Strathmore University
Institute for Mathematical Sciences

MFI 8205: Credit Risk Modelling
Default Risk Analysis: Applying a Logistic Regression Model to Australian Credit Approval Data

Kevin Alando: 171671
Sharon Okemwa: 12345
Victoriana Aluanga: 12345
Sheila Tonui: 12345

April 13, 2024

Contents

0.1 Introduction

Credit risk assessment remains a pivotal challenge for financial institutions and the advent of advanced statistical techniques has profoundly enhanced the predictive capabilities of credit risk models, facilitating more informed decision-making processes. This study aims to replicate and apply the logistic regression framework of Costa e Silva et al.'s (2020) model for consumer default risk analysis using an Australian credit approval dataset from the UCI machine learning repository.

Credit scoring systems, an integral component of credit risk management, evaluate the likelihood of default—a critical aspect influencing the approval or denial of credit applications. Historically, logistic regression has been one of the cornerstones in credit risk modeling due to its interpretability and efficacy in handling binary outcomes, such as loan default or repayment. Our replication extends this traditional approach to an Australian context, utilizing data characteristics pertinent to the Australian market and consumer behavior. The dataset used derived from the UCI repository (ID: 143), comprises credit card applications, with 690 instances and a variety of features: 6 numerical and 8 categorical.

This paper is structured to follow the methodology of the Portuguese study closely. It begins with an introduction to logistic regression and proceeds to detail the data description, variable selection, model building, and validation processes. Through this meticulous approach, we seek not only to validate the findings of the original study in a different geographical context but also to uncover the peculiarities and insights specific to the Australian financial environment, contributing to a broader understanding of consumer credit risk.

0.2 Logistic Regression

Logistic regression stands as a statistical model widely acknowledged for its applicability to binary classification problems, such as predicting whether a credit application will result in default or no default. This model's objective is to find the best fitting and most probable model to describe the relationship between the dichotomous characteristic of the target variable and a set of independent variables (predictors or features). It expresses the log-odds of the default probability as a linear combination of the independent variables, thereby enabling the assessment of the effect of each predictor on the likelihood of default, holding all other predictors constant. In the context of the Australian dataset, the logistic regression model will consider 14 independent variables, both numerical and categorical, as potential predictors of credit default.

When the target variable Y follows a Bernoulli distribution with parameter μ, the generalized linear model (GLM) employs the logit function as its canonical link function, defining a logistic regression model. In this model, when Y_i follows a Bernoulli distribution, denoted as $Y_i \sim \text{Ber}(\mu_i)$, then μ_i equals $P(Y_i = 1)$.

The binary variable `Default` represents Y, such that $Y = 1$ indicates default, and $Y = 0$ otherwise. The probability of default (PD) in the logistic regression model is determined by a set of explanatory variables X as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\beta X}}. \tag{1}$$

To estimate the regression coefficients of the GLM models, the maximum likelihood method is used. The implementation provided by the `statsmodels` library in Python is utilized. The estimates for β are obtained as a solution of a system of likelihood equations, that is solved using the Newton-Raphson algorithm, which is an iterative method that uses Fisher's information matrix. Note that several methods may be used to estimate the coefficients of a GLM model (e.g., Bayesian methods and M-estimation).

0.3 Data Description

The Australian Credit Approval dataset is a collection used for credit card application decisions, featuring 690 instances and 15 attributes, including a class attribute. The dataset contains a mix of continuous and categorical variables, with 6 numerical and 8 categorical attributes. The categorical attributes have been encoded with numerical labels to simplify their use in statistical algorithms.

Notably, the attribute names and values in this dataset have been anonymized to protect individuals' privacy. Some attributes, like A4, originally labeled with categorical data (e.g., 'p', 'g', 'gg'), have been converted to numerical labels (1, 2, 3) for analytical convenience.

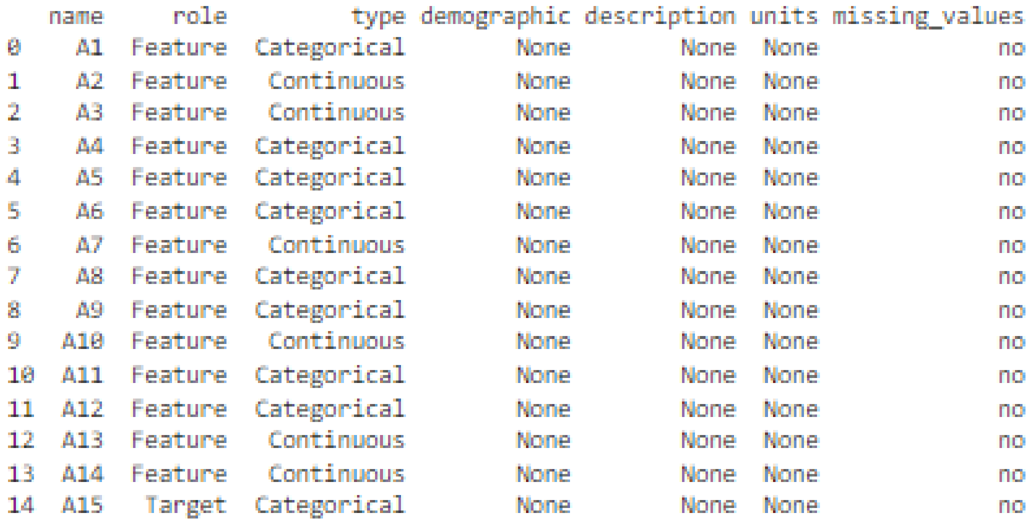
The dataset also includes some missing values, about 5% across various attributes, which have been imputed with the mode for categorical and the mean for continuous attributes to maintain integrity for statistical analysis.

The class attribute, A15, indicates the credit approval decision with two classes: positive (+) for approval and negative (-) for denial, with the dataset showing a distribution of 44.5% positives and 55.5% negatives.

Regarding preprocessing, several crucial steps were undertaken to ensure the integrity and usability of the data. Firstly, the data was reviewed for missing values and treated.

Secondly, categorical variables were encoded to facilitate their use in the logistic regression analysis. Given the anonymized and encoded nature of the variables, traditional methods such as one-hot encoding for nominal data and ordinal encoding for ordinal data were applied where appropriate. This transformation is essential for converting categorical variables into a format that can be provided as input to the logistic regression model.

Lastly, the dataset was partitioned into training and testing sets to allow for the evaluation of the model's predictive performance. The partitioning was performed in such a way as to maintain a balance between the two classes of the target variable, ensuring that the model is trained and tested on representative samples of the data.



```
   name     role         type demographic description units missing_values
0    A1  Feature  Categorical        None        None  None             no
1    A2  Feature   Continuous        None        None  None             no
2    A3  Feature   Continuous        None        None  None             no
3    A4  Feature  Categorical        None        None  None             no
4    A5  Feature  Categorical        None        None  None             no
5    A6  Feature  Categorical        None        None  None             no
6    A7  Feature   Continuous        None        None  None             no
7    A8  Feature  Categorical        None        None  None             no
8    A9  Feature  Categorical        None        None  None             no
9   A10  Feature   Continuous        None        None  None             no
10  A11  Feature  Categorical        None        None  None             no
11  A12  Feature  Categorical        None        None  None             no
12  A13  Feature   Continuous        None        None  None             no
13  A14  Feature   Continuous        None        None  None             no
14  A15   Target  Categorical        None        None  None             no
```

Figure 1: Data Description

0.4 Logistic Regression Model

0.4.1 Testing for interaction between Variables

1. **Visual Exploratory Data Analysis**

    - Heatmaps

The fist step taken was to do a visual EDA on the dataset, for this heatmaps of correlation coefficients were used to identify any obvious interactions or correlations between variables.

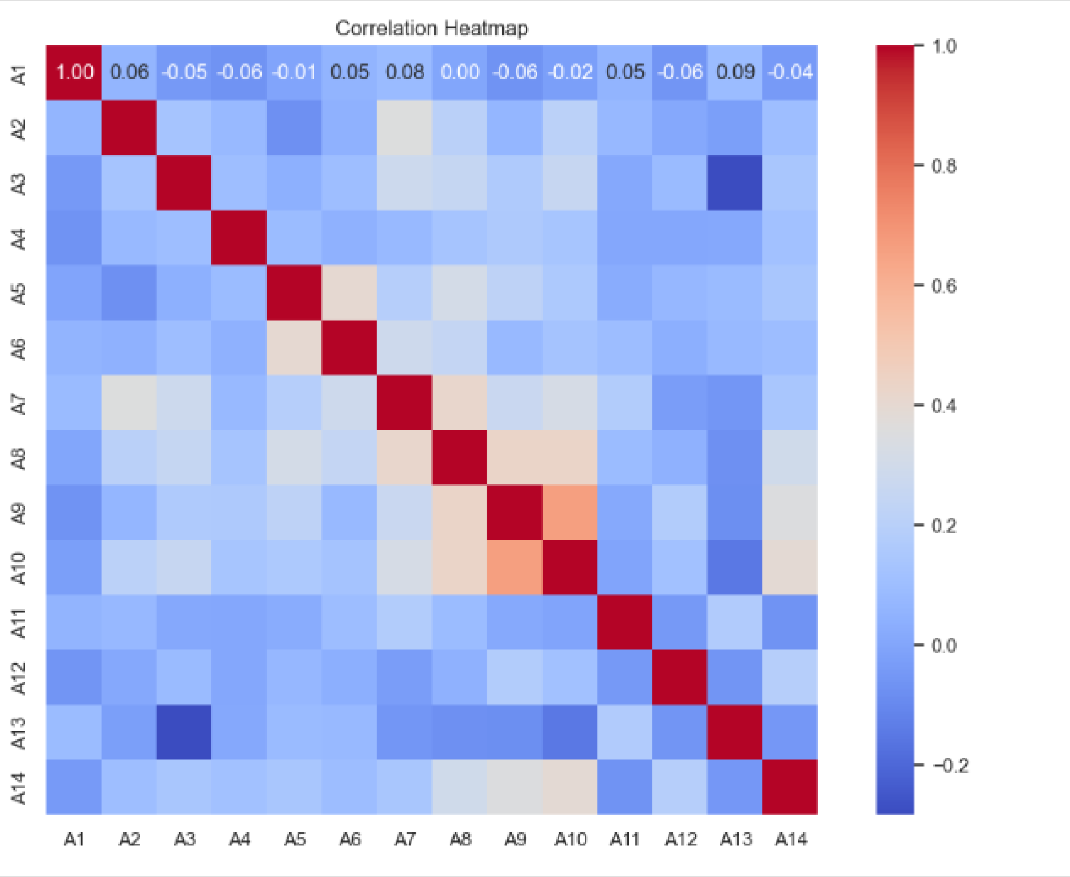


Figure 2: Correlation Heatmap

From the heatmap it appears A1 has very little linear correlation with any other variable, A9 and A10 indicate some level of positive correlation. A5 and A6 have a noticeable negative correlation with several variables.

2. **Variance Inflation Factor (VIF)**

   We proceeded to calculate the VIF for each feature to assess multicollinearity. The Variance Inflation Factor (VIF) helps to quantify the extent of correlation between one predictor and the other predictors in a model. It provides an index that measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be equal to 1.

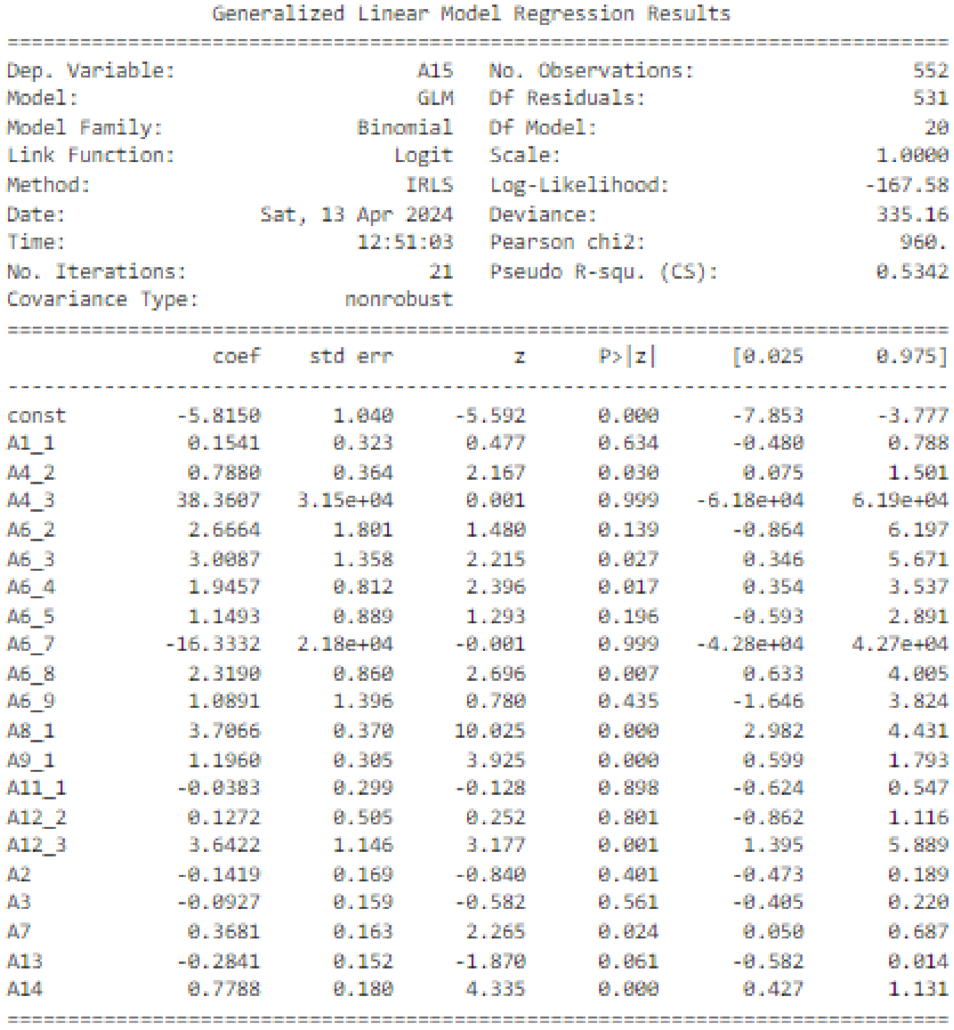
   High VIF values indicate that the feature can be linearly predicted from the other features with substantial accuracy, which might mask the true interaction effects. We were particularly interested in features, A9, A10, A5 and A6 which showed noticable correlation with Several variables. Based on the results from the VIF A5 and A6 showed multicollinearity with multiple other feature variables hence we dropped features A5 and A10 to correct for multicollinearity before building the model. After removing A5, the VIF scores for most of the remaining variables were below the threshold, which suggests that the feature adjustments were successful in mitigating multicollinearity. *(Iterations provided in the code)*

0.4.2 Building a logistic regression model

For building the logistic regression model, a simple random sample of 80% of the records was considered. First, a logistic regression model was fit to the sample of 552 records, and then this model was applied to the entire original dataset, consisted of 690 records, to predict the variable `Default`.

0.4.3 Model Estimates

The coefficients of the model were then estimated using MLE obtaining the regression results in Figure 3.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                    A15   No. Observations:                  552
Model:                            GLM   Df Residuals:                      531
Model Family:                Binomial   Df Model:                           20
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -167.58
Date:                Sat, 13 Apr 2024   Deviance:                       335.16
Time:                        12:51:03   Pearson chi2:                     960.
No. Iterations:                    21   Pseudo R-squ. (CS):             0.5342
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.8150      1.040     -5.592      0.000      -7.853      -3.777
A1_1           0.1541      0.323      0.477      0.634      -0.480       0.788
A4_2           0.7880      0.364      2.167      0.030       0.075       1.501
A4_3          38.3607   3.15e+04      0.001      0.999   -6.18e+04    6.19e+04
A6_2           2.6664      1.801      1.480      0.139      -0.864       6.197
A6_3           3.0087      1.358      2.215      0.027       0.346       5.671
A6_4           1.9457      0.812      2.396      0.017       0.354       3.537
A6_5           1.1493      0.889      1.293      0.196      -0.593       2.891
A6_7         -16.3332   2.18e+04     -0.001      0.999   -4.28e+04    4.27e+04
A6_8           2.3190      0.860      2.696      0.007       0.633       4.005
A6_9           1.0891      1.396      0.780      0.435      -1.646       3.824
A8_1           3.7066      0.370     10.025      0.000       2.982       4.431
A9_1           1.1960      0.305      3.925      0.000       0.599       1.793
A11_1         -0.0383      0.299     -0.128      0.898      -0.624       0.547
A12_2          0.1272      0.505      0.252      0.801      -0.862       1.116
A12_3          3.6422      1.146      3.177      0.001       1.395       5.889
A2            -0.1419      0.169     -0.840      0.401      -0.473       0.189
A3            -0.0927      0.159     -0.582      0.561      -0.405       0.220
A7             0.3681      0.163      2.265      0.024       0.050       0.687
A13           -0.2841      0.152     -1.870      0.061      -0.582       0.014
A14            0.7788      0.180      4.335      0.000       0.427       1.131
==============================================================================
```

Figure 3: Regression

- Interpretation
  - The `const` (intercept) has a large negative coefficient, suggesting that when all other variables are at zero, the log odds of the outcome $A15$ being 1 is negatively impacted.
  - Features such as $A6\_2$, $A6\_7$, $A9\_1$, $A10$, $A13$, and $A14$ have p-values suggesting they are statistically significant at a typical alpha level of 0.05.
  - Several categorical variables have high coefficients ($A6\_2$, $A6\_3$, $A9\_1$), which suggests they have a strong influence on the outcome when their category is present versus the baseline category.

  - Continuous variables $A2$, $A3$, $A7$, $A10$, $A13$, and $A14$ seem to have a significant effect on the outcome, with $A10$ and $A14$ having positive relationships with the dependent variable.

- Conclusion
  - The model indicates that some predictors have a statistically significant relationship with the outcome variable $A15$.
  - The presence of significant predictors implies that these factors are relevant in explaining the variance of the outcome.
  - The pseudo R^2 value suggests that the model explains a fair proportion of the variance in the outcome variable, but there is room for improvement.

0.5 Model Validation

Before using this model to estimate the probability of a client of the bank defaulting, the model has to be validated through a series of statistical tests, and the assumptions of the model have to be verified. For this we conducted the goodness of fit test and the residual analysis.

0.5.1 Goodness of fit test

To test the Goodness of fit of the model we conducted the Hosmer-Lemeshow Test to assess whether the observed event rates match expected event rates in subgroups of the model population.

The basic idea is to divide the dataset into deciles based on predicted probabilities. Within each group, the test compares the number of observed events to the number of expected events. A significant test result suggests that the model does not adequately fit the data, which could indicate model misspecification, insufficiently captured non-linearities, or other issues affecting the model's predictive power.

Hosmer-Lemeshow Test:

The Hosmer-Lemeshow test statistic, C, is computed as follows:

$$C = \sum_{g=1}^{G} \left(\frac{(O_{g1} - E_{g1})^2}{E_{g1}} + \frac{(O_{g0} - E_{g0})^2}{E_{g0}} \right) \tag{2}$$

Where:

- G is the number of groups (commonly 10).
- O_{g1} and O_{g0} are the observed number of events and non-events in group g, respectively.
- E_{g1} and E_{g0} are the expected number of events and non-events in group g, respectively, calculated from the model probabilities.

Results:

The results of the Hosmer-Lemeshow goodness-of-fit test are as follows:

- Hosmer-Lemeshow test statistic: $C = 0.5835539279945197$
- P-value: $p = 0.9997606021090188$

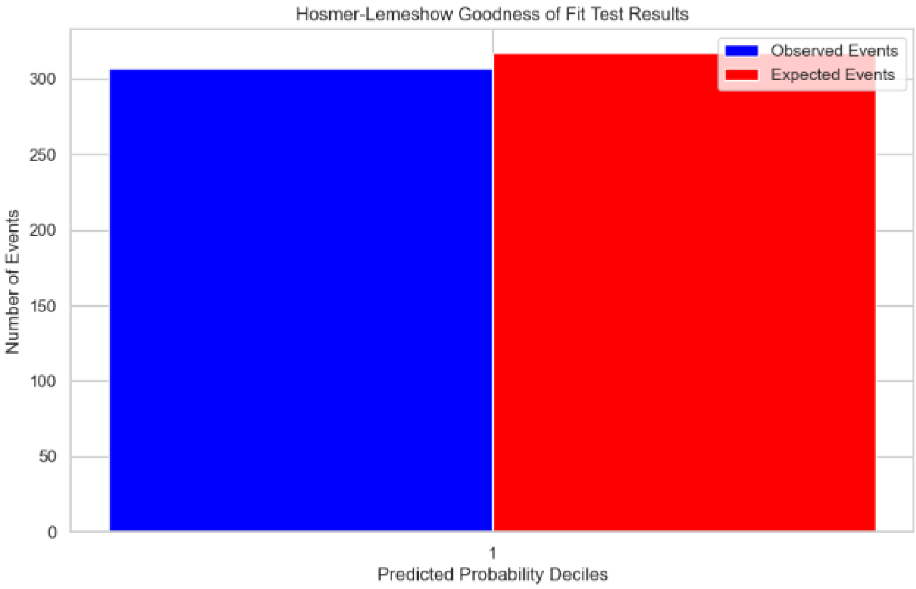


Figure 4: HL-test

The Hosmer-Lemeshow statistic value is quite low, which indicates that there is a small discrepancy between the observed and model-predicted values across the deciles of risk calculated by the model. The p-value of approximately 0.9998, is far above the common threshold, indicating that there is a very high probability that the discrepancies between the observed and predicted probabilities are due to random chance rather than a systematic error in the model.

The results suggest that the logistic regression model fits the data quite well. The low Hosmer-Lemeshow test statistic and the high p-value together indicate that there isn't a significant difference between the observed outcomes and the outcomes predicted by the model across the defined groups. This suggests that the model's assumptions and predictions are appropriate for the data.

Thus, from this goodness-of-fit test,we conclude that the model is adequate for the data at hand, and there's no evidence of poor fit that might otherwise lead us to reconsider the model's specifications.

0.5.2 Residual Analysis

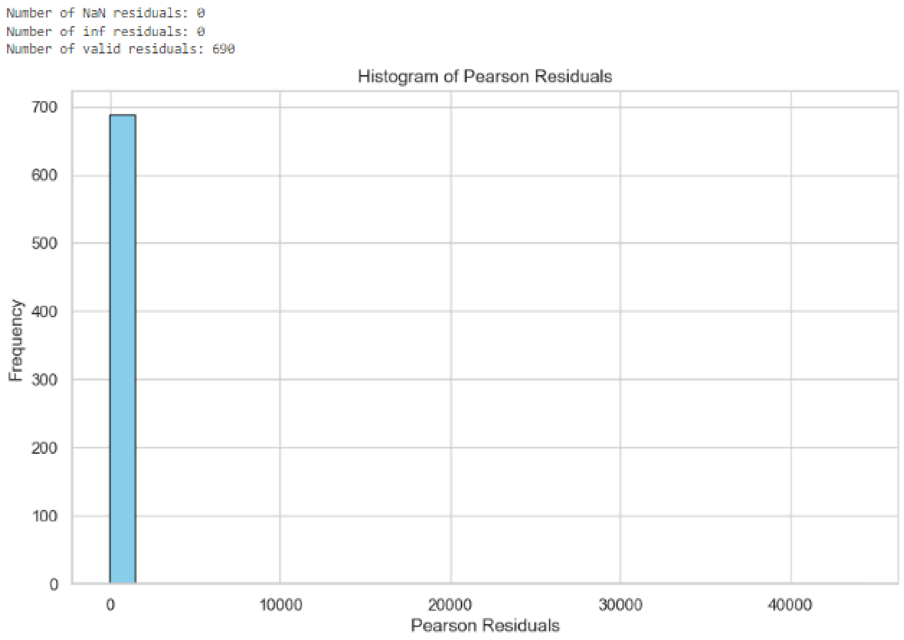
We further validated the model by studying the residuals by compute the Pearson residuals and the deviance residuals and plotting the results.

1. **Pearson residuals:**

   The Pearson residual for each observation was calculated as follows:

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{3}$$

   where y_i is the observed response (0 or 1), and $\hat{p}_i$ is the predicted probability of the response being 1. Once the residuals were calculated, we then plotted the histogram to visualize their distribution, which should ideally resemble a normal distribution if the model fits well.



```
Number of NaN residuals: 0
Number of inf residuals: 0
Number of valid residuals: 690
```

Figure 5: Pearson Residuals

From the observation, the vast majority of residuals are clustered around zero thus we also plotted the residuals on a logarithmic scale, to visualize the spread of residuals that include extreme values.

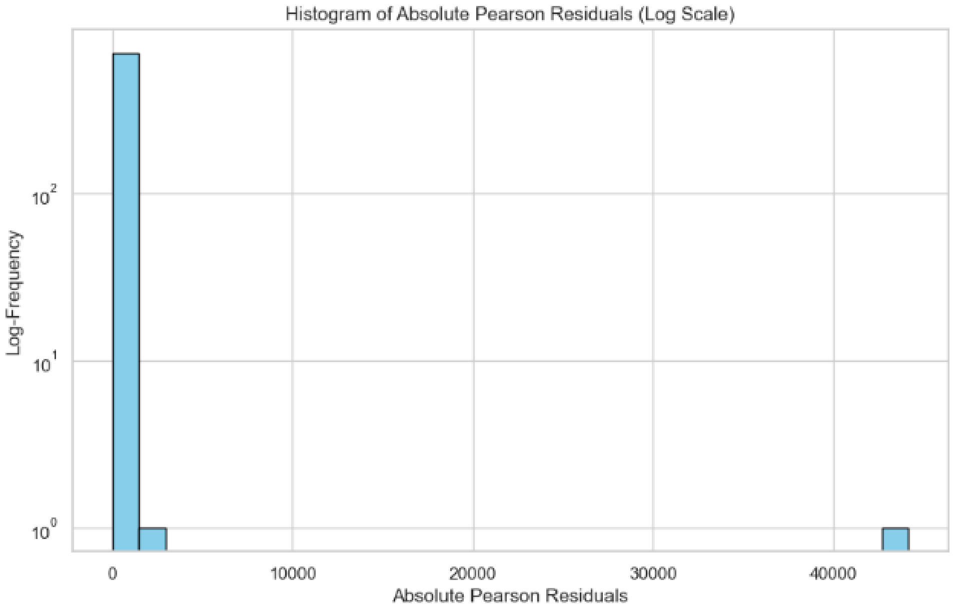


Figure 6: Absolute Pearson Residuals

From the plot, we observed:

- The majority of the residuals are concentrated in the first bin, close to zero, which indicates that for many observations, the predicted probabilities are close to the actual observed values.
- There are a few residuals with much larger magnitudes, as evidenced by the bins far to the right on the x-axis. These could be indicative of model misfit for particular observations

or outliers in the data.

Conclusion:

- **Concentration of Small Residuals:** The tall bar at the left side suggests that most residuals are small, which is a good sign indicating that the model predictions are generally close to the observed values.
- **Presence of Outliers:** The bins on the far right indicate that there are a few cases with very large residuals. These are cases where the model's predictions are very different from the actual values and could be outliers or instances where the model is not performing well.

2. **Deviance residuals:**

Deviance residuals are useful in examining the fit of the model at individual points. These residuals are particularly informative because they can indicate discrepancies between observed and predicted values in terms of the likelihood-based measure.

The Deviance residuals were computed from the model's fitted values and the observed data $i$using equation 4:

$$r_i = \text{sign}(y_i - \hat{p}_i)\sqrt{-2\left[y_i \ln\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right]} \tag{4}$$

where y_i is the observed response (0 or 1), and $\hat{p}_i$ is the predicted probability of $y_i = 1$.

Results:

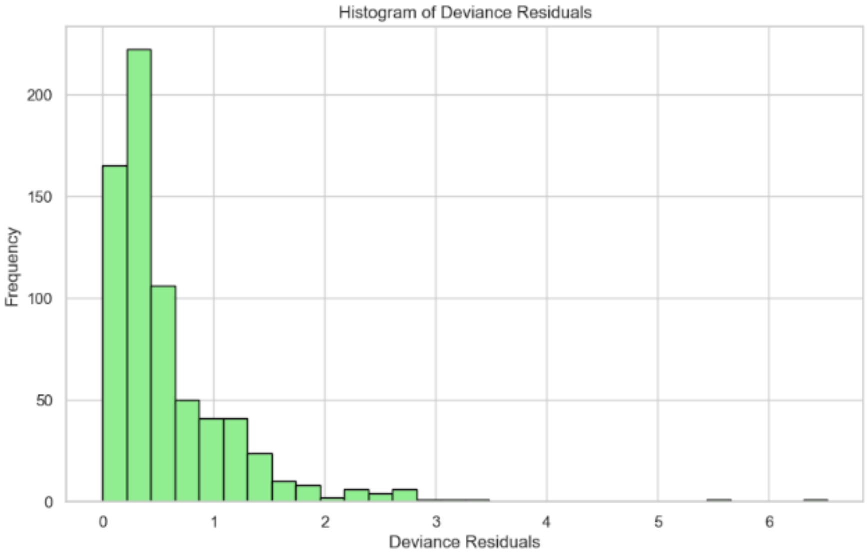


Figure 7: Deviance Residuals

From the histogram of deviance residuals we can observe that:

- **Distribution Shape:** The majority of deviance residuals are concentrated near zero. This suggests that for many observations, the model's predicted probabilities are reasonably close to the observed outcomes.
- **Frequency of Larger Residuals:** There are fewer residuals as the value increases, with very few residuals in the higher range (towards 5 and 6 on the x-axis). This tailing-off indicates that there are some observations for which the predicted probabilities deviate more substantially from the observed values.

Conclusion:

The concentration of residuals around zero indicates that the model is making accurate predictions for a significant proportion of the dataset. The long right tail, although containing fewer observations, points to the existence of some outliers or instances where the model's predictions are notably inaccurate. These may be due to inherently unpredictable variability in the data, model misspecifications, or influential points that the model fails to capture accurately.

0.6 Conclusions

In this study, we aimed to extend the logistic regression framework used by Costa e Silva et al. (2020) to the context of Australian credit approval, utilizing a dataset that includes both numerical and categorical predictors. Through rigorous model building and validation processes, our findings suggest the following:

1. **Model Fit and Predictive Ability:**

   Our logistic regression model, developed using a maximum likelihood estimation approach, revealed a collection of statistically significant predictors. Variables such as A6_2, A6_7, A9_1, A10, A13, and A14 were identified as significant, suggesting they play a notable role in the prediction of credit default. The model's constant (intercept) indicated a strong negative influence when all other variables are at zero, which impacts the log odds of a positive credit outcome. This set of variables provided a model that exhibited a reasonable goodness of fit, as indicated by a pseudo R-squared value that, while suggesting room for improvement, still accounts for a fair proportion of the variance observed.

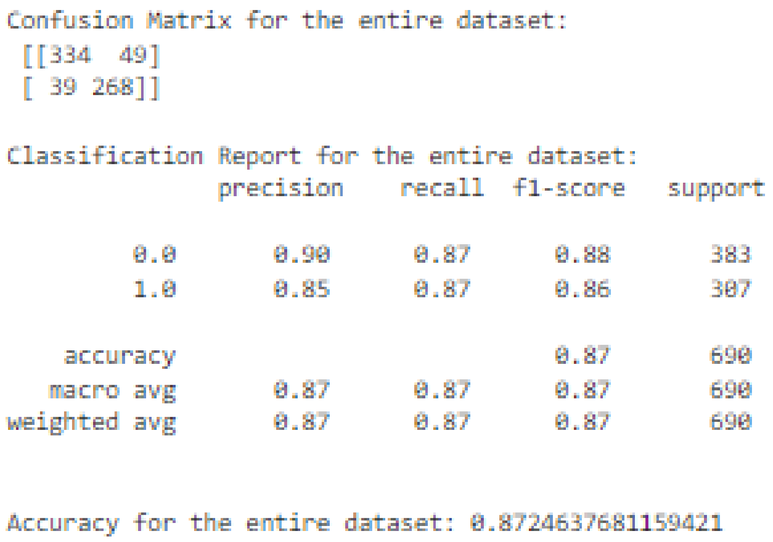
2. **Goodness of Fit:**

   The Hosmer-Lemeshow test produced a low test statistic and a high p-value, indicating no significant departure from a good fit. This result suggests that the discrepancies between the observed and model-predicted values are likely due to random variation rather than a systematic issue with the model. Hence, the model's performance is considered satisfactory according to this test.

3. **Residual Analysis:**

   Both Pearson and Deviance residuals were examined to assess the model's performance further. The histogram of Pearson residuals showed most values clustering around zero, indicating accurate predictions for the majority of the dataset. However, a small number of large residuals suggested the presence of outliers or instances where the model's predictions diverge significantly from actual outcomes. Similarly, the distribution of Deviance residuals confirmed this observation, with a concentration of residuals near zero but with a tail indicating the existence of outliers or influential data points.

Overall Conclusion:



```
Confusion Matrix for the entire dataset:
 [[334  49]
 [ 39 268]]

Classification Report for the entire dataset:
              precision    recall  f1-score   support

         0.0       0.90      0.87      0.88       383
         1.0       0.85      0.87      0.86       307

    accuracy                           0.87       690
   macro avg       0.87      0.87      0.87       690
weighted avg       0.87      0.87      0.87       690


Accuracy for the entire dataset: 0.8724637681159421
```

Figure 8: Classification Report

The models classification report also provides several key metrics:

- **Precision** for the negative class (0) is 0.90, meaning the model is correct 90% of the time when it predicts the negative class. For the positive class (1), the precision is slightly lower at 0.85.
- **Recall** (or sensitivity) for the negative class is 0.87, and for the positive class, it's higher at 0.87, indicating the model is equally good at identifying both classes.
- The **F1-score**, which is the harmonic mean of precision and recall, is 0.88 for the negative class and 0.86 for the positive class, reflecting a strong balance between precision and recall.
- The **accuracy** of the model is 0.87, meaning it correctly predicts the outcome 87% of the time across both classes.

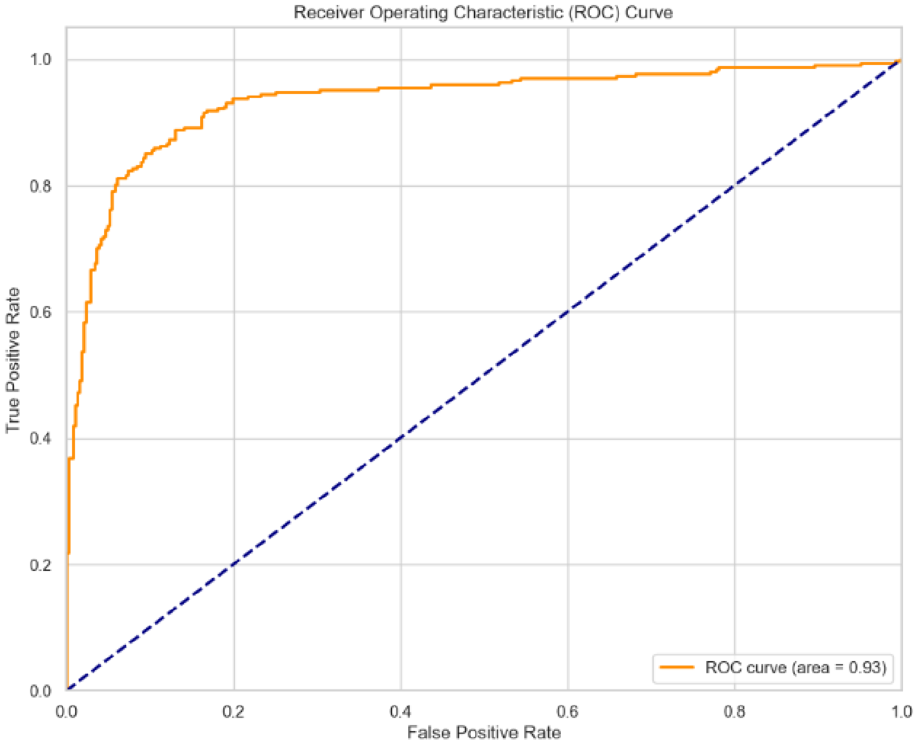


Figure 9: ROC Curve

The ROC curve is a graphical representation of a model's diagnostic ability. The area under the curve (AUC) is 0.93, which is close to 1, indicating a very good level of distinction between the positive and negative classes. This means that the model can differentiate between the classes with high confidence, as the curve stays significantly above the line of no discrimination.

Overally, the model demonstrates a significant level of accuracy in predicting credit approval outcomes, with certain predictors identified as having substantial impacts on the likelihood of default. The goodness of fit and residual analysis collectively suggest that while the model is generally well-fitted, attention should be given to the few cases where predictions are less accurate, potentially due to outliers or specific cases that the model does not capture well. These insights underline the importance of continuous model evaluation and the potential need for model refinement to better capture the nuances of credit risk within the Australian context. Based on the classification report and the ROC curve we can conclude that the model demonstrates a strong predictive ability, with a particular strength in distinguishing between the positive and negative classes. The high AUC value corroborates this, and the consistent precision, recall, and F1-scores across both classes suggest that the model is well-calibrated and reliable in its predictions.