Checking amount
* Saving amount
* Credit score
* Home loan
* Personal loan
* Term
* Education loan
* Amount

I would recommend that the lending team focused on using the above listed variables and models (the forward selection model has the highest testing AUC, however the three models engaged performed relatively similar in the AUC testing) in predicting the right customers to loan money to base on their likelihood to pay back the loan and not default on their loan.

This method (using a model to predict whether a loan customer would potentially default on their loan or not) would ensure a safer way (statistically more curated way) for Mammoth Bank to give out more loans while maintaining very low loan default rates. This is because Mammoth bank would only be giving loans to customers with the highest likelihood of paying back and not defaulting on their loans.