

|  |  |  |
| --- | --- | --- |
| **Module Code** | **:** | CT127-3-2-PFDA – PROGRAMMING FOR DATA ANALYSIS |
| **Intake Code** | **:** | APU2F2402CS(DA)/APD2F2402CS(DA) |
| **Lecturer Name** | **:** | MINNU HELEN JOSEPH |
| **Hand in Date** | **:** | 3RD JUNE 2024, MONDAY, 11:59PM |
| **Lab No.** | **:** | LAB 10 |
| **Group No.** | **:** | GROUP 32 |
| **Group Leader** | **:** | CHURILOV MIKHAIL |

|  |  |
| --- | --- |
| **Student ID** | **Student Name** |
| TP067853 | AZELEA GLORY NG ZI-LIN |
| TP072847 | CHURILOV MIKHAIL |
| TP074470 | JANE LEE |
| TP068019 | RAVIN A/L KANAGARAJAN |

Table of Contents

[1.0 Introduction 4](#_Toc168343255)

[1.1 Data Description 4](#_Toc168343256)

[1.2 Assumption 4](#_Toc168343257)

[1.3 Hypothesis 4](#_Toc168343258)

[1.4 Objectives 4](#_Toc168343259)

[2.0 Data Preparation 5](#_Toc168343260)

[2.1 Data Import 5](#_Toc168343261)

[2.2 Cleaning / Pre-Processing 5](#_Toc168343262)

[2.3 Data Validation 19](#_Toc168343263)

[3.0 Data Analysis 20](#_Toc168343264)

[3.1 Azelea Glory Ng Zi-Lin – The impact of Outstanding Debt on having Lower Credit Score 20](#_Toc168343265)

[3.1.1 Impact of Count of Poor Credit Scores by Outstanding Debt 20](#_Toc168343266)

[3.1.2 Impact of Proportion of Poor Credit Scores by Outstanding Debt 21](#_Toc168343267)

[3.1.3 Gradient Boosting Model: Predicting Low Credit Scores Based on Outstanding Debt 23](#_Toc168343268)

[3.1.4 Model Performance Metrics for Gradient Boosting 25](#_Toc168343269)

[3.1.5 Conclusion 29](#_Toc168343270)

[3.2 Mikhail Churilov – The impact of Delayed Payments on having Lower Credit Score 30](#_Toc168343271)

[3.2.1 Logistic Regression Model Output of Days of late payments and the number of later payments 30](#_Toc168343272)

[3.2.2 Centered Logistic Regression Model Output Of delay from due date and number of delayed payments 32](#_Toc168343273)

[3.2.3 The Impact of Payment Delay and Number of Delayed Payments on Probability of Low Credit Score 34](#_Toc168343274)

[3.3 Jane Lee – The impact of Payment Behavior on having Lower Credit Score 36](#_Toc168343275)

[3.3.1 What is the impact of different payment behaviors on credit scores? 36](#_Toc168343276)

[3.3.2 How does the combination of salary range and payment behavior influence average credit scores? 38](#_Toc168343277)

[3.3.3 How effective is the logistic regression model in predicting credit scores based on payment behavior and other financial factors? 40](#_Toc168343278)

[3.3.4 Extra Features 42](#_Toc168343279)

[3.3.5 Conclusion 43](#_Toc168343280)

[3.4 Ravin A/L Kanagarajan – The impact of Credit History age on having Lower Credit Score 44](#_Toc168343281)

[3.4.1 What is the impact of Poor count of Credit score 44](#_Toc168343282)

[3.4.2 What is the impact proportion of Poor Credit Scores by Credit History Duration 45](#_Toc168343283)

[3.4.3 How effective is the logistic regression modelling on predicting the probability of low credit score by credit history length 46](#_Toc168343284)

[3.4.4 Probability of Low Credit score by Credit History Length 48](#_Toc168343285)

[4.0 Findings 50](#_Toc168343286)

[5.0 Recommendations 51](#_Toc168343287)

[6.0 Limitations and Future Directions 52](#_Toc168343288)

[7.0 Conclusion 53](#_Toc168343289)

[8.0 Workload Matrix 54](#_Toc168343290)

[9.0 References 54](#_Toc168343291)

# 1.0 Introduction

## 1.1 Data Description

In this dataset it talks about the customer details and their credit behaviors. It has a total of 28 columns and 100001 rows in this dataset. All different columns have their own data type including characters, integer, or numeric values. The customer details in this dataset include their unique­\_identities, name, age, ssn (social security number) and occupation. Other than personal information, it also has each of the customer’s credit behaviors inside the dataset including number of accounts, outstanding debt, credit utilisation ratio and the credit history age. Additionally, it includes the monthly financial data on the EMI payments, in-hand salaries, and investments. At the same time, it has captured the qualitative aspects of customer payment behavior and credit scores by categorizing with Good, Standard and Poor. By having this dataset, it is crucial for analysts to examine the factors that influence the credit scores and understand the customer’s financial habits.

## 1.2 Assumption

-The dataset contains missing values

-The dataset contains inconsistency data

-The dataset contains special character and symbols

-The dataset contains outlier

-The data type is not appropriate

-The dataset contains illogical data

## 1.3 Hypothesis

Customers who have a higher number of delayed payments, a longer credit history age, a lower payment behavior score, and higher outstanding debt are likely to have a lower credit score.

## 1.4 Objectives

-The impact of credit history age on having lower credit score.

-The impact of payment behavior on having lower credit score.

-The impact of delayed payments on having lower credit score.

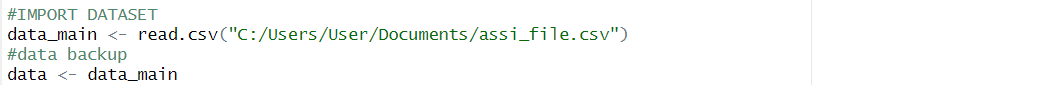
-The impact of outstanding debt on having lower credit score.

2.0 Data Preparation

2.1 Data Import

A white background with black dots

Description automatically generated



*Figure 2.1: Importing the Dataset*

Figure 2.1 shows the first few steps we’ve done, which involves importing the dataset. Before importing the dataset, we load essential libraries such as tidyr and caret. These libraries provide an extensive collection of R packages designed for data science tasks, including data manipulation, visualization, and modelling. The read.csv function is used to read the dataset from a CSV file. The dataset is stored in a variable named data, making it accessible for subsequent data processing and analysis.

Importing the dataset is a crucial first step in any data analysis project. It sets the foundation for all subsequent data processing, cleaning, and analysis tasks. Ensuring that the data is correctly imported and accessible allows for efficient and accurate data manipulation, which is vital for obtaining reliable results and insights from the analysis.

2.2 Cleaning / Pre-Processing

Lowercase all column names



*Figure 2.2.1: Lowercase All Column Names*

Cleaning of all columns

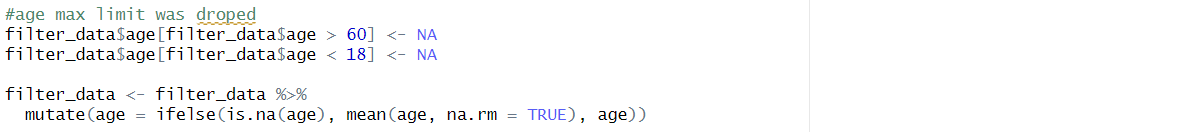
**AGE:**

A computer code on a white background

Description automatically generated

A white background with black text

Description automatically generated



*Figure 2.2.2: Cleaning age Column*

In the age column, the range must be between 18 and 60 and if is not inside this value will be changed to NA values. The reason that the age must be between 18 and 60 is because, in a logical way, below 18 usually is impossible to earn much money and over 60 is usually a retired age.

**SSN:**

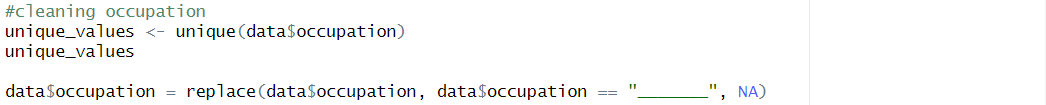
A computer screen with text

Description automatically generated with medium confidence

*Figure 2.2.3: Cleaning ssn Column*

Figure 2.2.3 defines a function to validate SSNs in a dataset by checking if they match the standard format of "XXX-XX-XXXX". It replaces any invalid SSNs with NA, ensuring that the dataset only contains valid SSNs or missing values. This cleaning process is crucial for maintaining data integrity and ensuring that subsequent analyses or operations on the dataset are performed on valid data.

**OCCUPATION:**



*Figure 2.2.4: Cleaning occupation Column*

Figure 2.2.4 does a cleaning operation on the occupation column of the dataset. It first identifies and prints the unique values in the occupation column to inspect the distinct entries present in the data. Then, it replaces all instances of the placeholder value "\_\_" with NA. This process ensures that the dataset has NA values for missing or placeholder occupations, which is essential for maintaining data integrity and preparing the dataset for further analysis.

**ANNUAL\_INCOME:**

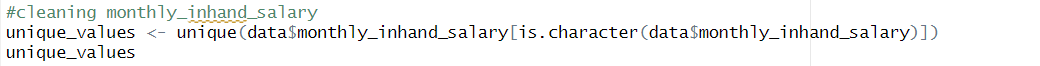
A computer code on a white background

Description automatically generated

*Figure 2.2.5: Cleaning annual\_income Column*

Figure 2.2.5 processes the annual\_income column in the dataset to ensure it contains valid numeric values. It removes any non-numeric characters from each entry, checks if the remaining string contains at least one digit, and converts valid entries to numeric type. Invalid entries, which are either empty or contain no digits after cleaning, are replaced with NA. This cleaning process ensures that the annual\_income column contains only valid numeric values, facilitating further analysis and operations on the dataset.

**MONTHLY\_INHAND\_SALARY:**



*Figure 2.2.6: Cleaning monthly\_inhand\_salary Column*

Figure 2.2.6 shows how we extract unique values from the column that are of character type and prints them to the console for inspection. This process helps identify any non-numeric entries in the monthly\_inhand\_salary column, which can then be further processed or cleaned to ensure data integrity. Identifying unique character values is useful for detecting and handling anomalies or inconsistent data entries before performing more complex cleaning or analysis tasks.

**TYPE\_OF\_LOAN:**



*Figure 2.2.7: Cleaning type\_of\_loan Column*

Figure 2.2.7 extracts and prints the unique values present in this column. By identifying the distinct entries, this process aids in detecting any anomalies or inconsistencies that need to be addressed in subsequent data cleaning steps.

**NUM\_BANK\_ACCOUNTS:**

A close up of a computer screen

Description automatically generated



*Figure 2.2.8: Cleaning num\_bank\_acc Column*

Figure 2.2.8 is designed to clean the num\_bank\_accounts column in a dataset by addressing both character type anomalies and numerical outliers. The first snippet identifies and inspects unique character values in the num\_bank\_accounts column, then replaces any negative values with NA. The second snippet replaces values that are greater than or equal to 15 with NA, ensuring that only reasonable counts of bank accounts are retained.

**NUM\_OF\_LOAN:**

A computer code on a white background

Description automatically generated

A computer code with numbers

Description automatically generated with medium confidence

*Figure 2.2.9: Cleaning num\_of\_loan Column*

Figure 2.2.9 is designed to clean the num\_of\_loan column in a dataset by addressing non-numeric characters, converting valid numeric parts to integers, and filtering out extreme values. The first snippet extracts unique values and processes each element to remove non-numeric characters and convert the remaining numeric part to an integer, replacing invalid entries with NA. The second snippet provides summary statistics, visualizations, and filters out extreme values (loans greater than or equal to 15), replacing them with NA.

**NUM\_OF\_DELAYED\_PAYMENT:**

A computer code with black text

Description automatically generated with medium confidence

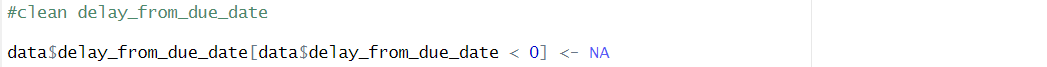
A screen shot of a computer code

Description automatically generated

*Figure 2.2.10: Cleaning num\_of\_delayed\_payment Column*

Figure 2.2.10 is designed to clean the num\_of\_delayed\_payment column in a dataset by addressing non-numeric characters, converting valid numeric parts to integers, and filtering out extreme values. The first snippet extracts unique values, processes each element to remove non-numeric characters, and converts the remaining numeric part to an integer, replacing invalid entries with NA. It then sorts the unique values in ascending order. The second snippet provides summary statistics, visualizations, and filters out extreme values (delayed payments greater than 60), providing a subset with reasonable numbers of delayed payments for further analysis.

**DELAY\_FROM\_DUE\_DATE:**



*Figure 2.2.11: Cleaning delay\_from\_due\_date Column*

Figure 2.2.11 is designed to clean the delay\_from\_due\_date column in a dataset by addressing invalid values, specifically negative values. Negative values in the context of delays are typically invalid as delays are generally non-negative. By replacing these invalid negative values with NA, the code ensures that the delay\_from\_due\_date column contains only valid entries.

**CHANGED\_CREDIT\_LIMIT:**

A computer code with text

Description automatically generated

*Figure 2.2.12: Cleaning changed\_credit\_limit Column*

Figure 2.2.12 processes the changed\_credit\_limit column in the dataset to ensure it contains valid numeric values. It removes any non-numeric characters from each entry, checks if the remaining string contains at least one digit, and converts valid entries to numeric type. Invalid entries, which are either empty or contain no digits after cleaning, are replaced with NA. This cleaning process ensures that the changed\_credit\_limit column contains only valid numeric values, facilitating further analysis and operations on the dataset.

**CREDIT\_MIX:**

A close-up of a logo

Description automatically generated

*Figure 2.2.13: Cleaning credit\_mix Column*

Figure 2.2.13 first extracts and prints the unique values in the credit\_mix column to inspect the distinct entries present in the data. Then, it replaces all instances of the placeholder value "\_" with NA. Finally, it counts the number of NA values in the credit\_mix column, providing a summary of the missing values after the cleaning process. This cleaning process ensures that the credit\_mix column contains valid entries.

**OUTSTANDING\_DEBT:**

A computer code on a white background

Description automatically generated

*Figure 2.2.14: Cleaning outstanding\_debt Column*

Figure 2.2.14 processes the outstanding\_debt column in the dataset to ensure it contains valid numeric values. It removes any non-numeric characters from each entry, checks if the remaining string contains at least one digit, and converts valid entries to numeric type. Invalid entries, which are either empty or contain no digits after cleaning, are replaced with NA. This cleaning process ensures that the outstanding\_debt column contains only valid numeric values.

**CREDIT\_HISTORY\_AGE:**

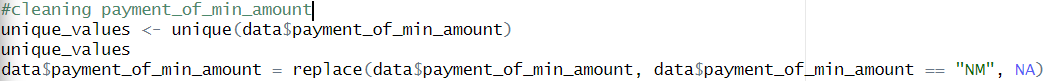
A computer screen shot of a code

Description automatically generated

*Figure 2.2.15: Cleaning credit\_history\_age Column*

Figure 2.2.15 processes the credit\_history\_age column in the dataset to convert it from a string format representing years and months into a numeric format representing the total number of months. The custom function convert\_time\_str\_to\_months performs the conversion, and the sapply function applies this conversion to each element in the column. Finally, the column is renamed to credit\_history\_months to accurately describe the transformed data. This cleaning and transformation process ensures that the credit\_history\_age column is in a consistent numeric format.

**PAYMENT\_OF\_MIN\_AMOUNT:**



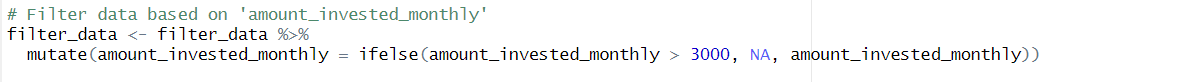
*Figure 2.2.16: Cleaning payment\_of\_min\_amount Column*

Figure 2.2.16 shows a cleaning operation on the payment\_of\_min\_amount column of a dataset. It first extracts and prints the unique values in the payment\_of\_min\_amount column to inspect the distinct entries present in the data. Then, it replaces all instances of the placeholder value "NM" with NA.

**AMOUNT\_INVESTED\_MONTHLY:**

A computer code with black text

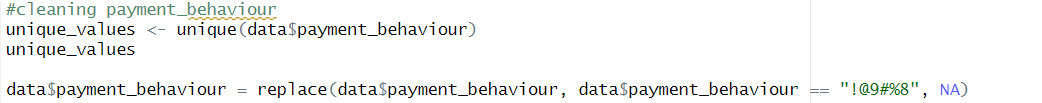
Description automatically generated with medium confidence



*Figure 2.2.17: Cleaning amount\_invested\_monthly Column*

Figure 2.2.17 addresses non-numeric characters, converting valid numeric parts to numeric type, and filtering out extreme values. The first snippet extracts numeric values, converts them to numeric type, and replaces invalid entries with NA. The second snippet filters out extreme values (amounts greater than 3000), replacing them with NA.

**PAYMENT\_BEHAVIOUR:**



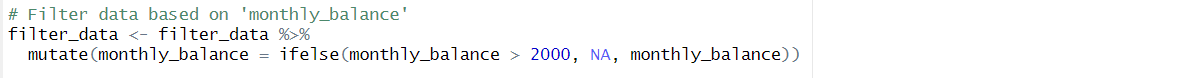
*Figure 2.2.18: Cleaning payment\_behaviour Column*

Figure 2.2.18 first extracts and prints the unique values in the payment\_behaviour column to inspect the distinct entries present in the data. Then, it replaces all instances of the erroneous value "!@#$%" with NA.

**MONTHLY\_BALANCE:**

A computer code on a white background

Description automatically generated



*Figure 2.2.19: Cleaning monthly\_balance Column*

Figure 2.2.19 designed to clean the monthly\_balance column in a dataset by addressing non-numeric characters, converting valid numeric parts to numeric type, and filtering out extreme values. The first snippet extracts numeric values, converts them to numeric type, and replaces invalid entries with NA. The second snippet filters out extreme values (monthly balances greater than 2000), replacing them with NA. These cleaning steps ensure that the monthly\_balance column contains valid and reasonable numeric values.

**CREDIT\_SCORE:**



A computer screen shot of a computer code

Description automatically generated

*Figure 2.2.20: Cleaning credit\_score Column*

Figure 2.2.20 first snippet extracts and prints the unique values in the credit\_score column to inspect the distinct entries present in the data. The second snippet transforms the credit\_history\_age column from a string format representing years and months to a numeric format representing the total number of months using a custom function. The column is then renamed to credit\_history\_months to accurately describe the transformed data. These cleaning and transformation processes ensure that the columns contain valid entries.

**INTEREST\_RATE:**

A computer code with text

Description automatically generated with medium confidence

*Figure 2.2.21: Cleaning interest\_rate Column*

Figure 2.2.21 performs summary and filtering operations on the interest\_rate column of a dataset. It first generates a summary of the interest\_rate values and then creates a subset of the data where interest\_rate is less than 50. A boxplot is generated for this subset to visually inspect the data distribution. The code then replaces extreme values (interest rates greater than or equal to 50) with NA in the original dataset and provides a summary of the modified interest\_rate column. These steps help in cleaning the data by removing outliers, ensuring that the interest\_rate column contains valid and reasonable values.

**NUM\_CREDIT\_INQUIRIES:**

A computer code with numbers

Description automatically generated with medium confidence

*Figure 2.2.22: Cleaning num\_credit\_inquiries Column*

Figure 2.2.22 performs summary and filtering operations on the num\_credit\_inquiries column of a dataset. It first generates a summary of the num\_credit\_inquiries value and then creates a subset of the data where num\_credit\_inquiries are less than 20. A boxplot and a histogram are generated for both the original and the filtered data to visually inspect the data distribution. The code then replaces extreme values (num\_credit\_inquiries greater than or equal to 20) with NA in the original dataset and provides a summary of the modified num\_credit\_inquiries column.

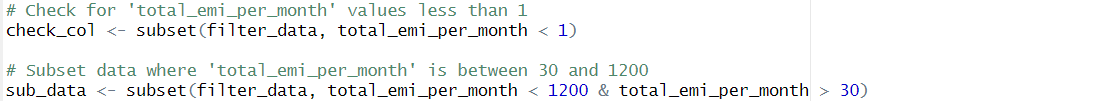
**NUM\_CREDIT\_CARD:**

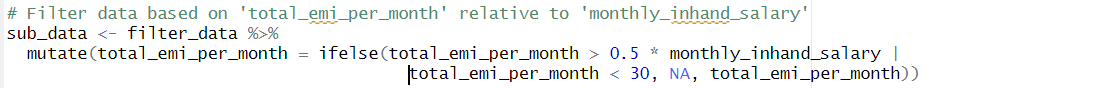


*Figure 2.2.23: Cleaning num\_credit\_card Column*

Figure 2.2.23 performs a filtering operation on the num\_credit\_card column of a dataset. It replaces extreme values (num\_credit\_card greater than or equal to 20) with NA.

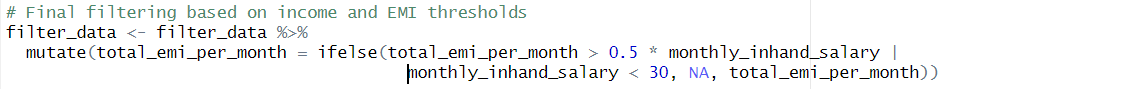
**TOTAL\_EMI\_PER\_MONTH:**





A white background with text

Description automatically generated



*Figure 2.2.24: Cleaning total\_emi\_per\_month Column*

Figure 2.2.24 performs a comprehensive series of cleaning and filtering operations on the total\_emi\_per\_month column of a dataset to ensure data integrity and facilitate accurate analysis. Initially, the code identifies and isolates rows where the total\_emi\_per\_month value is less than 1, helping to pinpoint potentially erroneous data entries that might have been recorded incorrectly or serve as placeholders.

Next, the dataset is filtered to include only rows where the total\_emi\_per\_month is between 30 and 1200. This range is chosen to focus the analysis on EMI values that are considered more realistic and manageable, excluding extreme outliers that could skew the analysis. Visual tools such as boxplots and histograms are then used to inspect the distribution of total\_emi\_per\_month values both before and after filtering. These visualizations help in understanding the spread and central tendency of the data, as well as identifying any remaining outliers.

The code further refines the dataset by applying a relative threshold based on the monthly\_inhand\_salary. It replaces values in the total\_emi\_per\_month column that exceed 50% of the monthly\_inhand\_salary or are less than 30 with NA. This ensures that the EMI values are proportionate to the individual’s income, maintaining financial realism.

A specific subset of customers with exceptionally high EMI payments is identified using advanced statistical methods. The filter selects rows where the total\_emi\_per\_month exceeds three times the interquartile range (IQR) above the third quartile (75th percentile). This step helps in isolating outliers for further context analysis, providing insights into potentially high-risk customer segments.

Finally, the code applies a final filtering step to ensure consistency and data integrity. It rechecks and replaces values in the total\_emi\_per\_month column that are greater than 50% of the monthly\_inhand\_salary or less than 30 with NA. This process eliminates outliers and ensures that the data is reliable.

Replace NA values using Median

A screenshot of a computer program

Description automatically generated

*Figure 2.2.25: Replacing NA Using Median*

Figure 2.2.25 performs a series of cleaning operations to handle missing values in a dataset. Initially, it fills missing values in the age column with the median age. Then, it defines a list of columns that require processing to handle missing values. A custom function, fill\_na\_mean, is applied to each of these columns to replace NA values with the mean of the respective column. Finally, the summary of the dataset is generated to verify that the cleaning process was successful.

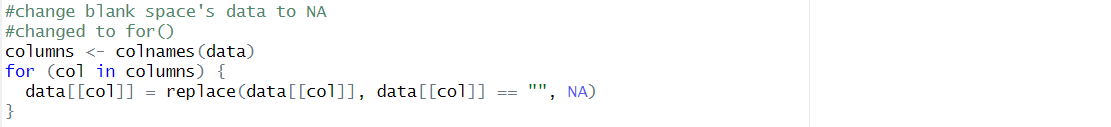
Remove Unused Attributes



*Figure 2.2.26: Drop Name from dataset*

In Figure 2.2.26, the name attribute has been removed from the dataset. Due to the fact that name in not useful in the analysis, the main of the attribute that used is the customer\_id to identify the customer so we dropped the attribute.

Change Blank Space to NA



*Figure 2.2.27: Change Blank Space’s Data to NA*

Figure 2.2.27 replaces blank spaces in a dataset with NA values, ensuring that missing data is correctly handled. Initially, it retrieves the names of all columns in the dataset. Then, it iterates over each column, replacing any blank spaces with NA. This process standardizes the representation of missing values, making the dataset more consistent and suitable for further analysis.

Use Mode to replace NAs

A screenshot of a computer code

Description automatically generated

*Figure 2.2.28: Fill in NA values using Mode*

Figure 2.2.28 had replace some of the NA values by using Mode. By replacing NA values with Mode function, is to ensure that missing data will be filled with the most common occurrence in the particular column, this will help to maintain the integrity of the dataset and reduces the impact of missing values.

Remove rows that has more than 4

A screenshot of a computer program

Description automatically generated

*Figure 2.2.29: Removing Rows That Has More Than 4*

Figure 2.2.29 performs data visualization and filtering operations to handle rows with missing values. Initially, a donut chart is created to visualize the distribution of NA values per row using the ggplot2 package. This chart provides a clear visual representation of the number of rows with varying counts of NAs. A histogram is then generated to further illustrate the frequency distribution of NA values per row. Finally, the dataset is filtered to remove rows that exceed a threshold of 4 NA values, ensuring that the remaining data is completer and more suitable for further analysis. These steps collectively help in understanding and addressing the presence of missing values in the dataset, thereby enhancing data quality and reliability for subsequent analyses.

A colorful circle with text

Description automatically generated

*Figure 2.2.30: Donut Chart of Distribution of NA Values Per Row*

2.3 Data Validation

A screenshot of a computer screen

Description automatically generated

*Figure 2.3.1: Summary of data after Pre-processing*

Based on Figure 2.3.1, it shows the range and central tendency of each attribute after data has been cleaned. Missing values (NA) have also been removed to maintain the integrity of the dataset.

A close-up of a white background

Description automatically generated

*Figure 2.3.2: Structure of data after pre-processing*

Figure 2.3.2 presents the structure of the dataset after the pre-processing steps. This figure confirms that the data has been cleaned successfully and is properly structured, and makes it suitable for subsequent analytical procedures.

3.0 Data Analysis

3.1 Azelea Glory Ng Zi-Lin – The impact of Outstanding Debt on having Lower Credit Score

3.1.1 Impact of Count of Poor Credit Scores by Outstanding Debt

A graph of a credit score

Description automatically generated

*Figure 3.1.1: Count of Poor Credit Scores by Outstanding Debt*

In this bar chart in Figure 3.1.1, the amount of consumers whose credit scores fell under the poor rating has been categorised according to the range of the outstanding amount they need to pay. A bar plot has been used with vertical axis representing each range of the outstanding debt where on the count of poor credit scores has been plotted on the vertical axis. As predicted, the percentage of consumers with poor credit scores is highest among the consumers with outstanding balance between $1500 and $2500.

A peak can be seen in the mid-range debt. This means that borrowers within the bracket of $1500 and $2500 have poor credit score rating hence its high proportion. This could be because of the inability to manage finances well, or they are broke and cannot control their debts adequately. A decline can also be seen beyond $2500. The reduction in the number of people with poor credit ratings for debts above $2500 could mean that the great many of these high amount debts do not exist or that those with these credits are handling their credits better. Suppose it may also imply that borrowers having more debt have better plan or some other resources of financial structure.

This figure is rather important as it shows the range of debt, within which people are most likely to experience worsening of credit scores, thus it corresponds to the focus area of credit counselling and financial support agencies.

3.1.2 Impact of Proportion of Poor Credit Scores by Outstanding Debt

A graph of a credit score

Description automatically generated

*Figure 3.1.2: Proportion of Poor Credit Scores by Outstanding Debt*

The bar chart in Figure 3.1.2 displays the comparison of poor credit score in relation to the outstanding debt range. The vertical axis signifies percent poor credit scores, and the horizontal axis illustrates the average of outstanding debts. Evaluating the data, it is necessary to state that the largest percentage of individuals with a poor credit score is observed in the debt interval of $1500-$2000, which is 46.5%.

High proportion in specific debt range can be observed. By evaluating the credit ranks of individuals belonging to the $1500 to $2000 debt range, it was found that roughly half of the consumers in this range have unfavourable credit records. This is helpful as it assesses the severity of debt beyond which an individual is most likely to be damaging their credit scores further. A decreasing proportion can also be seen with higher debt. Above $2500, the percentage of people with poor credit scores drops as the amount that is borrowed increases. This can imply that higher levels of debt can reflect efficient credit usage or perhaps even better management of credit or finance in general given better income or financial resource availability.

Knowing the distribution of poor credit score by debt amount assists in determining poor credit score from various debt classes that are most prone to risk, hence, may be useful in intervention strategies.

3.1.3 Gradient Boosting Model: Predicting Low Credit Scores Based on Outstanding Debt

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

*Figure 3.1.3: Gradient Boosting Model for Predicting Low Credit Scores Based on Outstanding Debt*

Figure 3.1.3 consists of code and result about applying a gradient boosting model to forecast the probability of having a low credit score while considering outstanding debt. The code shows the construction of the gradient boosting model, the model’s assessment, and the prediction and visualisation of the model’s performance.

The code creates a gradient boosting model, which aims at estimating the likelihood of a low credit score given the outstanding debt amount. First of all, there is the specification of the initial hypothesis, while the null hypothesis (H0) stated that outstanding deposit has no relationship with poor credit score probability. Again, the gbm function from the gbm package is used and the model is constructed to comprise the credit\_score\_binary as the response variable and outstanding\_debt as the predictor variable. The chosen model then uses the following parameters to be trained on the filter\_data dataset; distribution “bernoulli” and 100 trees. We train the model on the training data and use the trained model to predict probabilities on the training set and store the predicted probabilities in the variable ‘predicted\_prob\_gbm’. These are then presented using ggplot2 in order to estimate and present the probabilities in the form of a scatter plot, with outstanding debt on the X-axis, and the predicted probability on the Y axis, where a blue line is used to plot the least square regression line.

Thereafter, the code checks the model performance on the training and testing datasets that form part of the data partitioning. Thus, for the given class, performance evaluation is made on both the training set and the test set by constructing the model with the outputs that are mathematically fitted to the training data and then predict on the sets. These predictions help in determining the performance measures with reference to accuracy, precision, recall rates, and F1 score. The model reaches a 77.63% accuracy, meaning that such a model would be highly reliable in identifying users with low credit rating scores. The sensitivity (recall) of 90.69% is highly suggestive that the model based on the results found can be effective in determining clients with low credit ratings. However, it reached a value of 49.06% and proves the existence of some number of false positive results. The precision is 79.58%, and the F1 score is 0. 85, which equally achieves good results in terms of its precision as well as the recall. The general findings highlighted in this study also highlight the significance of effective management of credit balances that enables one to have a good credit rating and offer a good indication to banks and other institutions on how to develop mechanisms of credit risk management.

The gradient boosting model demonstrates a clear positive correlation between outstanding debt and the probability of having a low credit score. This correlation is visualised in the scatter plot (Figure 3.1.4.2), where the blue trend line shows that higher levels of outstanding debt increase the likelihood of a poor credit score. The model's performance metrics further highlight its effectiveness in predicting credit risk based on debt levels. High accuracy and recall indicate the model's reliability and its strength in identifying at-risk individuals. However, the moderate specificity suggests that the model can be further refined to reduce false positives. These insights emphasise the importance of managing outstanding debt to maintain a good credit score and provide valuable guidance for financial institutions in developing risk mitigation strategies.

3.1.4 Model Performance Metrics for Gradient Boosting

A screenshot of a computer

Description automatically generated

*Figure 3.1.4.1: Confusion Matrix and Performance Metrics for Gradient Boosting Model*

The table in Figure 3.1.4.1 shows the confusion matrix and the other performance indicators for the gradient boosting, model developed for prediction of low credit scores by using the outstanding debts. The confusion matrix gives a clear distinction and description of every test that the model has conducted, the true positives (TP), true negatives (TN), false positive (FP) and the false negative (FN).

Using the model described above, an overall accuracy of 77.63% is possible on this database which literally means that the proposed algorithm has a capability of providing the right classification for over three fourths of the instances. This high accuracy can therefore be attributed to the model’s almost precise ability in determining credit score. For the test set, the overall accuracy was 90.69%, which denotes that the model was able to perform well in identifying poor credit risky customers, a vital component in risk management. Low false positive rates mean that the model accurately captures a huge number of actual poor credit scores preventing a leakage of people who most likely warrant monitoring.

However, the specificity of 49.06% show that there is a relatively high level of false positives meaning that credit scoring models label good credit risks as bad credit risks. This implies that there is still some room to elevate the bar from what has been observed in relation to the number of false positives. Another key issue is thus to minimise false positives to prevent those who responsibly handle credit facilities from being penalised and target only those who are really in need of intervention.

The precision of 79.58% provides an insight into the robustness of the model in identifying the subjects with poor credit scores, implying that for every hundred poor credit scores predicted, seventy-nine are likely to be true. The high precision value is critical, as it shows the rate of accurate positive prediction of the model. For the evaluated precision and recall, the F1 score turned out to be 0. 85 balances precision and recall. This metric gives an overall picture of how effectively the model is performing. The F1 score is especially valuable in the settings when false positives and false negatives impact the result.

Additionally, the Kappa statistic of 0.4334 indicates a moderate agreement between the observed and predicted classifications, accounting for the possibility of agreement occurring by chance. The Kappa statistic is a better way of determining the performance of the model that should not be limited to accuracy rates.

These performance metrics are very important to evaluate the model and to find areas that need to be improved. The high sensitivity and high accuracy are the strengths of the model, which enable the identification of clients who are most likely to default, with credit risk assessment being the primary focus. However, the moderate specificity has indicated that further development would highly improve the model in identifying the accurately accurate credit scores between the good and poor credit scores. They may contain more variables within the model or use a higher degree of elaborateness that results in higher test specificity and would in a way identify higher risk and lower risk individuals.

A graph of a graph

Description automatically generated

*Figure 3.1.4.2: Impact of Outstanding Debt on Probability of Low Credit Score (Gradient Boosting)*

In Figure 3.1.4.2, it displays the working of outstanding debt on the variable of predicted probability of low credit calibrated with the help of the gradient boosting technique. Each point on the scatter plot refers to an individual prediction, while the blue line reflects the general trend by connecting points that have a high value of outstanding debt with a high probability of a low credit score.

From the plot, the relationship between highlighted variables is clear – the higher the outstanding debt is, the higher the probability of obtaining a low credit score is. In particular, when it comes to outstanding debt that had been accumulated over the period of time, it shows the dangers of receiving a poor credit score. This is expressed by the incline of the blue trend line symbolising that the probability of having a low credit score rises with the level of borrowing. These perceptions underscore the need for efficient and effective control of credit ratios in order to ensure that people do not default on their obligations.

The strength of the model is demonstrated by the fact that this relationship is well explained which suggests the validity of the model when used for credit risk estimation based on debt levels particularly for small enterprises. Furthermore, bars are densely clustered at specific debt figures, meaning that people take lots of debt at these values. This concentration proves that a large amount of people owes within these amount brackets, and it might illustrate normal borrowing at expense or common practices among debtors.

In addition, the dispersion of the points shows the presence of a great deal of risk for credit score when it is obtained for specific levels of debt. These fluctuations indicate that part of the reason why credit scores are reduced or raised could actually be because of accumulation of debt. For instance, variables like their income, employment status, and credit payment history could as well affect the credit scores.

Figure 3.1.4.2 is crucial for understanding the direct impact of outstanding debt on credit scores. The positive correlation highlighted in the plot emphasises the need for individuals to manage their debt levels carefully. For financial institutions, these insights can be valuable in refining their risk assessment models and developing targeted strategies to mitigate credit risk, such as offering debt management programs, or financial counselling services to high-risk borrowers.

3.1.5 Conclusion

The first analysis is to determine how outstanding debt affects a person’s probability of having a low credit score. With this, it can be seen clearly that people with average credit balances of $1500-$2500 have the highest probability of having a poor credit rating, which is a sign of a very high risk in this range of debts. On the other hand, the proportion of clients with poor credit ratings drops sharply for individuals with debts above $2500, which might imply that such users are more financially responsible to only possess better credit ratings.

The analysis performed in this study is further supported by the Gradient Boosting Model where it reveals a positive relationship between outstanding debts and the likelihood of ranking low in credit scores. The efficiency of the model can be seen through the high level of accuracy, which stands at 77.63% as well as sensitivity at 90.69%, which make it quite reliable in terms of pointing out clients with poor credit ratings. But the moderate level of specificity is merely 49.06% which denotes to increase specificity to minimise the number of false positives’ that have been determined by the model.

In conclusion, it can be seen that the issue of average outstanding balance has contributed significantly to the increase or decrease of credit score as indicated in this analysis. People who undertake proper control of their debts have a better chance of keeping their credit scores healthy, and those with uncontrolled debts have higher chances of poor credit scores.

3.2 Mikhail Churilov – The impact of Delayed Payments on having Lower Credit Score

3.2.1 Logistic Regression Model Output of Days of late payments and the number of later payments

A close-up of a computer screen

Description automatically generated

*Figure 3.2.1: Logistic Regression Analysis using glm*

The provided output shows the results of a logistic regression analysis performed using the glm function in R. The goal of the analysis is to model the binary outcome variable credit\_score\_binary using the predictors delay\_from\_due\_date, num\_of\_delayed\_payment, and their interaction term.

The logistic regression model is specified with the formula credit\_score\_binary ~ delay\_from\_due\_date \* num\_of\_delayed\_payment, indicating that the binary outcome variable credit\_score\_binary is modeled as a function of delay\_from\_due\_date, num\_of\_delayed\_payment, and their interaction term. The model uses the binomial family and data from h2\_data.

The table of coefficients includes estimates, standard errors, z-values, and p-values for each predictor in the model. The intercept has an estimate of -3.08433 with a standard error of 0.22051, yielding a highly significant z-value of -13.987 and a p-value of less than 0.0000000000000002. This intercept represents the log-odds of the outcome when all predictors are zero. The delay\_from\_due\_date predictor has an estimate of 0.48073 with a standard error of 0.08391, a z-value of 5.729, and a highly significant p-value of less than 0.000000010082518999, indicating that as the delay from the due date increases, the log-odds of having a poor credit score also increase. The num\_of\_delayed\_payment predictor has a negative estimate of -0.45945 with a standard error of 0.08585, a z-value of -5.352, and a highly significant p-value of less than 0.0000087004184552, suggesting that a higher number of delayed payments is associated with a decrease in the log-odds of having a poor credit score. The interaction term delay\_from\_due\_date:num\_of\_delayed\_payment has a positive estimate of 0.25059 with a standard error of 0.03090, a z-value of 8.109, and a highly significant p-value of less than 0.000000000000000509, indicating that the combined effect of delay from the due date and the number of delayed payments on the log-odds of having a poor credit score is greater than the sum of their individual effects.

The significance codes indicate the levels of significance for each coefficient, with \*\*\* for p < 0.001, \*\* for p < 0.01, \* for p < 0.05, for p < 0.1, and for p >= 0.1. The dispersion parameter for the binomial family is taken to be 1, indicating a standard logistic regression model. The null deviance is 46838 on 38533 degrees of freedom, and the residual deviance is 41559 on 38530 degrees of freedom, with an Akaike Information Criterion (AIC) of 41567, which measures the relative quality of the model for a given set of data.

The Variance Inflation Factor (VIF) values assess multicollinearity among the predictors. The VIF for delay\_from\_due\_date is 19.20848, for num\_of\_delayed\_payment it is 15.05338, and for the interaction term delay\_from\_due\_date:num\_of\_delayed\_payment it is 48.09518. These high VIF values indicate a high level of multicollinearity, suggesting that these predictors are highly correlated and may affect the stability of the coefficient estimates.

3.2.2 Centered Logistic Regression Model Output Of delay from due date and number of delayed payments

A close-up of a document

Description automatically generated

*Figure 3.2.2: Logistic Regression Analysis with Centered Predictors*

The provided output shows the results of a logistic regression analysis with centered predictors. Centering predictors is a common practice to reduce multicollinearity, especially when interaction terms are included in the model. The goal of this analysis is to model the binary outcome variable credit\_score\_binary using the centered predictors delay\_from\_due\_date\_centered, num\_of\_delayed\_payment\_centered, and their interaction term.

The logistic regression model is specified with the formula credit\_score\_binary ~ delay\_from\_due\_date\_centered \* num\_of\_delayed\_payment\_centered, indicating that the binary outcome variable credit\_score\_binary is modeled as a function of delay\_from\_due\_date\_centered, num\_of\_delayed\_payment\_centered, and their interaction term. The model uses the binomial family and data from h2\_data.

The table of coefficients includes estimates, standard errors, z-values, and p-values for each predictor in the model. The intercept has an estimate of -1.09488 with a standard error of 0.01427, yielding a highly significant z-value of -76.746 and a p-value of less than 0.0000000000000002. This intercept represents the log-odds of the outcome when all predictors are zero (after centering). The delay\_from\_due\_date\_centered predictor has an estimate of 1.11033 with a standard error of 0.02227, a z-value of 49.865, and a highly significant p-value of less than 0.0000000000000002, indicating that as the delay from the due date increases (after centering), the log-odds of having a poor credit score also increase. The num\_of\_delayed\_payment\_centered predictor has a positive estimate of 0.25009 with a standard error of 0.02603, a z-value of 9.606, and a highly significant p-value of less than 0.0000000000000002, suggesting that a higher number of delayed payments (after centering) is associated with an increase in the log-odds of having a poor credit score. The interaction term delay\_from\_due\_date\_centered:num\_of\_delayed\_payment\_centered has a positive estimate of 0.25059 with a standard error of 0.03090, a z-value of 8.109, and a highly significant p-value of less than 0.000000000000000509, indicating that the combined effect of delay from the due date and the number of delayed payments on the log-odds of having a poor credit score is greater than the sum of their individual effects.

The significance codes indicate the levels of significance for each coefficient, with \*\*\* for p < 0.001, \*\* for p < 0.01, \* for p < 0.05, for p < 0.1, and for p >= 0.1. The dispersion parameter for the binomial family is taken to be 1, indicating a standard logistic regression model. The null deviance is 46838 on 38533 degrees of freedom, and the residual deviance is 41559 on 38530 degrees of freedom, with an Akaike Information Criterion (AIC) of 41567, which measures the relative quality of the model for a given set of data.

The Variance Inflation Factor (VIF) values assess multicollinearity among the predictors. The VIF for delay\_from\_due\_date\_centered is 1.352754, for num\_of\_delayed\_payment\_centered it is 1.384491, and for the interaction term delay\_from\_due\_date\_centered:num\_of\_delayed\_payment\_centered it is 1.049424. The reduced VIF values (all below 2) indicate that centering the predictors has effectively mitigated multicollinearity.

3.2.3 The Impact of Payment Delay and Number of Delayed Payments on Probability of Low Credit Score

A graph showing a curve

Description automatically generated with medium confidence

*Figure 3.2.3: Impact of Payment Delay and Number of Delayed Payments on Probability of Low Credit Score*

The provided graph illustrates the combined impact of payment delay from the due date and the number of delayed payments on the probability of having a low credit score. The x-axis represents the delay in payment from the due date, measured in some unit of time (e.g., days or months), ranging from 0 to 4, indicating the length of time payments are delayed. The y-axis represents the predicted probability of having a low credit score, with values ranging from 0 to 0.6, indicating the likelihood that an individual will have a low credit score based on their payment delay and the number of delayed payments. Each colored line in the graph represents a different number of delayed payments, as indicated by the legend on the right side of the graph, helping to distinguish the impact of different numbers of delayed payments on the probability of a low credit score.

The trend across all lines shows an increasing pattern, indicating that as the payment delay from the due date increases, the probability of having a low credit score also increases. The impact of the number of delayed payments is evident from the position and slope of the lines, with lines representing higher numbers of delayed payments positioned higher on the graph, suggesting a higher probability of a low credit score for a greater number of delayed payments. The lines for different numbers of delayed payments start close together at lower delays but diverge more significantly as the payment delay increases. This divergence indicates that the impact of payment delay on the probability of a low credit score becomes more pronounced with a higher number of delayed payments.

In summary, the graph effectively demonstrates the relationship between payment delay from the due date, the number of delayed payments, and the probability of having a low credit score. It shows that both increasing payment delays and a higher number of delayed payments are associated with a higher probability of a low credit score. The colored lines provide a clear visualization of how the number of delayed payments modifies the impact of payment delays, illustrating the combined effect of these factors on credit score probabilities. This information is crucial for understanding credit risk and can inform strategies for managing and mitigating the risk of low credit scores.

3.3 Jane Lee – The impact of Payment Behavior on having Lower Credit Score

3.3.1 What is the impact of different payment behaviors on credit scores?

A graph of credit scores

Description automatically generated

*Figure 3.3.1: Distribution of Credit Scores by Payment Behaviors*

This is the first analysis about the analysis of Credit Scored by Payment Behavior. In this Figure 3.3.1, the bar chart has visualized the distribution of credit scores across different payment behaviors. The x-axis represents the various payment behavior categories, which total of 6 and the y-axis shows the count of individuals within each category. Besides that, Credit Scores are categorized into three groups, which was Good (green), Poor (purple) and Standard (yellow).

The 6 categories of Payment Behaviors:

1 Low\_spent\_Large \_value\_payments

2 Low\_spent\_Medium\_value\_payments,

3 Low\_spent\_Small\_value\_payments

4 High\_spent\_Large\_value\_payments

5 High\_spent\_Medium\_value\_payments

6 High\_spent\_Small\_value\_payments

Based on the Figure 3.3.1 can clearly see that the Low\_spent categories generally show a more stable and predictable distribution of credit scores compared to High\_spent categories. The stability of the Low\_spent category, can notice that the Low\_spent\_Medium\_value\_payments and Low\_spent\_Small\_value\_payments have a higher count of individual with the Standard and Good credit scores. In the Low\_spent categories, it suggests that lower spending is associated with better and more stable credit scores, especially on the small and medium value payments. Individuals in these categories are less likely to experience financial strain, and the results will be more consistent credit scores.

For High\_spent categories, it shows more instability and variation in the credit scores. In observing all three High\_spent categories, the instability in these categories suggests that frequent high-value expenditures can lead to unpredictable financial outcomes, with a significant risk of lowering credit scores. Despite the presence of some good credit scores, the overall pattern indicates that high spending behaviors are more likely to result in financial instability compared to low spending behaviors. The substantial number of Poor credit scores had highlighted the financial risks and challenges associated with managing large expenditures.

In conclusion, the analysis of Figure 3.3.1 reveals the spending behavior significantly impacts the credit scores. For Low spending behaviors are generally associated with a more stable and better credit scores. Especially in small and medium value payments that contribute to maintaining consistent credit health. On the contrary, high spending behaviors are linked with more instability and a higher risk of poor credit scores. High spending, especially in large value payments often leads to financial strain and unpredictable credit outcomes.

3.3.2 How does the combination of salary range and payment behavior influence average credit scores?

**Influence of Salary Range and Payment Behavior on Average Credit Scores**

A graph of credit scores

Description automatically generated

*Figure 3.3.2: Heatmap of Credit Scores by Salary Range and Payment Behavior*

In this Figure 3.3.2 presents the heatmap that visualizes the Average Credit Scores across different Monthly In-hand Salary Ranges and Payment Behaviors. The x-axis represents the Monthly In-hand Salary Range, ranging from 0 to 16000, and the y-axis represents the 6 various payment behaviors. The color gradient indicates the Average Credit Score, blue represents the lower average scores, and pink represents the higher average score.

The 6 categories of Payment Behaviors:

1 High\_spent\_Small \_value\_payments

2 High\_spent\_Medium\_value\_payments

3 High\_spent\_Large \_value\_payments

4 Low\_spent\_Small \_value\_payments

5 Low\_spent\_Medium\_value\_payments,

6 Low\_spent\_Large \_value\_payments

Explanation of each Payment Behavior performance:

1 High\_spent\_Small\_value\_payments:

-Pattern: It shows the most variation with low (blue) average credit scores

-Stability: The presence of blue in higher salary ranges indicates that individual with higher incomes but poor in spending management on small value items can still end up with the lowest credit score.

2 High\_spent\_Medium\_value\_payments:

-Pattern: It is similar to small value payments, with a predominantly purple gradient

-Stability: Indicates that moderate financial outcomes with no significant variation

3 High\_spent\_Large \_value\_payments:

-Pattern: Shows a balanced distribution, with a tendency towards purple shades

-Stability: Slight variations towards lower scores in higher salary ranges, indicating potential financial strain

4 Low\_spent\_Small \_value\_payments:

-Pattern: It shows the most variation with high (pink) average credit scores

-Stability: The presence of pink in lower salary ranges suggests that individuals managing small expenditures within their means can maintain higher credit scores

5 Low\_spent\_Medium\_value\_payments:

-Pattern: Predominantly pink, indicating higher average credit scores

-Stability: It is suggested that medium value expenditures when kept it infrequent, contribute to better financial stability

6 Low\_spent\_Large \_value\_payments:

-Pattern: Mostly purple with some variations, indicating average credit scores

-Stability: Suggests moderate financial stability with occasional financial strain

Figure 3.3.2 illustrates that Payment Behavior, combined with Monthly In-hand Salary, significantly impacts Average Credit Scores. In this diagram the most notable is the variation within the Low\_spent\_Small\_value\_payments and High\_spent\_Small\_value\_payments category. This dual presence of high and low average credit scores suggests that small expenditures can lead to both financial stability and instability, depending on the overall financial management and income levels.

Overall, the heatmap reveals that controlled and infrequent expenditures, particularly in small and medium value payments, are associated with better credit scores and financial stability. On the other hand, high spending behaviors, especially without effective management, can lead to financial instability and lower credit scores. Financial strategies should focus on promoting controlled spending habits to maintain or improve credit scores.

3.3.3 How effective is the logistic regression model in predicting credit scores based on payment behavior and other financial factors?

**Regression Model Performance for Predicting Credit Scores**

A screenshot of a computer

Description automatically generated

*Figure 3.3.3: The Confusion Matrix and Statistics for Logistic Regression Model*

The analysis of Figure 3.3.3 reveals the strengths and limitations of the Logistic Regression Model in Predicting Credit Scores. This section presents the output of the logistics regression model applied to the dataset to predict the Credit Scores based on Payment Behavior, Number of Delayed Payments, Outstanding Debt, and Monthly In-hand Salary. The confusion matrix and associated statistics provide a comprehensive evaluation of the model’s performance.

In Confusion Matrix and Statistics if the Accuracy is higher, then it means the model performs well. In this case, the Accuracy of Logistics Regression Model is 0.8286, which also means 82.86%, is count of a very high accuracy. High accuracy indicates that the model performs well overall, which is crucial for making reliable predictions about credit scores based on payment behavior and other factors.

For sensitivity, the value in this Logistics Regression Model is 0.9723, which also means the model correctly identified 97.23% of the actual positive instances. High sensitivity is critical for identifying individuals at risk of poor credit scores, ensuring that most of those who truly have poor credit scores are correctly identified by the model. This high sensitivity is particularly valuable for risk management, as it minimizes the chances of missing individuals who are at risk of financial instability.

The main purpose of Kappa is to measure the agreement between the observed accuracy and the expected accuracy. If the value is 1 means it was a perfect agreement, however 0 means no agreement beyond chance. The value of Kappa in this Logistic Regression Model are 0.1466 suggests only slight agreement beyond chance. This indicates that while the model has high accuracy, its performance compared to random chance is modest. This highlights the need for further model improvement.

The McNemar's Test P-Value being extremely low (<2e-16) indicates that there is a significant difference between the predicted and actual classifications. This suggests that the model's predictions are not simply random, reinforcing the validity of the model's accuracy and sensitivity metrics. A significant McNemar's Test P-Value indicates that the model has a meaningful impact on prediction, validating the observed high accuracy and sensitivity.

In conclusion, the logistics regression model provides valuable insights into how payment behavior influences credit scores. The high sensitivity is particularly beneficial for identifying individuals at risk of poor credit scores, aligning well with the objective of understanding the impact of payment behavior on credit scores. The high accuracy and significant McNemar’s Test P-Value further reinforce the model’s reliability and effectiveness in predicting credit scores based on the analyzed factors.

3.3.4 Extra Features

A graph of credit scores

Description automatically generated

*Figure 3.3.4.1: Heatmap that Highlights the Highest and Lowest Average Credit Scores*

To identify the graph more easily, an additional feature has been added to the heatmap to highlight the highest and lowest average credit scores. As can be seen in Figure 3.3.4.1, it represents the blue representing lower average credit scores and pink representing higher average scores. At the same time, the lowest average credit score is marked with a green dot and red dot is marked for the highest credit scores.

A screenshot of a computer

Description automatically generated

*Figure 3.3.4.2: Generalized Linear Model Coefficients for Predicting Credit Scores*

The Generalized Linear Model (GLM) is specifically for logistic regression model, it is used to predict the binary outcomes. In this context, the binary outcomes are whether the credit score is “Good” (coded as 1) or “Not Good” (coded as 0). This approach is crucial in understanding how various factors including payment behavior, the number of delayed payments, outstanding debt and monthly in-hand salary influence the likelihood of having a good credit score.

3.3.5 Conclusion

In summary, the third objective of the study is to understand the impact of different payment behaviors, salary ranges and other financial factors on the likelihood of having lower credit score. The first analysis results that low spending behaviors (small and medium value payments) are associated with more stable and better credit scores. In contrast, high spending behaviors, especially on large value payments often lead to financial strain and a higher risk of poor credit scores.

The second analysis can see that controlled and infrequent expenditures are linked to better stability, while high spending without effective management can result in financial instability and lower credit scores.

Other than that, the third analysis suggests that effective financial management and controlled spending habits are crucial for maintaining or improving credit scores, which means the significant impact of payment behavior on credit health.

3.4 Ravin A/L Kanagarajan – The impact of Credit History age on having Lower Credit Score

3.4.1 What is the impact of Poor count of Credit score

A graph showing different colored squares

Description automatically generated

*Figure 3.4.1: Count of Poor Credit Score by Credit History Duration*

The graph presents a breakdown of poor credit scores by credit history duration. The X-axis shows credit history length in months, with each bar representing a specific range (e.g., 0-50 months). The Y-axis depicts the count of poor credit scores within each range.

The bars use different colors for visual distinction, not specific meaning. Interestingly, the data reveals a pattern. People with short credit histories (0-50 months) have a moderate number of poor scores. However, the highest concentration of poor credit scores falls within the medium credit history ranges (50-200 months). There's a slight decrease in the 200-250 month range, followed by a significant drop in the count of poor scores for those with very long credit histories (over 300 months).

In summary, this graph suggests a link between credit history length and the likelihood of having a poor credit score. People in the 50-200 month range appear most susceptible to poor credit scores, while those with very short or very long histories show a lower risk. This information can be valuable in understanding credit risk based on credit history duration.

3.4.2 What is the impact proportion of Poor Credit Scores by Credit History Duration

A graph of different colored bars

Description automatically generated

*Figure 3.4.2: Proportion of Poor Credit Scores by Credit History Duration*

This graph dives into the relationship between credit history length and the proportion of poor credit scores. The X-axis tracks credit history duration in months, with each bar representing a specific range. The Y-axis, however, focuses on the percentage of poor credit scores within each range.

The graph uses color to differentiate the bars, but these colors hold no specific meaning. The data itself reveals a fascinating trend. People with short credit histories (0-50 months) have a roughly 40% chance of having a poor score. However, the risk jumps significantly for those in the medium range (50-200 months), with proportions exceeding 45% and even reaching 50%. There's a slight dip in the 150-200 month range, but the risk remains high.

The good news is that the proportion of poor credit scores drops noticeably for those with longer histories (over 200 months). It goes down to around 35% for the 200-250 month range and then falls even further to 20-25% for even longer histories. In conclusion, this graph suggests a strong connection between credit history length and creditworthiness. People in the 50-200 month range are most likely to have poor credit scores, while those with very short or very long histories appear to be at a lower risk. This information is valuable for understanding credit risk based on credit history duration.

3.4.3 How effective is the logistic regression modelling on predicting the probability of low credit score by credit history length

A screenshot of a computer screen

Description automatically generated

*Figure 3.4.3: Confusion Matrix and Statistics for KNN*

The KNN model's performance offers a mixed bag of results. On the one hand, it achieves a seemingly reasonable overall accuracy of 71.02%, indicating it can correctly classify credit scores in just over 70% of the cases. This might suggest some level of competency at distinguishing between good and poor credit scores. However, a deeper dive into the metrics reveals some concerning limitations.

One of the model's strengths lies in its ability to identify true positives, correctly classifying individuals with poor credit scores. This is evident in the high sensitivity of 88.04%. The model appears adept at catching most instances of poor creditworthiness. However, this effectiveness comes at a cost. The model struggles significantly with true negatives, reflected in the low specificity of 29.91%. This translates to a high number of false positives, where the model incorrectly classifies individuals with good credit scores as having poor credit.

Imagine the potential consequences of such misclassification. An otherwise creditworthy borrower might be denied a loan or offered unfavorable terms due to a mistaken poor credit score prediction. This highlights a critical weakness in the model's performance. While it's good at catching true positives, there's a concerning possibility of it unfairly labeling individuals with poor credit scores when they don't deserve it.

To address this imbalance and improve the model's overall effectiveness, further refinement might be necessary. This could involve adjusting key parameters like the number of neighbors considered in the classification process (k). Additionally, exploring alternative algorithms altogether, such as logistic regression or support vector machines, could potentially lead to a model with a better balance between identifying true positives and true negatives. By addressing these shortcomings, we can strive for a more reliable model that can accurately assess creditworthiness without unfairly penalizing individuals with good credit scores.

3.4.4 Probability of Low Credit score by Credit History Length

A graph of a line and a line

Description automatically generated

*Figure 3.4.4: Predicted Probability of Low Credit Score by Credit History Length (kNN)*

The graph displays the predicted probability of a low credit score as a function of credit history length using the k-Nearest Neighbors (kNN) method. The X-Axis represents the length of credit history. The values are scaled, meaning the original data has been transformed to a specific range to facilitate analysis. The Y-Axis represents the predicted probability that an individual has a low credit score based on their credit history length. The probability values range from 0 to 1.

Each black dot in the graph represents an individual data point. These dots indicate the predicted probability of having a low credit score for a specific credit history length. The concentration of dots at the top of the graph (around a probability of 1.0) and at the bottom (around a probability of 0.0) shows the distribution of predictions.

The blue line represents a trend line or the expected relationship between credit history length and the predicted probability of having a low credit score. The trend line has a negative slope, indicating an inverse relationship between credit history length and the predicted probability of a low credit score.

Interpretation:

High Probability for Short Credit History: On the left side of the graph (shorter credit history lengths), the predicted probability of having a low credit score is high (close to 1). This suggests that individuals with shorter credit histories are more likely to have poor credit scores.

Low Probability for Long Credit History: On the right side of the graph (longer credit history lengths), the predicted probability of having a low credit score is low (close to 0). This implies that individuals with longer credit histories are less likely to have poor credit scores.

k-Nearest Neighbors (kNN) Method: kNN is a non-parametric classification method. It classifies an instance based on the majority class among its 'k' nearest neighbors in the feature space. In this context, the model predicts the probability of having a low credit score based on the length of the credit history.

The graph illustrates that longer credit histories are generally associated with a lower probability of having a low credit score, aligning with the understanding that longer credit histories often reflect more stable and reliable credit behavior. The kNN method helps identify this pattern by using the nearest neighbors' data points to make predictions about the probability of low credit scores based on credit history length. The negative trend line reinforces the notion that an increase in credit history length decreases the likelihood of having a low credit score.

4.0 Findings

By examining different forms of **outstanding debts and their effects on credit scores**, Azelea enhanced the understanding of credit scores and how they work. In this context, to identify consumers with low credit scores, the Gradient Boosting Model was adopted for formulating predictions based on outstanding debt. self-studies indicated that outstanding debt is inversely proportional to credit scores whereby institutions with higher outstanding debt record low credits. The accuracy and attempts such as precision, recall, and F1 measure, among others, were reasonable for a model to perform for prediction, though not very specific for precision.

The outcomes of the correlations were modelled using Logistic regression, and the **impact of the delayed payments on credit scores** was examined by Mikhail in his work. The study revealed the probability of delays and days of delay as key variables that impact credit scores regarding the percentage of delayed payments. As depicted in the logistic regression model, there is a positive significant relationship between payment frequency and amount of time, and poor credit score that reiterates the importance of timely payment to healthy credit score.

Jane aimed to check **how various** **payment behaviours affects credit scores**. The research findings revealed that there is a positive correlation between low and moderate level of spending and better credit ratings on the other hand negative credits only indicate more instable spending level. High spending, especially on the large value payments, end up placing much pressure resulting to poor credits and other related effects. The model selected as the logistic regression model had a moderate performance in terms of accuracy in estimate credit scores given the payment behaviour providing credit as discouraging but important lesson about the essence of proper financial planning.

Ravin’s study involved **how credit history age affects credit scores**. To determine this, the number of available credit history was correlated with the credit scores, and it was evidenced that, in general, a higher credit score is connected with the usage of credit for a longer period, which implies a proper credit management. However, the results of the logistic regression analysis have shown that a long credit history is not exactly beneficial, but it is essential to have a proper management of the credit over the period.

5.0 Recommendations

Based on this comprehensive analysis dataset, there are several recommendations to improve the credit risk management and decision-making processes in the banking sector:

1 **Implement Regular Data Audits**:

-By having data audits is to ensure the data quality and completeness of the dataset, this includes check for missing, inconsistencies and outliers’ values, and address these issues promptly to maintain the data quality and dataset completeness.

2 **Expand the Dataset with Additional Variables**:

-To have a holistic view of influencing factors can lead to have a more accurate credit risk management, the dataset can include employment history, education levels and macroeconomic indicators.

3 **Adopt Advanced Data Analytics Techniques**:

-To enhance predictive accuracy, could use deep learning, ensemble methods and real-time data integration. For instance, employing techniques like SHAP (SHapley Additive exPlanations) values can improve the explicability of complex models.

4 **Develop Longitudinal Analysis Capabilities**:

-By establishing systems to collect and analyze data over multiple time periods, banks could track the changes in credit behaviors and scores to reveal the trends and patterns for proactive risk management.

By implement these recommendations, banks can significantly improve the credit risk management process, customer satisfaction, and maintain a competitive edge in the financial industry. Other than that, it could also let the bank have better insights in decision making.

6.0 Limitations and Future Directions

The analysis conducted in this study has several limitations that should be acknowledged.

1 **Dataset Quality and Completeness**:

-Although data cleaning and pre-processing steps had been done it in the dataset, but data that was inaccuracies, missing or inherent biases in the dataset could affect the results.

2 **Limited Scope of Variables**:

-In this dataset has excluded some significant influences such as employment history, education level or macroeconomic variables. By excluding these data could limit the comprehensiveness and predict the power of the models.

3 **Static Nature of the Dataset**:

-The dataset that analyzing is just a static snapshot in time, but credit scores and behavior are dynamic. Using longitudinal dataset will offer a more robust analysis of trends and behaviors.

To address these limitations and enhance future analyses, several directions can be pursued.

1 **Incorporate Additional Variables**:

-Include the employment records, education history and macroeconomic in dataset can broaden the scope and improve the comprehensiveness of the analysis.

2 **Explore Advanced Modeling Techniques**:

-Utilize the deep learning or ensemble methods can improve the predictive accuracy and robustness. Therefore, to help the stakeholders to understand the complex model better.

3 **Conduct Longitudinal Studies**:

-Collect and analyze the data in multiple periods of time can understand more the credit behaviors and scores evolve in revealing trends, seasonal effect and long-term patterns that no evident in a cross-sectional dataset.

7.0 Conclusion

The main objective in this study is to explore the factors that influence credit scores among the bank customers by using advanced data analytics techniques. Before going through the analysis part, due to having several problems in this dataset, we need to do data preparation to make the dataset tidier and more effective.

Data preparation had involved importing the dataset, cleaning the data by handling missing and duplicated data and transforming the data to appropriate data formats. The purpose of doing these steps is to ensure the data is reliable and valid for the further analysis part. Besides that, data exploration is also implemented before doing any analysis. It could provide descriptive and visualizations to highlight the distribution of key variables.

The objective-based analysis is to focus on understanding the relationships between demographic attributes and the credit behaviors and scores. For instance, based on our analysis about the relationship between the credit behaviors and a credit history age, the result reveals that the relationship is moderate positive in correlation. Based on these ways to do every objective analysis, the result can let analyst understand more about the customer behavior and credit scores.

Azelea's analysis using a Gradient Boosting Model found that higher outstanding debt leads to lower credit scores. Mikhail's logistic regression model showed that frequent and prolonged payment delays significantly worsen credit scores. Jane's research revealed that moderate spending levels correlate with better credit scores, while high spending leads to poor credit score. Ravin's study indicated that longer credit histories generally result in higher scores, but effective management over time is essential. These findings highlight the importance of managing debt, timely payments, prudent spending, and long-term credit management for maintaining good credit scores.

Besides, based on all of the analyzing process we observed that some limitations that can be improved. Firstly, the dataset does not include some relevant factors, it may constrain the dataset scope. Besides, analyzing should implement some additional data sources and advanced modeling techniques, so the analyze result may be more appropriate and complete. Due to credit behavior is a dynamic data, it should be monitoring the data continuously to maintain the predictive accuracy.

In conclusion, based on the analysis, it can demonstrate the effectiveness of comprehensive data analytics in understanding the credit score determinants and could provide valuable insights for stakeholders in the banking sector.

# 8.0 Workload Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tasks | Azelea Glory Ng Zi-Lin  TP067853 | Churilov Mikhail  TP072847 | Jane Lee  TP074470 | Ravin A/L Kanagarajan  TP068019 | Total |
| Introduction | 25% | 25% | 25% | 25% | 100% |
| Data Preparation | 25% | 25% | 25% | 25% | 100% |
| Data Analysis | 25% | 25% | 25% | 25% | 100% |
| Conclusion | 25% | 25% | 25% | 25% | 100% |

9.0 References

Awan, A. A. (2023, June 14). K-Nearest Neighbors (KNN) classification with R tutorial. https://www.datacamp.com/tutorial/k-nearest-neighbors-knn-classification-with-r-tutorial

Chugh, V. (2023, March 17). Logistic regression in R tutorial. https://www.datacamp.com/tutorial/logistic-regression-R

GeeksforGeeks. (2023, July 11). Boosting in R. https://www.geeksforgeeks.org/boosting-in-r/

Hoare, J. (2024, January 2). Gradient boosting explained - the coolest kid on the Machine Learning Block. https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/

Hotjar. (2024, January 12). *What are heat maps? A guide to heatmaps & how to use them*. What Are Heat Maps? A Guide to Heatmaps & How to Use Them. https://www.hotjar.com/heatmaps/