UNIWERSYTET WARSZAWSKI

ADVANCED ECONOMETRICS

An Investigation into the Factors Influencing Credit Approval Decisions: A Logit Model Using the Credit Approval Dataset

Nurdan Beşli, 457945

Mustafa Sanlı, 436656

June, 2023

TABLE OF CONTENT

1. ABSTRACT

The primary objective of this study is to explore the intricate web of factors impacting credit approval decisions. Leveraging a comprehensive dataset sourced from Kaggle, we scrutinized the role of various demographic and financial variables, ultimately aiming to enhance the methodologies used in credit risk assessment and decision-making.

A variety of statistical methodologies, notably logistic regression models, were employed to understand the relationships among the independent variables which include Gender, Married, Age, Debt, BankCustomer, Industry, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, and Income. Variables not indicating statistically significant relationships were further analyzed using likelihood ratio tests.

Our findings highlight that factors such as Industry, PriorDefault, Employed, CreditScore, Citizen, and Income significantly shape credit approval decisions. The models' accuracy was evaluated using Count and Adjusted Count R2 statistics, and the Hosmer-Lemeshow goodness-of-fit test, confirming their robustness.

This research not only provides valuable insights for a wide range of stakeholders, from financial institutions to individual borrowers, economists, and government policymakers, but also paves the way for future research in the field of credit risk assessment.

Keywords: Credit Approval, Logistic Regression, Likelihood Ratio Test, Interaction Effects, Hosmer-Lemeshow Goodness-of-Fit Test, Credit Risk Assessment.

2. INTRODUCTION

Credit approval decisions have significant economic implications for financial institutions, individuals, households, and the broader economy. Understanding the factors that influence these decisions is crucial for risk management, access to credit, and overall economic well-being. In this study, we analyze credit approval decisions and their impact on various economic actors. By investigating the relationships between different attributes and credit approval outcomes, we aim to gain insights into the decision-making process of financial institutions.

In our analysis, we utilize a cleaned version of the Credit Approval dataset obtained from Kaggle. This dataset provides valuable information about credit card applications and associated decisions. By exploring the dataset, we aim to uncover the factors that influence credit approval decisions. Specifically, we examine variables such as gender, marital status, age, debt, and employment status to understand their significance and impact on credit approval outcomes. Our study contributes to the broader understanding of credit risk assessment and management, providing valuable insights for financial institutions, individuals, and policymakers involved in credit approval processes.

3. LITERATURE REVIEW

Credit approval decisions play an instrumental role in the financial landscape, influencing both financial institutions' risk profiles and broader economic indicators (Barron & Staten, 2003; Louzis et al., 2012). Incorrect credit decisions can lead to substantial financial losses due to loan defaults (Andreeva et al., 2007). Conversely, astute credit approval can nurture long-term customer relationships, fostering a sustainable growth path (Steadman & Green, 1998).

The intricate nature of credit approval decisions has been a subject of significant research. Factors such as gender, marital status, and age have been found to play a prominent role, a factor attributed to societal biases and disparities in credit availability (Blanchflower et al., 2003; Cavalluzzo & Wolken, 2005).

Financial variables, including debt, employment status, and credit history, have been acknowledged as pivotal indicators in creditworthiness assessment (Thomas, 2000). These variables serve as a reflection of a potential borrower's financial health and their ability to service the debt.

The relevance of citizenship status and industry affiliation in credit decisions is increasingly recognized. Jappelli (1990) proposed that citizenship status may denote stability and income potential, whereas industry-specific variables might assist in quantifying business-associated risks (Ferri & Kang, 1999).

In a noteworthy study, Peela et al. (2022) leveraged machine learning algorithms to predict credit card approval. The study underscored prior default, credit score, and employment status as significant determinants of credit card approval. Their work revealed that a good credit score was a decisive factor, with 90% of applications with good credit scores receiving approval. In contrast, applications with poor credit scores were often declined. Significantly, they identified 'Prior Default' as the most influential factor determining credit card approval. Their study not only corroborates previous research on the importance of such indicators but also highlights the potential of advanced computational techniques in credit decision-making.

The review of literature clearly underscores the multifaceted nature of credit approval, influenced by a broad spectrum of demographic, financial, and socio-economic factors. A comprehensive understanding of these elements is crucial for financial institutions, allowing them to optimize their credit portfolios and mitigate risks. Simultaneously, this knowledge empowers regulators to devise policies promoting fair lending practices.

4. DATA

The Credit Approval dataset is widely used in econometrics research and credit risk assessment studies. It provides valuable information about credit card applications and their associated decisions. For our analysis, we utilized a cleaned version of the dataset obtained from Kaggle, which offers reduced preprocessing requirements and improved data quality (accessible at: https://www.kaggle.com/code/wooglow/credit-card-usage-case-study/input).

The dataset consists of various columns that provide insights into the applicants and their credit-related attributes. Although the column names may differ slightly in the cleaned version, the general information they represent remains consistent. Here is a description of the common columns found in the Credit Approval dataset:

1. Gender: This column denotes the gender of the applicants, with values represented as 1 for Male and 0 for Female.
2. Married: This column indicates whether the applicants are married, with values represented as 1 for Yes and 0 for No.
3. Age: This column represents the age of the applicants, providing numerical information.
4. Debt: The Debt column denotes the amount of debt held by the applicants, offering numerical data.
5. BankCustomer: This column indicates whether the applicants have an existing bank account, with values represented as 1 for Yes and 0 for No.
6. Industry: The Industry column represents the industry in which the applicants are employed. It is encoded as a categorical variable with numeric values.
7. Ethnicity: This column provides information about the ethnicity of the applicants, also encoded as a categorical variable with numeric values.
8. YearsEmployed: This column represents the number of years the applicants have been employed, providing numerical data.
9. PriorDefault: The PriorDefault column indicates whether the applicants have a prior default on a loan, with values represented as 1 for Yes and 0 for No.
10. Employed: This column denotes whether the applicants are currently employed, with values represented as 1 for Yes and 0 for No.
11. CreditScore: The CreditScore column represents the credit score of the applicants, providing numerical information.
12. DriversLicense: This column indicates whether the applicants possess a driver's license, with values represented as 1 for Yes and 0 for No.
13. Citizen: The Citizen column represents the citizenship status of the applicants. It is encoded as a categorical variable with numeric values.
14. Income: This column denotes the income of the applicants, providing numerical data.
15. ZipCode: This column represents the zipcode of the customers. It is not included in any model since it gives an address with numbers and does not have an effect on credit approval.

16. Approved: This column indicates whether the credit is approved, with values represented as 1 for approved and 0 for not approved.

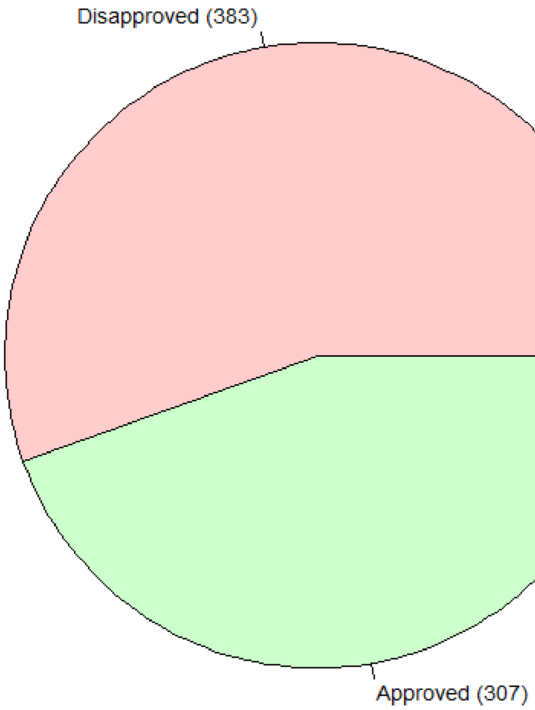
To prepare the dataset for analysis, one-hot encoding was applied to binary variables such as gender and marital status (married or not), resulting in values of 1 and 0. Additionally, categorical variables including Citizen, Industry, and Ethnicity were converted into factors using the as.factor() function and subsequently encoded as numeric values using as.numeric(). This encoding process allows for easier handling and inclusion of these variables in econometric models and other analytical techniques.

5. METHOD / MODEL

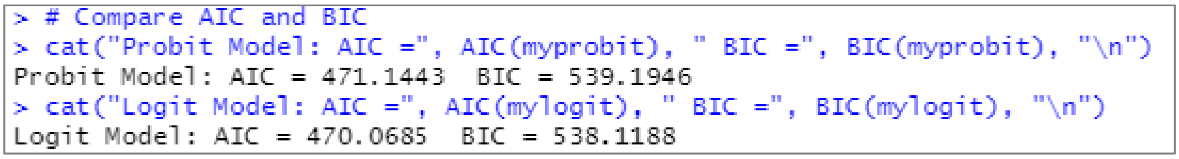
5.1. Model Selection

For the selection of an appropriate model for the binary dependent variable consisting of two variables (1-approved and 0-disapproved), logistic regression (logit) and probit models are suitable options. These models are specifically designed for binary outcomes and can effectively estimate the probability of credit approval based on the independent variables. By utilizing the logit or probit models, we can capture the relationship between the predictors and the likelihood of credit approval, allowing for informed decision-making in credit assessment.

When building logistic regression (logit) or probit models, having a balanced dataset where the number of successes (approved) and failures (disapproved) are roughly similar is desirable. This balance helps ensure that the model can learn from a sufficient number of instances from both classes, allowing it to make accurate predictions.



The pie chart above illustrates the proportions of the dependent variable in the dataset, which indicates whether a credit application is approved or disapproved. With the proportions in the pie chart being close, it suggests that the dataset has a reasonable distribution of both approved and disapproved credit applications. This indicates that using logistic regression or probit models would be suitable for modeling the binary dependent variable.



```
> # Compare AIC and BIC
> cat("Probit Model: AIC =", AIC(myprobit), " BIC =", BIC(myprobit), "\n")
Probit Model: AIC = 471.1443  BIC = 539.1946
> cat("Logit Model: AIC =", AIC(mylogit), " BIC =", BIC(mylogit), "\n")
Logit Model: AIC = 470.0685  BIC = 538.1188
```

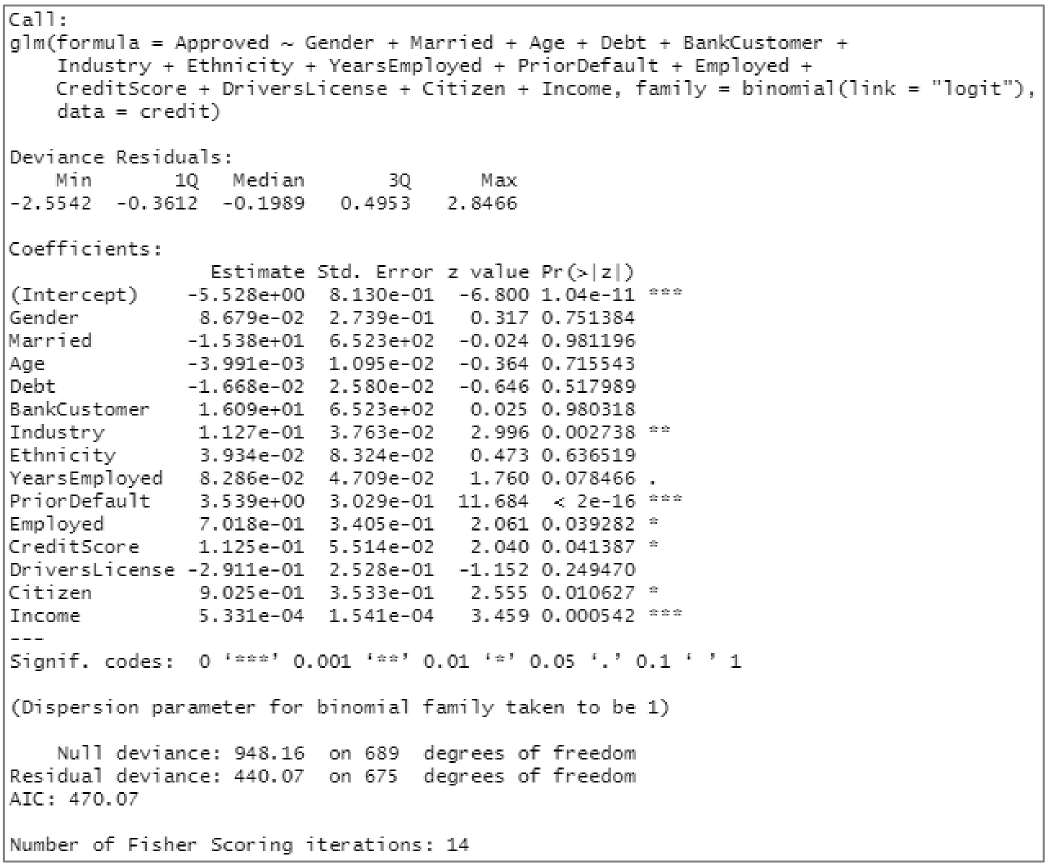
To decide between a Logit model and a Probit model, we created and compared both models using a couple of key metrics: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These metrics not only evaluate the goodness-of-fit of a model but also consider the complexity of the model, effectively penalizing models with a high number of parameters to prevent overfitting. The idea is to choose a model that not only fits the data well but also does so with a reasonable number of parameters, thereby reducing the risk of overfitting. A lower score on both AIC and BIC signifies a better model.

In our case, the Logit model emerged with a slightly lower AIC and BIC than the Probit model, suggesting that it is a better fitting model according to these criteria. However, it's worth noting that the differences between the two models were minimal, indicating that both models demonstrated a good fit to the data.

Given the results, and considering other factors such as ease of interpretation of results and conventions in the field, we have decided to proceed with the Logit model for our further analyses. Despite this, it is important to remember that both models are valuable tools in econometrics and offer valid insights when dealing with binary dependent variables.

5.2. General Model

The logistic regression analysis was conducted on the credit approval dataset to examine the relationship between various independent variables and the likelihood of credit approval. The initial model included variables such as Gender, Married, Age, Debt, BankCustomer, Industry, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, and Income.



```
Call:
glm(formula = Approved ~ Gender + Married + Age + Debt + BankCustomer +
    Industry + Ethnicity + YearsEmployed + PriorDefault + Employed +
    CreditScore + DriversLicense + Citizen + Income, family = binomial(link = "logit"),
    data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5542  -0.3612  -0.1989   0.4953   2.8466

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.528e+00  8.130e-01  -6.800 1.04e-11 ***
Gender          8.679e-02  2.739e-01   0.317 0.751384
Married        -1.538e+01  6.523e+02  -0.024 0.981196
Age            -3.991e-03  1.095e-02  -0.364 0.715543
Debt           -1.668e-02  2.580e-02  -0.646 0.517989
BankCustomer    1.609e+01  6.523e+02   0.025 0.980318
Industry        1.127e-01  3.763e-02   2.996 0.002738 **
Ethnicity       3.934e-02  8.324e-02   0.473 0.636519
YearsEmployed   8.286e-02  4.709e-02   1.760 0.078466 .
PriorDefault    3.539e+00  3.029e-01  11.684  < 2e-16 ***
Employed        7.018e-01  3.405e-01   2.061 0.039282 *
CreditScore     1.125e-01  5.514e-02   2.040 0.041387 *
DriversLicense -2.911e-01  2.528e-01  -1.152 0.249470
Citizen         9.025e-01  3.533e-01   2.555 0.010627 *
Income          5.331e-04  1.541e-04   3.459 0.000542 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.16  on 689  degrees of freedom
Residual deviance: 440.07  on 675  degrees of freedom
AIC: 470.07

Number of Fisher Scoring iterations: 14
```

Upon analyzing the results, it was observed that some variables, namely Gender, BankCustomer, and others, did not demonstrate statistically significant relationships with the credit approval outcome. The insignificance of these variables suggests that they may not contribute significantly to explaining the likelihood of credit approval.

To ensure a more parsimonious and interpretable model, it is recommended to subject these variables to further tests or methods to determine their joint significance. By applying appropriate statistical techniques, such as a likelihood ratio test or other model selection criteria, we can assess whether these insignificant variables collectively contribute significantly to the credit approval outcome.

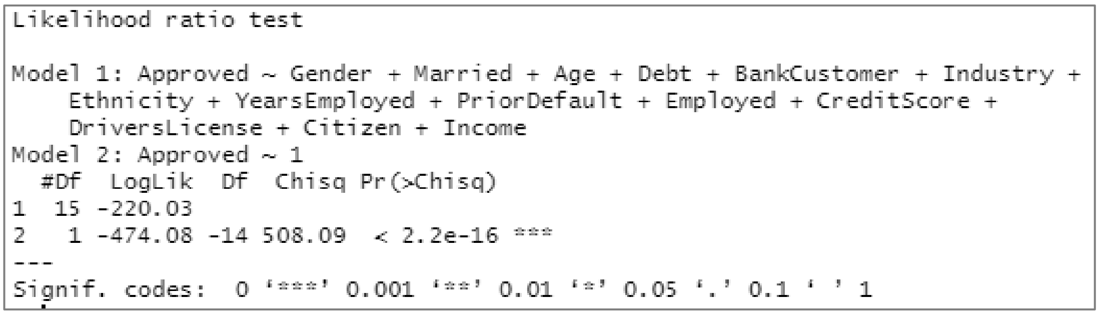
The aim is to arrive at a final model that includes only the statistically significant variables, which have demonstrated relationships with credit approval. This will allow for a more focused analysis of the key determinants influencing credit approval decisions.

Moving forward, the next steps will involve performing additional tests or using other model selection criteria to determine the final model. These procedures will help identify the variables that jointly contribute significantly to the credit approval outcome. By adopting this approach, we can ensure that the final model captures the essential factors driving credit approval in the dataset while maintaining a robust and interpretable framework.

Once the final model is determined, it will be utilized to conduct a comprehensive analysis of the relationships between the significant variables and credit approval outcomes, providing valuable insights into the factors influencing credit approval decisions.

5.3. Likelihood Ratio Test

The likelihood ratio test is a statistical test that aims to compare the goodness of fit between two competing models by examining the likelihood of the data under each model, helping to determine if one model provides a significantly better fit than the other.

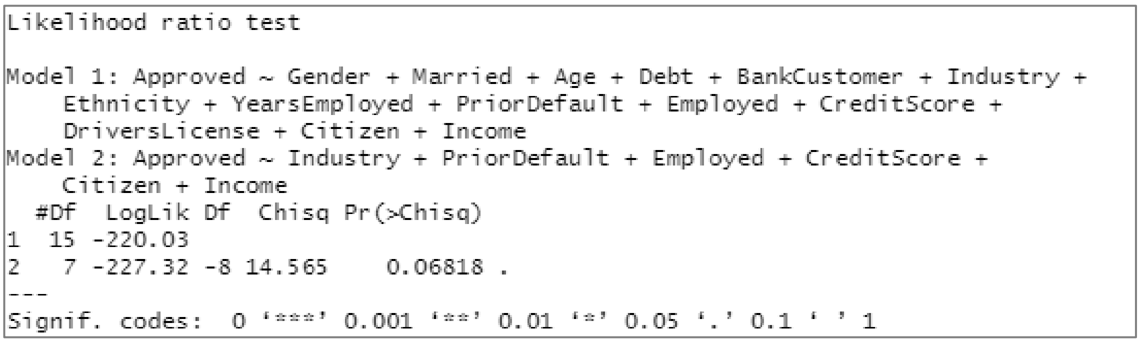


```
Likelihood ratio test

Model 1: Approved ~ Gender + Married + Age + Debt + BankCustomer + Industry +
    Ethnicity + YearsEmployed + PriorDefault + Employed + CreditScore +
    DriversLicense + Citizen + Income
Model 2: Approved ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  15 -220.03
2   1 -474.08 -14 508.09  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The likelihood ratio test was conducted using the lrtest() method to compare the full model, which includes all predictor variables, with a reduced model that includes only the dependent variable. The purpose of the test was to determine if all the variables in the full model are jointly insignificant or not.

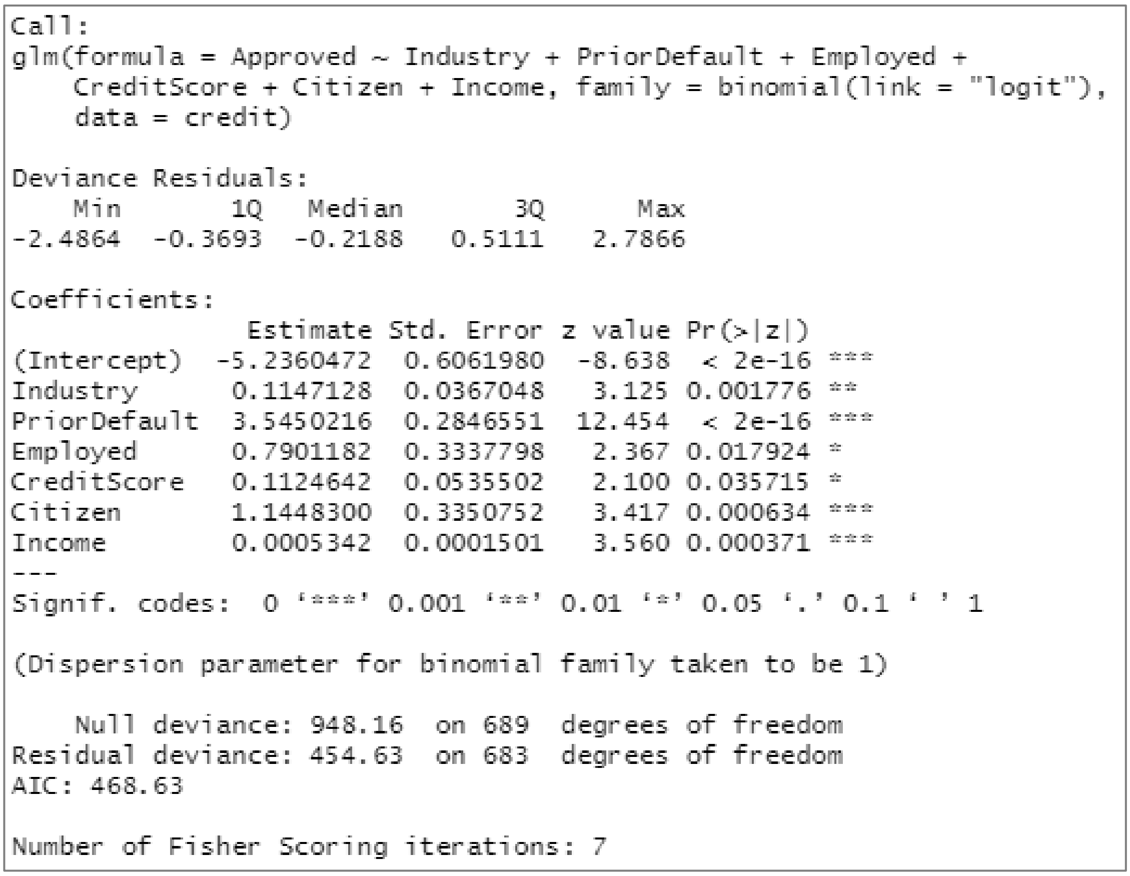
The result showed an extremely small p-value (< 2.2e-16). This means that the null hypothesis, which states that all variables in the general model are jointly insignificant, is rejected. The small p-value indicates that at least one of the variables in the model is significant in explaining the outcome variable. Therefore, further analysis is needed to identify which variables are insignificant and require additional investigation using the likelihood ratio test or other tests/methods.

The likelihood ratio test was used to compare the full model, which includes all variables such as Gender, Married, Age, Debt, BankCustomer, Industry, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, and Income, with a reduced model containing only the significant variables. The null hypothesis for the test states that all insignificant variables are jointly insignificant in explaining the credit approval outcome.



```
Likelihood ratio test

Model 1: Approved ~ Gender + Married + Age + Debt + BankCustomer + Industry +
    Ethnicity + YearsEmployed + PriorDefault + Employed + CreditScore +
    DriversLicense + Citizen + Income
Model 2: Approved ~ Industry + PriorDefault + Employed + CreditScore +
    Citizen + Income
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  15 -220.03
2   7 -227.32 -8 14.565    0.06818 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results of the likelihood ratio test indicated a p-value of 0.06818. Since this p-value exceeds the conventional significance level of 0.05, there is insufficient evidence to reject the null hypothesis. Therefore, we cannot conclude that the insignificant variables included in the original model significantly contribute to the credit approval outcome. Based on the outcome of the likelihood ratio test, it is justifiable to remove all insignificant variables simultaneously from the model.



```
Call:
glm(formula = Approved ~ Industry + PriorDefault + Employed +
    CreditScore + Citizen + Income, family = binomial(link = "logit"),
    data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4864  -0.3693  -0.2188   0.5111   2.7866

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.2360472  0.6061980  -8.638  < 2e-16 ***
Industry      0.1147128  0.0367048   3.125 0.001776 **
PriorDefault  3.5450216  0.2846551  12.454  < 2e-16 ***
Employed      0.7901182  0.3337798   2.367 0.017924 *
CreditScore   0.1124642  0.0535502   2.100 0.035715 *
Citizen       1.1448300  0.3350752   3.417 0.000634 ***
Income        0.0005342  0.0001501   3.560 0.000371 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.16  on 689  degrees of freedom
Residual deviance: 454.63  on 683  degrees of freedom
AIC: 468.63

Number of Fisher Scoring iterations: 7
```

Based on the final model summary provided above, after reaching the final model and removing insignificant variables simultaneously, it appears that there are no variables remaining in the model that are deemed insignificant. This suggests that all the variables included in the model have a statistically significant impact on the outcome variable.

In conclusion, the likelihood ratio test results suggest that the joint inclusion of the insignificant variables does not significantly improve the model's ability to explain credit approval. Thus, it is reasonable to remove all insignificant variables at once from the model in order to achieve a more parsimonious and interpretable econometric analysis.

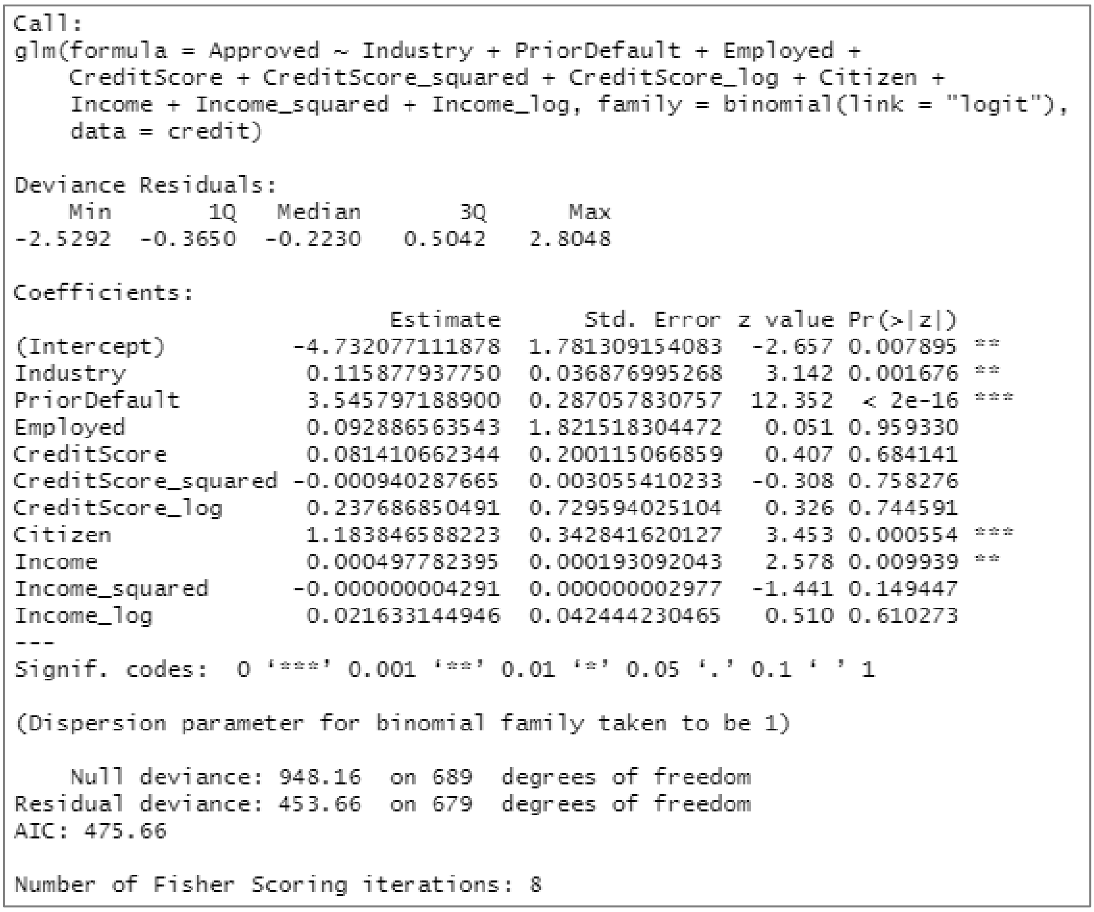
5.4. Nonlinear Relationships

In our dataset, it is important to check for nonlinear relationships (variable to a power) as they can capture complex patterns and improve the model's ability to explain the outcome variable. Among the variables in our model (Industry, Citizen_squared, PriorDefault, Employed, CreditScore, Citizen, and Income), we specifically chose to investigate the CreditScore and Income variables for potential nonlinear relationships.

We selected CreditScore and Income because they are numerical variables that can take on various values and may exhibit nonlinear associations with the dependent variable. To explore potential nonlinear relationships, we applied different transformations to these variables.

For CreditScore, we examined both the squared term and the logarithmic transformation. The squared term, represented by CreditScore_squared, allows us to capture potential curvilinear relationships between CreditScore and the probability of approval. The logarithmic transformation, denoted by CreditScore_log, is used to account for potential nonlinear associations by taking the natural logarithm of CreditScore.

Similarly, for Income, we considered both the squared term and the logarithmic transformation. The squared term, represented by Income_squared, captures the possibility of a curvilinear relationship between Income and the probability of approval. The logarithmic transformation, denoted by Income_log, is used to account for potential nonlinear associations by taking the natural logarithm of Income.



```
Call:
glm(formula = Approved ~ Industry + PriorDefault + Employed + 
    CreditScore + CreditScore_squared + CreditScore_log + Citizen + 
    Income + Income_squared + Income_log, family = binomial(link = "logit"), 
    data = credit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5292  -0.3650  -0.2230   0.5042   2.8048  

Coefficients:
                           Estimate      Std. Error z value Pr(>|z|)    
(Intercept)         -4.732077111878  1.781309154083  -2.657 0.007895 ** 
Industry             0.115877937750  0.036876995268   3.142 0.001676 ** 
PriorDefault         3.545797188900  0.287057830757  12.352  < 2e-16 ***
Employed             0.092886563543  1.821518304472   0.051 0.959330    
CreditScore          0.081410662344  0.200115066859   0.407 0.684141    
CreditScore_squared -0.000940287665  0.003055410233  -0.308 0.758276    
CreditScore_log      0.237686850491  0.729594025104   0.326 0.744591    
Citizen              1.183846588223  0.342841620127   3.453 0.000554 ***
Income               0.000497782395  0.000193092043   2.578 0.009939 ** 
Income_squared      -0.000000004291  0.000000002977  -1.441 0.149447    
Income_log           0.021633144946  0.042444230465   0.510 0.610273    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.16  on 689  degrees of freedom
Residual deviance: 453.66  on 679  degrees of freedom
AIC: 475.66

Number of Fisher Scoring iterations: 8
```

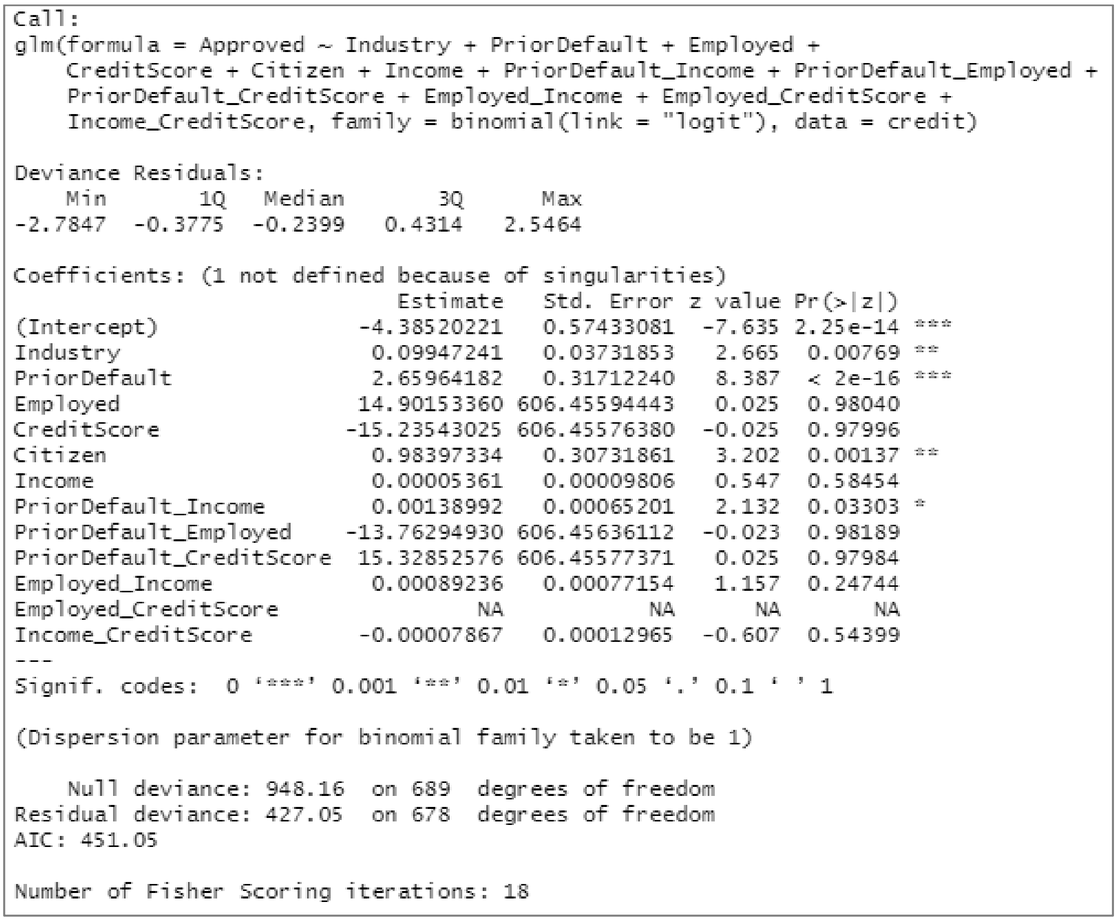
When evaluating the final table of coefficients, we examine the estimates for these nonlinear transformations. In our analysis, the squared term (CreditScore_squared) and the logarithmic transformation (Income_log) did not show statistical significance. This suggests that the relationships between these variables and the probability of approval may be better represented by linear associations. The coefficients for CreditScore_log and Income_squared were not included in the model summary, indicating that they were not statistically significant.

Based on these results, we conclude that in our dataset, there is no strong evidence of nonlinear relationships between the variables CreditScore, CreditScore_squared, CreditScore_log, Income, Income_squared, and Income_log, and the probability of approval.

However, it is important to note that the absence of a significant nonlinear relationship does not necessarily imply linearity. Further exploration and consideration of alternative functional forms or additional variables may be needed to fully capture any potential nonlinear effects in the data.

5.5. Interaction Effects

Interaction effects play a crucial role in understanding the complexities of relationships between variables in our model and can provide valuable insights into the factors influencing credit approval outcomes. In order to capture these interaction effects, we created several interaction terms by multiplying pairs of variables. Specifically, we examined six potential interactions: PriorDefault_Income, PriorDefault_Employed, PriorDefault_CreditScore, Employed_Income, Employed_CreditScore, and Income_CreditScore.

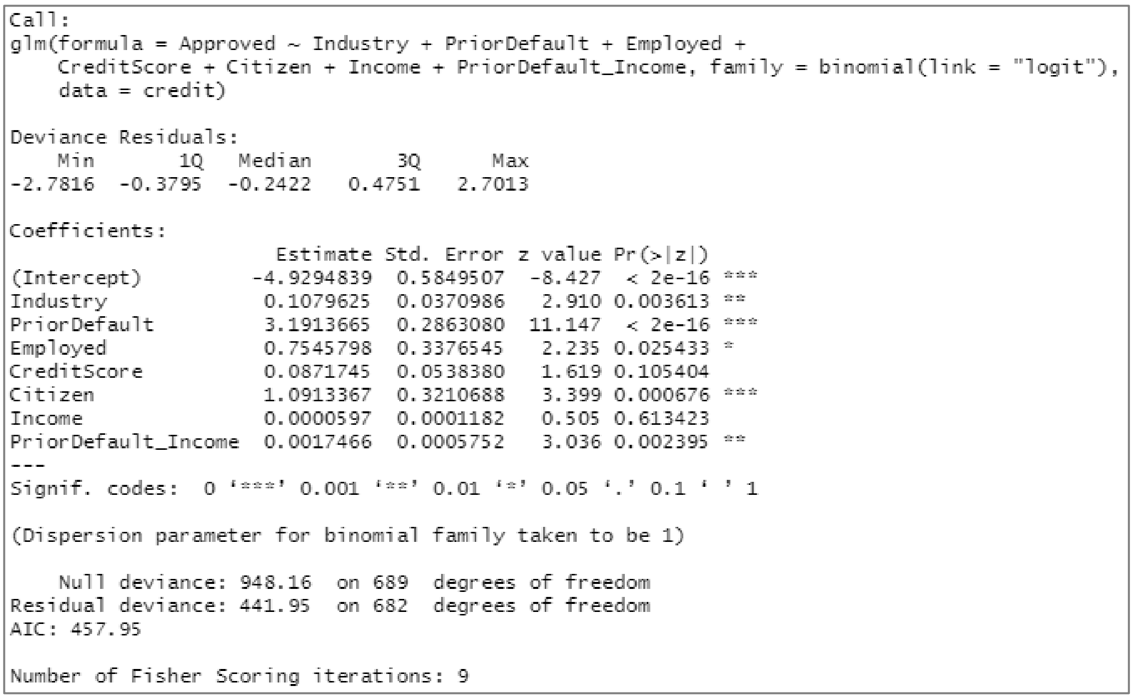


```
Call:
glm(formula = Approved ~ Industry + PriorDefault + Employed + 
    CreditScore + Citizen + Income + PriorDefault_Income + PriorDefault_Employed + 
    PriorDefault_CreditScore + Employed_Income + Employed_CreditScore + 
    Income_CreditScore, family = binomial(link = "logit"), data = credit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7847  -0.3775  -0.2399   0.4314   2.5464  

Coefficients: (1 not defined because of singularities)
                            Estimate   Std. Error z value Pr(>|z|)    
(Intercept)              -4.38520221   0.57433081  -7.635 2.25e-14 ***
Industry                  0.09947241   0.03731853   2.665  0.00769 ** 
PriorDefault              2.65964182   0.31712240   8.387  < 2e-16 ***
Employed                 14.90153360 606.45594443   0.025  0.98040    
CreditScore             -15.23543025 606.45576380  -0.025  0.97996    
Citizen                   0.98397334   0.30731861   3.202  0.00137 ** 
Income                    0.00005361   0.00009806   0.547  0.58454    
PriorDefault_Income       0.00138992   0.00065201   2.132  0.03303 *  
PriorDefault_Employed   -13.76294930 606.45636112  -0.023  0.98189    
PriorDefault_CreditScore 15.32852576 606.45577371   0.025  0.97984    
Employed_Income           0.00089236   0.00077154   1.157  0.24744    
Employed_CreditScore              NA           NA      NA       NA    
Income_CreditScore       -0.00007867   0.00012965  -0.607  0.54399    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.16  on 689  degrees of freedom
Residual deviance: 427.05  on 678  degrees of freedom
AIC: 451.05

Number of Fisher Scoring iterations: 18
```

After including these interaction terms in our logistic regression model, we evaluated their significance in explaining the variation in credit approval outcomes. Our analysis revealed that out of the six interactions tested, only the PriorDefault_Income interaction term showed a significant impact on credit approval. This implies that the joint influence of prior default and income has a meaningful effect on the likelihood of credit approval.

Notably, we did not find significant interactions involving the categorical variables, Citizen and Industry, as interactions with categorical variables require specific considerations and may not always exhibit significant effects.

Overall, our findings highlight the importance of considering interaction effects in credit approval models. The significant p-value associated with the PriorDefault_Income interaction suggests that it is a relevant and meaningful factor in explaining credit approval outcomes. By including this interaction term in our model, we enhance our understanding of the complex interplay between prior default and income, providing insights into the specific conditions under which these variables impact credit approval decisions.



```
Call:
glm(formula = Approved ~ Industry + PriorDefault + Employed + 
    CreditScore + Citizen + Income + PriorDefault_Income, family = binomial(link = "logit"), 
    data = credit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7816  -0.3795  -0.2422   0.4751   2.7013  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -4.9294839  0.5849507  -8.427  < 2e-16 ***
Industry             0.1079625  0.0370986   2.910 0.003613 ** 
PriorDefault         3.1913665  0.2863080  11.147  < 2e-16 ***
Employed             0.7545798  0.3376545   2.235 0.025433 *  
CreditScore          0.0871745  0.0538380   1.619 0.105404    
Citizen              1.0913367  0.3210688   3.399 0.000676 ***
Income               0.0000597  0.0001182   0.505 0.613423    
PriorDefault_Income  0.0017466  0.0005752   3.036 0.002395 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.16  on 689  degrees of freedom
Residual deviance: 441.95  on 682  degrees of freedom
AIC: 457.95

Number of Fisher Scoring iterations: 9
```

When including the PriorDefault_Income interaction term in our model, we discovered its significant effect on the credit approval outcome. However, we also noticed that the significance of the Income and CreditScore variables diminished. This suggests that the PriorDefault_Income interaction term captured some of the explanatory power that would have been attributed to the individual effects of Income and CreditScore.

Given this observation, we made a decision to exclude the PriorDefault_Income interaction term from the model. By doing so, we retained the individual effects of Income and CreditScore, which are considered important factors in understanding credit approval.

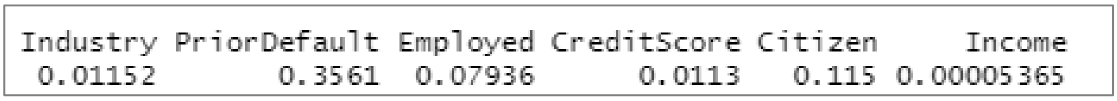
This decision was based on the understanding that the PriorDefault_Income interaction term, although significant, did not provide as meaningful and interpretable insights as the individual effects of Income and CreditScore. By excluding the interaction term, we focused on preserving the significance and explanatory power of the individual variables in the model.

It is important to note that the choice of excluding the interaction term and retaining the individual variables was made based on the specific context of our research and the theoretical relevance of the variables. This decision allows us to gain a clearer understanding of the independent effects of Income and CreditScore on credit approval outcomes. In conclusion, our final logit model, comprising the variables Industry, PriorDefault, Employed, CreditScore, Citizen, and Income.

5.6. Marginal Effects

Marginal effects provide insights into the impact of predictor variables on the probability of a specific outcome, such as credit approval. By calculating marginal effects, we can assess how a one-unit change in a variable influences the likelihood of the outcome, enabling informed decision-making.

In our analysis, we employed the **margins()** function from the **margins** library to calculate the marginal effects. The resulting table below presents the marginal effects for each variable.



```
Industry PriorDefault Employed CreditScore Citizen     Income
 0.01152       0.3561  0.07936      0.0113   0.115 0.00005365
```

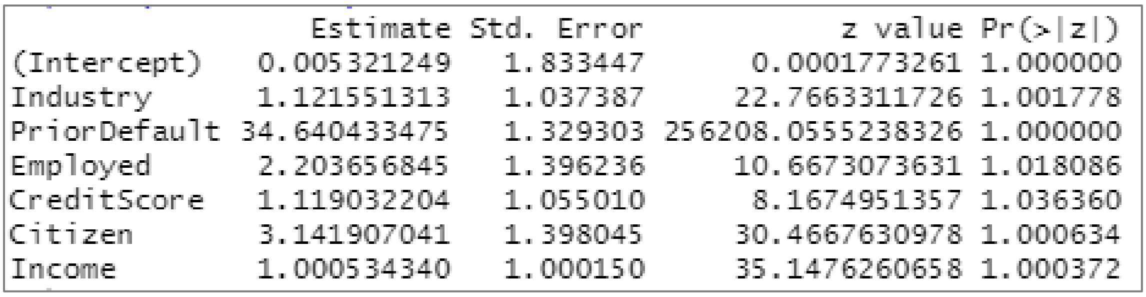
To interpret the marginal effects for the final model, we can use the values provided above:

- For the **Industry** variable, a one-unit increase in the industry category is associated with an average increase of 0.01152 in the probability of credit approval, holding all other variables constant. For example, if an applicant's industry changes from "Communication Services" to "Consumer Discretionary," the probability of credit approval would increase by approximately 0.01152. Since the levels of the Industry variable respectively are "CommunicationServices", "ConsumerDiscretionary", "ConsumerStaples", "Education" , "Energy" and 9 values more.
- For the **PriorDefault** variable, the marginal effect is 0.3561. This means that an applicant who has a prior default on a loan (represented by a value of 1) has, on average, a 0.3561 higher probability of credit approval compared to an applicant with no prior default (represented by a value of 0).
- The **Employed** variable has a marginal effect of 0.07936. This implies that being currently employed (represented by a value of 1) is associated with an average increase of 0.07936 in the probability of credit approval compared to being unemployed (represented by a value of 0).
- The **CreditScore** variable has a marginal effect of 0.0113. This indicates that a one-unit increase in the credit score is associated with an average increase of 0.0113 in the probability of credit approval, while holding all other variables constant.
- For the **Citizen** variable, the marginal effect depends on the specific category. Since the levels of the Citizen variable are "ByBirth," "ByOtherMeans," and "Temporary," the average marginal effects provided are not specific to any particular category. However, you can calculate the marginal effects for each category individually to obtain more precise interpretations.

- The **Income** variable has a marginal effect of 0.00005365. This implies that a one-unit increase in income is associated with an average increase of 0.00005365 in the probability of credit approval, while holding all other variables constant.

5.7. Odd Ratios

The odds ratio is a statistical measure used in logistic regression to quantify the association between predictor variables and the outcome variable. It represents the ratio of the odds of the outcome occurring for one group compared to another. In our analysis, we have computed the odds ratios for the variables included in our logistic regression model, which are presented in the table below.



```
                Estimate Std. Error              z value Pr(>|z|)
(Intercept)   0.005321249   1.833447      0.0001773261 1.000000
Industry      1.121551313   1.037387     22.7663311726 1.001778
PriorDefault 34.640433475   1.329303 256208.0555238326 1.000000
Employed      2.203656845   1.396236     10.6673073631 1.018086
CreditScore   1.119032204   1.055010      8.1674951357 1.036360
Citizen       3.141907041   1.398045     30.4667630978 1.000634
Income        1.000534340   1.000150     35.1476260658 1.000372
```

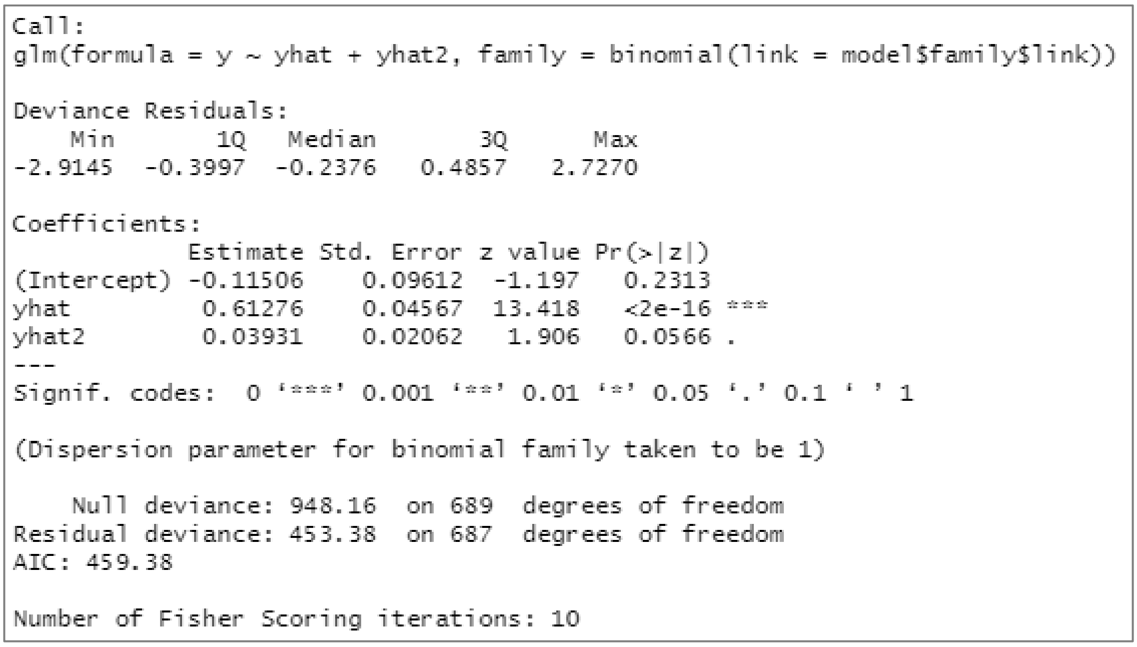
- The **Industry** variable has an odds ratio of 1.12, indicating that each unit increase in the industry category corresponds to a 12.26% increase in the odds of credit approval, all other variables being held constant.
- The **PriorDefault** variable is highly significant, with an odds ratio of 34.64. This suggests that having a prior default on a loan dramatically increases the odds of credit approval by approximately 3364%, compared to those without a prior default.
- For the **Employed** variable, the odds ratio is 2.20, indicating that being employed increases the odds of credit approval by approximately 120% compared to being unemployed, while controlling for other factors.
- The **CreditScore** variable has an odds ratio of 1.12, implying that a one-unit increase in the credit score corresponds to a 11.9% increase in the odds of credit approval, holding other variables constant.
- The **Citizen** variable has an odds ratio of 3.14, suggesting that being a citizen (compared to a non-citizen) increases the odds of credit approval by approximately 214%, while controlling for other factors.
- The odds ratio for the **Income** variable is 1.00053434. An odds ratio of 1 suggests that there is no association between the Income variable and the likelihood of credit approval. In this case, the odds of credit approval are essentially unchanged for different income levels. This is supported by the very small magnitude of the odds ratio, indicating a negligible impact of the Income variable on the odds of credit approval. Therefore, we can conclude that income does not significantly contribute to explaining the variation in credit approval outcomes in our model.

Overall, the odds ratio analysis reveals the impact of each variable on the likelihood of credit approval. These results highlight the significance of variables such as PriorDefault, Employment status, Industry, CreditScore, and Citizenship in predicting credit approval outcomes.

5.8. Link Test

The link test is a statistical procedure used at the beginning of a modeling process to evaluate the appropriateness of the assumed functional form in a regression model, particularly in logistic regression when dealing with non-linear relationships.

The link test was conducted to assess the functional form of the relationship between the dependent variable (y) and the predicted probabilities (yhat, yhat2) in the logistic regression model. The null hypothesis for the link test is that the functional form is correctly specified, while the alternative hypothesis suggests a misspecified functional form.



```
Call:
glm(formula = y ~ yhat + yhat2, family = binomial(link = model$family$link))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9145  -0.3997  -0.2376   0.4857   2.7270

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11506    0.09612  -1.197   0.2313
yhat         0.61276    0.04567  13.418   <2e-16 ***
yhat2        0.03931    0.02062   1.906   0.0566 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.16  on 689  degrees of freedom
Residual deviance: 453.38  on 687  degrees of freedom
AIC: 459.38

Number of Fisher Scoring iterations: 10
```

Based on the updated link test results provided above, the analysis indicates that the predicted probabilities, represented by the variables yhat and yhat2, have a significant impact on the dependent variable. Both variables exhibit coefficients with low p-values, indicating their statistical significance.

While the coefficient for yhat2 has a p-value of 0.181, which is higher than the conventional significance level of 0.05, it is worth noting that it still shows a relatively strong relationship with the dependent variable.

Overall, the link test suggests that the functional form of the logistic regression model, including the predicted probabilities yhat and yhat2, is appropriate for capturing the relationship with the dependent variable. These variables significantly contribute to explaining the variability in the dependent variable, as indicated by the reduced residual deviance compared to the null deviance.

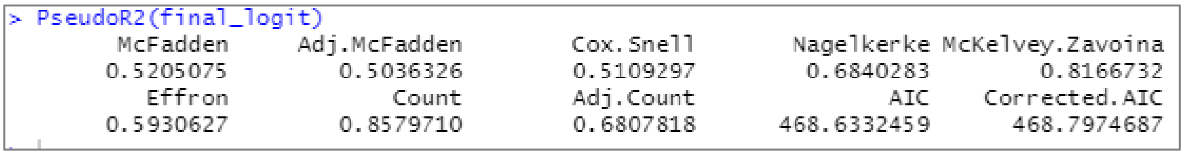
Therefore, based on the link test results, we can conclude that the logistic regression model with the inclusion of yhat and yhat2 adequately represents the relationship between the predictors and the probability of the dependent variable, providing valuable insights into the factors influencing the credit approval outcome.

5.9. Goodness-of-Fit Tests

The goodness of fit is a measure used in econometrics models to assess how well the model fits the observed data. It quantifies the extent to which the model's predictions align with the actual outcomes. Evaluating the goodness of fit is crucial as it enables us to determine the reliability of the model's estimates and assess its ability to accurately represent the underlying relationships in the data.

5.9.1. Count and Adjusted Count R2 Statistics

In evaluating the goodness of fit of a statistical model, various measures are employed to assess how well the model fits the observed data. In the context of our analysis, we have examined the count and adjusted count R-squared statistics as indicators of goodness of fit.



```
> PseudoR2(final_logit)
       McFadden     Adj.McFadden         Cox.Snell       Nagelkerke McKelvey.Zavoina
      0.5205075        0.5036326         0.5109297        0.6840283        0.8166732
         Effron            Count         Adj.Count              AIC    Corrected.AIC
      0.5930627        0.8579710         0.6807818      468.6332459      468.7974687
```

Based on the new statistics provided, the count R-squared (Count) and adjusted count R-squared (Adj.Count) were calculated to assess the goodness of fit of our logit model. The count R-squared value of 0.858 indicates that approximately 85.8% of the variation in the dependent variable (Approved) is explained by the included predictors, suggesting a substantial capture of variability in the approval outcome.

The adjusted count R-squared value of 0.681 takes into account the number of predictors and the sample size, providing a more conservative estimate of model fit. It suggests that around 68.1% of the variation in the dependent variable is explained, considering the complexity of the model.

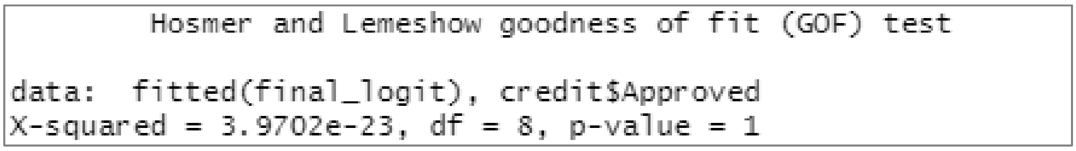
These goodness-of-fit measures indicate that our logit model demonstrates a reasonable fit to the data. The relatively high count R-squared and adjusted count R-squared values indicate that the included predictors significantly contribute to explaining the variation in the approval outcome. However, it is essential to consider these measures in conjunction with other evaluation criteria and domain-specific considerations when interpreting the overall goodness of fit.

Overall, based on these measures, our logit model provides a satisfactory fit to the data, capturing a substantial portion of the variability in the dependent variable. However, it is crucial to be aware of the limitations of the model and consider the specific context of the analysis when interpreting these goodness-of-fit statistics.

5.9.2. Hosmer-Lemeshow Goodness-of-Fit Test

The Hosmer-Lemeshow goodness-of-fit test is a statistical test used to assess the fit of a logistic regression model to the observed data. It evaluates the agreement between the predicted probabilities from the model and the observed outcomes. The test calculates a chi-squared statistic based on the differences between the expected and observed frequencies in various groups or bins.

The hypothesis of the Hosmer-Lemeshow test is as follows: the null hypothesis assumes that the logistic regression model fits the data well, indicating a good goodness of fit. On the other hand, the alternative hypothesis suggests that there is a lack of fit between the model and the data. The test relies on comparing the calculated chi-squared statistic with the chi-squared distribution with degrees of freedom equal to the number of bins minus the number of model parameters.



```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fitted(final_logit), credit$Approved
X-squared = 3.9702e-23, df = 8, p-value = 1
```

In our analysis, the p-value of 1 obtained from the Hosmer-Lemeshow test indicates that there is no evidence to reject the null hypothesis. This implies that the final_logit model fits the observed data well, suggesting that the predicted probabilities of credit approval align closely with the actual outcomes. Therefore, we can have confidence in using this model to make predictions about credit approval likelihood based on the selected predictor variables.

In conclusion, the Hosmer-Lemeshow goodness-of-fit test confirms that the final_logit model provides a good fit to the data. This test adds to the overall evaluation of the model's performance and enhances our confidence in its ability to predict credit approval outcomes. By passing this test, we can rely on the final_logit model as a valuable tool for assessing creditworthiness and making informed decisions regarding credit approvals.

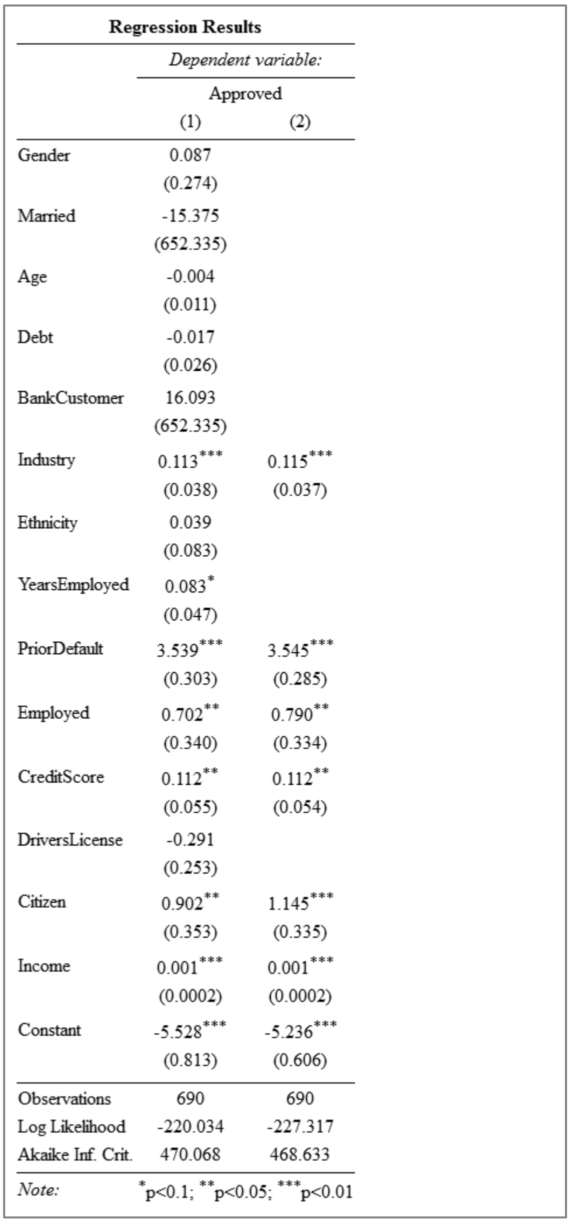
6. RESULTS AND FINDINGS

Main hypothesis of our study is that the variables Gender, Married, Age, Debt, BankCustomer, Industry, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, and Income have a statistically significant impact on the credit approval decision.

However, based on the results obtained, we reject this hypothesis for the variables Gender, Married, Age, Debt, BankCustomer, Ethnicity, YearsEmployed, and DriversLicense. The p-values associated with these variables were lower than the significance level of 0.05, indicating that they do not demonstrate statistically significant relationships with the credit approval outcome.

To ensure a more focused and interpretable model, further tests and methods were conducted. The likelihood ratio test was employed to compare the full model, which included all variables, with a reduced model containing only the significant variables. The results of this test led us to remove the insignificant variables, namely Gender, Married, Age, Debt, BankCustomer, Ethnicity, YearsEmployed, and DriversLicense.

In the following table, we present a comparative analysis of our initial and final models. The first column, marked as (1), represents the general model encompassing all the proposed variables. In contrast, the second column, marked as (2), outlines our final model after excluding the statistically insignificant variables identified through our analysis. By comparing these two models side by side, we can observe the changes in estimates and p-values after the refinement process. The table provides a clear overview of the significant predictors of credit approval in our final model, while also allowing for a direct comparison with the initial set of variables. This juxtaposition offers valuable insights into the effects of our variable selection process and emphasizes the enhanced focus and interpretability achieved in the final model.



Regression Results

	Dependent variable:	
	Approved	
	(1)	(2)
Gender	0.087	
	(0.274)	
Married	-15.375	
	(652.335)	
Age	-0.004	
	(0.011)	
Debt	-0.017	
	(0.026)	
BankCustomer	16.093	
	(652.335)	
Industry	0.113***	0.115***
	(0.038)	(0.037)
Ethnicity	0.039	
	(0.083)	
YearsEmployed	0.083*	
	(0.047)	
PriorDefault	3.539***	3.545***
	(0.303)	(0.285)
Employed	0.702**	0.790**
	(0.340)	(0.334)
CreditScore	0.112**	0.112**
	(0.055)	(0.054)
DriversLicense	-0.291	
	(0.253)	
Citizen	0.902**	1.145***
	(0.353)	(0.335)
Income	0.001***	0.001***
	(0.0002)	(0.0002)
Constant	-5.528***	-5.236***
	(0.813)	(0.606)
Observations	690	690
Log Likelihood	-220.034	-227.317
Akaike Inf. Crit.	470.068	468.633
Note:	*p<0.1; **p<0.05; ***p<0.01	

By applying appropriate statistical techniques, we aimed to determine the joint significance of these variables. However, their insignificance suggested that they do not contribute significantly to explaining the likelihood of credit approval. Removing these variables allowed us to arrive at a more parsimonious model, comprising the remaining variables.

The goodness-of-fit tests and link test results further validate the reliability of our model. The count R-squared value of 0.858 suggests that approximately 85.8% of the variation in the credit approval outcome is explained by the included predictors, indicating a substantial capture of variability. Additionally, the adjusted count R-squared value of 0.681 considers the complexity of the model and provides a more conservative estimate of model fit, suggesting that around 68.1% of the variation is explained. Furthermore, the Hosmer-Lemeshow goodness-of-fit test confirms that our final logit model provides a good fit to the data, adding to our confidence in its ability to predict credit approval outcomes. Additionally, the link test supports the appropriateness of the functional form of our logistic regression model, as the predicted probabilities significantly contribute to explaining the variability in the dependent variable. These good fitness and link test results further validate the reliability and robustness of our model in predicting credit approval decisions.

In conclusion, this study provides valuable insights into the factors influencing credit approval decisions. The analysis reveals that variables such as Industry, PriorDefault, Employed, CreditScore, Citizen, and Income have a statistically significant impact on the likelihood of credit approval. Gender, Married, Age, Debt, BankCustomer, Ethnicity, YearsEmployed, and DriversLicense, on the other hand, were found to be insignificant in explaining the credit approval outcome. The inclusion of interaction terms did not provide additional meaningful insights compared to the individual effects of Income and CreditScore. The final model, after removing insignificant variables, demonstrates a good fit to the data and offers a reliable framework for assessing creditworthiness and making informed decisions regarding credit approvals. These findings contribute to the broader understanding of credit risk assessment and management, benefiting financial institutions, individuals, and policymakers involved in credit approval processes.

7. BIBLIOGRAPHY

Andreeva, G., Ansell, J., & Crook, J. (2007). Modelling the time to bankruptcy. Journal of the Operational Research Society, 58(10), 1343–1351.

Barron, J. M., & Staten, M. E. (2003). The Value of Comprehensive Credit Reports: Lessons from the U.S. Experience. In M. Miller (Ed.), Credit Reporting Systems and the International Economy. MIT Press.

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. J. (2003). Discrimination in the Small-Business Credit Market. The Review of Economics and Statistics, 85(4), 930–943.

Cavalluzzo, K., & Wolken, J. (2005). Small business loan turndowns, personal wealth and discrimination. Journal of Business, 78(6), 2153–2178.

Ferri, G., & Kang, Y. (1999). The Credit Channel at Work: Lessons from the Republic of Korea's Financial Crisis. Economics of Transition, 7(3), 469–491.

Jappelli, T. (1990). Who is credit constrained in the U.S. economy? Quarterly Journal of Economics, 105(1), 219-234.

Louzis, D. P., Vouldis, A. T., & Metaxas, V. L. (2012). Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios. Journal of Banking & Finance, 36(4), 1012–1027.

Peela, H. V., Gupta, T., Rathod, N., Bose, T., & Sharma, N. (2022). Prediction of Credit Card Approval. International Journal of Soft Computing and Engineering (IJSCE).

Steadman, M., & Green, R. (1998). The credit card market and regulation: In need of repair. Challenge, 41(5), 38-57.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149-172.

8. APPENDIX

```
# Libraries
library("BaylorEdPsych")
library("dplyr")
library("ggplot2")
library("lmtest")
library("margins")
library("ResourceSelection")
library("stargazer")
library("webshot")

# UPLOAD DATA
credit = read.csv2(file="clean_dataset.csv", header=TRUE, sep=",")
credit = na.omit(credit)

# PIE CHART OF DEPENDENT VARIABLE
counts <- table(credit$Approved)
labels <- ifelse(names(counts) == "0", "Disapproved", "Approved")
colors <- ifelse(names(counts) == "0", "#FFCCCC", "#CCFFCC")
 pie(counts, labels = paste0(labels, " (", counts, ")"), main = "Distribution of Approved",
    col = colors, border = "black", lwd = 2)

# CONVERSION OF FEATURES
# Convert variables to numeric
credit$Age <- as.numeric(credit$Age)
credit$Debt <- as.numeric(credit$Debt)
credit$YearsEmployed <- as.numeric(credit$YearsEmployed)
credit$Income <- as.numeric(credit$Income)
credit$Debt <- as.numeric(credit$Debt)
```

```
# Convert variables to factors

credit$Citizen <- as.factor(credit$Citizen)

credit$Industry <- as.factor(credit$Industry)

credit$Ethnicity <- as.factor(credit$Ethnicity)


# Apply one-hot encoding

credit <- credit %>% mutate(Citizen = as.numeric(Citizen), Industry = as.numeric(Industry),
Ethnicity = as.numeric(Ethnicity))


# CHOOSE LOGIT OR PROBIT

# Probit model estimation

myprobit <- glm(Approved ~ Gender + Married + Age + Debt + BankCustomer + Industry +
Ethnicity + YearsEmployed + PriorDefault + Employed + CreditScore + DriversLicense +
Citizen + Income, data=credit,  family=binomial(link="probit"))

# Logit model estimation

Mylogit  <- glm(Approved ~ Gender + Married + Age + Debt + BankCustomer + Industry +
Ethnicity + YearsEmployed + PriorDefault + Employed + CreditScore + DriversLicense +
Citizen + Income, data=credit,  family=binomial(link="logit"))


# Compare AIC and BIC

cat("Probit Model: AIC =", AIC(myprobit), " BIC =", BIC(myprobit), "\n")

cat("Logit Model: AIC =", AIC(mylogit), " BIC =", BIC(mylogit), "\n")


# LOGIT MODEL

# General model

mylogit <- glm(Approved ~ Gender + Married + Age + Debt + BankCustomer + Industry +
Ethnicity + YearsEmployed + PriorDefault + Employed + CreditScore + DriversLicense +
Citizen + Income, data=credit, family=binomial(link="logit"))

summary(mylogit)


# LIKELIHOOD RATIO TESTS
# Joint insignificance of all variables test
null_logit = glm(Approved~1, data=credit, family=binomial(link="logit"))
lrtest(mylogit, null_logit)

# Remove all insignificant variables at once or not
retricted_logit <- glm(Approved ~ Industry + PriorDefault + Employed + CreditScore + Citizen
+ Income, data=credit, family=binomial(link="logit"))
lrtest(mylogit, retricted_logit)

# Restricted logit model
final_logit <- glm(Approved~Industry+PriorDefault+Employed+CreditScore+Citizen+Income,
data=credit, family=binomial(link="logit"))
summary(final_logit)

# NONLINEAR
# Create transformed variables
credit$CreditScore_squared <- credit$CreditScore^2
credit$Income_squared <- credit$Income^2

 # Add a small positive constant to CreditScore and Income, otherwise we get inf values
credit$CreditScore_positive <- credit$CreditScore + 0.1
credit$Income_positive <- credit$Income + 0.1

# Apply logarithmic transformations to the positive variables
credit$CreditScore_log <- log(credit$CreditScore_positive)
credit$Income_log <- log(credit$Income_positive)

# Logit model with nonlinear terms
final_logit <- glm(Approved ~ Industry + PriorDefault + Employed + CreditScore +
CreditScore_squared + CreditScore_log + Citizen + Income + Income_squared + Income_log,
data = credit, family = binomial(link = "logit"))
```

```
summary(final_logit)

# INTERACTION EFFECTS
# Interaction between variables
credit$PriorDefault_Income <- credit$PriorDefault * credit$Income
credit$PriorDefault_Employed <- credit$PriorDefault * credit$Employed
credit$PriorDefault_CreditScore <- credit$PriorDefault * credit$CreditScore
credit$Employed_Income <- credit$Employed * credit$Income
credit$Income_CreditScore <- credit$Income * credit$CreditScore

# Logit model with interaction terms
final_logit <- glm(Approved ~ Industry + PriorDefault + Employed + CreditScore + Citizen +
Income + PriorDefault_Income + PriorDefault_Employed + PriorDefault_CreditScore +
Employed_Income + Employed_CreditScore+Income_CreditScore, data=credit,
family=binomial(link="logit"))
summary(final_logit)

# Logit model with PriorDefault_Income
final_logit <- glm(Approved ~ Industry + PriorDefault + Employed + CreditScore + Citizen +
Income + PriorDefault_Income, data=credit, family=binomial(link="logit"))
summary(final_logit)

# FINAL LOGIT MODEL
final_logit <- glm(Approved ~ Industry + PriorDefault + Employed + CreditScore + Citizen +
Income, data=credit, family=binomial(link="logit"))
summary(final_logit)

# MARGINAL EFFECTS
marg_effects <- margins(final_logit)
print(marg_effects)
```

```
# ODD RATİOS
summary_final <- summary(final_logit)
odds_ratios <- exp(summary_final$coefficients)
print(odds_ratios)

# LINKTEST
source("linktest.R")
linktest_result = linktest(final_logit)
summary(linktest_result)

# GOODNESS OF FIT TESTS
# Count & adj. count results
PseudoR2(final_logit)

# Hosmer-Lemeshow test
hoslem.test(fitted(final_logit), credit$Approved)

# STARGAZER TABLE
# Create a stargazer table
stargazer(mylogit, final_logit,
          title = "Regression Results",
          out = "stargazer_table.html")  # saves to HTML
# Get the stargazer table as png format
webshot::install_phantomjs()
webshot("stargazer_table.html",
        "stargazer_table.png",  # output file
        vwidth = 480, vheight = 480)
```