Table of Contents

[1.0 Introduction 3](#_Toc204551951)

[1.1 Data Description 3](#_Toc204551952)

[1.2 Assumptions 3](#_Toc204551953)

[1.3 Hypothesis and Objectives 4](#_Toc204551954)

[2.0 Data Preparation 5](#_Toc204551955)

[2.1 Data Import 5](#_Toc204551956)

[2.2 Data Cleaning 6](#_Toc204551957)

[2.3 Data Validation 17](#_Toc204551958)

[3.0 Data Analysis 19](#_Toc204551959)

[3.1 To examine the connection between loan duration and credit class (Eugene Tan Ting Siang) 19](#_Toc204551960)

[3.1.1 Analysis 1: General Information of Loan Duration and Credit Class 19](#_Toc204551961)

[3.1.2 Analysis 2: What is the Distribution of Loan Duration by Credit Class? 22](#_Toc204551962)

[3.1.3 Analysis 3: What is the Correlation Between Loan Duration and Credit Class? 24](#_Toc204551963)

[3.1.4 Analysis 4: Is the Difference in Mean of Loan Duration for Good and Bad Credit Class Statistically Significant? 25](#_Toc204551964)

[3.1.5 Analysis 5: Why do Some Customers with Short Loan Duration have Bad Credit Class and Customers with Long Loan Duration have Good Credit Class? 27](#_Toc204551965)

[3.1.6 Analysis 6: Can the Loan Duration be Used to Predict Credit Class of Customers? Will the Prediction Improve When Considering Other Variables? 31](#_Toc204551966)

[3.1.7 Analysis 7: What Can Customers Do to Achieve Good Credit Class? 35](#_Toc204551967)

[3.2 To identify the impact of employment status and saving status to credit class. (Sweetha Pramasivam) 37](#_Toc204551968)

[3.2.1 Analysis 1: How does the distribution of saving status vary across different employment statuses for each credit class? 37](#_Toc204551969)

[3.2.2 Analysis 2: What proportion of customers with different saving statuses exist within each employment status, and how does this relate to their overall population size? 42](#_Toc204551970)

[3.2.3 Analysis 3: How does saving status change across employment levels, and how is it affected by imputing missing values? 46](#_Toc204551971)

[3.2.4 Analysis 4: What is the distribution of employment status across different credit classes (good/bad)? 51](#_Toc204551972)

[3.2.5 Analysis 5: How does the relationship between employment status and credit class vary across different saving statuses? 56](#_Toc204551973)

[3.2.6 Extra Features 59](#_Toc204551974)

[3.2.7 Analysis Conclusion 59](#_Toc204551975)

[3.3 To explore the impact of credit amount and installment commitment to credit class. (Hong Xiang Lin) 61](#_Toc204551976)

[3.3.1 Analysis 3-1: Assessing the Relationship Between Credit Amount, Installment Commitment and Credit Class. 61](#_Toc204551977)

[3.3.2 Analysis 3-2: Investigating the Interaction of Credit Amount and Installment Commitment in Influencing Credit Class. (Hong Xiang Lin) 66](#_Toc204551978)

[3.3.3 Analysis 3-3: What is the predictive power of credit amount and installment commitment combined compared to using credit amount alone? (Hong Xiang Lin) 72](#_Toc204551979)

[3.3.4 Analysis 3-4: What decision rules can be applied to optimize credit class outcomes based on credit amount and installment commitment? – Hong Xiang Lin 79](#_Toc204551980)

[3.3.5 Extra Features 83](#_Toc204551981)

[3.3.6 Conclusion 83](#_Toc204551982)

[3.3.7 Recommendation 83](#_Toc204551983)

[3.4 To investigate the relationship between loan purposes and credit class (Eng Hui Ern) 84](#_Toc204551984)

[3.4.1 Analysis 1: Is there any relationship between loan purposes and credit class? 84](#_Toc204551985)

[3.4.2 Analysis 2: Is installment commitment a strong predictor for credit class? 87](#_Toc204551986)

[3.4.3 Analysis 3: What Are the External Factors That Interact with Installment Commitment to Influence Credit Class? 90](#_Toc204551987)

[3.4.4 Extra Features 95](#_Toc204551988)

[4.0 Conclusion 96](#_Toc204551989)

[4.1 Findings 96](#_Toc204551990)

[4.2 Recommendations 96](#_Toc204551991)

[4.3 Limitations and Future Directions 97](#_Toc204551992)

[5.0 Workload Matrix 98](#_Toc204551993)

[6.0 References 99](#_Toc204551994)

# **1.0 Introduction**

The objective of this project is to explore how various factors influence credit classification, with a focus on loan-related variables and borrower characteristics. Variables such as loan purpose, loan duration, credit amount, instalment commitment, employment status, and saving status are analysed using methods such as mode, hot deck, predictive mean matching and miss forest. This analysis aims to uncover the relationships between these factors and credit class, providing insights into how each variable impacts creditworthiness. By leveraging data analysis and machine learning, the findings can guide financial institutions in assessing risk, making informed lending decisions, and tailoring financial products to specific borrower profiles.

## **1.1 Data Description**

This dataset contains 6,000 rows and 21 columns; credit risk categorization is the problem it solves. It provides features that summarize consumer financial information, demographics, and loan-related data. The important features in this dataset include checking status and savings status, example; "<0" represents no checking. Next, loan data such as duration, credit amount, and purpose, employment and work status reflect length of employment and job type. Demographics like age, no dependents, and housing status. Credit class is binary; hence, the classes of good or poor borrowers of their creditworthiness. There are several variables containing missing values in the example dataset attributes, such as other\_payment\_plans, credit\_history, and credit\_amount.

## **1.2 Assumptions**

1. Duration is counted in months.
2. “all paid” category in the credit\_history column is same as “no credits/all paid”.
3. Employment is counted in years.
4. Installment\_commitment is counted in years.
5. Residence\_since is counted in years.
6. Credit\_amount, installment\_commitment, residence\_since, age, existing\_credits, and num\_dependents should only be recorded as integers.

## **1.3 Hypothesis and Objectives**

**Hypothesis:**

The loan duration, saving status, employment, credit amount, instalment commitment, and loan purpose will impact the credit class.

**Objectives:**

1. To examine the connection between the loan duration and credit class.
2. To identify the impact of employment status and saving status to credit class.
3. To explore the impact of credit amount and installment commitment to credit class.
4. To investigate the relationship between loan purposes and credit class.

# **2.0 Data Preparation**

## **2.1 Data Import**

Code:

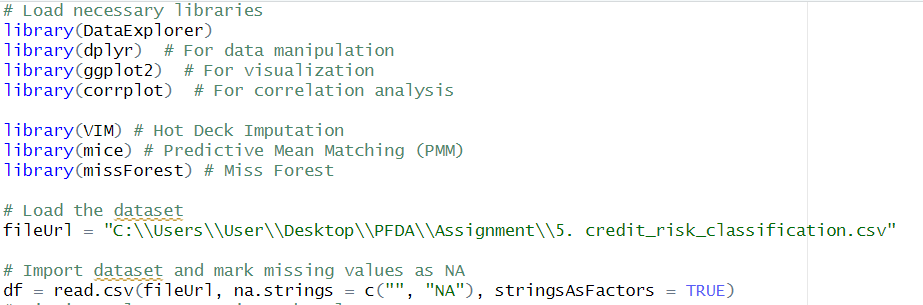


Figure 1: R Script for Importing Necessary Packages and Dataset

Before the data preparation phase, we must first load all necessary libraries. These libraries contain functions that will be used in the data cleaning process later on. Then, we will import the dataset as a data frame and mark all empty strings or “NA” strings as NA values.

## **2.2 Data Cleaning**

Code:

A white background with black text

Description automatically generated

Figure 2: R Script for Checking Missing Values

Output:

A screenshot of a computer program

Description automatically generated

Figure 3: Output for Total of NA Values in Each Column

A screenshot of a computer

Description automatically generated

Figure 4: Structure of Each Column

A white background with black text

Description automatically generated

Figure 5: Summary of Each Column

A graph with numbers and text

Description automatically generated

Figure 6: Visualization of Percentage of NA Values in Each Column

In these codes, we will explore the data and find out how many NA values are present for each column in the dataset. The colSums() and plot\_missing() functions are the ones responsible for calculating number the missing values. Then, there is the str() function which displays the internal structure of each column. Lastly, the summary() function allows us to view the statistical measurements of continuous variables, like the mean and median values, and the sum of each different category for categorical variables.

A computer screen shot of a program

Description automatically generated

Figure 7: R Script for Histogram Function

A screen shot of a computer code

Description automatically generated

Figure 8: R Script for Bar Chart Function

A screenshot of a computer code

Description automatically generated

Figure 9: R Script for Visualizing Columns After Imputation

A few of the Outputs:

A graph of a number of age

Description automatically generated

Figure 10: Histogram of Age

A graph with green squares

Description automatically generated

Figure 11: Bar Chart of Job

Next, we will declare functions to plot histograms for continuous variables, and bar graphs for categorical variables. This will allow us to see the distribution of the variables before imputation. Then, the visualize\_column() function will be called after imputing the missing values so we can compare the new distribution with the previous distribution.

A computer code with text

Description automatically generated with medium confidence

Figure 12: R Script for Mode Imputation Function

After that, we will declare functions for the method we intend to use for imputing missing values. The first method is mode imputation, where we will replace missing values with the most frequent value in a specific categorical variable.

A computer code with black text

Description automatically generated with medium confidence

Figure 13: R Script for Hot Deck Function

Next for the Hot Deck Imputation function replaces missing values in a specified column using similar cases from the dataset, preserving the distribution and relationships within the data.

A screenshot of a computer code

Description automatically generated

Figure 14: R Script for Predictive Mean Matching (PMM) Function

The Predictive Mean Matching function imputes missing values using Predictive Mean Matching, generating realistic values based on similar observed data while preserving the original data structure and relationships. The code first creates a subset of the data and uses mice to impute missing values in the subset. Then, it will update the original dataset with the values from the subset.

A screen shot of a computer code

Description automatically generated

Figure 15: R Script for MissForest Function

The MissForest algorithm iteratively predicts missing values for each column using random forests. It has multiple iterations so that the random forest model will be trained with better quality data and correctly predict missing data (Morgan, 2020). Therefore, it will take some time to complete but it has excellent predictive power, which is the reason why we chose to use it.

The code first checks if any columns are completely empty, and it will generate an error message if it is so. Then, it will create a different dataset with the target column and predictors to proceed with the imputation. After it is done, it will extract the values and replace them with the missing values in the original dataset.

A white background with text

Description automatically generated

Figure 16: R Script for Applying Imputation

Mode imputation is used for columns that have low cardinality (only a few unique values) and where the most frequent value is the most logical replacement. As for hot deck imputation, we use it for columns that are more likely to impact credit class so that we can retain their distribution and avoid bias. Then, PMM is used for numerical variables to ensure that a reliable range is generated. Lastly, MissForest is used for variables with complex interactions, for example credit\_history which might depend on other factors in the dataset. A histogram or bar chart will be generated after each column has been cleaned to show the new distribution.

|  |  |
| --- | --- |
| **Before Imputation** | **After Imputation** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

*Figure 17: Comparison of Distribution of Variables Before and After Imputation*

A screen shot of a computer

Description automatically generated

Figure 18: R Script for Additional Cleaning in credit\_history Column

|  |  |
| --- | --- |
| **Before Cleaning** |  |
| **After Cleaning** |  |

*Figure 19: Comparison of credit\_history column before and after cleaning*

An additional step is needed to clean the credit\_history column, where there are two categories that should be grouped together: “all paid” and “no credit/all paid”.

A computer code with text

Description automatically generated

Figure 20: R Script for Rounding Variables to 0 Decimal Places

Besides that, we also have to round some numerical variables in the dataset so that they do not have decimal places. Some variables like age or number of dependents should not contain decimal places as it is noisy data, so we will fix them using rounding.

## **2.3 Data Validation**

Code:

A screenshot of a computer

Description automatically generated

Figure 21: R Script for Data Validation

Output:

A screen shot of a computer code

Description automatically generated

Figure 22: Number of NA Values in Each Column

A screenshot of a computer screen

Description automatically generated

Figure 23: Summary of Each Column After Imputation

A close-up of words

Description automatically generated

Figure 24: Number of Rows After Imputation: 6000

A screenshot of a graph

Description automatically generated

Figure 25: Percentage of Missing Values for Each Column

After the data cleaning process comes the data validation process. In this section, we have verified that all columns have their missing values cleaned using the colSums() and plot\_missing() functions. We also did not remove any rows during imputation, so the number of rows after cleaning remains at 6000.

# **3.0 Data Analysis**

## **3.1 To examine the connection between loan duration and credit class (Eugene Tan Ting Siang)**

### **3.1.1 Analysis 1: General Information of Loan Duration and Credit Class**

|  |  |
| --- | --- |
| Type of Analysis | Descriptive Analysis |
| Independent Variable(s) | Loan Duration (continuous data) |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | Boxplot |

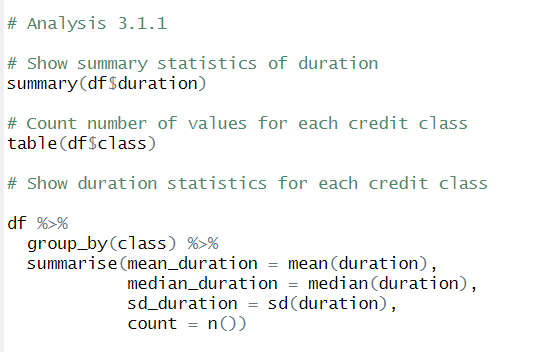


Figure 3.1.1.1: Code to Analyse General Information of Loan Duration and Credit Class

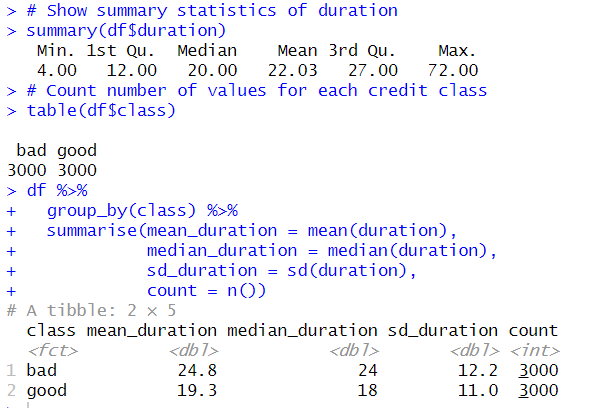


Figure 3.1.1.2: Output of General Information on Loan Duration and Credit Class

From the results, we can see that there are a total of **6000 rows** in the dataset, with a 50:50 split of good credit class and bad credit class. Besides that, we can also see that the loan duration ranges from 4 months, which is the lowest value, to 72 months, which is the highest value. The mean value for loan duration is also higher than the median value, indicating that loan duration has a **positively skewed distribution**. When we classify the data into bad and good credit classes, we can see that bad credit classes have a **larger average and median duration** value than good credit classes, suggesting that bad credit classes have a longer loan duration than good credit classes. The standard deviation for loan duration in bad credit classes is also larger, meaning that the **data is more spread out and has more variance** than loan durations with good credit classes.

A close-up of a math equation

Description automatically generated

Figure 3.1.1.3: Code for Boxplot of Loan Duration by Credit Class

A graph of a credit class

Description automatically generated

Figure 3.1.1.4: Boxplot of Loan Duration by Credit Class

By using a boxplot, we are able to identify outliers in the loan duration column. From the graph, it can be concluded that **good credit class has more outliers** than bad credit class in terms of loan duration. However, bad credit class has a more extreme duration value of 72 months (maximum value of duration). Although these extreme values may affect our analysis, we will still retain these values as they are logically correct since loans can be used for many purposes, including car loans that usually result in longer loan durations.

### **3.1.2 Analysis 2: What is the Distribution of Loan Duration by Credit Class?**

|  |  |
| --- | --- |
| Type of Analysis | Descriptive Analysis |
| Independent Variable(s) | Loan Duration (continuous data) |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | Histogram, Density Plot |

A computer code with black and white text

Description automatically generated

Figure 3.1.2.1: Code for Histogram and Density Plot of Duration Grouped By Class

A graph of a loan duration

Description automatically generated

Figure 3.1.2.2: Histogram of Loan Duration by Credit Class

By using a histogram, we are able to examine how the duration values are distributed for good and bad credit classes. It can be observed from the graph that loan durations with a good credit class **peak at shorter durations** of around 12 months, and it has the highest frequency of around 580. Besides that, there is an overlap between the two credit classes in the range of 12–24 months, suggesting some **bad and good loans share similar durations**. This will have to be analysed further.

Apart from a histogram, the graph also has a density plot overlayed on top to display trends of duration for each credit class. From the plot, we can see that short loan durations (< 10 months) are rare for both classes, **but more so for bad credit class**. This indicates that having short loan durations has a higher probability of achieving a good credit class. Beyond 30 months, the **frequency of good credit class declines significantly**, while bad credit class remain frequent, implying that customers with a bad credit class tend to have extended loan periods. In order to investigate this further, we will proceed with a correlation analysis.

### **3.1.3 Analysis 3: What is the Correlation Between Loan Duration and Credit Class?**

|  |  |
| --- | --- |
| Type of Analysis | Descriptive Analysis |
| Independent Variable(s) | Loan Duration (continuous data) |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | Correlation Analysis (Point-Biserial Correlation) |

A computer screen shot of a computer code

Description automatically generated

Figure 3.1.3.1: Code for Point-Biserial Correlation Analysis

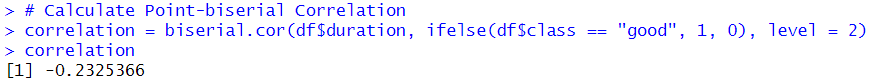


Figure 3.1.3.2: Output of Point-Biserial Correlation Analysis

We use a point biserial correlation analysis for the duration and credit class columns because we are examining the correlation between a continuous data and a categorical data (Statistics Resources: Point Biserial, 2024). The result we retrieved from the analysis is **-0.2325366**, and the correlation being a small negative value suggests a **weak inverse correlation** between the 2 variables. In other words, as loan duration increases, the likelihood of it belonging to good credit class decreases slightly. This implies that customers with longer loan durations are more likely to be classified as bad credit, whereas customers with shorter loan durations are more likely to be classified as good credit.

Some possible explanations for the conclusion can be deduced. For example, customers with longer loan durations may struggle with repayments due to high interest rates, thus leading to a higher probability of being classified as bad credit. Conversely, customers with shorter loan durations might reflect smaller loan amounts, so their credit class is more likely to be good.

### **3.1.4 Analysis 4: Is the Difference in Mean of Loan Duration for Good and Bad Credit Class Statistically Significant?**

|  |  |
| --- | --- |
| Type of Analysis | Descriptive Analysis |
| Independent Variable(s) | Loan Duration (continuous data) |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | T-test |

We will utilize the t-test to determine whether credit class is actually affected by loan duration, and it is not just a coincidence. The t-test compares the difference in mean duration between good and bad credit classes to find out if the two variables are statistically significant. For this test, our null hypothesis (H0) would be that the difference in means between bad and good credit class is equal to 0, whereas our alternative hypothesis (H1) would be that the difference in means between bad and good credit class is not equal to 0.

A black and white text

Description automatically generated

Figure 3.1.4.1: Code for t-test (duration & class)

A white background with black text

Description automatically generated

Figure 3.1.4.2: Output of t-test (duration & class)

From the results of the t-test, we can observe a p-value that is extremely small **(< 2.2e – 16)**. This is much lower than the significance level, assuming that significance level 𝛼 = 0.05. Therefore, we can reject the null hypothesis and conclude that the difference in means between bad and good credit class is **not equal to 0**, thus proving that they are statistically significant. Apart from that, the t statistic, which measures the difference of mean between the 2 groups, has a rather large value of **18.5**. This indicates that there is a large difference between the 2 sample sets, which represents further evidence against the null hypothesis.

### **3.1.5 Analysis 5: Why do Some Customers with Short Loan Duration have Bad Credit Class and Customers with Long Loan Duration have Good Credit Class?**

|  |  |
| --- | --- |
| Type of Analysis | Diagnostic Analysis |
| Independent Variable(s) | Loan Duration (continuous data),  Employment Duration (categorical data) |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | Mosaic Plot, Chi-Square Test |

In order to analyse this, we will have to investigate other variables that might explain the relationship between loan duration and credit class. Mosaic plots are the perfect way to visualize how these variables interact with each other.

A computer screen shot of a program

Description automatically generated

Figure 3.1.5.1: Code for Mosaic Plot (employment vs class - loan duration)

A screenshot of a computer screen

Description automatically generated

Figure 3.1.5.2: Mosaic Plot for Employment vs Class for Short Loan Durations

A screenshot of a graph

Description automatically generated

Figure 3.1.5.3: Mosaic Plot for Employment vs Class for Long Loan Durations

Here are 2 mosaic plots that show the relationship between credit class and employment for short and long loan durations. An assumption is made where **durations that are less than the first quartile (25%) are considered short loan durations, whereas durations that are more than the third quartile (75%) are considered long loan durations.** From the first chart (short loan durations), we can see that the width of the good credit class is significantly larger than the width of the bad credit class, while the opposite is true for the second chart (long loan duration). This means that that the **proportion of good credit class is much more than bad credit class for short loan durations**, but the opposite applies to long loan durations with bad credit class loans taking precedence. Apart from that, a comparison of the 2 plots shows that customers with good credit class tend to have more balanced representation in employment categories, whereas **bad credit class more dominated by shorter employment durations**. In the second graph, it even shows that majority of those who have long loan durations, but good credit class have been employed for **at least 7 years**. Thus, it can be observed that even though some customers have short loan durations, their credit class may still be considered as bad due to short employment durations. A possible explanation for this is that since most of those customers have not been employed for very long, their salary might not be enough to repay their loans, resulting in a bad credit class. On the other hand, some customers have good credit class despite the fact that they have longer loan durations. These customers are mostly financially secure and have a stable employment, so they are able to keep up with the monthly payments and long repayment periods.

A close-up of a computer code

Description automatically generated

Figure 3.1.5.4: Code for Chi-Squared Test (Employment & Class)

A computer code with blue text

Description automatically generated

Figure 3.1.5.5: Output of Chi-Squared Test (Employment & Class)

Just to confirm that there is association between employment duration and credit classification, we use the Chi-squared test. We do not use the t-test in this scenario because we are dealing with 2 categorical variables. From the result, we can see that the **p-value is a lot lower** than the significance value of 0.05, so it means that employment duration is a significant factor of determining credit class. Therefore, we can conclude that employment duration is one of the reasons why some customers with short loan durations have bad credit class and vice versa. We can also draw another conclusion that customers are encouraged to have a secure employment before taking loans, especially long duration loans if they want to maintain a good credit class. This whole process can be repeated with other variables to analyse their relationships and how they affect credit class as well.

### **3.1.6 Analysis 6: Can the Loan Duration be Used to Predict Credit Class of Customers? Will the Prediction Improve When Considering Other Variables?**

|  |  |
| --- | --- |
| Type of Analysis | Predictive Analysis |
| Independent Variable(s) | Loan Duration (continuous data) & others |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | Logistic Regression, Confusion Matrix |

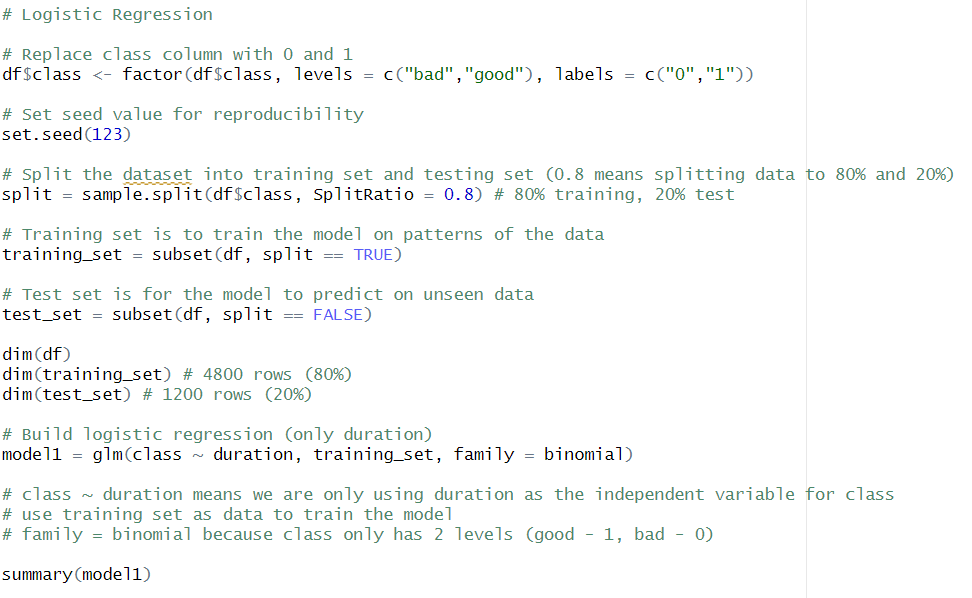


Figure 3.1.6.1: Code for Logistic Regression (Class and Duration)

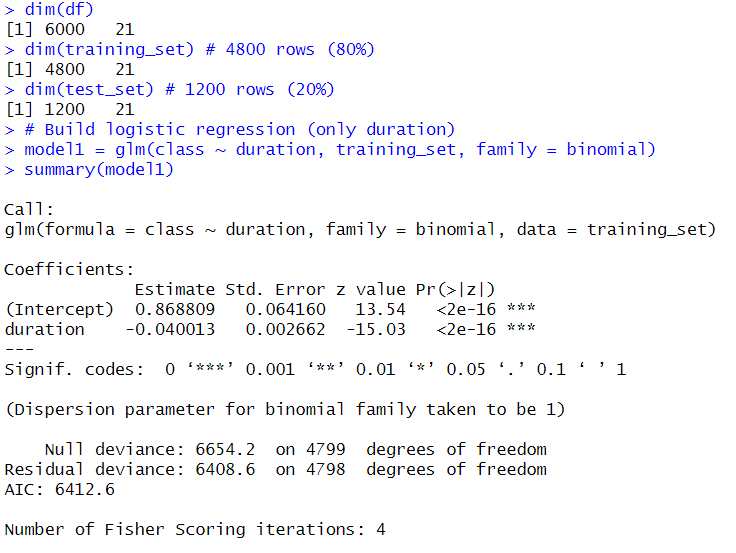


Figure 3.1.6.2: Output of Logistic Regression (Class and Duration)

In this first code, we use a logistic regression model on the training set to quantify the association between duration and class (Peng et al., 2002). Based on the output, we can retrieve the coefficient estimate for duration of **-0.040013**, which means that as duration increases, the likelihood of class being good decreases. In addition, the p-values for both coefficients are **less than 0.05**, indicating that they are statistically significant and that there is strong evidence that duration affects the outcome of class. These results align with the findings in previous analyses.

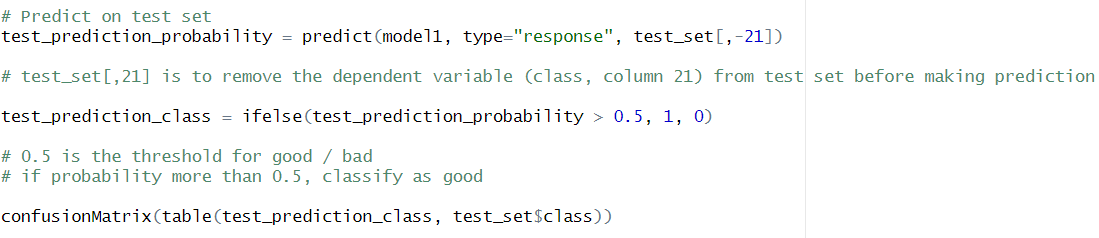


Figure 3.1.6.3: Code for Prediction and Confusion Matrix (Duration and Class)

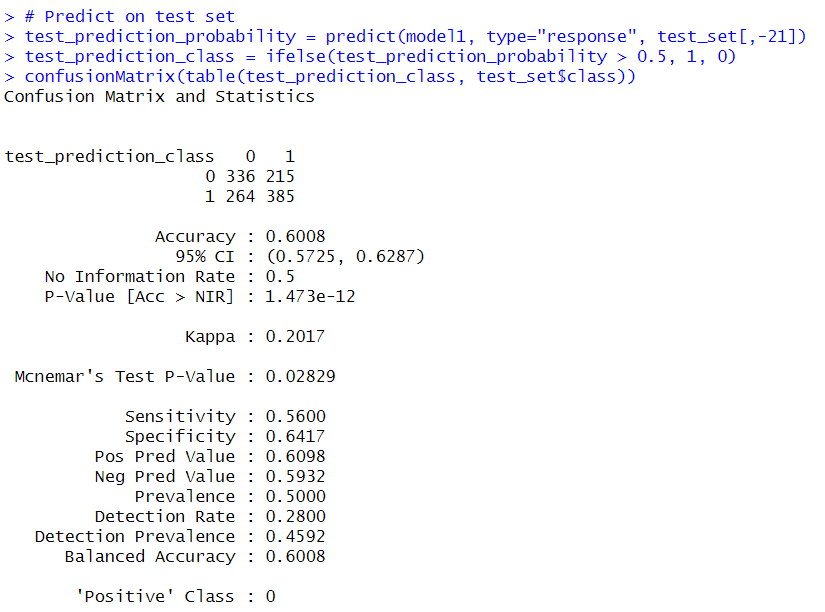


Figure 3.1.6.4: Output of Confusion Matrix (Duration and Class)

By using a confusion matrix, we are able to define the performance of the classification algorithm (Singh & Singh, 2021). From the generated matrix, we can interpret that 336 entries have been correctly predicted to be in the bad class, while 385 entries have been correctly predicted to be in the good class. However, there are **215 false negatives** (type II error), where entries were supposed to be in the good credit class but were predicted to be in the bad credit class. Likewise, there are also **264 false positives** (type I error), where customers were misclassified as having bad credit class when they actually had good credit class. This resulted in a rather low **accuracy of 60.08%,** proving that the prediction model is quite weak. Besides that, the **kappa statistic of** **0.2017** is close to 0, denoting that the model's predictions are only slightly better than random guessing and has **poor agreement** with the actual data. Another thing to note is that the specificity value is higher than the sensitivity value, suggesting that the model struggles more with identifying those with good credit class than bad credit class. Thus, to improve the model’s predictive performance, we should use all independent variables in the logistic regression instead of just relying on the duration variable.

A screen shot of a computer code

Description automatically generated

Figure 3.1.6.5: Code for Logistic Regression, Prediction and Confusion Matrix (All Variables)

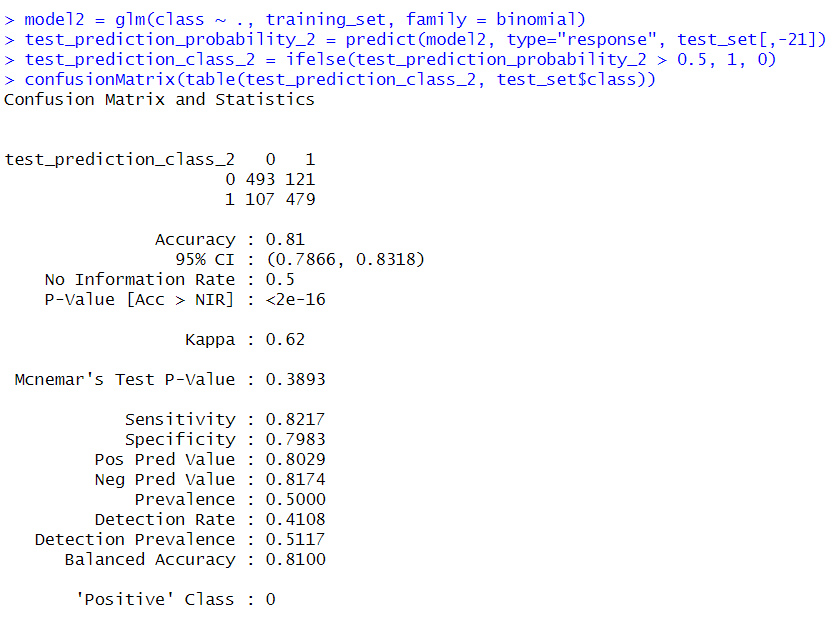


Figure 3.1.6.6: Output for Confusion Matrix (All Variables)

Through incorporating all other independent variables, we can see that the model has a much **higher accuracy (81%).** The number of **type I and II errors have also decreased**, and the kappa statistic now represents a **moderate agreement** with the actual data. As a result, we can conclude that just using duration as a predictive measure may not be sufficient, and including other independent variables can provide a more reliable forecast.

### **3.1.7 Analysis 7: What Can Customers Do to Achieve Good Credit Class?**

|  |  |
| --- | --- |
| Type of Analysis | Prescriptive Analysis |
| Independent Variable(s) | Loan Duration (continuous data) |
| Dependent Variable | Credit Class (categorical data) |
| Techniques Used | Decision Tree |

A black and blue text

Description automatically generated

Figure 3.1.7.1: Code for Decision Tree Using Duration Only

A diagram of a diagram

Description automatically generated

Figure 3.1.7.2: Decision Tree of Duration Only

A decision tree is used in this scenario to classify data based on certain conditions. The correlation analysis and regression analysis conducted previously has already indicated that shorter loans tend to result in better credit class, but we can use a decision tree to get more detailed insights. From the tree, it is clear that those who have a loan duration of **12 months or less** have the **highest probability** of achieving a good credit class. However, taking loans of at most **5 months** yield the best likelihood of a good credit class. On that account, customers that want good credit class are advised to take short loans with terms of **around 1 to 12 months**.

|  |  |
| --- | --- |
| **Extra Features** | **Justification** |
| Correlation Analysis (Point-Biserial Correlation) | Find correlation between categorical variable (class) and continuous variable (duration) |
| T-test | Test if the difference in mean of duration for 2 groups of class is statistically significant |
| Mosaic Plot | Analyse the relationship of employment and class (2 categorical variables) for short and long loan durations |
| Chi-Square Test | Test if the employment and class is statistically significant |
| Logistic Regression | Used to examine association of duration and class and predict test set based on data learned from the training set |
| Confusion Matrix | Visualise and summarize the results of the prediction |
| Decision Tree | Classify duration and find patterns |

## **3.2 To identify the impact of employment status and saving status to credit class. (Sweetha Pramasivam)**

### **3.2.1 Analysis 1: How does the distribution of saving status vary across different employment statuses for each credit class?**

|  |  |
| --- | --- |
| **Type** | **Descriptive Analysis** |
| Independent variables | Employment, Savings\_status |
| Dependent variables | Credit Class |
| Analysis techniques | Ridge Density Plot, Multinomial Logistic Regression |
| visualizations | Ridge Density Plot, Stacked Bar Chart of Predicted Savings |

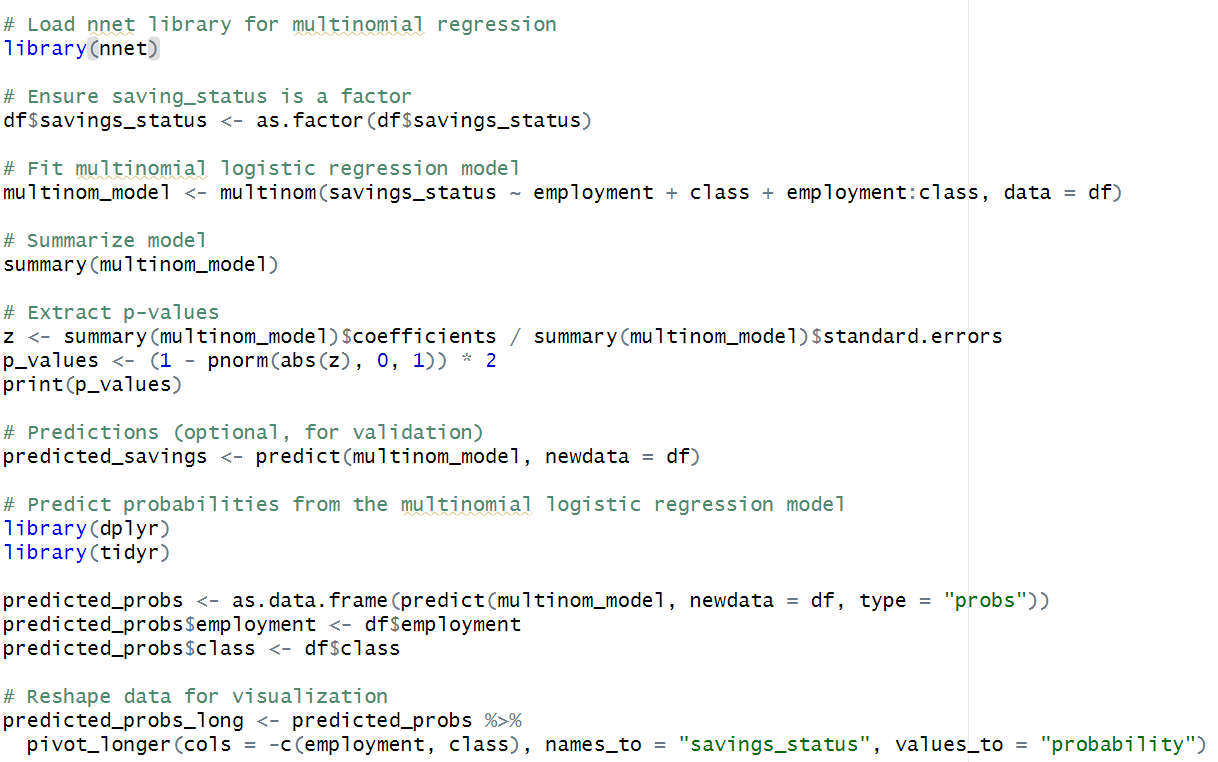
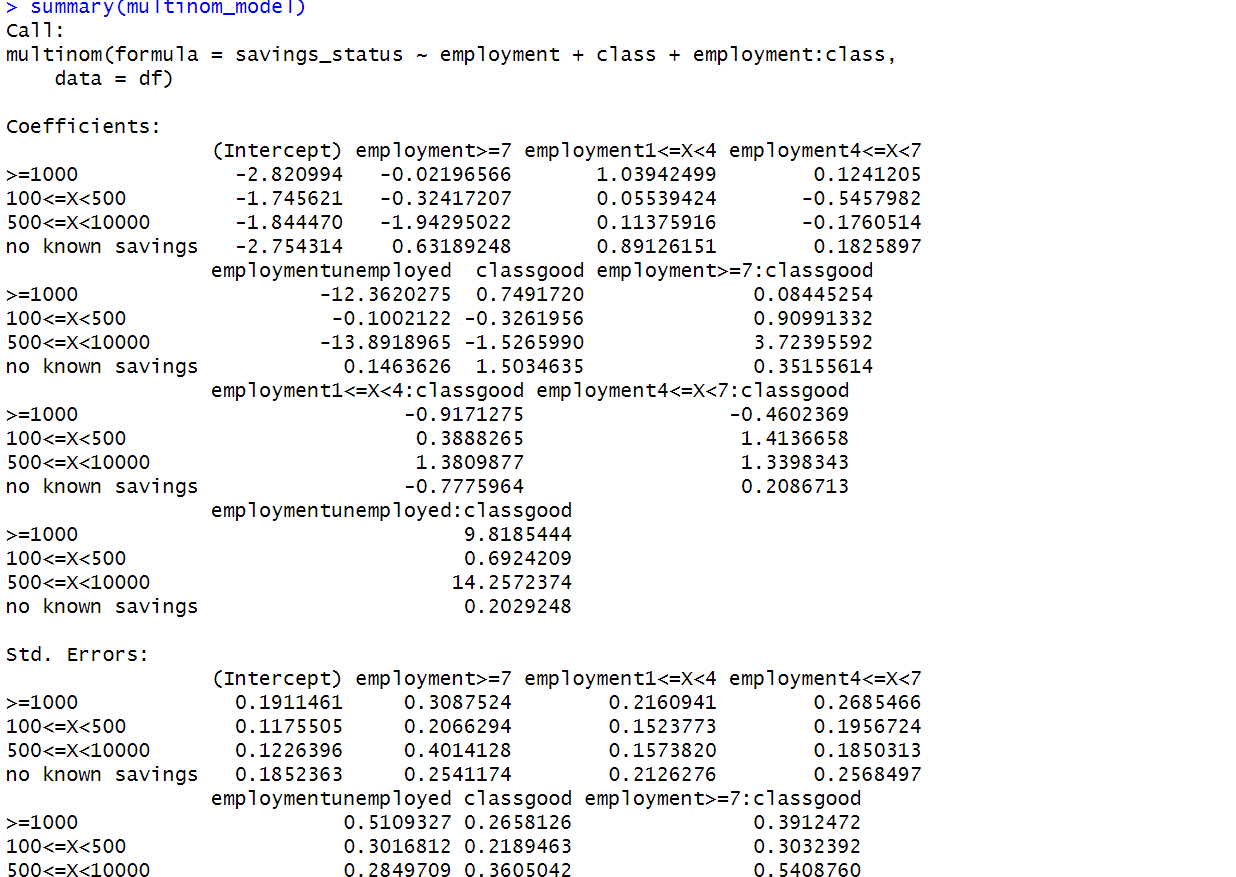
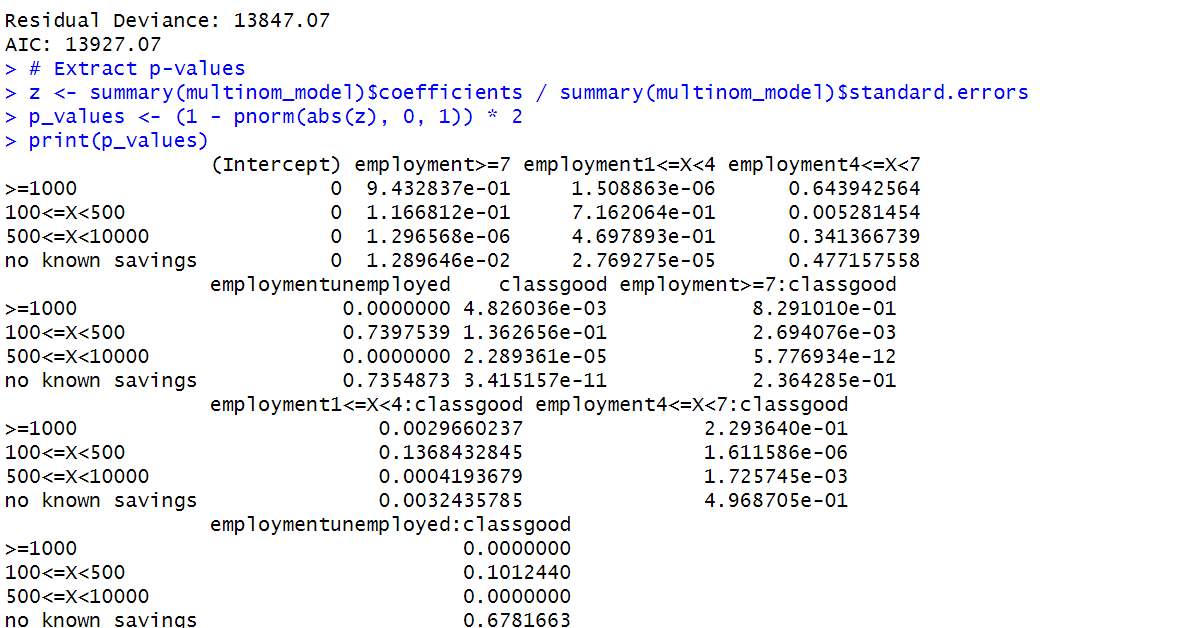
Figure 3.2.1.1: Code for Multinomial Logistic Regression

Figure 3.2.1.2: Output for Multinomial Logistic Regression  
Figure 3.2.1.3: Output for Multinomial Logistic Regression

Figure 3.2.1.4: Output for Multinomial Logistic Regression

Explanation:

The following multinomial logistic regression model explores the relationship between saving status-the dependent variable-and job status, credit class, and their interaction-independent variables. This model converged after 50 iterations with a residual deviance of 13,847.07 and an AIC of 13,927.07, indicating a good fit of the model to the data. Significant coefficients such as job levels and their relationships with credit class provide crucial predictors for various savings categories. The p-values indicate significant relationships, such as employment>=7 and employment-unemployed for some savings categories. Interaction terms, such as employment1<=X<4: classgood, point to the interaction of employment and credit class with respect to savings status. Predicted probabilities, as provided by the algorithm, are molded for visualization, enabling deep investigation of the probability distributions across different employment and credit types. Generally, this would provide the required analysis of characteristics that influence saving practices to enable financial or governmental actions based on such findings.

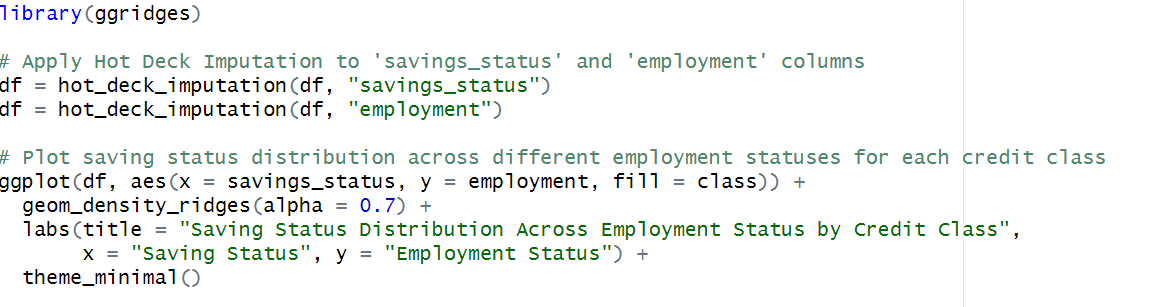
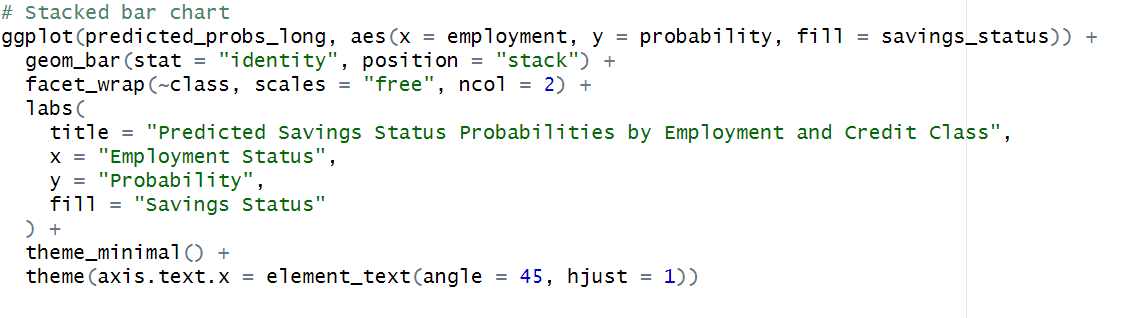
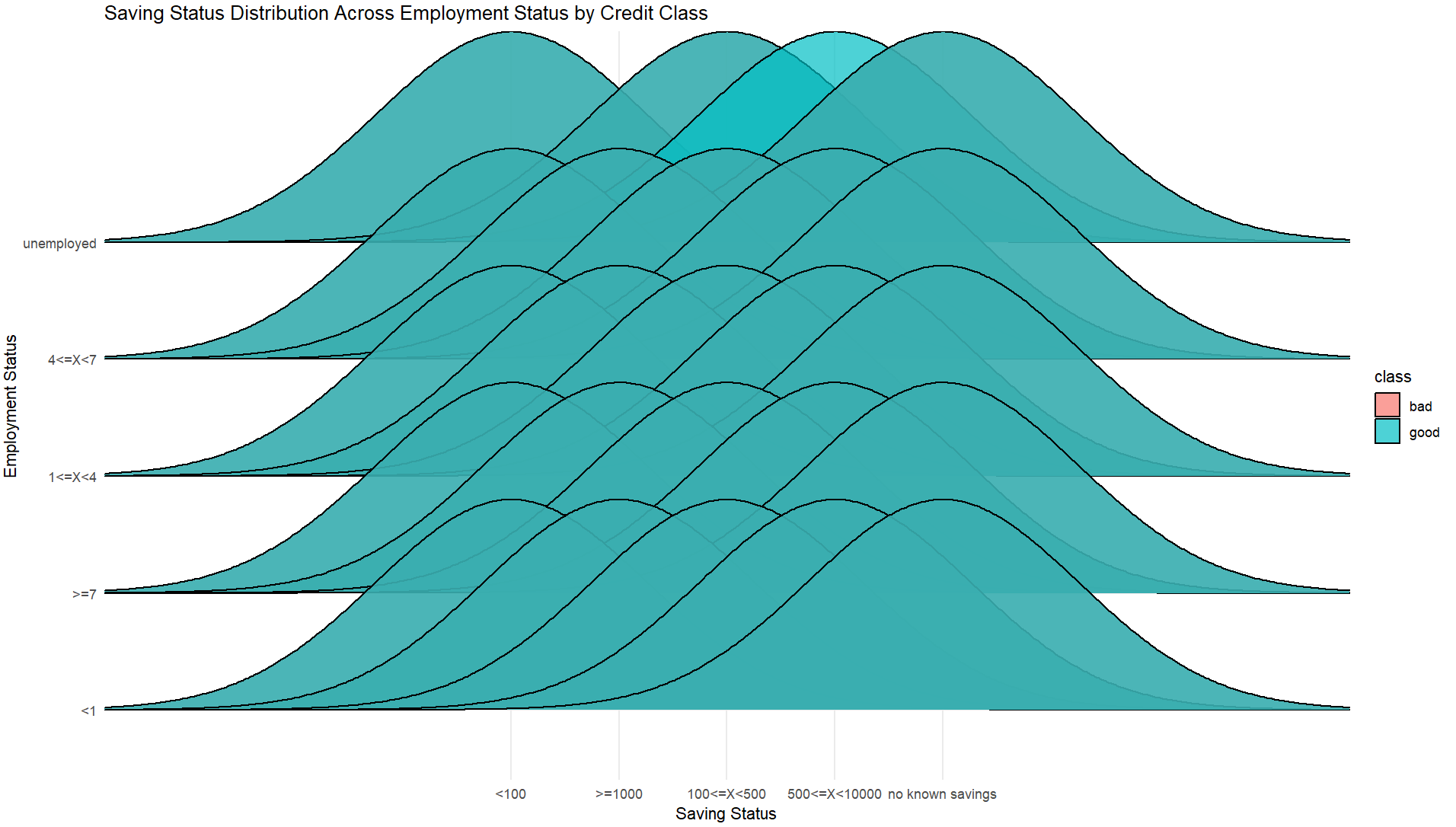
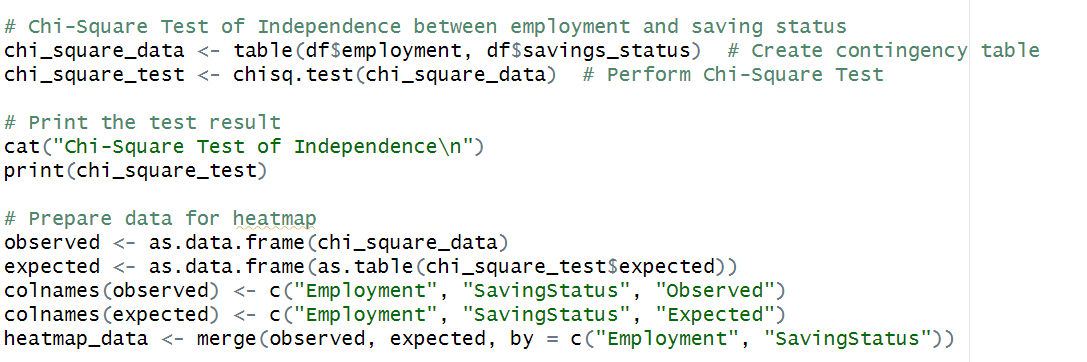
Figure 3.2.1.5: Code for Stacked Bar Plot  
Figure 3.2.1.6: Code for Stacked Bar Plot

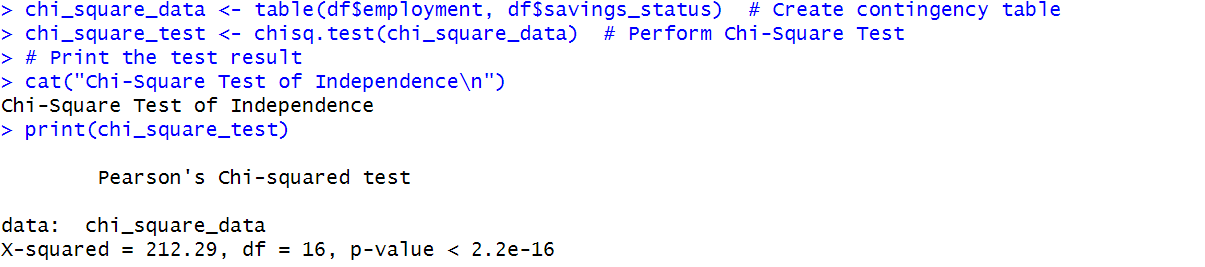
Figure 3.2.1.7: Output for Stacked Bar Plot

Figure 3.2.1.8: Output for Stacked Bar Plot

### **3.2.2 Analysis 2: What proportion of customers with different saving statuses exist within each employment status, and how does this relate to their overall population size?**

|  |  |
| --- | --- |
| **Type** | **Descriptive Analysis** |
| Independent variables | Employment, savings\_status |
| Dependent variables | - |
| Analysis techniques | Chi-Squared Test of Independance, Proportional Heatmap Visualization |
| visualizations | Mosaic Plot, HeatMap |

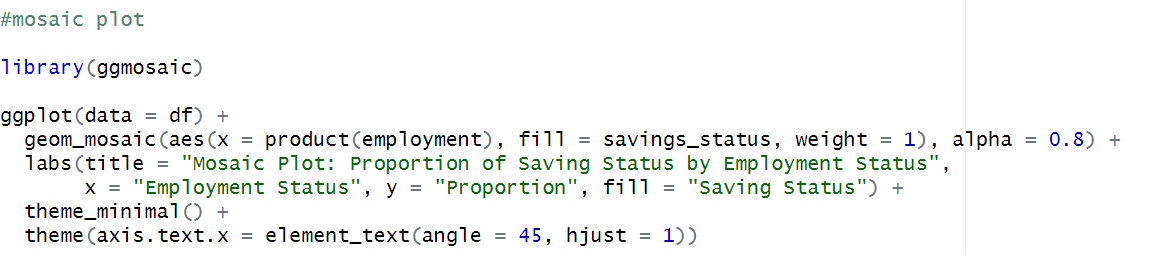
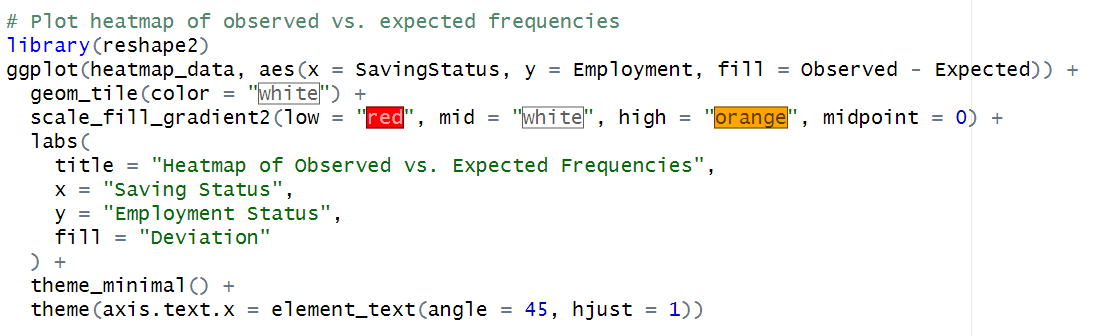
Figure 3.2.2.1: Code for Chi-Squared Test of Independance

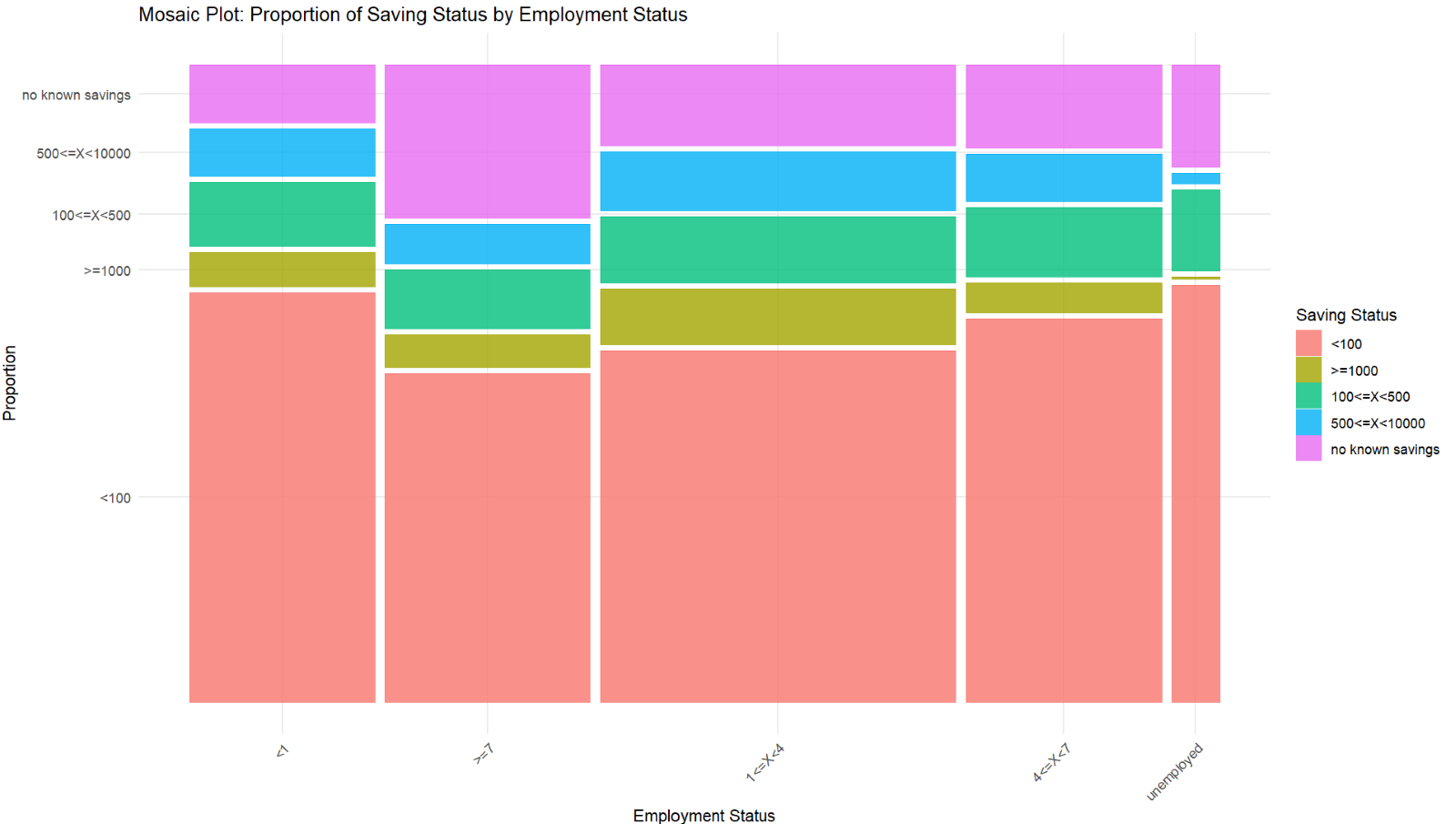
Figure 3.2.2.2: Output for Chi-Squared Test of Independance

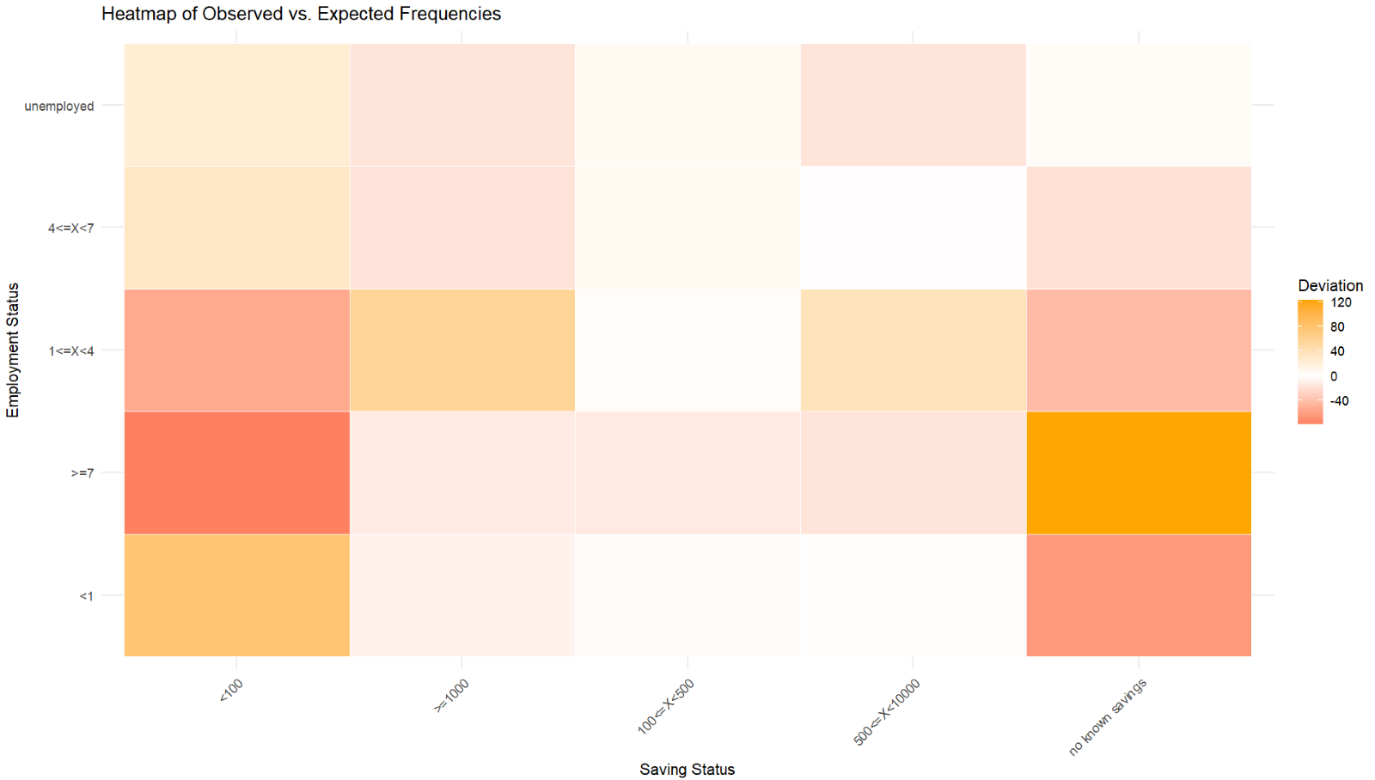
Explanation:

The Chi-Square Test of Independence determines whether the employment status and savings status are linked statistically. The test makes use of the contingency table of the two categorical variables, resulting in a chi-squared statistic of 212.29 with 16 degrees of freedom. The p-value (< 2.2e-16) is extremely low, much below any common significance level (e.g., 0.05). This indicates a strong statistical dependence of work status on savings status and hence negates the hypothesis of independence.

This portrays that work status significantly contributes to savings behavior, showing how employment levels change the categorization of savings. All these would be important information in financial planning and policy formulation and in making relevant targeted interventions in increasing savings across all employments categories. The value of chi-square is greater; hence, there were differences between observed and predicted frequency. This also emphasizes that the association is strong. Further investigation could reveal some pattern changes in savings habits in different employment groups.

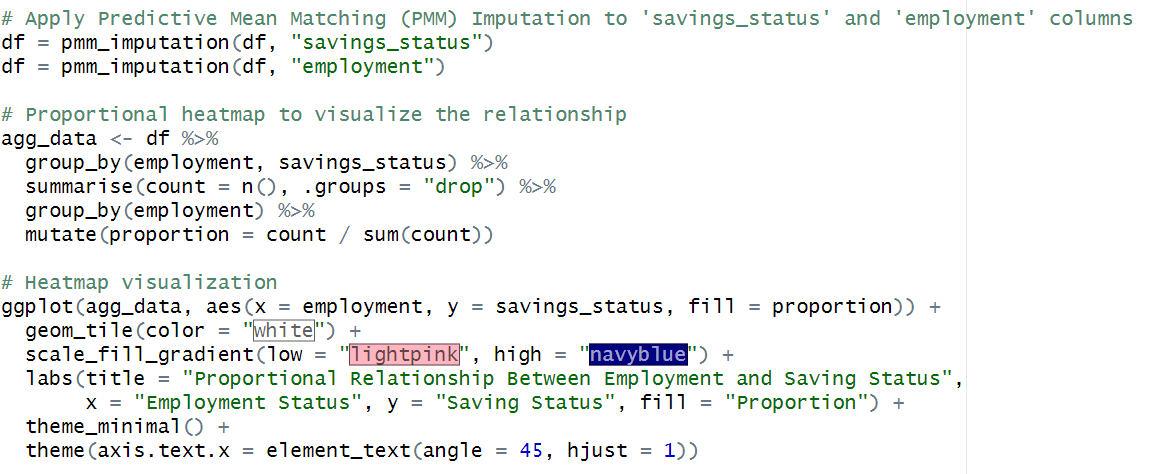
Figure 3.2.2.3: Code for Mosaic Plot  
Figure 3.2.2.4: Code for HeatMap

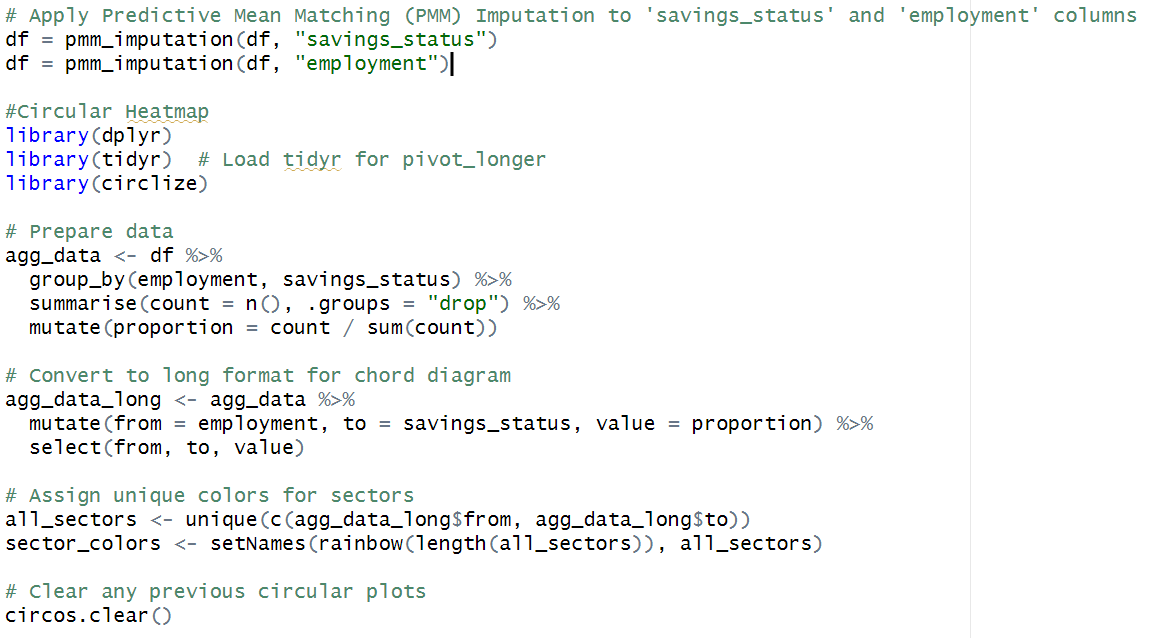
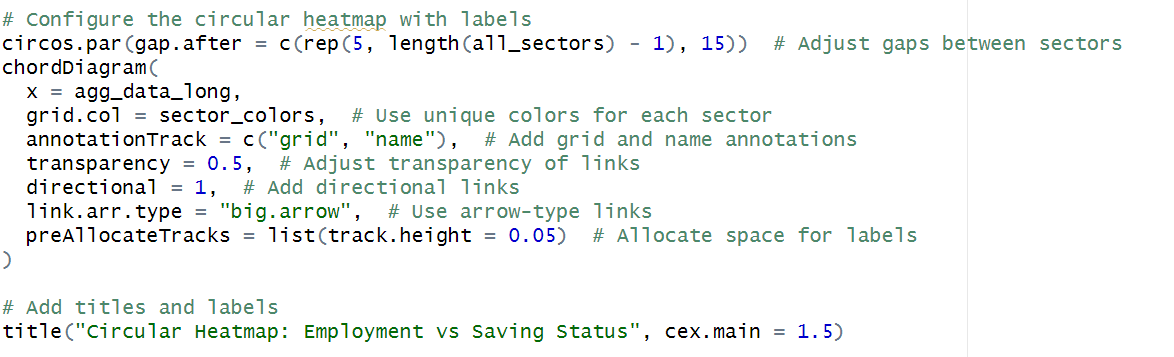
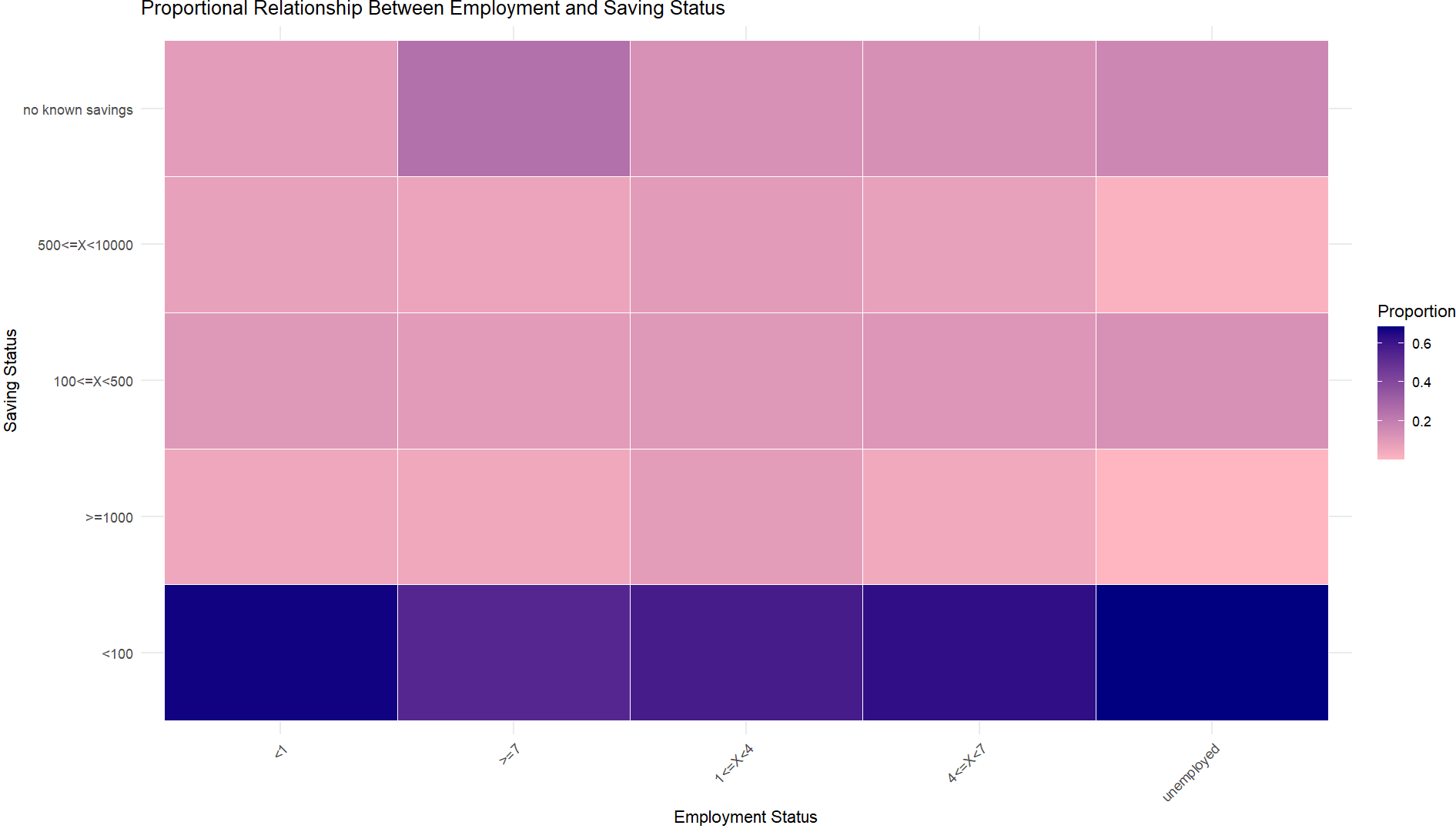
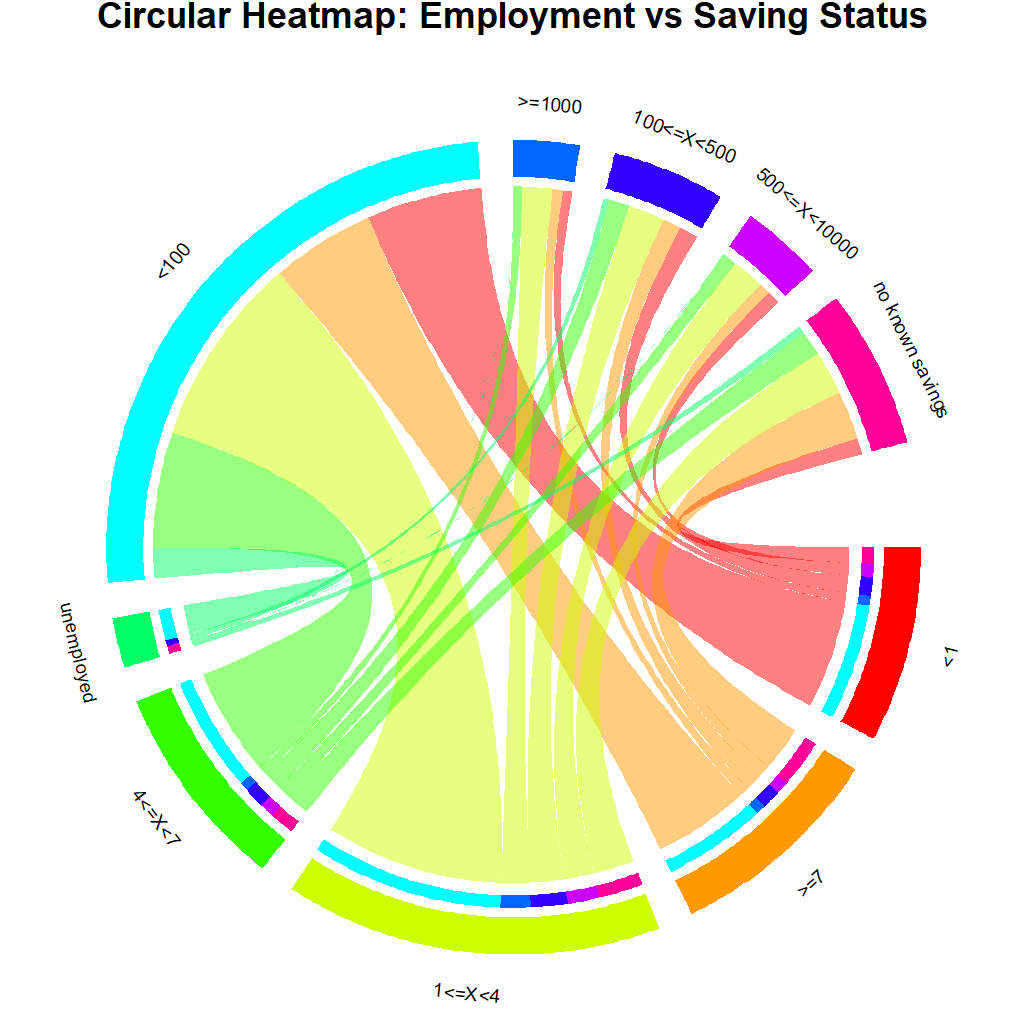
Figure 3.2.2.5: Output for Mosaic Plot

Figure 3.2.2.6: Output for Heatmap

**3.2.3 Analysis 3: How does saving status change across employment levels, and how is it affected by imputing missing values?**

|  |  |
| --- | --- |
| **Type** | **Descriptive Analysis** |
| Independent variables | Employment, Saving\_status |
| Dependent variables | - |
| Analysis techniques | Heatmap |
| visualizations | Heatmap of proportions and Circular Heatmap |

Figure 3.2.3.1: Code for Proportional Heatmap

Figure 3.2.3.2: Code for Circular Heatmap  
Figure 3.2.3.3: Code for Circular HeatmapFigure 3.2.3.4: Output for Proportional Heatmap  
Figure 3.2.3.5: Output for Circular Heatmap

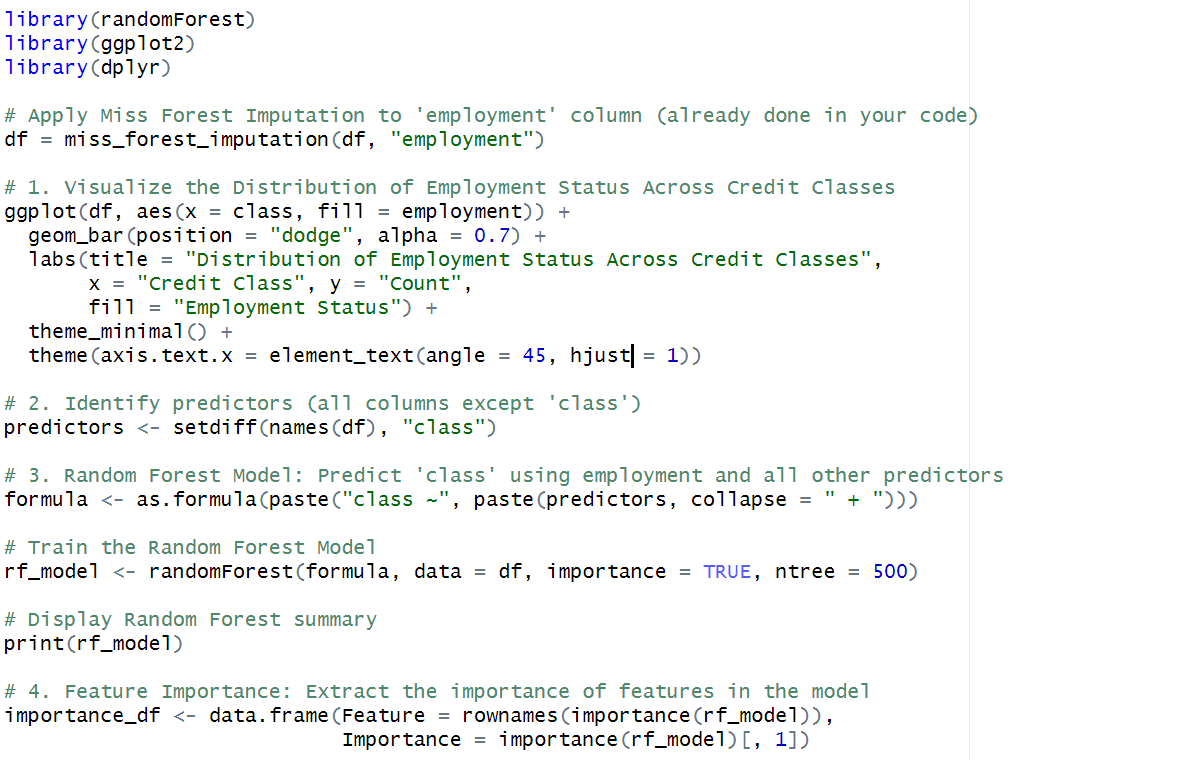
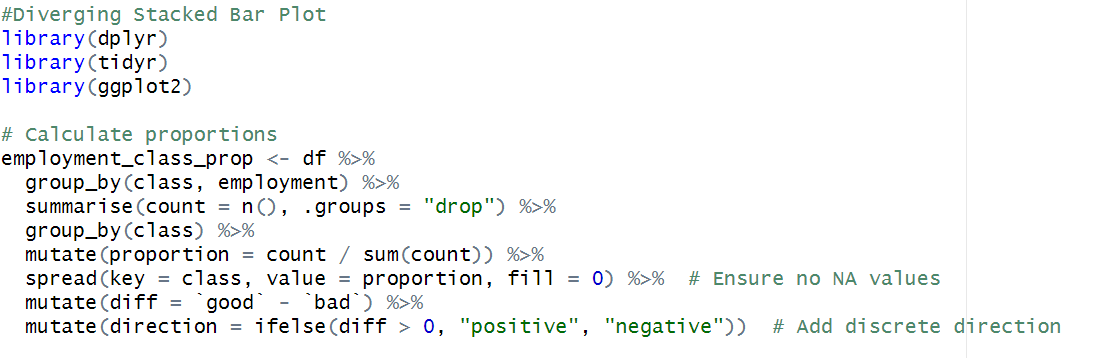
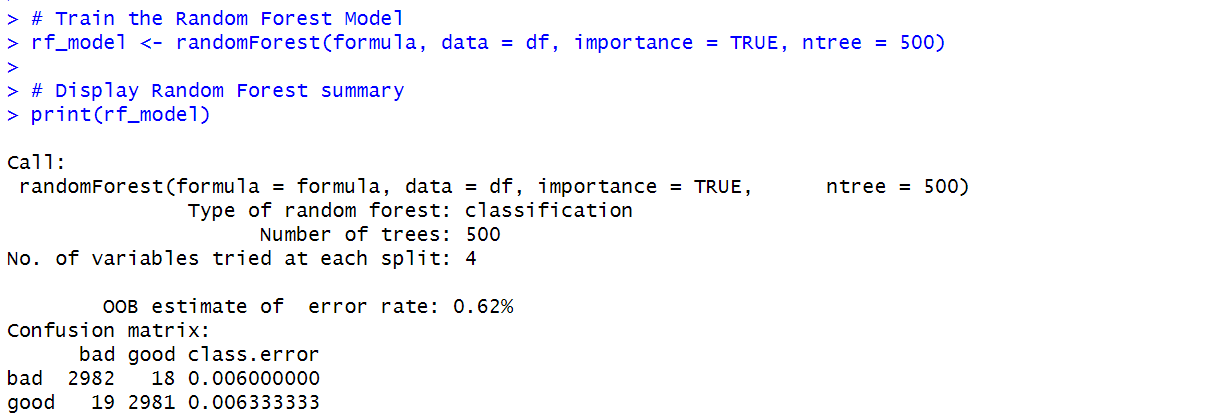
Explanation:

The round heat map illustrates the relationship between employment status and savings status, emphasizing the proportionate distribution of these categories. Each slice of the circle represents an employment category or savings status and is connected by directional links that depict the transitions or linkages between different groupings. The thickness of the linkages corresponds to proportional flow, with thicker links representing stronger relationships. This heatmap is based on a dataset where missing values for "employment" and "savings\_status" were imputed using Predictive Mean Matching (PMM). PMM guarantees realistic imputed values because the missing entries are matched to observed ones with similar anticipated values, thus making data more reliable for analysis. The aggregated data was sorted and normalized to find proportions, which made the visualization useful for comparison examination. This visualization depicts important patterns, including the most common savings categories for each employment group. For instance, unemployed workers are highly associated with "no known savings" or low ranges of saving, while employed workers with higher incomes may relate to higher categories of saving. Color-coded sectors and transparent, directional relationships enhance interpretability by showing how employment status directly impacts saving behavior. This technique is very effective in determining how imputed data matches the observed patterns and, therefore, confirms the imputation procedure. Because of the interactive and attractive design of the heat map, it helps in identifying disparities and correlations between categories, giving a clear view of the relationship between employment and financial behavior. Such insights might be crucial for politicians and financial organizations seeking to improve saving behavior among various groups of employment.

### 

### **3.2.4 Analysis 4: What is the distribution of employment status across different credit classes (good/bad)?**

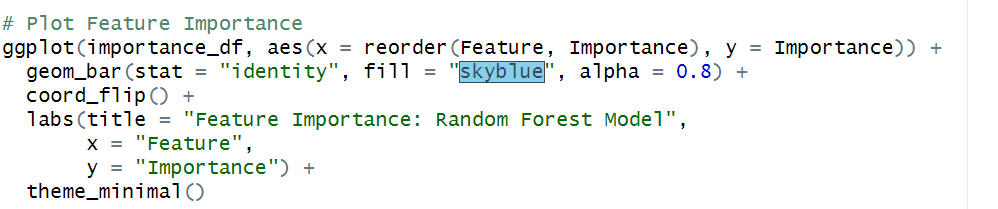
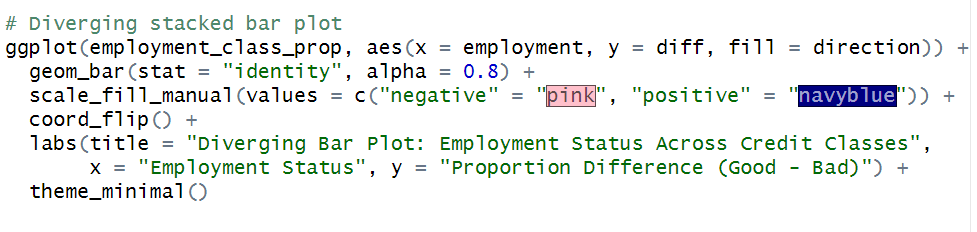
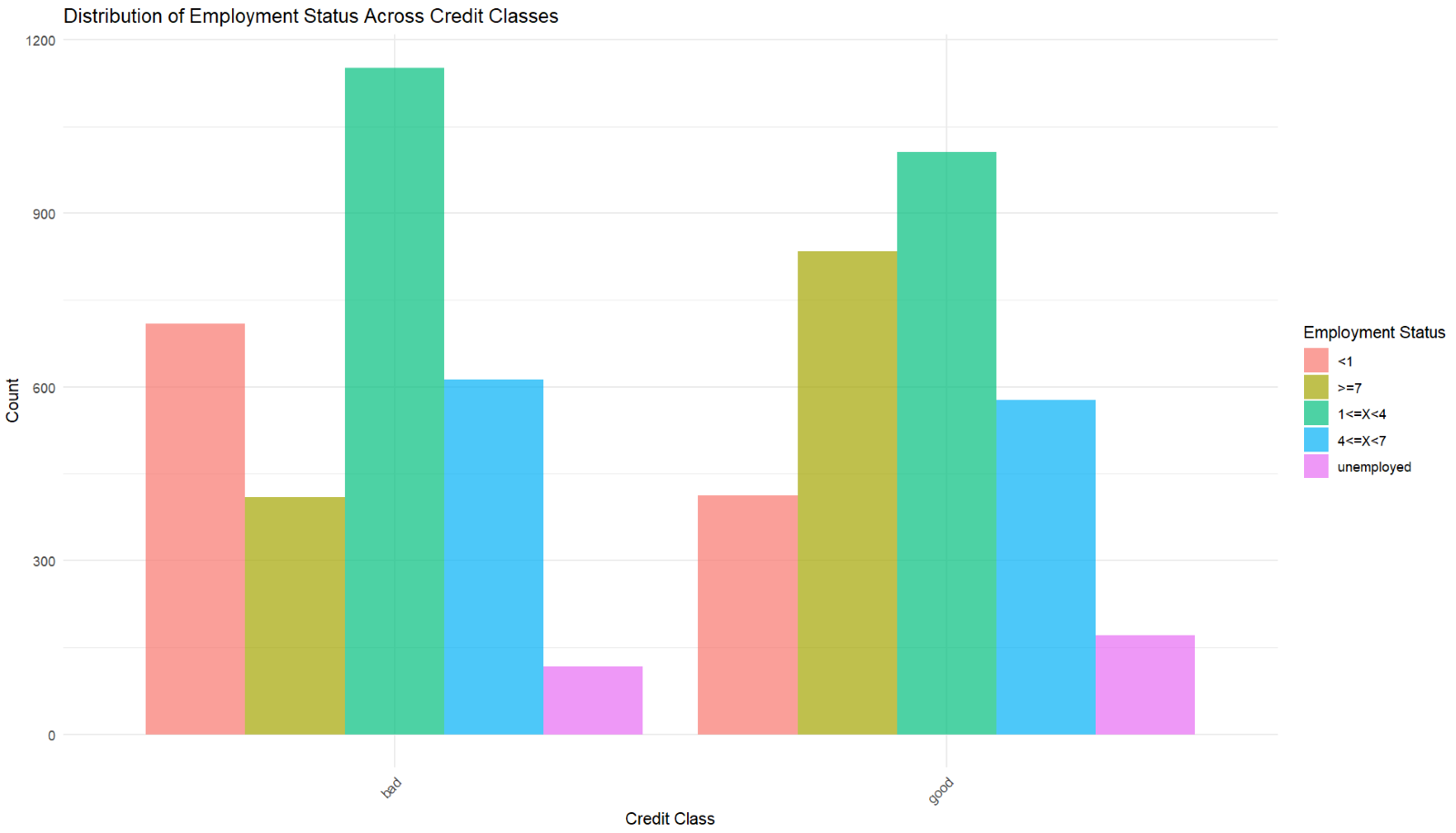
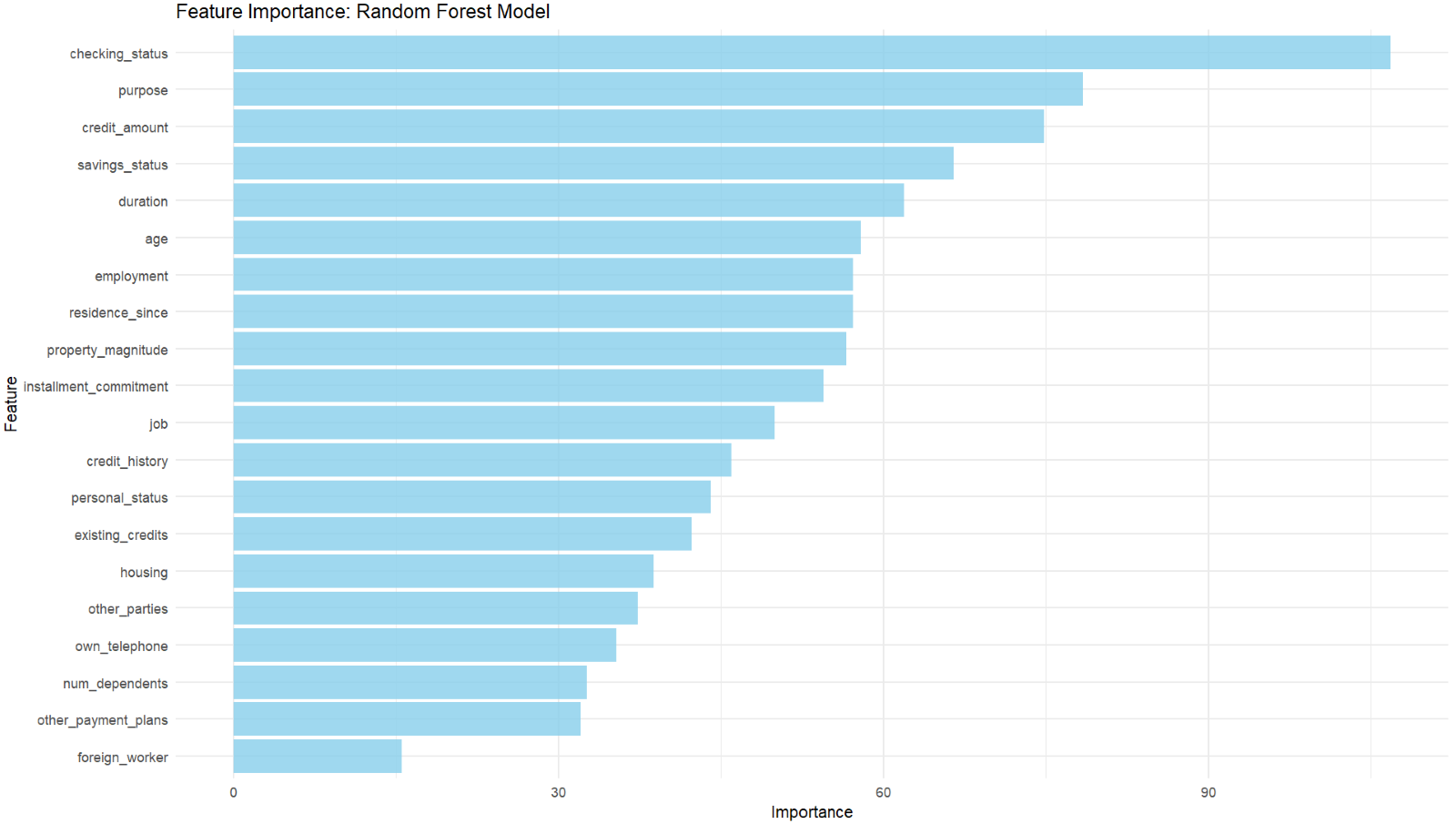
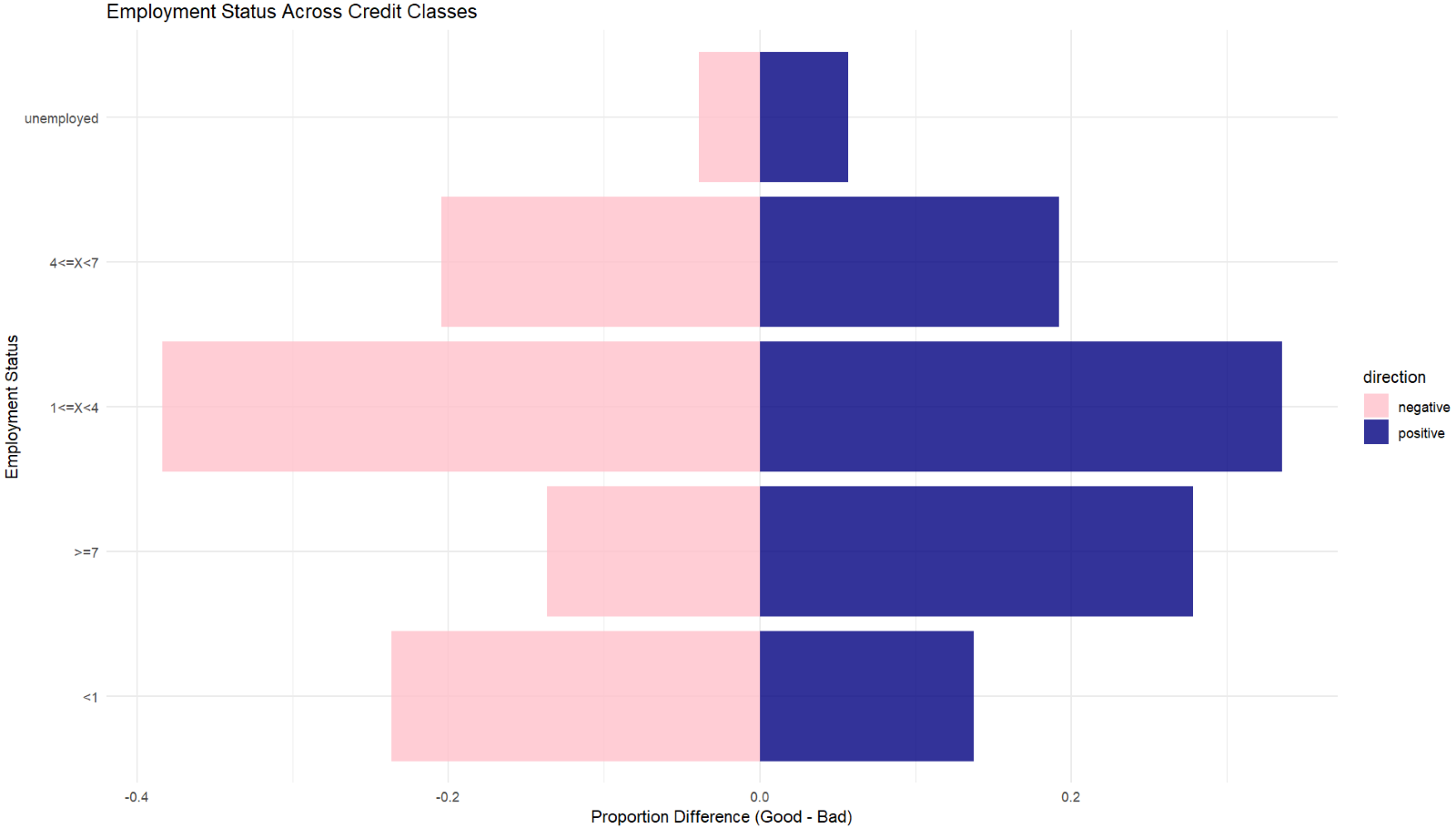
|  |  |
| --- | --- |
| **Type** | **Predictive Analysis** |
| Independent variables | Employment |
| Dependent variables | Credit Class |
| Analysis techniques | Random Forest Model, Diverging Stacked Bar Plot |
| visualizations | Diverging Stacked Bar Plot, Feature Importance Plot |

Figure 3.2.4.1: Code for Random Forest  
Figure 3.2.4.2: Code for Diverging Stacked Bar Plot  
  
Figure 3.2.4.3: Output for Random Forest  
:

Explanation:

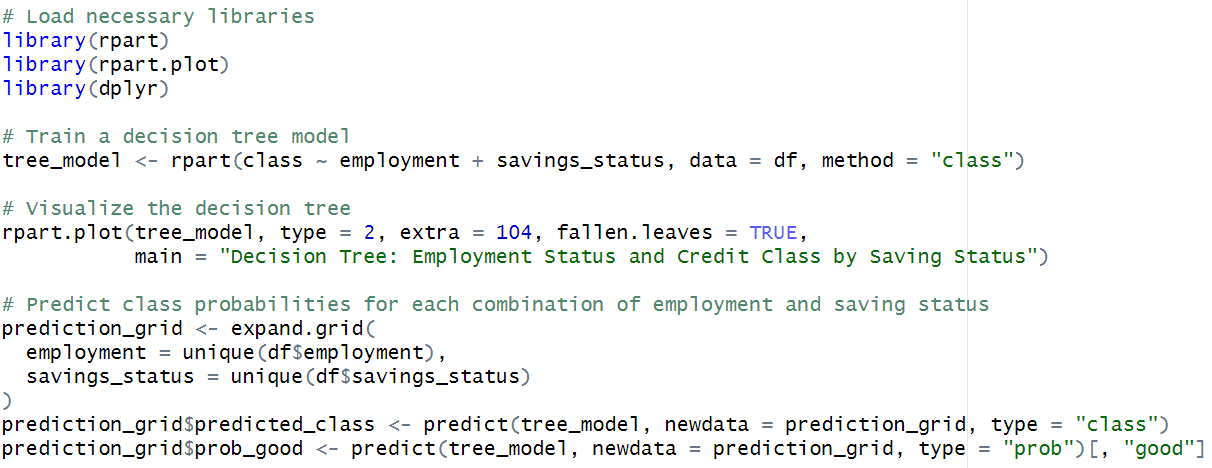
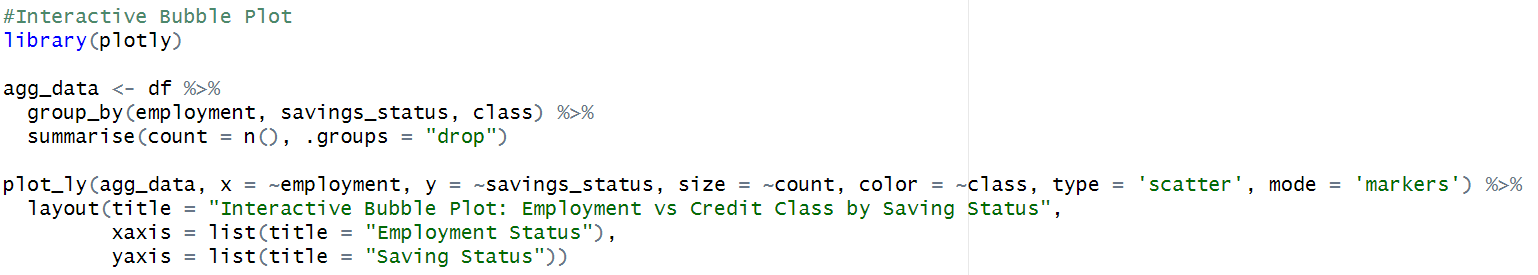
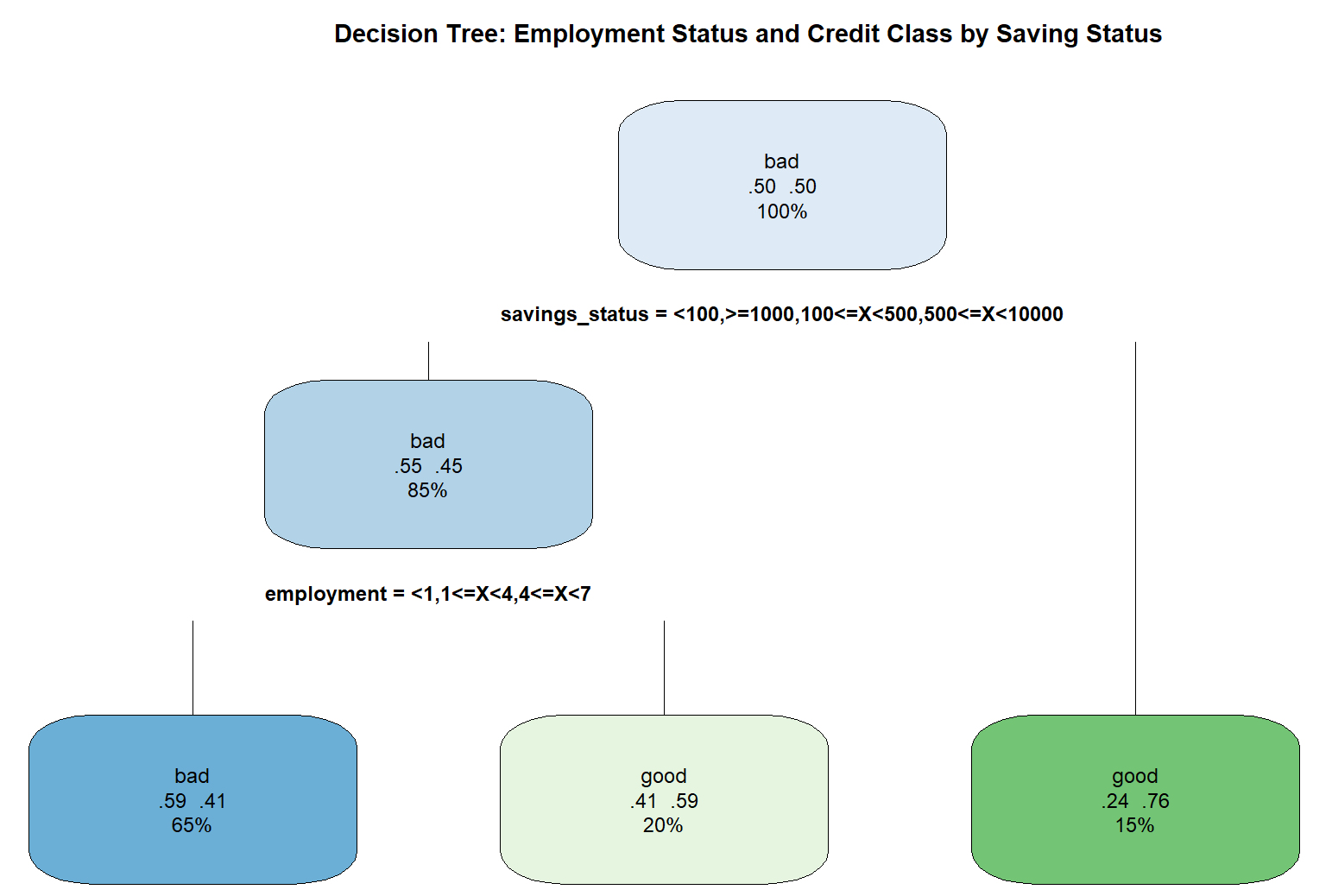
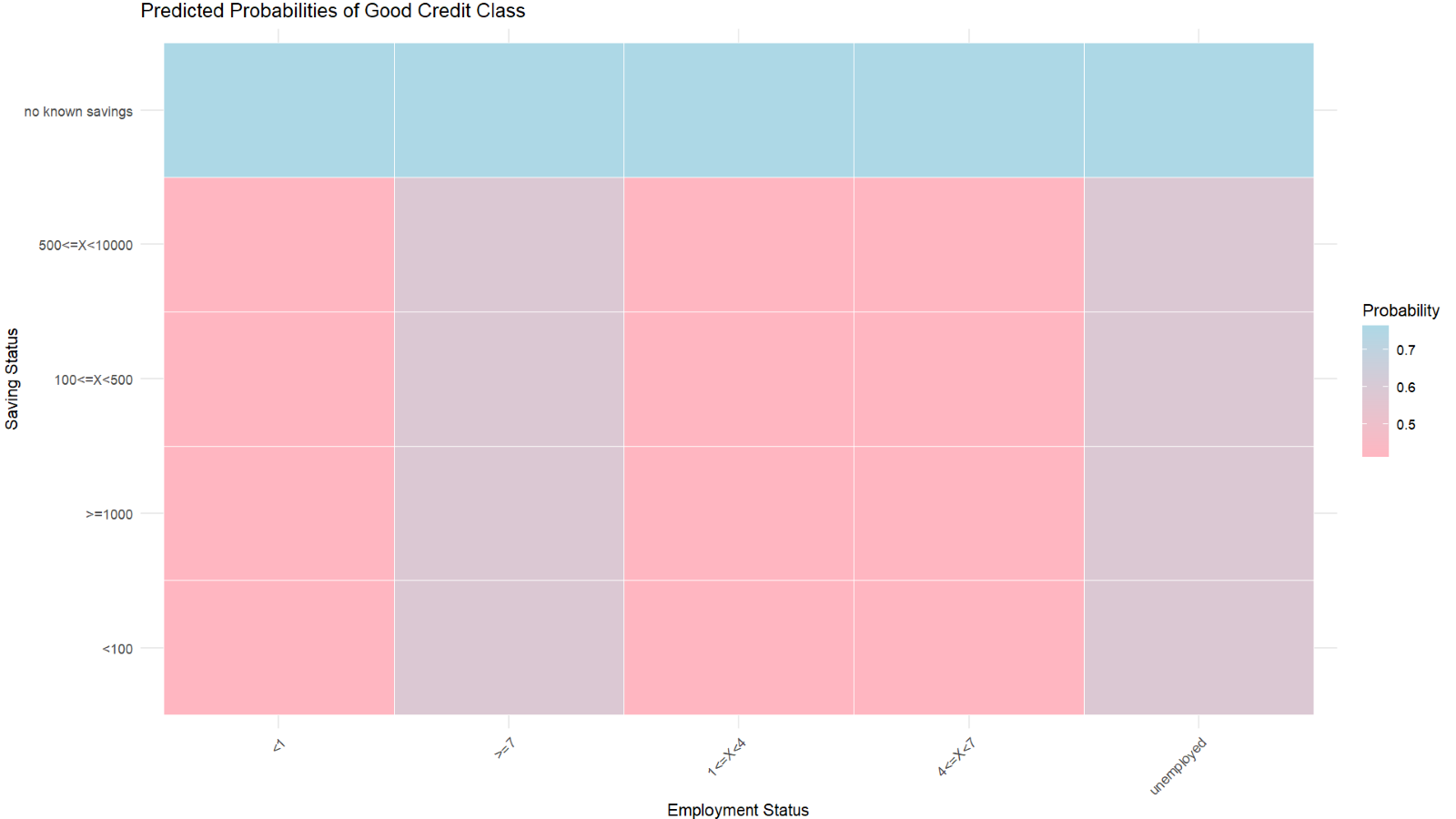
Missing value imputation, data visualization, and predictive modeling in this analysis are used to look at the association between a borrower's job status and class of credit. The paper has filled missing values within the column "employment" from the set with realistic estimates via Miss Forest Imputation for robustness. A bar chart will show the dispersion within each of the credit classes-generating good and bad-shedding light on employment impacts.

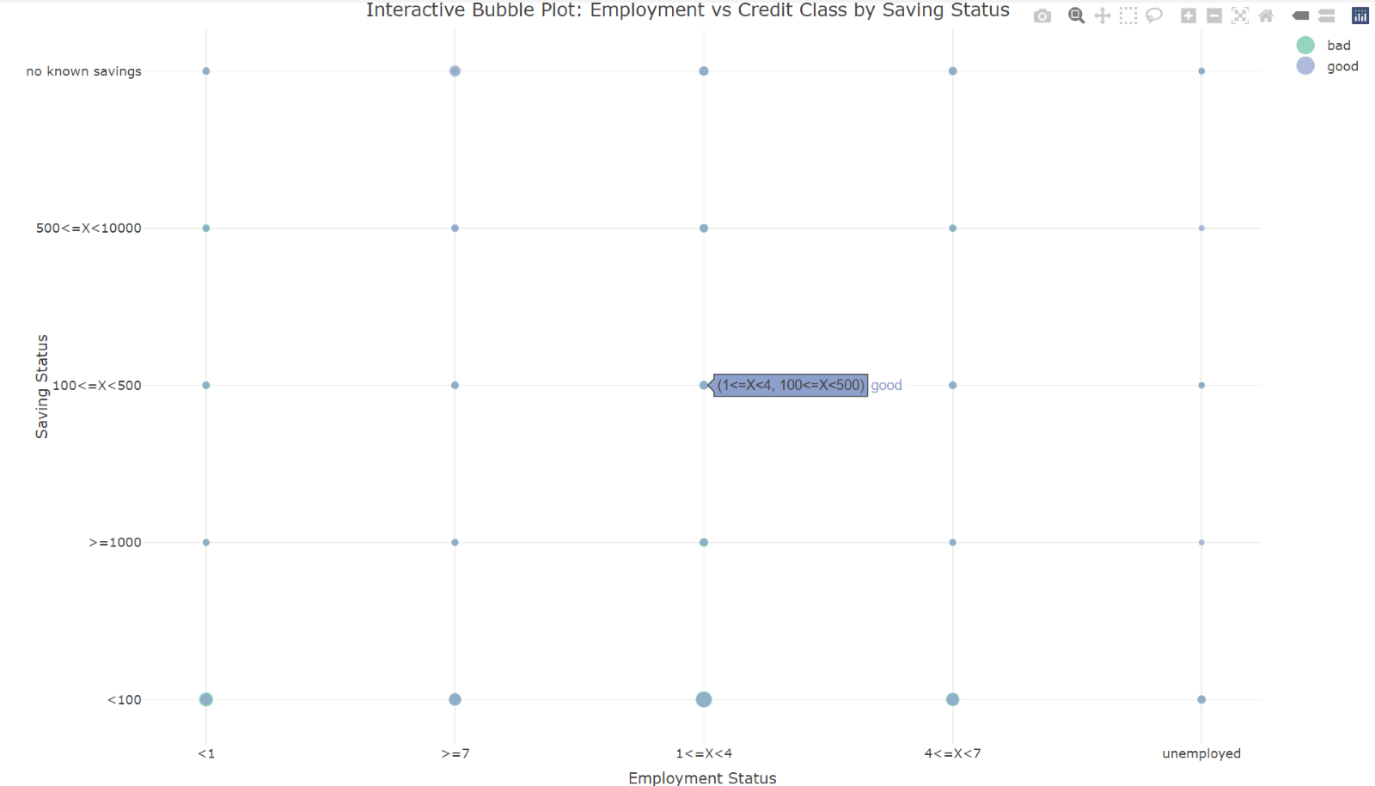
Next, the Random Forest model was used to predict credit class based on employment and other factors. The model developed on 500 trees, with four variables analyzed at each split, resulted in an impressive OOB error rate of 0.63%, which showed very accurate predictions. The accuracy confirms this, projecting only 17 "bad" cases as "good" and 21 "good" ones as "bad." It can be seen that this model is reliable and of high ability to distinguish creditworthiness based on the available predictors, including work status.

Figure 3.2.4.4: Code for Feature Importance Plot  
  
Figure 3.2.4.5: Code for Diverging Stacked Bar Plot  
  
Figure 3.2.4.6: Output for Distribution of Employment Across Credit Class  
  
Figure 3.2.4.7: Output for Feature Importance Random Forest Model  
  
Figure 3.2.4.8: Output for Diverging Stacked Bar Plot

### **3.2.5 Analysis 5: How does the relationship between employment status and credit class vary across different saving statuses?**

|  |  |
| --- | --- |
| **Type** | **Predictive Analysis** |
| Independent variables | Employment, Savings\_status |
| Dependent variables | Credit Class |
| Analysis techniques | Decision Tree Analysis (CART) |
| visualizations | Interactive Bubble Plot, Decision Tree Visualization, Heatmap of Predicted Probabilities |

Figure 3.2.5.1: Code for Decision Tree   
Figure 3.2.5.2: Code for Interactive Bubble Plot  
  
Figure 3.2.5.3: Code for Heatmap of Predicted Probabilities  
  
Figure 3.2.5.4: Output for Decision Tree  
  
Figure 3.2.5.5: Output for Heatmap of Predicted Probabilities

  
Figure 3.2.5.6: Output for Interactive Bubble Plot

Explanation

The decision tree in the graphic depicts the relationship between employment status, savings status, and credit class. It begins by segmenting individuals based on their savings status, which is the most significant predictor. Of those who had less savings (savings less than $100), the majority (85%) were classed as having "bad" credit, with work status further narrowing forecasts. Lower levels of work (e.g., ≤4 years) are associated with bad credit at 59 percent, whereas higher employment levels, such as 4≤X<7 years, increase one's likelihood of good credit to 59 percent.

Individuals with larger savings (say, $500 or greater) are more likely to have "good" credit, regardless of work status (76%). This means that the saving status exerts a stronger influence on creditworthiness, while work is relegated to a secondary role in the finer classification. The striking structure of this tree shows thresholds that are relevant for practical applications of this analysis of credit risk with the two variables under discussion.

### **3.2.6 Extra Features**

|  |  |
| --- | --- |
| **Extra Features** | **Justifications** |
| 1. Ridge Density Plot | Highlights Distribution of Saving Status Across Employment Type Based on Credit Classes. |
| 1. Multinomial Logistic Regression | Relates models between saving status, employment, and credit class, showing interaction effects. |
| 1. Chi-Squared Test of Independence | It assesses the association between employment and saving statuses for insights into dependency patterns. |
| 1. Proportional Heatmap Visualization | Visualizes deviations between observed and expected frequencies in saving-employment distributions. |
| 1. Mosaic Plot | Summarizes saving status proportions across employment categories for quick comparisons. |
| 1. Circular Heatmap | Depicts employment and saving status relationships in a compact and visually appealing format. |
| 1. Random Forest Model | Predicts credit class using employment and savings data, highlighting key influencing features. |
| 1. Feature Importance Plot | Identifies and ranks variables most influential in predicting credit class. |
| 1. Diverging Stacked Bar Plot | Illustrates the proportion difference of employment between good and bad credit classes. |
| 1. Decision Tree Analysis (CART) | Explores decision paths involving employment, saving status, and credit class relationships. |
| 1. Interactive Bubble Plot | Visualizes saving and employment trends interactively by credit class and population size. |

### **3.2.7 Analysis Conclusion**

The findings indicate a robust relationship between employment status, saving status, and credit class. Lower savings and fewer years of employment raise the risk of "bad" credit. Savings status is the most important determinant of creditworthiness, with employment ranking second. These findings are supported through a variety of statistical tests and predictive models that demonstrate the major impact of stable employment on financial behaviour. The insights identify actionable thresholds, providing clear knowledge about how work and savings influence credit risk.

**3.2.8 Analysis Recommendation**

Financial institutions and politicians could lower credit risk by promoting savings and job security. Financial literacy, savings incentives, job training, and career development could be ways of enhancing financial stability. Predictive models could be applied to stratify high-risk groups by credit providers; these can subsequently design appropriate financial products including low-interest loans. Heatmaps can be developed to show where targeted interventions need to be carried out. Research and continuous analysis of data are crucial to fine-tune strategies that evolve with changing economic conditions.

## **3.3 To explore the impact of credit amount and installment commitment to credit class. (Hong Xiang Lin)**

### **3.3.1 Analysis 3-1: Assessing the Relationship Between Credit Amount, Installment Commitment and Credit Class.**

|  |  |
| --- | --- |
| Type | Descriptive Analysis |
| Independent Variable | Credit Amount (Continuous Data), Installment Commitment (Categorical Data) |
| Dependent Variable | Class (Categorical Data) |
| Analysis Technique Used | * Multivariate Descriptive Statistics |
| Visualization | * Violin Plot with Boxplot Overlay * Bar Plot with Error Bars |

**Multivariate Descriptive Statistics (Extra Feature 1)**

A computer screen with text

Description automatically generatedCode:

*Figure 3.3.1.1 Code of Multivariate Descriptive Statistics*

A screen shot of a computer

Description automatically generatedOutput:

*Figure 3.3.1.2 Summary Table of Multivariate Descriptive Statistics*

This analysis used multivariate descriptive statistics to discover data patterns and initial exploration. In the **“bad” credit class**, the **highest mean of credit amount (6360)** is observed at the **installment commitment level 1**, with a **large standard deviation (4552)**, indicating **wide variability**. Conversely, the **“good” credit class's mean** **credit amount steadily decreased** from **3574 at level 1** to **2308 at level 4**, with lower variability. It should be noted that **larger and more variable credit amounts** are associated with a **“bad” credit class**, particularly at **lower installment commitment levels**, hinting at greater unpredictability for riskier borrowers.

**Violin Plot with Boxplot Overlay (Extra Feature 2)**

A computer code on a black background

Description automatically generatedCode:

*Figure 3.3.1.3 Code of Violin Plot with Boxplot Overlay*

Output:

*Figure 3.3.1.4 Violin Plot with Boxplot OverlayA graph of credit score

Description automatically generated with medium confidence*

The **violin plot with boxplot overlay** highlights data density and central tendencies of credit amount across installment commitment levels. The **“bad” credit class** shows **decreasing median and spread** as the installment commitment level increases, while the “**good” credit class has a generally lower, less variable credit amount** across all levels. Higher commitment levels resulting in lower credit amounts for both classes are most likely due to affordability constraints or lenders' risk aversion. Overall, “**bad” credit classes are associated with higher and more variable credit amounts**, especially at lower installment commitment levels, indicating high credit amounts may increase the risk of poor credit outcomes.

**Bar Plot with Error Bars (Extra Feature 3)**

A computer screen shot of a program code

Description automatically generatedCode:

*Figure 3.3.1.5 Code of Bar Plot with Error Bars*

A graph of a bar chart

Description automatically generated with medium confidenceOutput:

*Figure 3.3.1.6 Bar Plot with Error Bars*

The **bar plot with error bars** provides a comparative view of mean credit amounts with its variability across installment commitment levels. At **level 1**, the **“bad” credit class has a higher mean credit amount** (6360), which is noticeably **higher than the “good” credit class** (3574), with the **gap narrowing down as installment commitment levels increase**. Besides, the **error bars for “bad” credit class are wider at lower commitment levels**, revealing **greater variability**, which also decreases as levels increase. This suggests that higher installment commitment helps in stabilizing credit behaviour, reducing variability and differences across credit classes.

### **3.3.2 Analysis 3-2: Investigating the Interaction of Credit Amount and Installment Commitment in Influencing Credit Class. (Hong Xiang Lin)**

|  |  |
| --- | --- |
| Type | Diagnostic Analysis |
| Independent Variable | Credit Amount (Continuous Data), Installment Commitment (Categorical Data) |
| Dependent Variable | Class (Categorical Data) |
| Analysis Technique Used | * Logistic Regression |
| Visualization | * Partial Dependence Plot * Boxplot |

**Logistic Regression (Extra Feature 4)**

Code:

A computer screen shot of a program

Description automatically generated

*Figure 3.3.2.1 Code of Logistic Regression*

Output:

A computer screen with numbers and numbers

Description automatically generated

*Figure 3.3.2.2 Result of Logistic Regression*

A screen shot of a computer

Description automatically generated

*Figure 3.3.2.3 Odds Ratio Table*

The **logistic regression model** evaluates the influence of credit amount and installment on credit class. From the output, the **intercept** (1.850, p-value of 2 × 10^(-16)) points out that **the "good" credit class is more likely** than the **"bad" credit class** at baseline(credit amount = 0, installment commitment = Level 1). Additionally, the **negative coefficient for** **credit amount** (-0.0002422,p-value of 2 × 10^(-16)) implies **higher credit amounts lower the likelihood of a "good" credit class**, while **Installment Commitment Levels 2, 3 and 4** also **negatively affect the probability compared to Level 1**. Model metrics (AIC = 8052.3, residual deviance = 8036.6) indicate a relatively good fit model.

From the **odds ratio table**, the **intercept** (6.36) reveals that at baseline levels, **the odds of a “good” credit classification are 6.36 times higher than “bad” credit class**. The **odds ratio of credit amount** (1), aligns with its negative coefficient, signifying a **decrease in odds of “good” classification** as credit amount increases. Similarly, higher commitment levels significantly lower the odds of a “good” classification. This analysis highlights the nuanced relationship between credit amount, commitment levels and credit classes.

**Partial Dependence Plot (Extra Feature 5)**

Code:

A screen shot of a computer program

Description automatically generated

*Figure 3.3.2.4 Code of Partial Dependence Plot*

Output:

*Figure 3.3.2.5 Partial Dependence PlotA graph showing different colored lines

Description automatically generated*

The **Partial Dependence Plot** visualizes the interaction between credit amount and installment commitment on the probability of a **“good” credit classification**, highlighting nonlinear relationships. The plot reveals that the probability of a **“good” credit class decreases** sharply with increasing credit amounts at **Installment Commitment Level 1**, while Levels 2, 3, and 4 show consistently lower probabilities, regardless of credit amount. This emphasizes the significant interaction between installment commitment and credit amount, where the influence of credit amount diminishes as commitment levels rise. Overall, the highest predicted probability of a **“good” credit class** occurs with **lower credit amounts** and **lower installment commitments**.

**Boxplot**

Code:

A computer code on a black background

Description automatically generated

*Figure 3.3.2.6 Code of Boxplot*

Output:

*Figure 3.3.2.7 BoxplotA graph of credit amount by installment

Description automatically generated*

A **faceted boxplot** provides a clear comparison of credit amount distributions across installment commitment levels and credit classes. For the **“bad” credit class**, the median credit amount decreases from ~4300 at Level 1 to ~2300 at Level 4, with variability (Interquartile Range, IQR) reducing from ~2500–11000 to ~1250–3500. This highlights greater dispersion at lower commitment levels. The **“good” credit class** shows consistently lower medians, dropping from ~3500 at Level 1 to ~1750 at Level 4, with smaller variability (IQR ~2000–4000 at Level 1 to ~1250–2500 at Level 4). Both classes exhibit outliers across all levels. In summary, **higher installment commitments** correlate with **lower credit amounts** and **reduced variability**, indicating more stable credit behavior, particularly for the **“good” credit class**.

### **3.3.3 Analysis 3-3: What is the predictive power of credit amount and installment commitment combined compared to using credit amount alone? (Hong Xiang Lin)**

|  |  |
| --- | --- |
| Type | Predictive Analysis |
| Independent Variable | Credit Amount (Continuous Data), Installment Commitment (Categorical Data) |
| Dependent Variable | Class (Categorical Data) |
| Analysis Technique Used | * **Logistic Regression** for prediction**, Confusion Matrix** for evaluation and **AUC-ROC Analysis** |
| Visualization | * ROC Curves * Overlayed Boxplot |

**Logistic Regression (Extra Feature 4)**

Code:

A computer screen shot of a program code

Description automatically generated

*Figure 3.3.3.1 Code of Logistic Regression*

**Confusion Matrix (Extra Feature 6)**

Code:

A screen shot of a computer code

Description automatically generated

*Figure 3.3.3.2 Code of Confusion Matrix*

Output:

A screenshot of a computer program

Description automatically generated

*Figure 3.3.3.3 Result of Confusion Matrix 1*

A screenshot of a computer

Description automatically generated

*Figure 3.3.3.4 Result of Confusion Matrix 2*

**Area Under the Curve - Receiver Operating Characteristic Analysis, AUC-ROC Analysis (Extra Feature 7)**

Code:

A screen shot of a computer program

Description automatically generated

*Figure 3.3.3.5 Code of ROC Curve Analysis*

Output:

A screenshot of a computer

Description automatically generated

*Figure 3.3.3.6 Result of ROC Curve Analysis*

A screen shot of a computer program

Description automatically generated

*Figure 3.3.3.7 Summary of Model Performance*

In this analysis, **Logistic Regression** predicts credit class using credit amount and installment commitment, evaluated through a **confusion matrix** (classification accuracy) and **AUC-ROC** (discriminative ability).

**Model 1 (Credit Amount only)**: Accuracy = 55.83%, Sensitivity = 40.67%, Specificity = 71.00%, AUC = 0.555. The model **moderately predicts credit class** but **struggles to identify "good" credit cases due to low sensitivity**.

**Model 2 (Credit Amount + Installment Commitment)**: Accuracy = 53.44%, Sensitivity = 61.56%, Specificity = 45.33%, AUC = 0.583. This confirms that adding installment commitment can **improve sensitivity** and help with **"good" credit identification**, but it **reduces specificity**.

Overall, **Model 2's slight AUC improvement suggests enhanced predictive power**, though the sensitivity-specificity trade-off highlights a need for further refinement.

**ROC Curves (Extra Feature 8)**

Code:

A computer screen shot of text

Description automatically generated

*Figure 3.3.3.8 Code of ROC Curves*

A graph showing a graph

Description automatically generated with medium confidenceOutput:

*Figure 3.3.3.9 ROC Curves*

The **ROC curve** effectively visualizes the trade-off between sensitivity and specificity, enabling model comparison. **Model 1 (Credit Amount only, blue line) has an AUC of 0.555**, showing moderate discriminative ability. **Model 2 (Credit Amount + Installment Commitment, red line) slightly improves with an AUC of 0.583**, reflecting marginally better distinction between “good” and “bad” credit classes. While **Model 2's curve lies slightly above Model 1**, particularly in the middle range, both remain near the diagonal, indicating **limited predictive power**. The small AUC gain suggests that **adding installment commitment enhances performance slightly**, but further predictors or tuning are needed for substantial improvement.

**Overlayed Boxplot (Extra Feature 9)**

Code:

A computer screen shot of a code

Description automatically generated

*Figure 3.3.3.10 Code of Overlayed Boxplot*

Output:

*Figure 3.3.3.11 Overlayed BoxplotA graph of colored squares

Description automatically generated with medium confidence*

The overlayed boxplot combines data distribution with model predictions, showcasing the alignment of predicted probabilities with actual values. The color-coded points (ranging from blue to red) reflect the model’s predicted probabilities for "good" credit. For installment commitment Level 1 and 2, they are associated with **higher predicted probabilities for "good" credit** (indicated by red points) and **higher median credit amounts**, particularly for the "bad" credit class. While installment commitment Level 3 and 4 showed **lower predicted probabilities for "good" credit** (indicated by blue points) and **lower median credit amounts**, with less variability in both credit classes. Overall, the plot demonstrates that **higher installment commitments** and **lower credit amounts** correlate with **higher probabilities for "good" credit outcomes**, providing actionable insights for credit risk management.

### **3.3.4 Analysis 3-4: What decision rules can be applied to optimize credit class outcomes based on credit amount and installment commitment? – Hong Xiang Lin**

|  |  |
| --- | --- |
| Type | Prescriptive Analysis |
| Independent Variable | Credit Amount (Continuous Data), Installment Commitment (Categorical Data) |
| Dependent Variable | Class (Categorical Data) |
| Analysis Technique Used | * Decision Tree |
| Visualization | * Decision Tree Visualization * Bar Plot |

**Decision Tree Analysis (Extra Feature 10)**

Code:

A screen shot of a computer program

Description automatically generated

*Figure 3.3.4.1 Code of Decision Tree Analysis*

**Decision Tree Visualization**

Code:



*Figure 3.3.4.2 Code of Decision Tree Visualization*

Output:

*Figure 3.3.4.3 Decision Tree VisualizationA diagram of a credit class

Description automatically generated*

The decision tree is an ideal visualization tool for prescriptive analysis because it provides interpretable decision rules for optimizing credit outcomes based on credit amount and installment commitment.

At **Node 1(28%)**, borrowers with credit\_amount ≥ 3808 are mainly of **"bad" credit**, reflecting higher financial risk. At **Node 2 (22%)**, the cases where installment\_commitment ≥ 3 and 430 ≤ credit\_amount < 1382 are dominated by **"bad" credits**, indicating that higher commitment alone does not reduce risks. Then, at **Node 3 (1%)** with the smallest proportion, installment\_commitment ≥ 3 and credit\_amount < 430, the borrowers’ credit is **mostly "good"**, showing very low credit amounts with high commitments creating a safer profile. Next, the largest proportion of nodes, **Node 4 (30%)**, with installment\_commitment ≥ 3 and credit\_amount ≥ 1382, is associated with **"good" credit**, highlighting the stabilizing role of moderate credit and high commitments. Finally, borrowers with credit\_amount < 3808 and installment\_commitment < 3 at **Node 5 (19%)** are **mostly "good" in credit**, indicating that the lower the credit amounts and installment commitment levels, the lower the risk.

In short, the **higher the credit amounts, the greater the risk**, while **combining lower credit amounts with higher commitments improves the likelihood of a "good" credit class**. This balance is key to effective credit management.

**Bar Plot (Variable Importance Plot)**

Code:

A screen shot of a computer code

Description automatically generated

*Figure 3.3.4.4 Code of Variable Importance Plot*

A screen shot of a computer program

Description automatically generatedOutput:

*Figure 3.3.4.5 Result of Variable Importance*

*Figure 3.3.4.6 Variable Importance PlotA graph of a graph with text

Description automatically generated*

The **variable importance plot** quantifies the relative influence of each predictor, allowing us to prioritize variables in decision-making based on the data. The score of **credit amount** is **180.32**, which indicates that it is the **more influential variable in determining credit class outcomes**. This aligns with its dominance as the root-splitting criterion in the decision tree. Moreover, **installment commitment** with a score of **94.82** indicates that it also **plays an important role but is secondary to ‘credit\_amount’**, emphasizing its supportive yet impactful contribution to credit class predictions.

### **3.3.5 Extra Features**

1. Multivariate Descriptive Statistics
2. Violin Plot with Boxplot Overlay
3. Bar Plot with Error Bars
4. Logistic Regression
5. Partial Dependence Plot
6. Confusion Matrix
7. AUC-ROC Analysis
8. ROC Curves
9. Overlayed Boxplot
10. Decision Tree

### **3.3.6 Conclusion**

The analyses performed has demonstrated that credit amount and installment commitment impact the credit class significantly. Higher credit amounts were strongly associated with increased risk, just as seen in the “bad” credit classification, whereas higher installment commitment levels stabilized credit behaviour and improved the likelihood of “good” credit outcomes. The AUC-ROC Analysis can assure us that combining these predictors enhances the model performance, albeit marginally. On the other hand, the decision tree has provided further insights in revealing interpretable rules that highlight critical thresholds for credit amounts and installment commitments. To sum up, balancing credit and installment levels is crucial for optimizing credit class predictions and minimizing financial risks.

### **3.3.7 Recommendation**

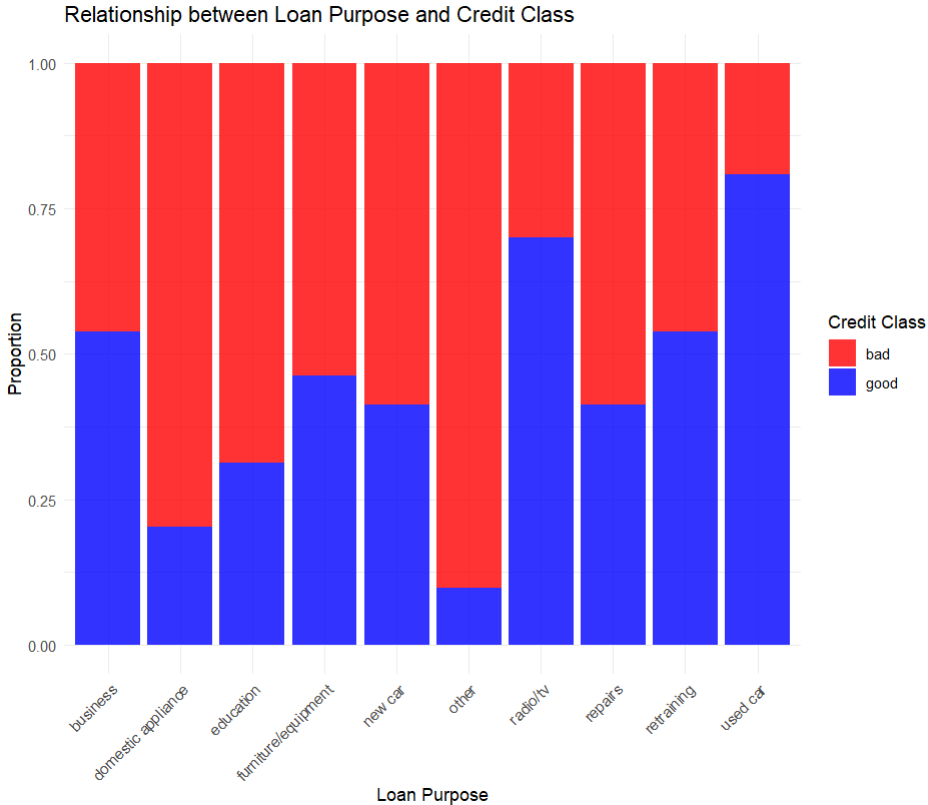
First, policy adjustments should prioritize stricter monitoring of cases with credit\_amount ≥ 3808, as these cases are strongly associated with a high risk of “bad” credit outcomes. In addition, it is recommended to focus on adopting a strategy that combines moderate credit amounts with higher installment commitments to reduce risks and stabilize credit behavior. Additionally, integrating decision tree rules and logistic regression insights into automated credit assessment systems can ensure consistent and accurate decision-making. Therefore, regular revision of credit policies and decision rules based on updated data is encouraged to improve forecast accuracy and minimize risks.

## **3.4 To investigate the relationship between loan purposes and credit class (Eng Hui Ern)**

The data analysis aims to explore key patterns and relationships within the dataset to address the research objectives. Several analytical techniques, including descriptive statistics and visualization methods, have been utilized. Descriptive statistics provide insights into data distribution, while visualization tools such as bar charts, scatter plots, and line graphs are employed to communicate findings effectively. Data manipulation techniques, implemented using tools like dplyr and similar packages, assist in cleaning and aggregating the data for meaningful analysis and visualization.

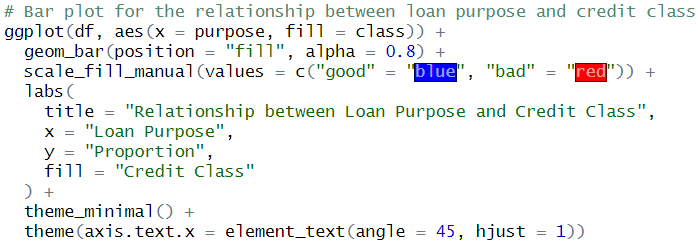
### **3.4.1 Analysis 1: Is there any relationship between loan purposes and credit class?**

|  |  |
| --- | --- |
| Type of Analysis | Diagnostic analysis |
| Independent Variable(s) | Loan purpose |
| Dependent Variable | Credit class |
| Techniques Used | Bar chart, chi-square test |

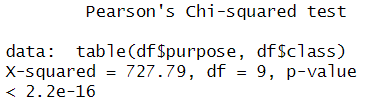


*Figure 3.4.1.1: Bar plot for the relationship between loan purpose and credit class.*

The bar plot is designed to provide an overview of the relationship between loan purposes and credit class. Each bar from the X-Axis (Loan Purposes) represents a specific loan purpose, such as "education", "car purchase", or "home improvement". Y-Axis (Proportion) displays the proportion of loan outcomes (good or bad) for each purpose. The bar heights are divided into two colours which are blue and red. Blue represents the proportion of loans categorized as "good" and red represents the proportion of loans categorized as "bad". If a specific loan purpose has a higher proportion of "bad" credit classes (more red), it indicates that this loan purpose might be riskier. Conversely, loan purposes with more "good" credit classes (more blue) are associated with lower risk. The chart clearly shows how the proportion of "good" and "bad" credit classifications varies across different loan purposes. For example, loans for domestic appliances and repairs tend to have a higher proportion of "bad" classifications (red). Loan purposes like business and furniture/equipment show a more balanced distribution between good (blue) and bad (red) classes. Some purposes, like used car loans, have a greater proportion of "good" credit classes.



*Figure 3.4.1.2: Bar plot script.*



*Figure 3.4.1.3: Results of the chi-square test.*

The test yielded a chi-squared statistic of 727.79 with 9 degrees of freedom and a p-value < 2.2e-16. Since the p-value is extremely small (less than 0.05), we reject the null hypothesis, indicating a significant association between loan purposes and credit class. In other words, the credit class is not distributed equally across loan purposes.

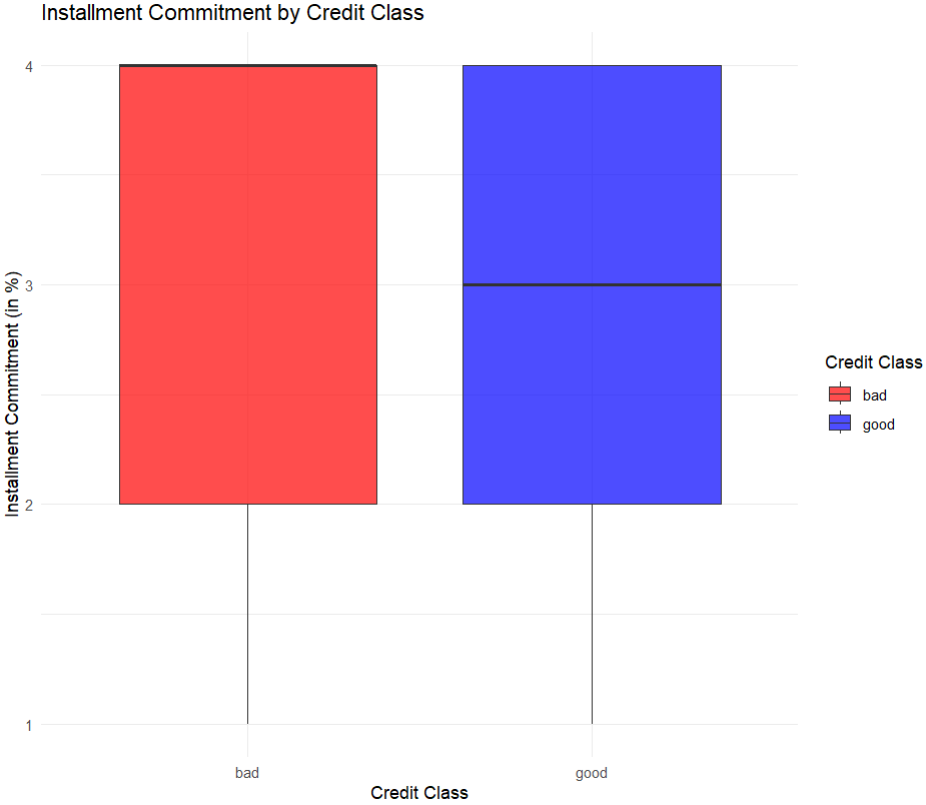


*Figure 3.4.1.4: Chi-square test script.*

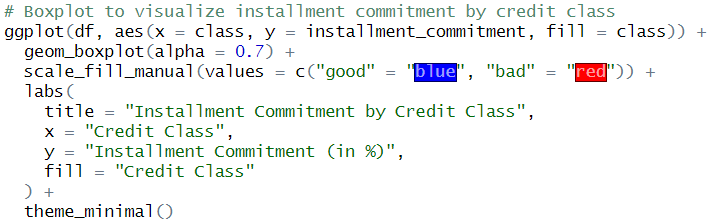
Loan purposes significantly influence the credit class distribution. Certain purposes, such as domestic appliances and repairs, are associated with a higher likelihood of "bad" classifications. Insights from this analysis can help financial institutions target high-risk loan purposes for improved credit risk assessment and tailored lending policies.

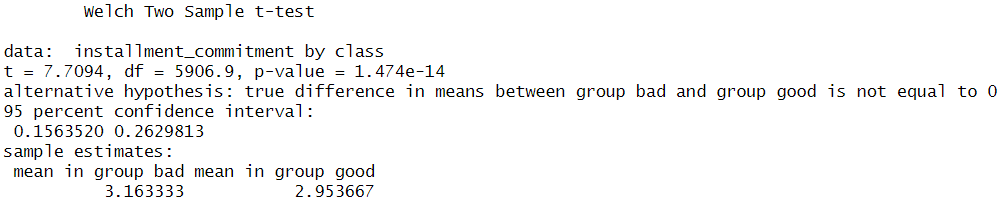
### **3.4.2 Analysis 2: Is installment commitment a strong predictor for credit class?**

|  |  |
| --- | --- |
| Type of Analysis | Predictive analysis |
| Independent Variable(s) | Installment commitment |
| Dependent Variable | Credit class |
| Techniques Used | Box plot, t-test |

*Figure 3.4.2.1: Boxplot to visualize installment commitment by credit class.*

The boxplot illustrates the distribution of installment commitment across different credit classes. The X-axis represents the credit class (good or bad), and the Y-axis shows the installment commitment percentage. Blue boxes represent the distribution for the "good" credit class, and red boxes represent the distribution for the "bad" credit class. The median installment commitment is visibly higher for the "bad" credit class compared to the "good" credit class. The interquartile range (IQR) for the "bad" class is narrower, suggesting less variability in installment commitment for loans categorized as "bad". This visualization suggests that installment commitment differs between credit classes, with higher commitments potentially contributing to a "bad" classification.

*Figure 3.4.2.2: Boxplot script.*

*Figure 3.4.2.3: T-test to determine the predictive strength.*

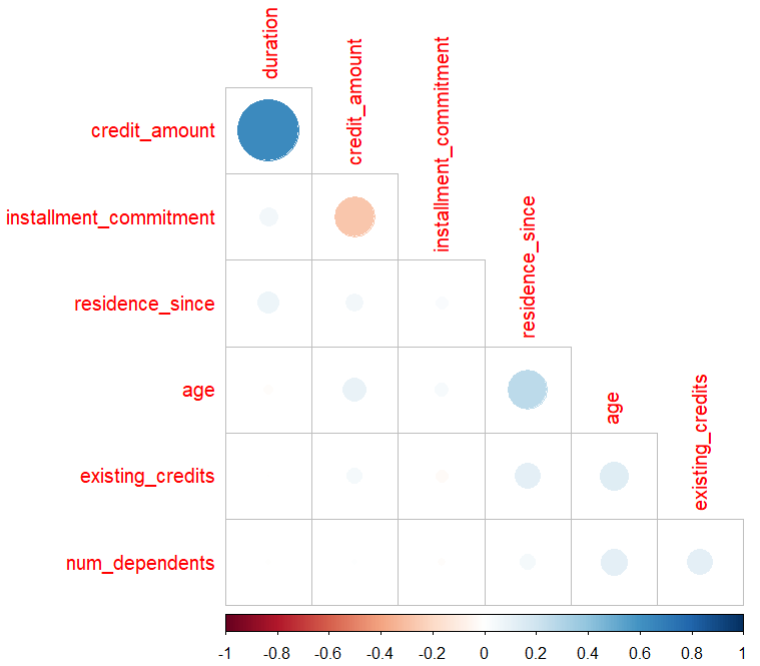
The t-test compares the mean installment commitment percentages between the "good" and "bad" credit classes. The test yielded a t-statistic of 7.71 and a p-value of < 0.001. Since the p-value is extremely small (less than 0.05), we reject the null hypothesis, indicating that the mean installment commitment significantly differs between the two credit classes. The results suggest that loans classified as "bad" have a higher mean installment commitment (3.16) compared to loans classified as "good" (2.95). This significant difference highlights the importance of installment commitment as a predictive factor in credit risk assessment.

*Figure 3.4.2.4: T-test script.*

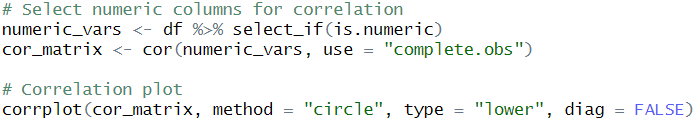
Installment commitment is a strong predictor for credit class, as evidenced by the significant differences in its distribution and mean values across classes. This insight highlights the importance of monitoring installment commitments as part of credit risk assessment strategies.

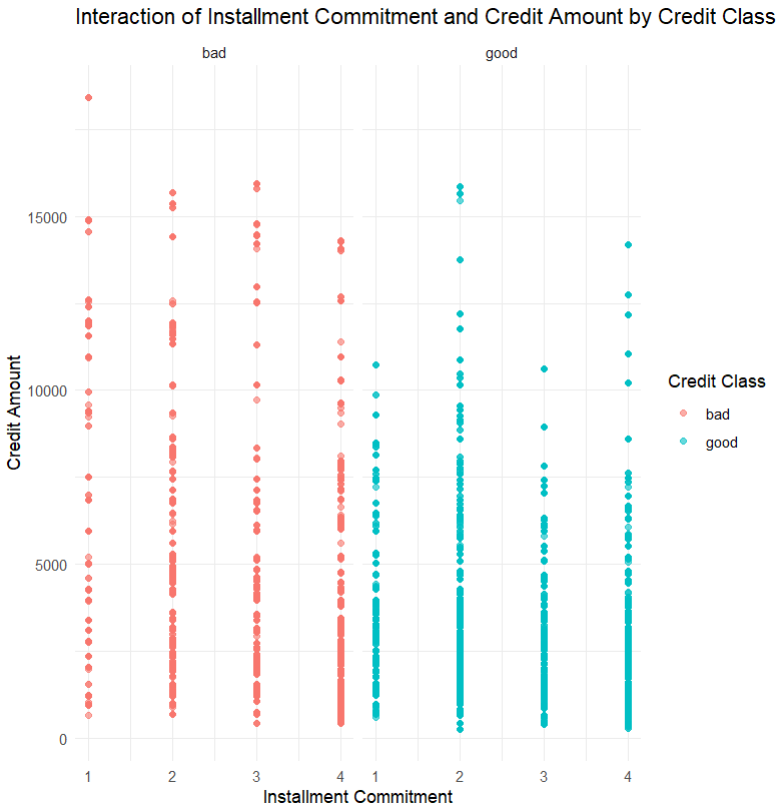
### **3.4.3 Analysis 3: What Are the External Factors That Interact with Installment Commitment to Influence Credit Class?**

|  |  |
| --- | --- |
| Type of Analysis | Descriptive analysis, dispersion analysis |
| Independent Variable(s) | Installment commitment, credit amount, age |
| Dependent Variable | Credit class |
| Techniques Used | Scatter plot |

*Figure 3.4.3.1: Correlation plot.*

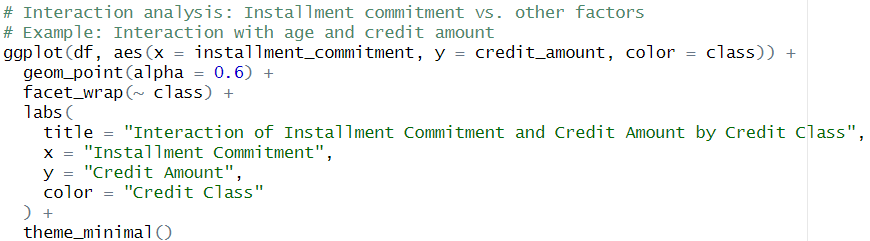
The correlation matrix visualizes relationships between numeric variables in the dataset, with installment commitment as a key focus. Darker or more saturated colors represent stronger correlations. Installment commitment shows moderate positive correlation with credit\_amount and weaker relationships with age. No extreme correlations (>0.9) suggest that the variables are not collinear, allowing for meaningful individual contributions to predictive models.

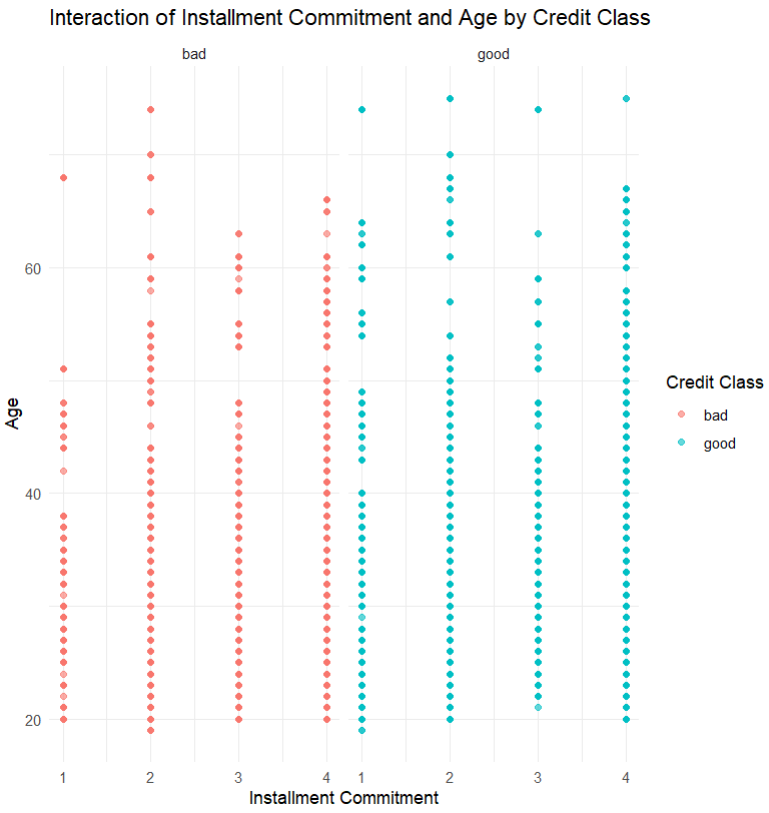
*Figure 3.4.3.2: Correlation plot script.*



*Figure 3.4.3.3: Scatter plot of the interaction of installment commitment and credit amount by credit class.*

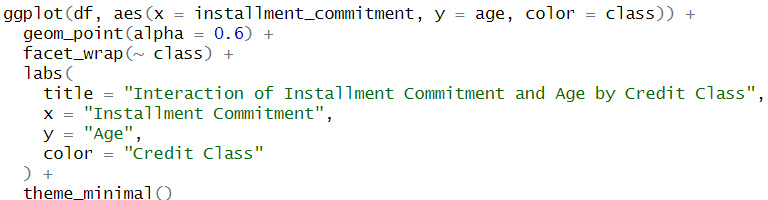
This scatter plot illustrates how installment commitment interacts with credit amount across credit classes. The X-axis represents installment commitment, and the Y-axis represents credit amount. Each point is colored by credit class (good in blue and bad in red). For the "bad" class, data points cluster at higher installment commitments and mid-to-low credit amounts. For the "good" class, points are more evenly distributed across credit amounts, with fewer instances of high installment commitment.

*Figure 3.4.3.4: Scatter plot of the interaction of installment commitment and credit amount by credit class script.*



*Figure 3.4.3.5: Scatter plot of the interaction of installment commitment and age by credit class.*

This scatter plot explores the interaction between installment commitment and age. The X-axis shows installment commitment, and the Y-axis shows age. "Bad" credit classifications (red) are more concentrated in the younger age range and higher installment commitments. "Good" credit classifications (blue) are spread more broadly across ages and lower installment commitments.

*Figure 3.4.3.6: Scatter plot of the interaction of installment commitment and age by credit class script.*

Correlation analysis highlights how installment commitment is moderately related to credit amount but less so with other variables. Interaction analysis reveals that the relationship between installment commitment and other variables, such as credit amount and age, differs significantly across credit classes. These findings suggest that external factors interact with installment commitment, further influencing credit classification.

### **3.4.4 Extra Features**

|  |  |
| --- | --- |
| Extra Features | Justification |
| Correlation matrix with visualization | The correlation matrix (calculated using cor()) along with the corrplot() visualization method provides an overview of the relationships between all numeric variables. This is a more comprehensive approach compared to simply analyzing pairwise correlations. |
| Faceted interaction plots | The faceted scatter plots (using facet\_wrap(~ class)) illustrate the interactions between installment commitment and other factors (e.g., credit amount, age) separately for each credit class. This approach allows for a side-by-side comparison of patterns within good and bad credit classes. |
| Custom color-coding for class labels | The use of custom color coding for credit classes (scale\_fill\_manual() and aes(color = class)) enhances visual clarity, making it easier to distinguish between good and bad credit classifications in both boxplots and scatter plots. |

# **4.0 Conclusion**

The analysis conducted in this study provides a detailed understanding of the various factors influencing credit classification. Key variables, such as loan duration, employment status, savings, credit amount, installment commitments, and loan purposes, were thoroughly examined for their impact on creditworthiness. The findings highlight critical patterns and relationships that can inform the practices of financial institutions and borrowers alike. By leveraging these insights, financial institutions can enhance their credit risk assessment models, while borrowers can make more informed financial decisions to improve their credit outcomes.

## **4.1 Findings**

The findings reveal that shorter loan durations are strongly associated with a higher likelihood of achieving good credit classification. Conversely, longer loan durations tend to correlate with an increased risk of bad credit classification. Employment and savings status also emerged as significant predictors; borrowers with stable employment and adequate savings are more likely to be classified as good credit, whereas those with unstable employment or low savings are at a higher risk of bad credit outcomes. Credit amount and installment commitments further demonstrated a notable influence, with higher credit amounts and lower installment commitments posing greater risks. Balanced commitments and moderate credit amounts were found to stabilize repayment behavior and improve credit classifications. Lastly, the study identified variations in risk based on loan purposes, where loans for domestic appliances or repairs were more likely to result in bad credit outcomes, while business loans and furniture equipment loans exhibited a more balanced distribution of credit classifications.

## **4.2 Recommendations**

Based on these findings, several recommendations can be made for both financial institutions and borrowers. For financial institutions, it is essential to enhance risk assessment practices by closely monitoring customers with long loan durations, high credit amounts, or low installment commitments. Automated credit risk assessment systems that incorporate insights from logistic regression and decision tree models should be adopted to improve predictive accuracy. Additionally, financial institutions could provide tailored financial advice to borrowers, encouraging them to select shorter loan durations and balanced repayment terms to mitigate risks.

Borrowers are encouraged to focus on improving their creditworthiness by maintaining stable employment and building adequate savings before applying for loans. Choosing loan purposes with historically better credit outcomes, such as business loans, and ensuring repayment terms that align with their financial capabilities can significantly improve their chances of achieving a good credit classification.

## **4.3 Limitations and Future Directions**

This study faced several limitations that should be addressed in future research. One of the primary constraints was the presence of missing values in the dataset, which were imputed using various techniques. Although effective, these imputations may have introduced biases, potentially affecting the accuracy of the models. The predictive models, including logistic regression and decision trees, provided valuable insights but were limited by moderate accuracy and trade-offs between specificity and sensitivity, as highlighted in the confusion matrix results. Furthermore, the dataset was specific to a particular context, which may limit the generalizability of the findings to different demographic or economic scenarios.

Future research should aim to expand the dataset to include a broader range of borrower demographics and economic contexts to improve the robustness of the findings. Advanced predictive models, such as ensemble learning techniques, could be explored to enhance accuracy and address limitations in existing methods. Incorporating real-time data into credit risk assessment processes could enable dynamic evaluation of creditworthiness, providing more timely and effective insights. Additionally, studying supplementary variables, such as credit history and financial education levels, could offer a more comprehensive understanding of credit risk factors, ultimately supporting better decision-making for both financial institutions and borrowers.

Total Word Count: 8108 words

# **5.0 Workload Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Components / Members | Eng Hui Ern | Eugene Tan Ting Siang | Hong Xiang Lin | Sweetha Pramasivam |
|  |  |  |  |
| Introduction | 15% | 25% | 15% | 45% |
| Data Preparation | 5% | 30% | 60% | 5% |
| Conclusion | 70% | 10% | 10% | 10% |
| Signature |  | Eugene | A black background with a black square  Description automatically generated with medium confidence |  |

# **6****.0 References**

Morgan, L. (2020, August). *MissForest - The best imputation algorithm*. Kaggle.com; Kaggle. https://www.kaggle.com/code/lmorgan95/missforest-the-best-imputation-algorithm

Bioinformatics, L. (3 June, 2022). *Dealing with Missing Data in R*. Retrieved from Youtube: https://www.youtube.com/watch?v=\_BFMS1IefzE&t=1218s

Choure, R. (8 May, 2023). *How to impute missing data using mice package in R programming*. Retrieved from Youtube: https://www.youtube.com/watch?v=C9qKSUTReds

Choure, R. (10 May, 2023). *missForest: Imputation of missing data using Random Forest approach in R programming*. Retrieved from Youtube: https://www.youtube.com/watch?v=yOTi-mbVZXE

Hadley Wickham, W. C. (n.d.). *Function reference*. Retrieved from ggplot2: https://ggplot2.tidyverse.org/reference/index.html

Technology, I. (7 February, 2022). *What is Random Forest?* Retrieved from Youtube: https://www.youtube.com/watch?v=gkXX4h3qYm4

*Chapter 6 The circos.heatmap() function*. (n.d.). Retrieved from jokergoo.github.io: https://jokergoo.github.io/circlize\_book/book/circos-heatmap.html

*Decision Trees in Machine Learning Using R*. (1 june, 2023). Retrieved from datacamp: https://www.datacamp.com/tutorial/decision-trees-R

*Interactive bubble chart*. (n.d.). Retrieved from r-graph-gallery: https://r-graph-gallery.com/bubble\_chart\_interactive\_ggplotly.html

*Intro to R Programming*. (14 November 2024). Retrieved from libguides.chapman.ed: https://libguides.chapman.edu/R/graphs

*Multinomial Logistic Regression | R Data Analysis Examples*. (2024). Retrieved from stats.oarc.ucla.edu: https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/

*Random Forest Feature Importance Chart using Python*. (n.d.). Retrieved from stackoverflow: https://stackoverflow.com/questions/44101458/random-forest-feature-importance-chart-using-python

Chao-Ying Joanne Peng, K. L. (2010). “An Introduction to Logistic Regression Analysis and Reporting.”. *The Journal of Educational Research, 96*, 3–14. doi:https://doi.org/10.1080/00220670209598786

Pushpa Singh, A. S. (2021). *Confusion Matrix - Diagnosing of Disease Using Machine Learning.* Retrieved from ScienceDirect: https://www.sciencedirect.com/topics/engineering/confusion-matrix#:~:text=A%20confusion%20matrix%20is%20a,performance%20of%20a%20classification%20algorithm

*Statistics Resources: Point Biserial.* (2024, 10 31). Retrieved from National University. : https://resources.nu.edu/statsresources/Pointbiserial

*T test | Practical Applications of Statistics in the Social Sciences | University of Southampton.* (n.d.). Retrieved from Southampton.ac.uk.: https://www.southampton.ac.uk/passs/gcse\_scores/bivariate\_analysis/t\_test.page