Binomial GLM of Loan Default Prediction Dataset

Basil Hawash 47592

1- Objective description:

The objective of this dataset, sourced from Coursera's Loan Default Prediction Challenge, is to provide a platform for addressing a key industry-related machine learning problem. Specifically, the dataset aims to challenge individuals in developing predictive models to identify those at high risk of defaulting on loans. By analyzing the provided data, the goal is to enhance the understanding and application of machine learning techniques in predicting loan defaults, ultimately contributing to the development of effective interventions and risk management strategies in the financial domain.

2- The chosen dataset:

https://www.kaggle.com/datasets/nikhil1e9/loan-default/data

3- Context:

Financial loan services, utilized by diverse entities such as major banks, financial institutions, and government loan programs, aim to minimize payment defaults and ensure timely repayments. To achieve this systematically, companies employ machine learning to predict individuals at high risk of default. This dataset, extracted from Coursera's Loan Default Prediction Challenge, presents an industry-relevant machine learning problem. With 255,347 rows and 18 columns, it challenges your modeling skills, offering hands-on experience in addressing the intricacies of loan default prediction.

4- Dependent variable:

Default, binary variable, indicating whether the loan was defaulted or not.

Binomial GLM of Loan Default Prediction Dataset

5- Independent variables:

Loan ID (Nominal Variable): a unique identifier for each loan

Age (Ratio Variable): The age of the borrower

Income (Ratio Variable): The annual income of the borrower

Loan Amount (Ratio Variable): The amount of money being borrowed

Credit Score (Ratio Variable): The credit score of the borrower

Months Employed (Ratio Variable): The number of months the borrower has been employed

Num Credit Lines (Ratio Variable): The number of credit lines the borrower has open

Interest Rate (Ratio Variable): The interest rate for the loan

Loan Term (ordinal variable): The term length of the loan in months

DTI Ratio (Ratio Variable): The Debt-to-Income ratio

Education (ordinal variable): The highest level of education attained by the borrower

Employment Type (nominal variable): The type of employment status of the borrower

Marital Status (nominal variable): The marital status of the borrower

Has Mortgage (binary variable): Whether the borrower has a mortgage

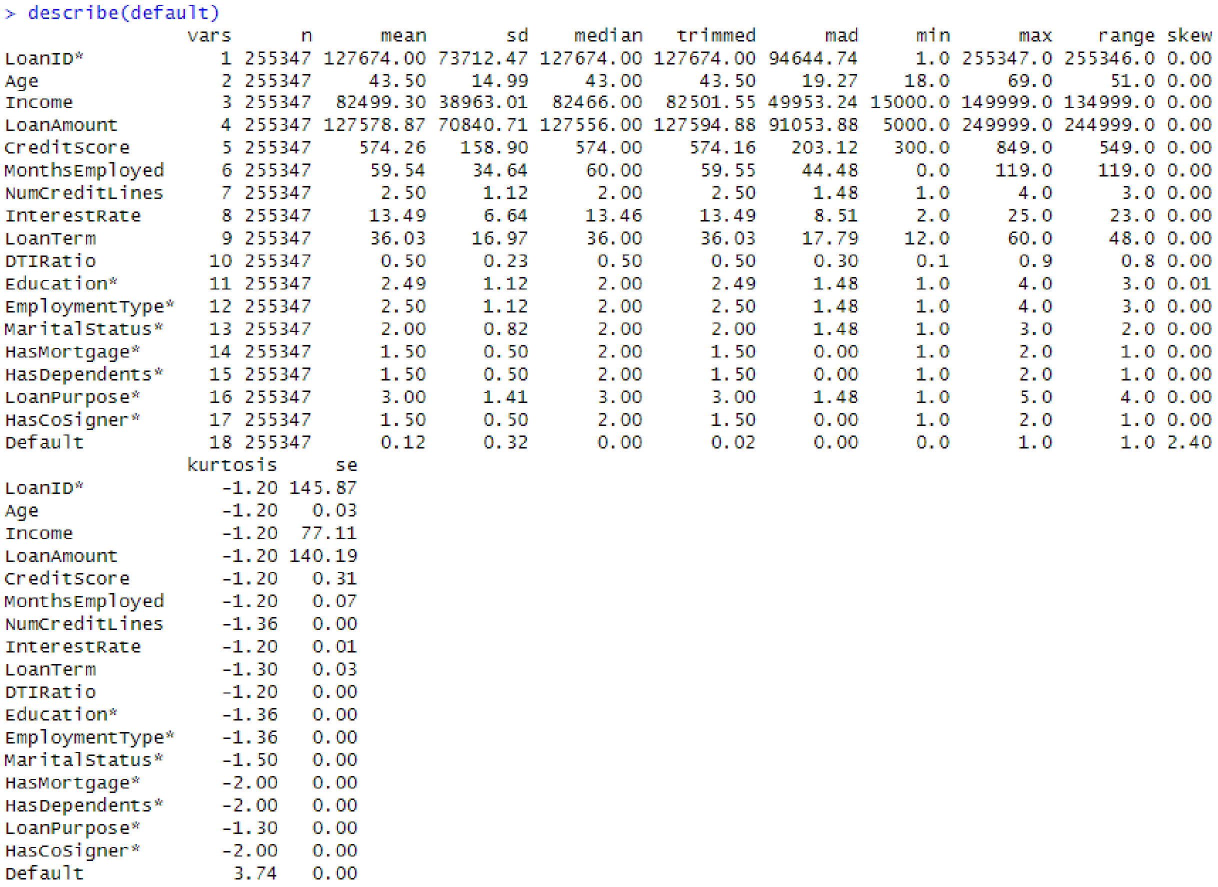
Has Dependents (binary variables): Whether the borrower has dependents

Loan Purpose (nominal variables): The purpose of the loan

Has Cosigner (binary variable): Whether the loan has a co-signer

Binomial GLM of Loan Default Prediction Dataset

6- Exploratory Data Analysis (EDA): I checked for NA values and used the describe function to perform initial analysis.



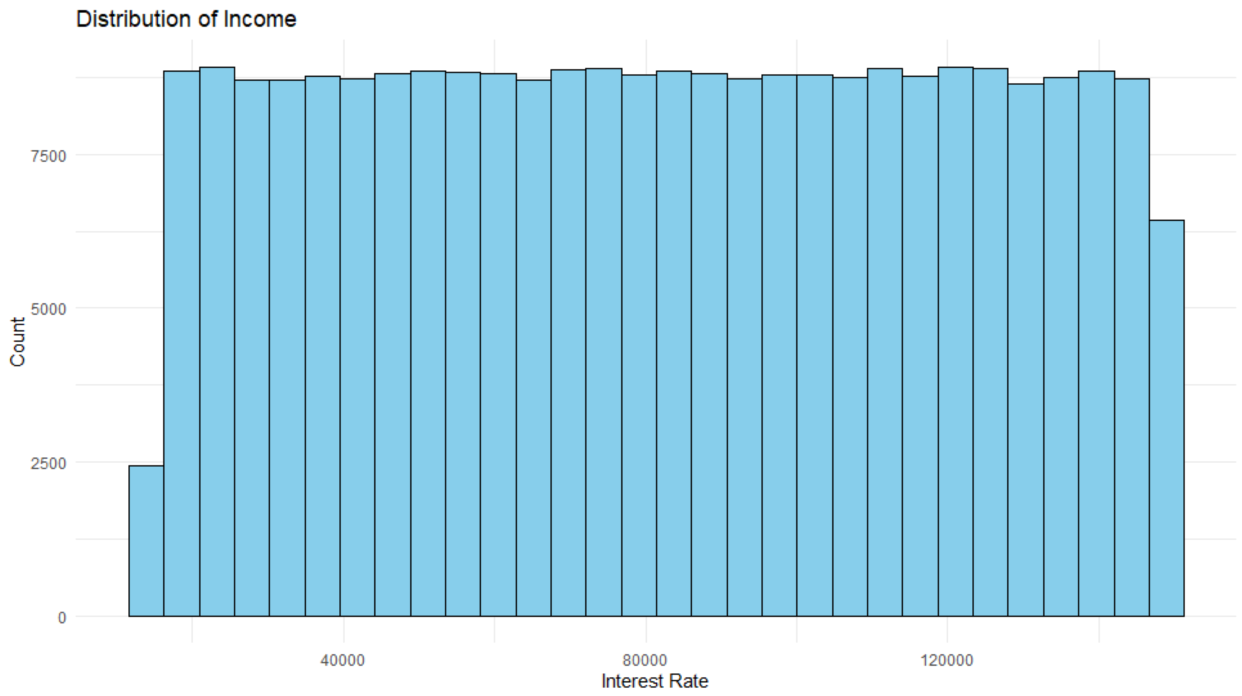
```
> describe(default)
                vars      n      mean       sd    median   trimmed      mad     min      max    range skew
LoanID*            1 255347 127674.00 73712.47 127674.00 127674.00 94644.74     1.0 255347.0 255346.0 0.00
Age                2 255347     43.50    14.99     43.00     43.50    19.27    18.0     69.0     51.0 0.00
Income             3 255347  82499.30 38963.01  82466.00  82501.55 49953.24 15000.0 149999.0 134999.0 0.00
LoanAmount         4 255347 127578.87 70840.71 127556.00 127594.88 91053.88  5000.0 249999.0 244999.0 0.00
CreditScore        5 255347    574.26   158.90    574.00    574.16   203.12   300.0    849.0    549.0 0.00
MonthsEmployed     6 255347     59.54    34.64     60.00     59.55    44.48     0.0    119.0    119.0 0.00
NumCreditLines     7 255347      2.50     1.12      2.00      2.50     1.48     1.0      4.0      3.0 0.00
InterestRate       8 255347     13.49     6.64     13.46     13.49     8.51     2.0     25.0     23.0 0.00
LoanTerm           9 255347     36.03    16.97     36.00     36.03    17.79    12.0     60.0     48.0 0.00
DTIRatio          10 255347      0.50     0.23      0.50      0.50     0.30     0.1      0.9      0.8 0.00
Education*        11 255347      2.49     1.12      2.00      2.49     1.48     1.0      4.0      3.0 0.01
EmploymentType*   12 255347      2.50     1.12      2.00      2.50     1.48     1.0      4.0      3.0 0.00
MaritalStatus*    13 255347      2.00     0.82      2.00      2.00     1.48     1.0      3.0      2.0 0.00
HasMortgage*      14 255347      1.50     0.50      2.00      1.50     0.00     1.0      2.0      1.0 0.00
HasDependents*    15 255347      1.50     0.50      2.00      1.50     0.00     1.0      2.0      1.0 0.00
LoanPurpose*      16 255347      3.00     1.41      3.00      3.00     1.48     1.0      5.0      4.0 0.00
HasCoSigner*      17 255347      1.50     0.50      2.00      1.50     0.00     1.0      2.0      1.0 0.00
Default           18 255347      0.12     0.32      0.00      0.02     0.00     0.0      1.0      1.0 2.40
                kurtosis     se
LoanID*            -1.20 145.87
Age                -1.20   0.03
Income             -1.20  77.11
LoanAmount         -1.20 140.19
CreditScore        -1.20   0.31
MonthsEmployed     -1.20   0.07
NumCreditLines     -1.36   0.00
InterestRate       -1.20   0.01
LoanTerm           -1.30   0.03
DTIRatio           -1.20   0.00
Education*         -1.36   0.00
EmploymentType*    -1.36   0.00
MaritalStatus*     -1.50   0.00
HasMortgage*       -2.00   0.00
HasDependents*     -2.00   0.00
LoanPurpose*       -1.30   0.00
HasCoSigner*       -2.00   0.00
Default             3.74   0.00
```

- The skewness and kurtosis values indicate the symmetry and peakedness of the distributions for each variable. Most variables have a skewness close to 0, indicating fairly symmetrical distributions. However, the Default variable's positive skew (2.40) reflects its imbalance (more 0s than 1s). Kurtosis values provide insight into the tails of the distributions; for example, the Default variable has a kurtosis of 3.74, indicating a leptokurtic distribution, which is more peaked than a normal distribution, reflecting the binary nature of this variable.

- The mean of Default (0.12) suggests that around 12% of the loans in this dataset default, indicating class imbalance, which is common in default prediction but something to be aware of during model training and evaluation.

Binomial GLM of Loan Default Prediction Dataset

- Distribution of Income



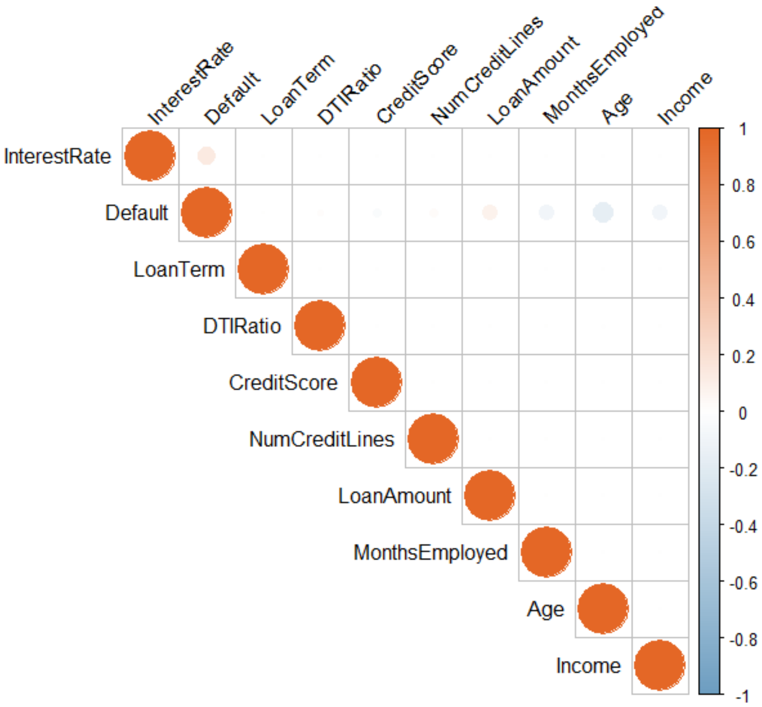
- The histogram shows a relatively uniform distribution of Income, with each bin having a similar count of observations. This is quite unusual for income data, which typically follows a right-skewed distribution with many observations at the lower end and few at the higher end.

- The uniformity might indicate a potential issue with the data. It could be that the data has been artificially binned or capped, or it may have undergone some form of transformation. This does not reflect the natural variation one would expect in an income distribution.

- There are no visible outliers in the distribution as it is uniform across the range. Outliers would typically appear as isolated bars on the extreme ends of the x-axis.

Binomial GLM of Loan Default Prediction Dataset

- Correlation Matrix



- The LoanAmount appears to be somewhat correlated with Income and DTIRatio, which could indicate multicollinearity if these variables are used together in a regression model. This needs to be investigated further, as multicollinearity can affect the stability and interpretation of the regression coefficients. LoanAmount will not be included in the model.

- Some variables show clear correlations with the target variable (Default), suggesting they may be good predictors in a logistic regression model for predicting defaults.

Binomial GLM of Loan Default Prediction Dataset

7- Model selection:

In the process of model selection, our choice leaned towards employing the Generalized Linear Model (GLM) with the binomial link function due to the binary nature of our target variable, which represents loan default status. Ensuring the adherence to the model assumptions is crucial for its effectiveness:

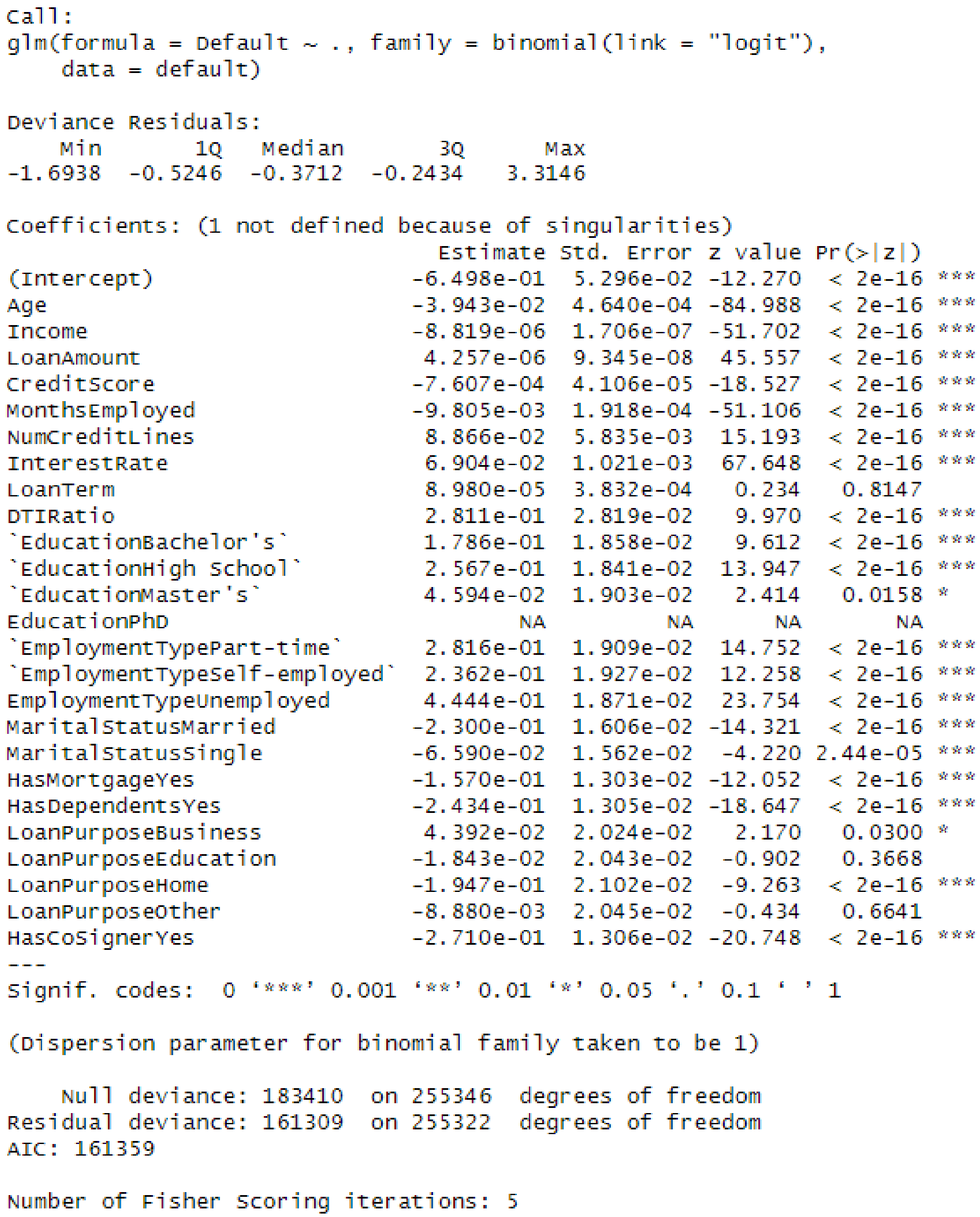
- Binary Response Variable: The binary nature of our response variable aligns with the requirement of the GLM with the binomial link function.
- Independence of Observations: Each observation corresponds to an individual customer, preserving the independence of observations.
- Absence of Multicollinearity: To uphold the assumption of no multicollinearity among explanatory variables, a thorough examination of correlations between them was conducted, we are not 100% sure tho.
- No Extreme Outliers: Rigorous efforts were made to identify and eliminate extreme outliers, ensuring that the dataset remains free from any potentially influential data points.
- Linear Relationship: Assumption of a linear relationship between all explanatory variables and the logit of the response variable is acknowledged, and post-model fitting checks will be conducted to verify this relationship.
- Sufficient Sample Size: With the dataset comprising over 25000 observations, it meets the requirement for a sufficiently large sample size as stipulated by the GLM assumptions.

These considerations collectively validate the suitability of the GLM with the binomial link function for our binary categorical loan default prediction model.

8- Predictor Selection:

The initial model included all variables offered in the dataset. I did this to test which variables are significant and which variables should be removed. Before building the model I divided the unique values of each categorical variable into a column of its own, this will help me more accurately narrow down the variables to fit in the model. I used the summary function to produce the following results:

Binomial GLM of Loan Default Prediction Dataset



```
Call:
glm(formula = Default ~ ., family = binomial(link = "logit"), 
    data = default)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6938  -0.5246  -0.3712  -0.2434   3.3146  

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -6.498e-01  5.296e-02 -12.270  < 2e-16 ***
Age                           -3.943e-02  4.640e-04 -84.988  < 2e-16 ***
Income                        -8.819e-06  1.706e-07 -51.702  < 2e-16 ***
LoanAmount                     4.257e-06  9.345e-08  45.557  < 2e-16 ***
CreditScore                   -7.607e-04  4.106e-05 -18.527  < 2e-16 ***
MonthsEmployed                -9.805e-03  1.918e-04 -51.106  < 2e-16 ***
NumCreditLines                 8.866e-02  5.835e-03  15.193  < 2e-16 ***
InterestRate                   6.904e-02  1.021e-03  67.648  < 2e-16 ***
LoanTerm                       8.980e-05  3.832e-04   0.234   0.8147    
DTIRatio                       2.811e-01  2.819e-02   9.970  < 2e-16 ***
`EducationBachelor's`          1.786e-01  1.858e-02   9.612  < 2e-16 ***
`EducationHigh School`         2.567e-01  1.841e-02  13.947  < 2e-16 ***
`EducationMaster's`            4.594e-02  1.903e-02   2.414   0.0158 *  
EducationPhD                          NA         NA      NA       NA    
`EmploymentTypePart-time`      2.816e-01  1.909e-02  14.752  < 2e-16 ***
`EmploymentTypeSelf-employed`  2.362e-01  1.927e-02  12.258  < 2e-16 ***
EmploymentTypeUnemployed       4.444e-01  1.871e-02  23.754  < 2e-16 ***
MaritalStatusMarried          -2.300e-01  1.606e-02 -14.321  < 2e-16 ***
MaritalStatusSingle           -6.590e-02  1.562e-02  -4.220 2.44e-05 ***
HasMortgageYes                -1.570e-01  1.303e-02 -12.052  < 2e-16 ***
HasDependentsYes              -2.434e-01  1.305e-02 -18.647  < 2e-16 ***
LoanPurposeBusiness            4.392e-02  2.024e-02   2.170   0.0300 *  
LoanPurposeEducation          -1.843e-02  2.043e-02  -0.902   0.3668    
LoanPurposeHome               -1.947e-01  2.102e-02  -9.263  < 2e-16 ***
LoanPurposeOther              -8.880e-03  2.045e-02  -0.434   0.6641    
HasCoSignerYes                -2.710e-01  1.306e-02 -20.748  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183410  on 255346  degrees of freedom
Residual deviance: 161309  on 255322  degrees of freedom
AIC: 161359

Number of Fisher Scoring iterations: 5
```

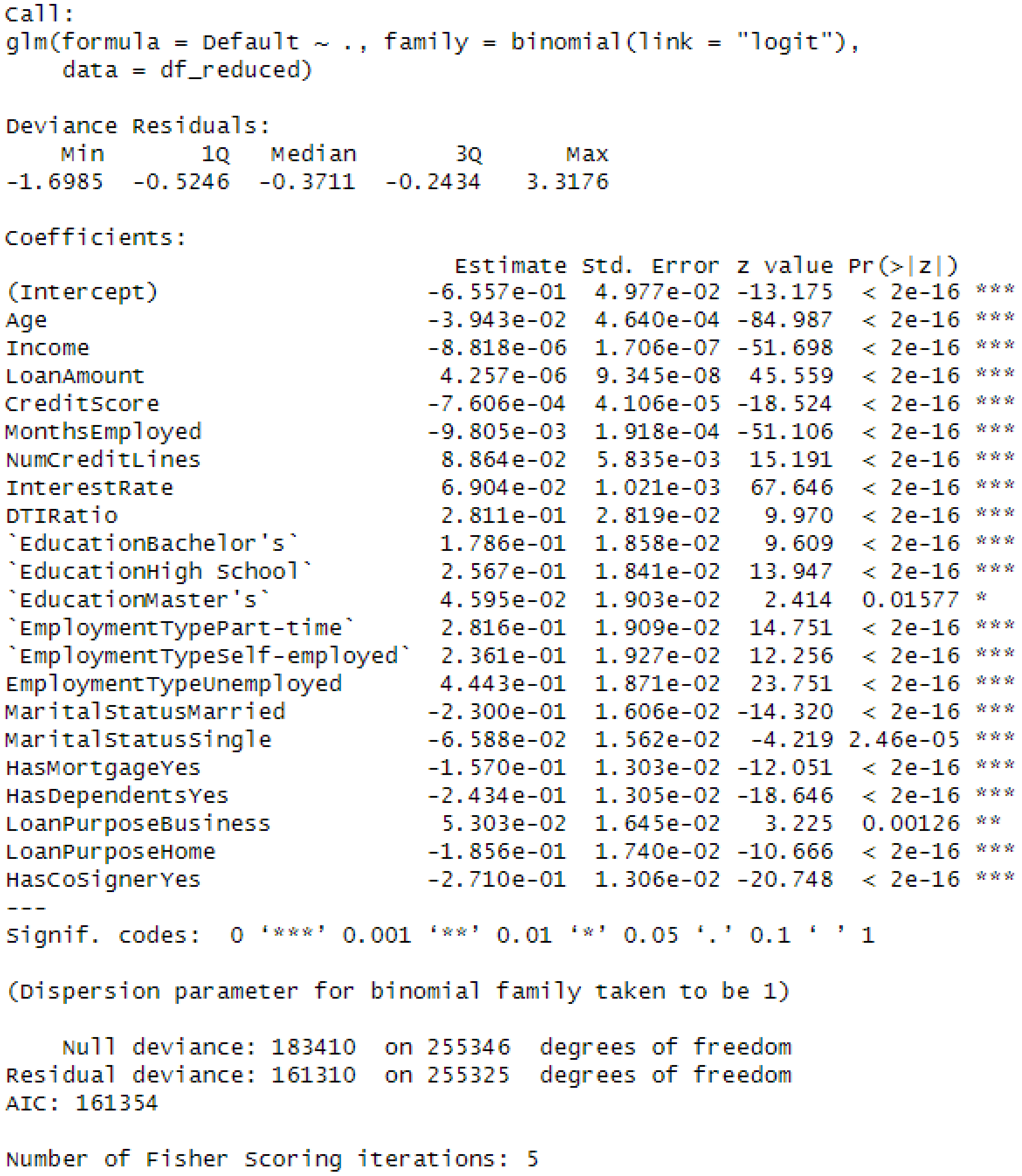
The following variables have a high p-value and are not significant to the model:

- Loan term
- Loan purpose other
- Education (PHD)

Binomial GLM of Loan Default Prediction Dataset

The initial model has an AIC score of 161359

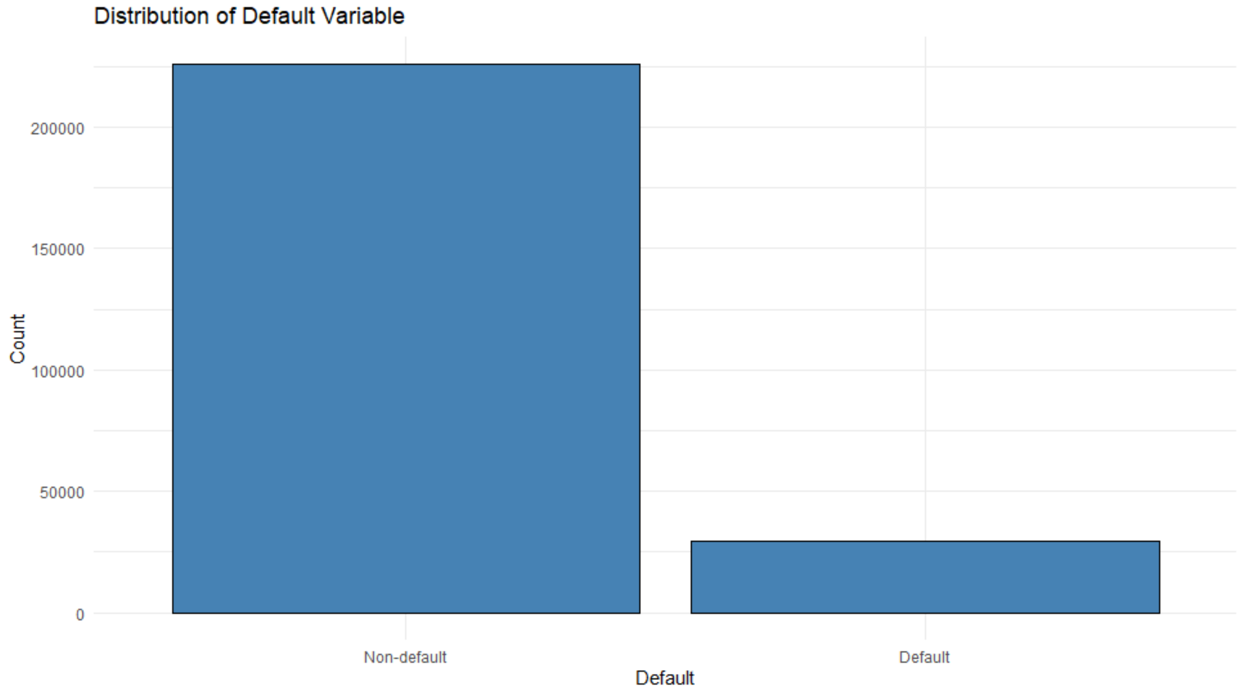
- The reduced model:



```
Call:
glm(formula = Default ~ ., family = binomial(link = "logit"), 
    data = df_reduced)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6985  -0.5246  -0.3711  -0.2434   3.3176  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -6.557e-01  4.977e-02 -13.175  < 2e-16 ***
Age                           -3.943e-02  4.640e-04 -84.987  < 2e-16 ***
Income                        -8.818e-06  1.706e-07 -51.698  < 2e-16 ***
LoanAmount                     4.257e-06  9.345e-08  45.559  < 2e-16 ***
CreditScore                   -7.606e-04  4.106e-05 -18.524  < 2e-16 ***
MonthsEmployed                -9.805e-03  1.918e-04 -51.106  < 2e-16 ***
NumCreditLines                 8.864e-02  5.835e-03  15.191  < 2e-16 ***
InterestRate                   6.904e-02  1.021e-03  67.646  < 2e-16 ***
DTIRatio                       2.811e-01  2.819e-02   9.970  < 2e-16 ***
`EducationBachelor's`          1.786e-01  1.858e-02   9.609  < 2e-16 ***
`EducationHigh School`         2.567e-01  1.841e-02  13.947  < 2e-16 ***
`EducationMaster's`            4.595e-02  1.903e-02   2.414  0.01577 *  
`EmploymentTypePart-time`      2.816e-01  1.909e-02  14.751  < 2e-16 ***
`EmploymentTypeSelf-employed`  2.361e-01  1.927e-02  12.256  < 2e-16 ***
EmploymentTypeUnemployed       4.443e-01  1.871e-02  23.751  < 2e-16 ***
MaritalStatusMarried          -2.300e-01  1.606e-02 -14.320  < 2e-16 ***
MaritalStatusSingle           -6.588e-02  1.562e-02  -4.219 2.46e-05 ***
HasMortgageYes                -1.570e-01  1.303e-02 -12.051  < 2e-16 ***
HasDependentsYes              -2.434e-01  1.305e-02 -18.646  < 2e-16 ***
LoanPurposeBusiness            5.303e-02  1.645e-02   3.225  0.00126 ** 
LoanPurposeHome               -1.856e-01  1.740e-02 -10.666  < 2e-16 ***
HasCoSignerYes                -2.710e-01  1.306e-02 -20.748  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183410  on 255346  degrees of freedom
Residual deviance: 161310  on 255325  degrees of freedom
AIC: 161354

Number of Fisher Scoring iterations: 5
```

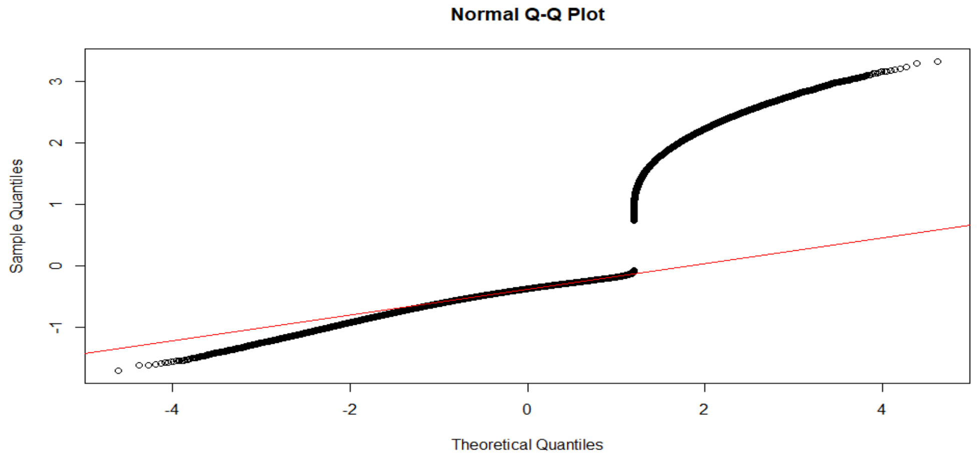
- Removing the variables with high p-values did not significantly improve the model which has an improved AIC Score. All outliers have been removed, there is not much more.

9- Model Fitting and Interpretation:



- This distribution shows that most of the loans in the datasets did not default. Using a more balanced dataset would help produce better results.

- QQ Plot of residuals



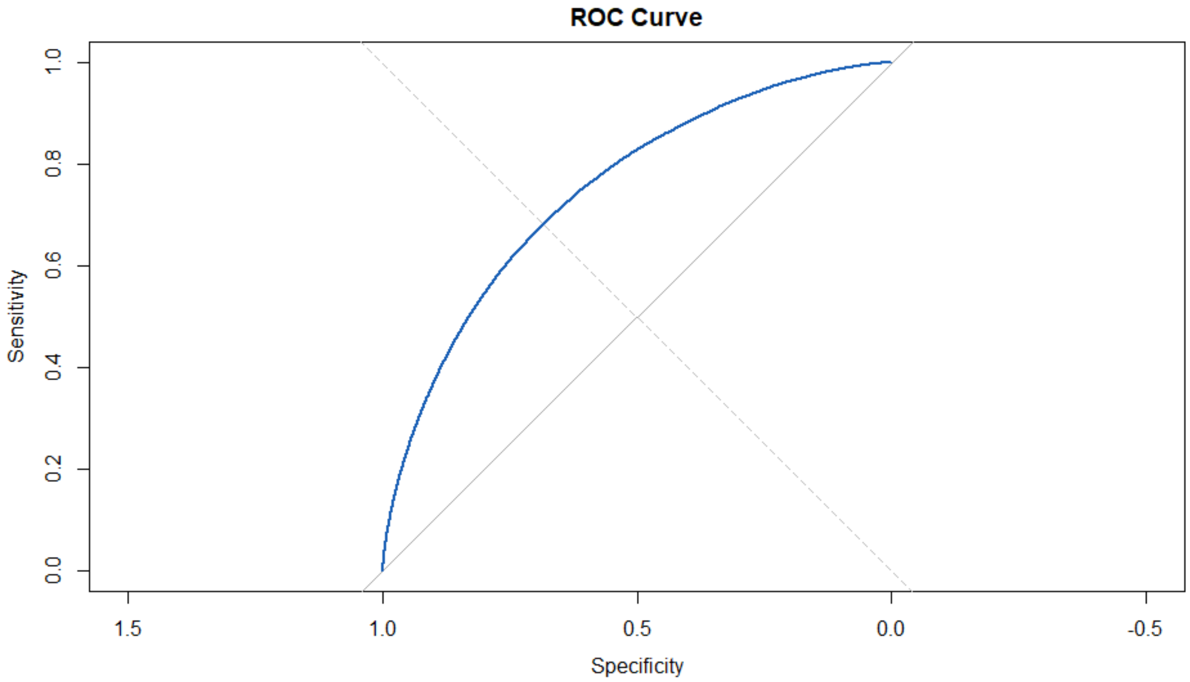
- The points in the center of the plot follow the red reference line quite closely. This indicates that the middle values of the dataset are normally distributed.

- Deviation in the Tails: The points at both ends of the plot, particularly on the right (upper) end, deviate sharply from the line. This indicates that the tails of the dataset do not conform to a normal distribution; they are heavier than expected under normality.

- Potential Outliers: The points that deviate from the line at the upper end of the Q-Q plot suggest the presence of outliers or extreme values that are larger than what would be expected in a normal distribution.

10- Result Visualization:

- I calculated the AUC Score to evaluate the model: UAC = 0.7479
- In addition to the AUC score, to evaluate the model, I plotted the ROC Curve. Unlike accuracy, which only reflects how well the model can classify, the ROC curve evaluates the model's ability to rank predictions correctly as probability scores.



Binomial GLM of Loan Default Prediction Dataset

- The ROC curve is above the diagonal line of no-discrimination (which represents a random guess) and appears to be closer to the top left corner, indicating a good level of discrimination.
- Model Performance: The ROC curve shows the trade-off between sensitivity (or true positive rate) and specificity (1 - false positive rate). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. In this case, the curve suggests that the model has a good ability to differentiate between the positive class and the negative class.
- Area Under the Curve (AUC): While the AUC value is not directly visible in the graph you provided, if we assume it is around 0.7479 as mentioned earlier, this value indicates that the model has acceptable discrimination ability. In practical terms, this means the model is considerably better than random guessing (which would have an AUC of 0.5) at distinguishing between the two classes.