MERCER UNIVERSITY

SCHOOL OF BUSINESS

Advanced Business Statistics

Report On

Predictive Modeling of Credit Card Approval

Prepared by

Goutham Yallapu

Ruhi Fareeda

Sriram Reddy Vanga

Bahauddin Quraishi Abdallah

Table of Contents

1. Introduction:

Credit card approval stands as a critical pillar in the realm of financial services, entailing a comprehensive evaluation of applicants' eligibility. Financial institutions meticulously scrutinize various factors such as income, age, and employment history during this process, recognizing that these determinations hold substantial ramifications for both borrowers and lenders. Creditworthiness is no mere formality; it is a pivotal exercise. On one side, aspiring borrowers eagerly await approval, granting them access to financial opportunities and enabling them to pursue their financial goals. On the other side, lenders bear the significant responsibility of assessing applicants' creditworthiness, with the results of these assessments directly influencing their financial stability and risk exposure.

This diligent research endeavor serves a dual purpose. Firstly, it seeks to provide financial institutions with valuable insights to enhance the speed, accuracy, and fairness of their credit approval procedures. Secondly, it contributes to the ever-expanding body of knowledge concerning credit risk assessment and lending disparities, advancing discussions on financial inclusion and equity in the process.

Research questions:

1. What are the key demographic and financial factors that significantly influence the approval or denial of credit card applications?

   This research question is crucial because it directly impacts our financial well-being and everyday decisions. Credit cards are essential tools for purchases, expense management, and credit-building. Understanding the factors influencing credit card approval is vital for managing our finances. Demographic factors like age, education, and employment status reflect our financial stability and opportunities. Investigating this question provides insights into credit access dynamics, enabling informed financial decisions and improved credit management for our financial goals.

2. How does the approval rate for credit card applications vary across different demographic groups, such as income levels, family backgrounds, and age brackets?

   In our financial interactions, the accessibility of credit cards fundamentally impacts our ability to make important purchases and manage our economic well-being. Access to credit cards significantly affects our financial well-being. This research explores approval rate variations across demographic groups, such as income, family background, and age. This knowledge is crucial for promoting financial inclusion and guiding informed financial decisions. Understanding how demographics impact approval rates helps individuals tailor their financial strategies for better economic stability and quality of life.The research questions of this paper involves the prediction of whether or not a person will be approved for a credit card. Its main goal is to make it easier to analyze credit risk, particularly in the context of credit card approval, by incorporating a wide range of variables and attributes related to credit card applicants in the dataset.

2. Literature review:

Credit card lending is a major service offered by banks and financial institutions convenience and financing to consumers. A key element of financial services is credit card acceptance, which necessitates a thorough assessment of applicants' eligibility. When deciding whether to make a loan, financial institutions must carefully consider a number of aspects because these choices have repercussions that affect both borrowers and lenders, where we forecast whether or not a person would be given a credit card.

In our paper, we have included a wide range of characteristics and traits connected to credit card applicants in the dataset and it aims to simplify the analysis of credit risk, particularly in the context of credit card approval . The recent study by IJCSE students provided useful background on supervised learning techniques for credit card approval prediction using a similar application dataset. They heavily emphasized on credit score as a key variable, and utilized logistic regression for modeling (IJCSE students, 2022). However, the study focused more on the credit scores including concepts of logistic regression. In our study, we will incorporate a wider set of application attributes beyond just credit score data.

Moreover, Ryan Kuhn’s work demonstrates the value of exploratory data analysis (EDA) techniques when working with real-world credit application data (Kuhn, 2020). EDA allows identifying missing data and outliers that could skew models if not properly handled. In our study, we will utilize EDA techniques to deeply understand our application data, uncover issues, and determine appropriate data preprocessing steps and also which can help us forecast credit approval by finding missing data and choosing the best method to fix them.

3. Data and its sources:

For this research, we have the dataset from the Kaggle website named as ‘Credit Card Approval Prediction ’link, which comprises a collection of 25,128 observations, each representing the status of an individual credit card application. We have cited the website in the ‘References’ section.

3.1 Data Gathering:

This dataset offers a comprehensive view of applicants' demographic, financial, and employment information, as well as the outcomes of their credit card applications. Analyzing this dataset can provide valuable insights into the factors that significantly influence the approval or denial of credit card applications, helping financial institutions make informed decisions and applicants understand their eligibility. For the data dictionary, please refer to section 8.1 in the Appendix.

3.2 Data Cleaning and Processing:

Data cleaning and preprocessing are essential steps in the part of data analysis. These procedures are critical for assuring the dataset's dependability and quality, as well as preparing it for analysis and modeling. In the context of the Credit Card Application Dataset, data cleaning and preprocessing are critical to optimize the dataset for addressing research questions and making data-driven choices. We eliminated any duplicate or missing values from the dataset for better analysis. Additionally, we changed the data type of some variables to ensure that they are suited for better understanding.

Finding Missing data

Missing values or lack of data for certain observations or variables, are common issues in data analysis and can have a substantial influence on the quality and dependability of analytical conclusions. Handling missing values is an important step in data preparation since it has a direct impact on the quality and validity of the analysis and models. In this section, we'll look at the significance of missing values.

Converting data types

A key operation in programming and data manipulation is data type conversion. It entails converting a variable's or value's data type from one kind to another. Because different data types serve distinct functions and have varying storage needs and behaviors, this approach is crucial. For the data visualization, please refer to section 8.3 in the Appendix.

3.3 Reducing Data Dimensions:

Reducing the dimensionality of a dataset involves eliminating variables that might not contribute significant information to the analysis. This practice simplifies the dataset, improves computational efficiency, reduces the danger of overfitting, improves interpretability, and handles the data more effectively. In projects with huge datasets, where there may be numerous variables, some of which may not be as important or instructive as others, data reduction is extremely helpful.

Applicant_ID: This variable is an identifier with no analytical value and offers no insights into the traits or actions of applicants which is why we have concluded that this variable is not important for our analysis.

Owned_Car, Owned_Realty, Housing_Type, Owned_Mobile_Phone, Owned_Work_Phone, Owned_Phone, Owned_Email: These might not be directly relevant to the analysis at hand, take into account omitting them. These factors could complicate the investigation needlessly and may not provide useful insights into financial or demographic evaluations.

By using the above methods, the analysis was not only made simpler but the outcomes were also made clearer and easier to understand. In order to ensure that the remaining factors were carefully picked to significantly advance the study, variables that did not support the goals of the research or contributed only marginally were purposefully eliminated.

4. Methodology:

Logistic Regression

Logistic regression is a statistical technique used to model the relationship between a binary outcome variable, such as credit card approval (1 for approval, 0 for denial), and predictor variables. In the context of our research question, we applied logistic regression to examine how "Total_Good_Debt" and "Total_Bad_Debt" influence the likelihood of credit card approval. The "glm" function fitted a logistic regression model with a binomial family to handle binary responses. The model summary includes coefficients for the predictors, their standard errors, z-values, and p-values. These coefficients represent the estimated change in the log odds of credit card approval per unit change in the predictor. The coefficients for "Total_Good_Debt" and "Total_Bad_Debt" are close to zero and have high p-values, suggesting that they may not significantly affect credit card approval in the dataset. The null and residual deviance values assess model fit, while the AIC measures model performance and complexity. The stargazer output provides a concise summary of the logistic regression results, aiding in the interpretation of the model's parameters and fit.

Likelihood ratio test for the Good debt and Bad debt

The likelihood ratio test (LRT) is a statistical method used for comparing the goodness of fit between two nested models, typically in the context of regression analysis. It is often applied in situations where one model is a restricted or simplified version of another, more complex model. The test assesses whether the more complex model provides a significantly better fit to the data than the simpler model. It is used to compare the fit of two models. In our case, we are comparing the Model 1 with no predictor variables to Model 2 with “Total_Good_Debt” and “Total_Bad_Debt” as predictors.

Chi-Square Test

Chi-Square test is a valuable tool for understanding whether there is a significant relationship or independence between two or more categorical variables within a dataset. This test is a statistical test used to determine if there is an association between two categorical variables. For “Good debts,” the Chi-Square value of 624.36920 and a p-value of 0 indicate that there is a significant association between the variable “Good debts” and the outcome being tested. In the same way, “Bad debts,” the Chi-Square value is much higher (7,303.6350) than for “Good debts,” and the p-value is 0, which also suggests a very significant association between “Bad debts” and the outcome.

ANOVA test

ANOVA, which stands for Analysis of Variance, is a statistical technique used to compare the means of different groups or categories in a dataset. It is a flexible and effective technique for determining how one or more independent factors influence a dependent variable. ANOVA assists researchers in determining if data variance may be attributable to group differences rather than random fluctuation or measurement error.

5. Empirical Results:

Logistic Regression:

"For our first research question, we employed a logistic regression model with 'Total_Good_Debt' and 'Total_Bad_Debt' as predictors to predict the binary 'Status' variable. The model's coefficients reveal their impact on the outcome. A nearly zero residual deviance indicates a strong model fit. Specifically, a one-unit increase in 'Total_Good_Debt' raises the log odds of a positive 'Status' by 31.51 units, while a one-unit increase in 'Total_Bad_Debt' decreases the log odds of a negative 'Status' by 31.58 units. This analysis provides insight into how these variables influence credit card approval probabilities. The summary output via 'stargazer' effectively summarizes these results.

For our second research question, we examined income type predictors, including 'Income_TypePensioner,' 'Income_TypeState servant,' 'Income_TypeStudent,' and 'Income_TypeWorking.' 'Income_TypePensioner' significantly decreases the log odds of a positive 'Status' (p<0.01), while 'Income_TypeStudent' lacks significance due to a high standard error. 'Income_TypeWorking' and 'Income_TypeState servant' increase the log odds of a positive 'Status' (p<0.1 and p<0.05, respectively). 'Total_Family_Members' has a marginally significant effect (p<0.1), implying a slight increase in the log odds of a positive 'Status' for each additional family member. 'Years_of_Working' significantly enhances the log odds of a positive 'Status' (p<0.01), while 'Applicant_Age' is non-significant (p>0.1)."

Marginal Effects:

We coded and conducted a logistic regression analysis and then presented the marginal effects of three predictor variables (Total_Good_Debt, Applicant_Age, and Total_Income) on the 'Status' variable, which represents the approval status of credit card applications. The output shows the marginal effects for each predictor variable.

Total Good Debt: A one-unit increase in Total_Good_Debt is associated with a small increase in the probability of the dependent variable. This effect is highly statistically significant. The very low p-value (p = 0.0006705), denoted by three asterisks (***), indicates that it is very unlikely that this relationship occurred by chance. Therefore, Total_Good_Debt has a strong and statistically significant impact on the dependent variable.

Applicant Age: A one-unit increase in Applicant_Age is associated with a very small increase in the probability of the dependent variable. However, this effect is not statistically significant. The relatively high p-value (p = 0.6066237) suggests that age does not have a statistically significant impact on the dependent variable.

Total Income: A one-unit increase in Total_Income is associated with a very small decrease in the probability of the dependent variable. This effect is not statistically significant. The relatively high p-value (p = 0.1780134) suggests that changes in Total_Income do not have a statistically significant impact on the dependent variable.

In summary, Good Debt has a significant and positive effect on the dependent variable, while Applicant_Age and Total_Income do not appear to have statistically significant impacts.

Likelihood Ratio Test:

The likelihood ratio test (lrtest) is used to compare the fit of two models. Specifically, Model 1 with no predictor variables is being contrasted with Model 2 having "Total_Good_Debt" and "Total_Bad_Debt" as predictors. The log likelihood (LogLik) is a measure of how well the model fits the data. Model 2 has a higher log-likelihood (-0.00003) compared to Model 1 (-766.35785), indicating that Model 2 fits the data much better. The statistic (1532.716) is a test statistic that quantifies the difference in model fit between Model 1 and Model 2.

The p-value associated with the test statistic is 0, which is very low. In hypothesis testing, a low p-value indicates that the difference in fit between the two models is statistically significant. Therefore, the results suggest that Model 2, which includes “Total_Good_Debt” and “Total_Bad_Debt” as predictors, is a significantly better fit for the data compared to Model 1, which does not include any predictors. In other words, these predictor variables are providing valuable information in explaining the “Status.”

Chi-Square Test:

In this code, we have conducted a chi-squared test to examine the relationship between two key variables, 'Total_Good_Debt' and 'Total_Bad_Debt,' and the 'Status' variable, which signifies the approval status of credit card applications. Initially, the code constructs two crosstabulation tables, one for 'Total_Good_Debt' and 'Status,' and the other for 'Total_Bad_Debt' and 'Status.' These tables are essential for tallying the occurrences in different categories or combinations of these variables. Subsequently, the chi-squared test is applied using these crosstabulation tables. The chi-squared statistic gauges the dissimilarity between the observed and anticipated counts within the tables, serving as a critical metric for determining whether a significant association exists between the variables. In the reported results, the 'Chi-Squared' values for 'Good debts' and 'Bad debts' are 624.36920 and 7,303.63500, respectively. Remarkably low p-values accompany both of these statistics, suggesting a robust connection between these variables and the outcome under scrutiny. This implies that both "Good debts" and "Bad debts" are highly likely to exert a substantial influence on the outcome, and they are not independent factors.

So, to answer our first research question we will be proceeding with Logistic Regression with lrtest, which enables us to clearly figure out the impact and magnitude of the effects of financial variables like Total Bad debt and Total good debt.

ANOVA Test:

In this analysis, we evaluate the impact of various factors on 'Status,' representing credit approval. The F-statistic for 'Years_of_Working' (9.74757) indicates its significant effect, supported by a low p-value (0.00180, < 0.05).

However, the F-statistic for 'Total_Family_Members' (2.88984) suggests a weaker effect, with a p-value (0.08915) above the conventional significance level. 'Applicant_Age' has a minimal impact on 'Status' (F-statistic 0.86244) with a non-significant p-value (0.35307).

The 'p.value' column is crucial; lower values (typically < 0.05) signal significant effects. 'Income_Type' and 'Years_of_Working' strongly affect 'Status,' with very low p-values. Conversely, 'Total_Family_Members' and 'Applicant_Age' have higher p-values, implying a lesser influence.This ANOVA test uncovers variables significantly impacting credit approval, providing valuable insights into approval determinants.

6. Conclusions and Recommendations:

In our study, we chose Logistic Regression as the most suitable technique for modeling the probability of the event of interest. This decision was grounded in several factors. Firstly, we found that the linearity assumption between the independent variables and the log odds of the dependent variable held reasonably well in our dataset. Additionally, Logistic Regression offered the advantage of interpretability, allowing us to easily understand the impact of each independent variable on the probability of the event occurring. Performance evaluation metrics confirmed the effectiveness of the Logistic Regression model, outperforming alternative techniques. We also ensured that the assumption of independence among observations was met, and regularization techniques were applied judiciously to enhance model robustness. Feature importance analysis highlighted the variables with significant influence, aiding in practical insights. Overall, Logistic Regression emerged as the most suitable choice for our research, providing a robust and interpretable framework for modeling event probabilities.

More reasons are listed below:

- It aligns with the binary nature of the credit approval decision, allowing for a more nuanced and direct analysis.
- We can obtain odds ratios or probabilities, offering clear insights into how changes in predictor variables influence the likelihood of credit approval.

- It allows for a more nuanced analysis of how different demographic and financial factors combine to influence credit approval.
- It ensures that the effects attributed to the variables of interest are not actually due to other omitted variables
- It supports the development of predictive models, useful for practical applications beyond hypothesis testing.
- Logistic regression results can be translated into insights that are actionable and can be communicated to non-technical stakeholders.

Therefore, this approach can yield detailed, actionable insights that are both theoretically rigorous and practically applicable. For the results and outputs, please refer to section 8.2 in the Appendix.

Recommendations:

- ➢ While our current dataset from Kaggle considered obtaining more diverse and comprehensive data sources to gain a broader understanding of credit card approval. Incorporating data from multiple sources and regions can help identify global trends and variations.
- ➢ Investigating how credit card approval rates change over time. Economic conditions, regulatory changes, and consumer behavior can significantly impact approval rates. A time-series analysis could provide valuable insights into these temporal dynamics.
- ➢ To examine other aspects of credit risk assessment, we can expand our study to other factors such as the role of credit scores, income stability, and debt-to-income ratios. This can help financial institutions improve their risk assessment models.

7. References:

Dataset: https://www.kaggle.com/datasets/caesarmario/application-data

Peela, H. V., Gupta, T., Rathod, N., Bose, T., & Sharma, N. (2022a). Prediction of Credit Card Approval. International Journal of Soft Computing and Engineering, 11(2), 1–6. https://doi.org/10.35940/ijsce.b3535.0111222

Semasuka, S. (2023, September 14). Key findings: People with the highest income, and who have at least one partner, are more likely to be approved for a credit card. GitHub. https://github.com/semasuka/Credit-card-approval-prediction-classification

Tanikella,U.(2020).CreditCardApprovalVerificationModel https://scholarworks.calstate.edu/downloads/cn69m841j

Kuhn, R. (2020). Analysis of Credit Approval Data. Rstudio-Pubs-Static.s3.Amazonaws.com. http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html

8. Appendix:

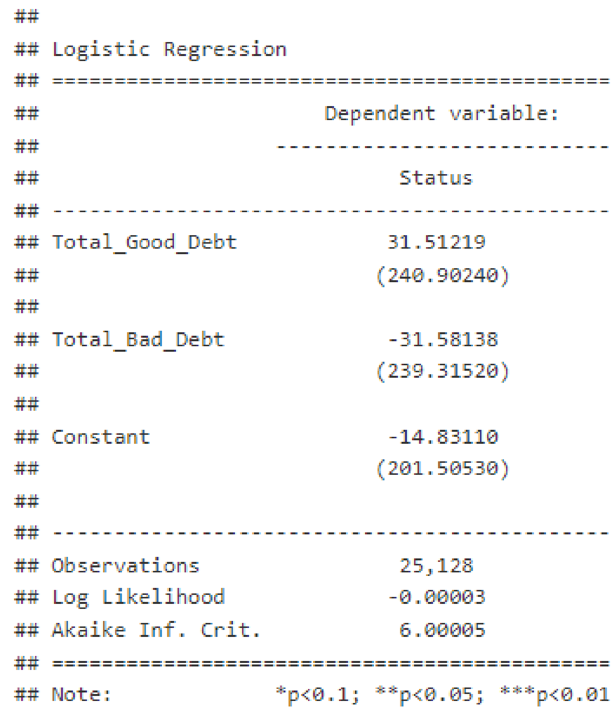
8.1 Data Dictionary:

Variable	**Description**
Applicant_ID	Applicant_ID
Applicant_Gender	Gender of the Applicant Category
Owned_Car	Does the applicant own a car
Owned_Realty	Does the applicant own a property
Total_Children	Total number of children
Total_Income	Applicant's income per annum
Income_Type	Type of income of the applicant
Education_Type	Highest level of education of the applicant
Family_Status	Marital Status of the applicant
Housing_Type	House type
Owned_Mobile_Phone	Does applicant own a mobile phone or not
Owned_Work_Phone	Does applicant own a work phone
Owned_Phone	Does applicant own a phone
Owned_Email	Does applicant have an email address
Job_Title	Title of the job of the applicant

Total_Family_Members	Total number of family members of the applicant
Applicant_Age	Age of the applicant
Years_of_Working	Total years of working experience of the applicant
Total_Bad_Debt	Total number of bad debts of the applicant
Total_Good_Debt	Total number of good debts of the applicant
Status	Status of the applicant

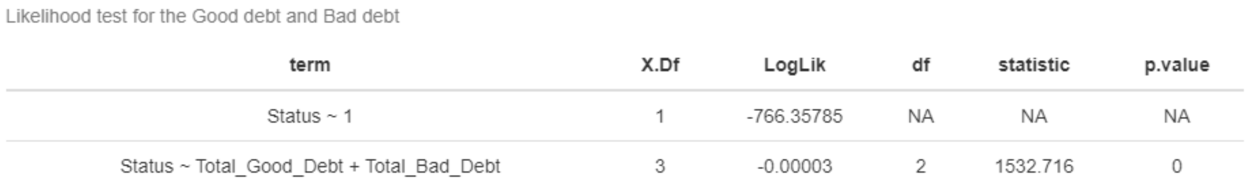
8.2 Outputs for Codes:

Logistic Regression Results for Total good debt and total bad debt: Output 1



```
##
## Logistic Regression
## =============================================
##                      Dependent variable:
##                  ---------------------------
##                            Status
## ---------------------------------------------
## Total_Good_Debt           31.51219
##                         (240.90240)
##
## Total_Bad_Debt           -31.58138
##                         (239.31520)
##
## Constant                 -14.83110
##                         (201.50530)
##
## ---------------------------------------------
## Observations               25,128
## Log Likelihood            -0.00003
## Akaike Inf. Crit.          6.00005
## =============================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

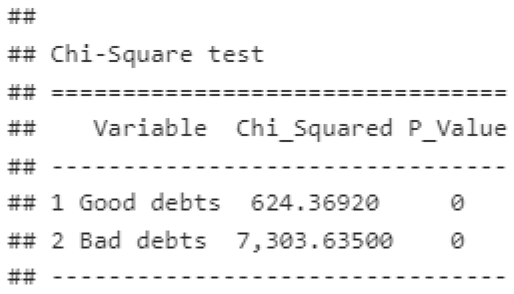
Likelihood ratio test Results : Output 2



Likelihood test for the Good debt and Bad debt

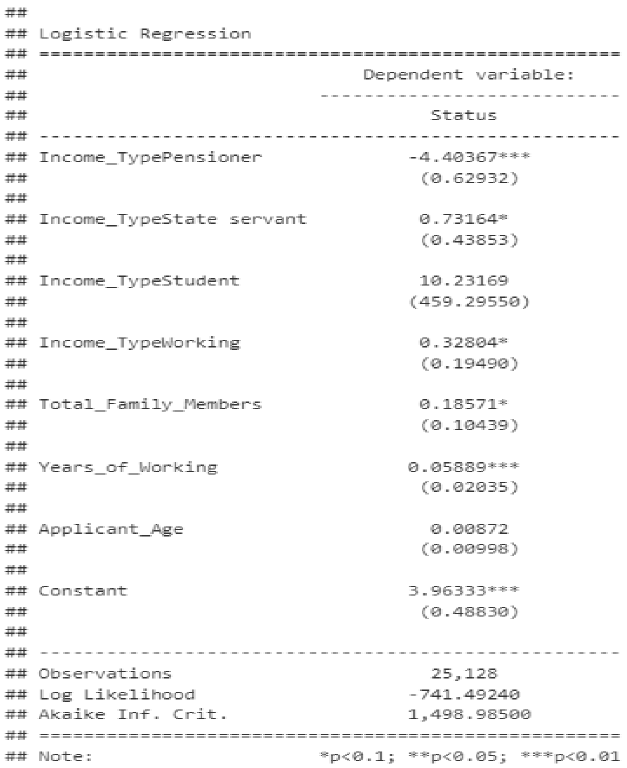
term	X.Df	LogLik	df	statistic	p.value
Status ~ 1	1	-766.35785	NA	NA	NA
Status ~ Total_Good_Debt + Total_Bad_Debt	3	-0.00003	2	1532.716	0

Chi - Squared Test Results : Output 3



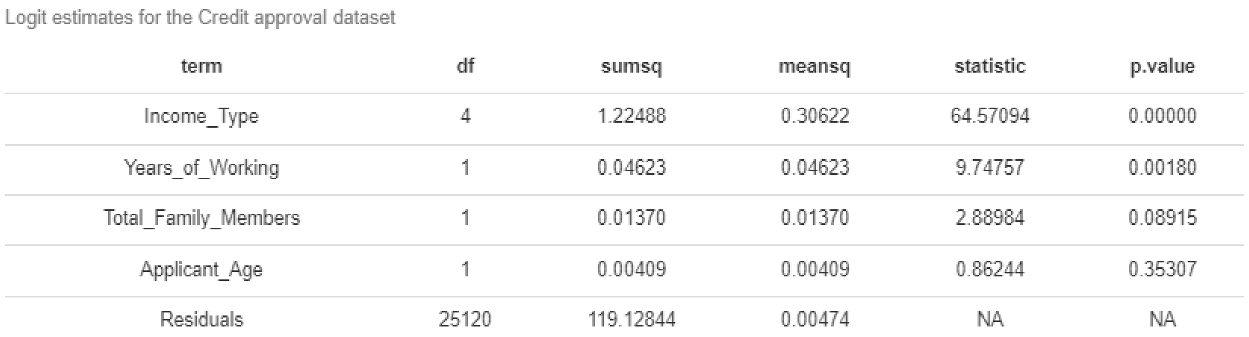
```
##
## Chi-Square test
## ================================
##    Variable  Chi_Squared P_Value
## --------------------------------
## 1 Good debts  624.36920     0
## 2 Bad debts  7,303.63500    0
## --------------------------------
```

Logistic Regression Results results for research question 2: Output 4



```
##
## Logistic Regression
## ====================================================
##                              Dependent variable:
##                          ---------------------------
##                                    Status
## ----------------------------------------------------
## Income_TypePensioner              -4.40367***
##                                    (0.62932)
##
## Income_TypeState servant           0.73164*
##                                    (0.43853)
##
## Income_TypeStudent                 10.23169
##                                   (459.29550)
##
## Income_TypeWorking                 0.32804*
##                                    (0.19490)
##
## Total_Family_Members               0.18571*
##                                    (0.10439)
##
## Years_of_Working                  0.05889***
##                                    (0.02035)
##
## Applicant_Age                       0.00872
##                                    (0.00998)
##
## Constant                          3.96333***
##                                    (0.48830)
##
## ----------------------------------------------------
## Observations                        25,128
## Log Likelihood                    -741.49240
## Akaike Inf. Crit.                 1,498.98500
## ====================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

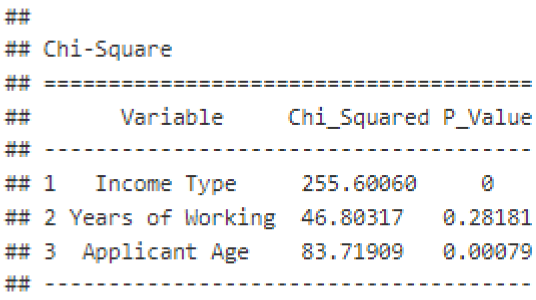
ANOVA test result Output 5:



Logit estimates for the Credit approval dataset

term	df	sumsq	meansq	statistic	p.value
Income_Type	4	1.22488	0.30622	64.57094	0.00000
Years_of_Working	1	0.04623	0.04623	9.74757	0.00180
Total_Family_Members	1	0.01370	0.01370	2.88984	0.08915
Applicant_Age	1	0.00409	0.00409	0.86244	0.35307
Residuals	25120	119.12844	0.00474	NA	NA

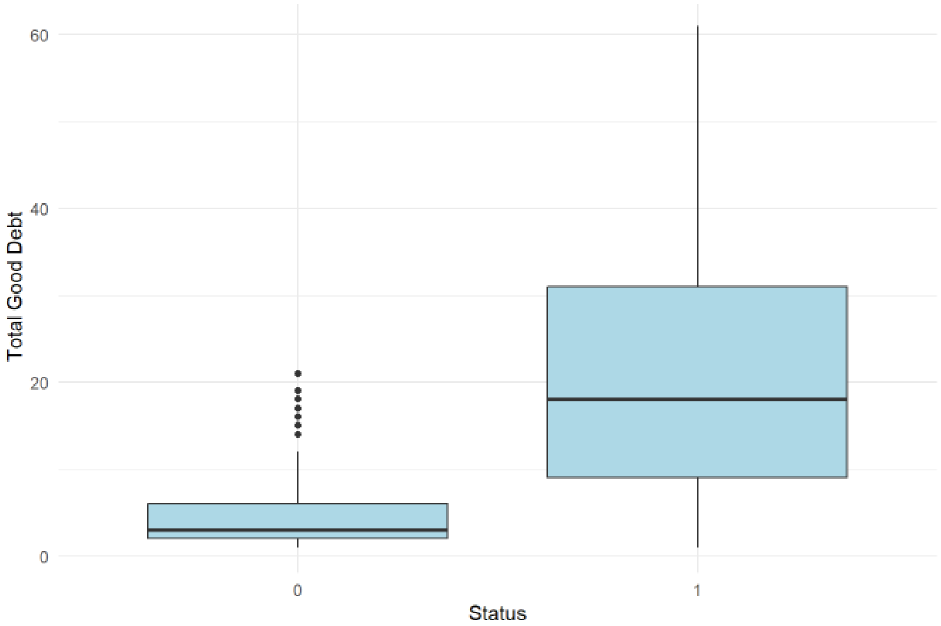
Chi - Squared Test Results : Output 6



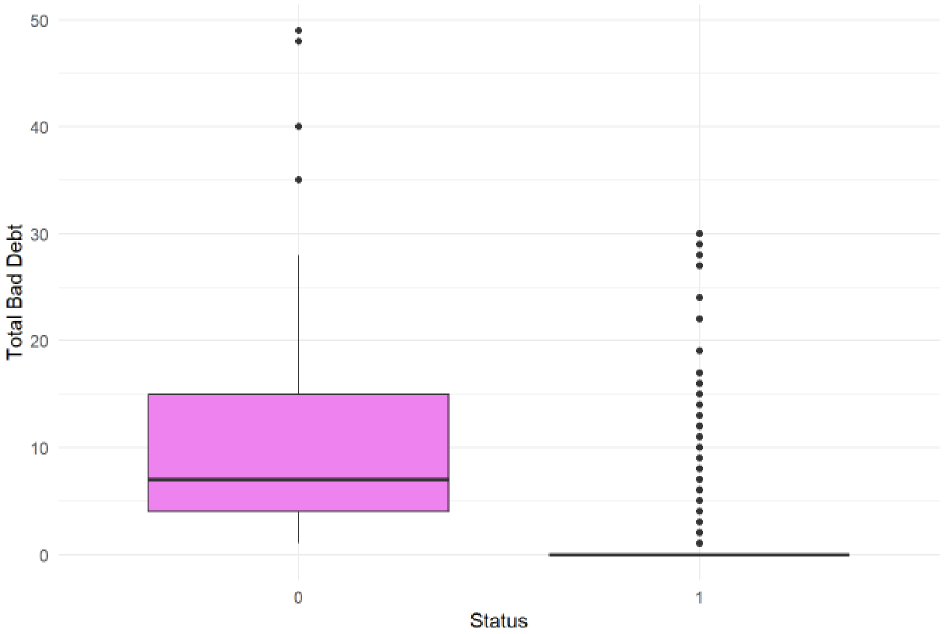
```
## 
## Chi-Square
## ======================================
##       Variable     Chi_Squared P_Value
## --------------------------------------
## 1   Income Type     255.60060     0   
## 2 Years of Working  46.80317   0.28181
## 3  Applicant Age    83.71909   0.00079
## --------------------------------------
```

8.3 Data Visualization:

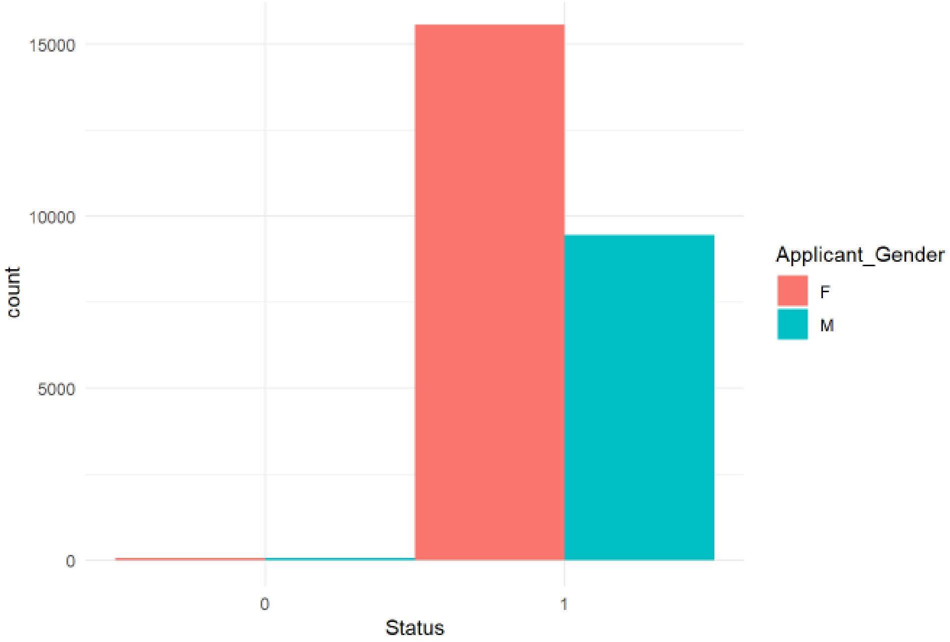
Plot 1: Box-Plot for Total Good Debt by Status



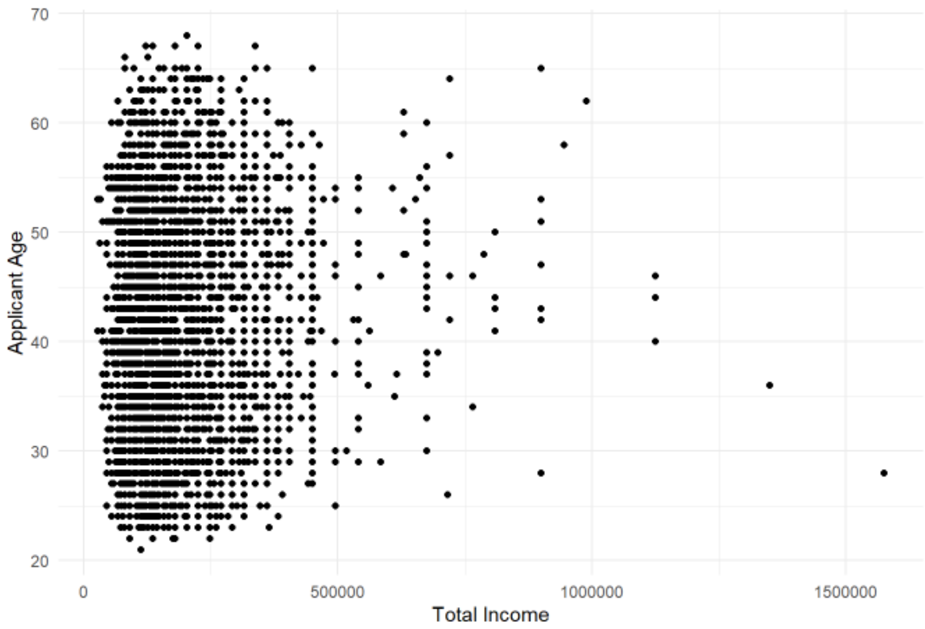
Plot 2: Box-Plot for Total Bad Debt by Status



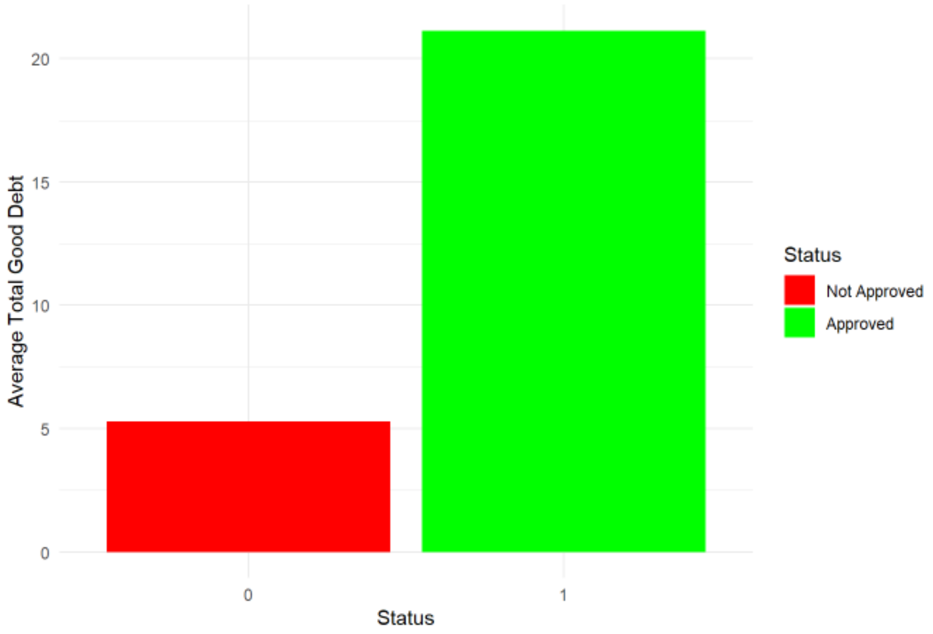
Plot 3: Distribution of Status by Applicant Gender



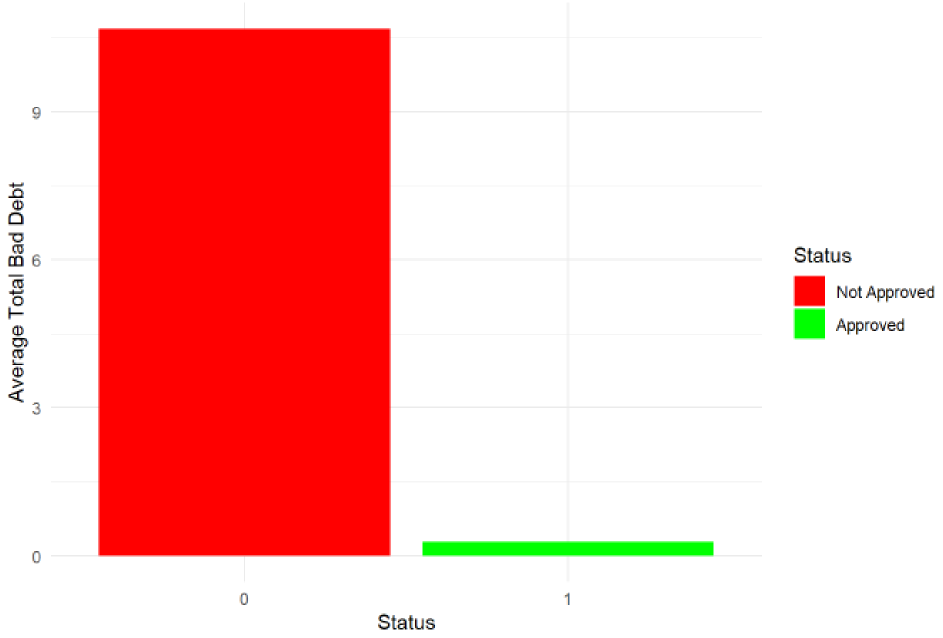
Plot 4: Scatter plot for Total Income and Applicant Age



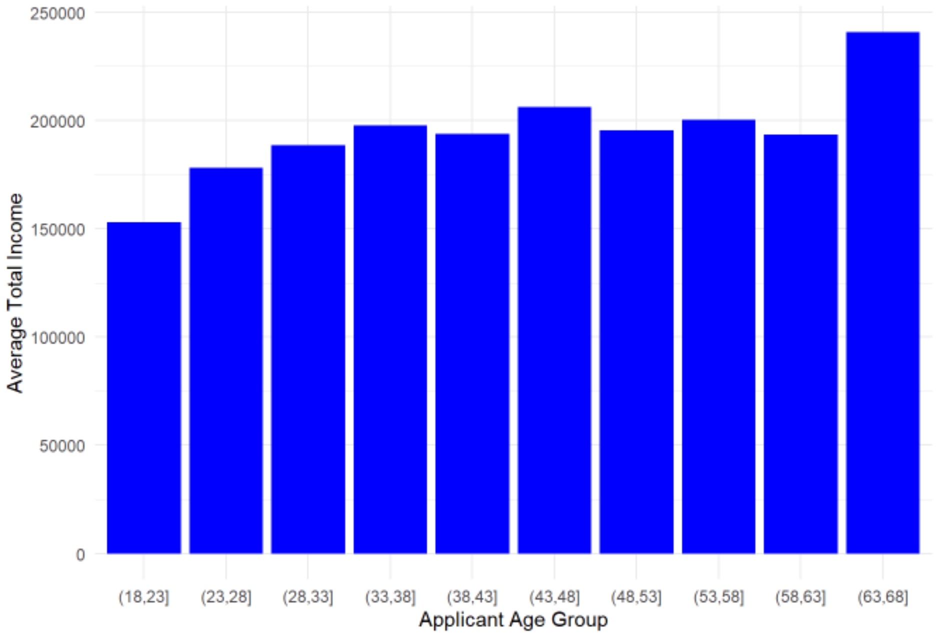
Plot 5: Average Total Good Debt by Status



Plot 6: Average Total Bad Debt by Status



Plot 7: Average Total Income by Applicant Age Group


Average Total Income
(18,23]
(23,28]
(28,33]
(33,38]
(38,43]
(43,48]
(48,53]
(53,58]
(58,63]
(63,68]
Applicant Age Group