Homework 3
Problem 3
Ayad Masud
733009045

I certify that I have personally done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment. ARM.

1 Part 1: Exploratory Data Analysis

Figure 1 is the covariance matrix for the features found in the dataset. Looking at the matrix we can see that the more correlated feature is a person's age and a person's year of employment experience. This is an obvious correlation. Another highly correlated feature is between a person's age and the length of their credit history. Some other correlated features include the loan amount correlated with the loan percent income, as well as loan status and loan interest rate. There are also negatively correlated features shown in the covariance matrix. For example, loan status and previous loan defaults are negatively correlated. As is a person's home ownership and loan status.

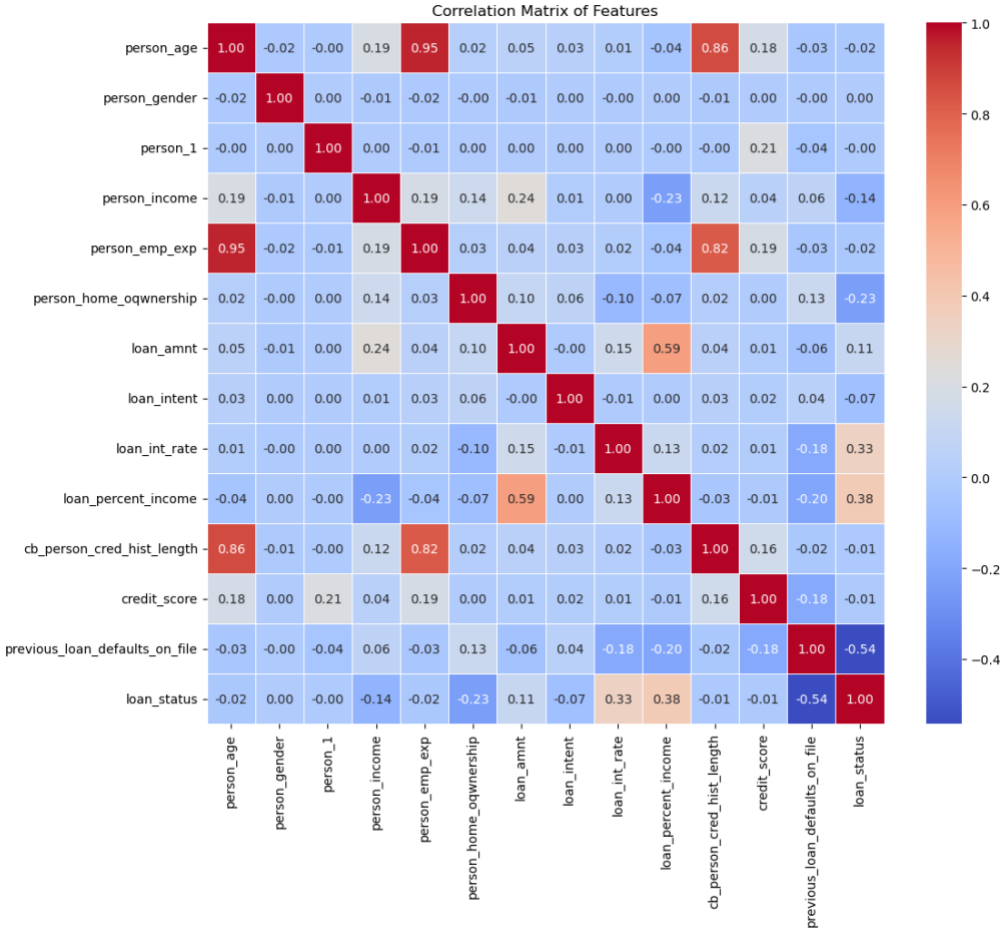

Figure 1. Covariance matrix of features in dataset.

2 Part 2: Estimating Classification Rate for K-Fold Values

Figure 2 shows the classification accuracy for different k-fold values. Each boxplot represents an iteration on that number of k-folds. The red line shows the average between the 10 runs which is around 88%

accuracy. As we iterated through the folds there is minimal variation in the classification accuracy indicating a fairly stable classification model.

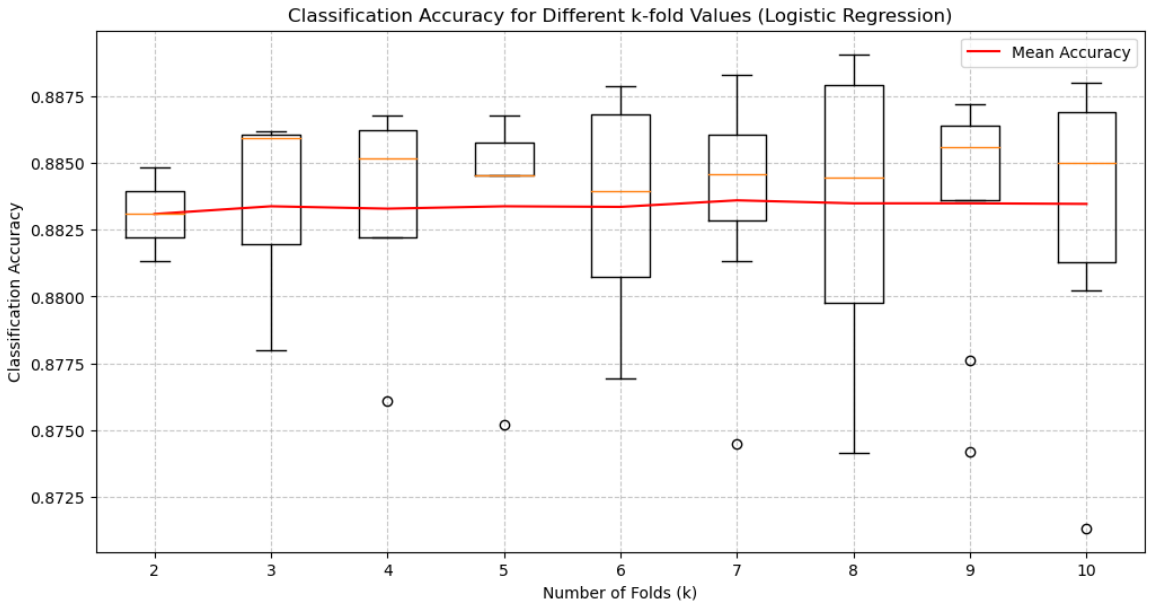

Figure 2. Classification rate on validation data.

3 Part 3: Regression Coefficients

Figure 3 shows the distribution of logistic regression coefficients across folds for different features. Loan interest rate has a positive coefficient of 4.29. This means that higher interest rates usually mean higher default risk. The loan percent income feature has a positive coefficient of 6.75, meaning that larger loan-to-income ratios are a strong indicator of default risk. The person home ownership feature has a negative coefficient of -3.56 which means that a person that owns a home has a lower default risk than a person that does not own a home. The previous loan defaults feature also has a negative coefficient of -7.59. This means that previous defaults predict future defaults and result in lower approval rates for loans.

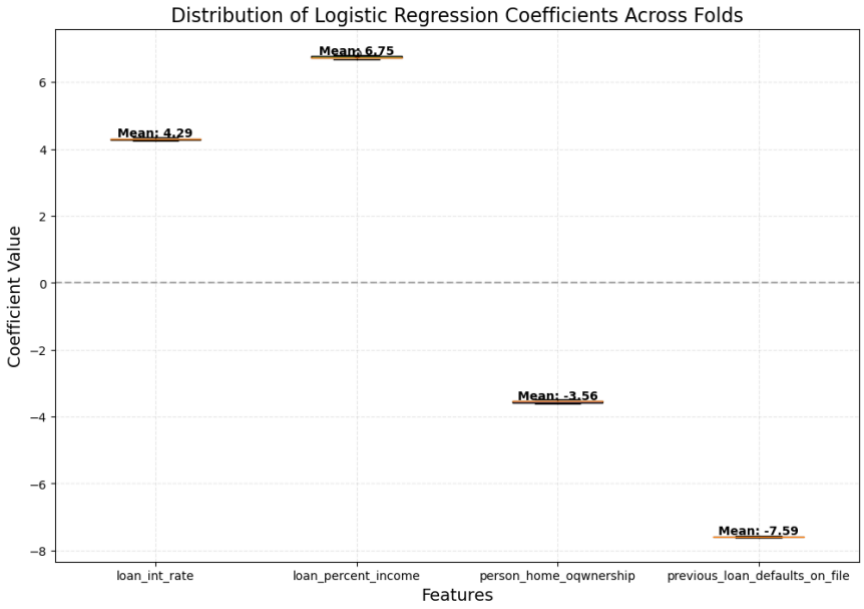

Figure 3. Distribution of logistic regression coefficients across folds.

4 Resources used to achieve this goal

Canvas: Homework template

Python Libraries: NumPy, pandas, matplotlib, seaborn, scikit-learn

5 References

scikit-learn. “Sklearn.linear_model.LogisticRegression — Scikit-Learn 0.21.2 Documentation.” *Scikit-Learn.org*, 2014, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Uday Malviya. “Bank_loan_data.” *Kaggle.com*, 2025, www.kaggle.com/datasets/udaymalviya/bank-loan-data.