Sean McLean

ALY 6020

Module 3

Mid-Week Assignment

Introduction

This paper focuses on the application of logistic regression to a dataset containing loan information and determining the likelihood of loan acceptance. The goal is to identify key variables that influence loan approval and evaluating the model's performance. A combination of methods is employed that were taught from this week's module and used to solve the three questions outlined in the assignment. The tools used in this assignment include logistic regression analysis and support vector machine regression that both are a part of binary classification.

Analysis

The dataset was first prepared by separating the target variable and the features, with the target variable for the analysis being 'personal loan,' indicating whether a loan was approved (1) or not (0). The features used for prediction were selected from the first 10 columns of the dataset. The dataset was split into training and test sets using an 80-20 split to ensure that the models could be trained on one subset and evaluated on a separate subset to assess their performance.

A logistic regression model was trained using the 'liblinear' solver, and the model was fitted to the training data which achieved an accuracy of 90.73% on the training set. This accuracy indicates that the model's performance in correctly predicting the loan approval status during training. An SVM regression with a linear kernel was also employed for the classification task with the model fitting to the same training data. The accuracy of 91.05% on the training set demonstrates a slightly better performance compared to the logistic regression model in terms of training accuracy. Both models were evaluated based on their accuracy on the training data with these metrics providing an initial understanding of how well each model performs in predicting loan approval.

Questions/Results

1. What were the three most significant variables?

The logistic regression analysis identified the following three most significant variables based on their coefficients:

1. **Income**: 0.036896
2. **CCAvg**: 0.001055
3. **Mortgage**: 0.000552

These variables are considered significant as they have the highest coefficients, indicating their strong influence on loan approval decisions. A correlation heatmap was created to provide evidence that the three variables had the most influence on the dependent variable. The other variable that had a more positive impact than other attributes was 'CD Account,' strongly pairing with the dependent variable and the variables 'security account' and 'credit card' (See Appendix A). These combinations suggest that these types of financial assets have a more positive effect on whether an individual or party is approved for a loan.

2. Of those three, which had the most negative influence on loan acceptance?

Among the variables considered, the 'age' attribute had the most negative influence on loan acceptance, with a coefficient of -0.000233. This possibly suggests that as age increases, the likelihood of loan approval slightly decreases. A bar chart of the variables shows that age is at the bottom in terms of its negative impact (See Appendix B).

3. How accurate was the model overall and what was the precision rate?

The accuracy of the logistic regression model is quite high at 91 percent, indicating a robust overall performance. The precision for loan approvals, however, is relatively low at 0.53, which suggests that while the model is generally good at predicting rejections, it is less reliable in correctly identifying approved loans. To clarify the model's accuracy, a confusion matrix was executed that shows that 882 of the 910 cells are true positives with the remaining 28 cells being false positives, totaling a 96 percent accuracy rating (See Appendix C).

Conclusion

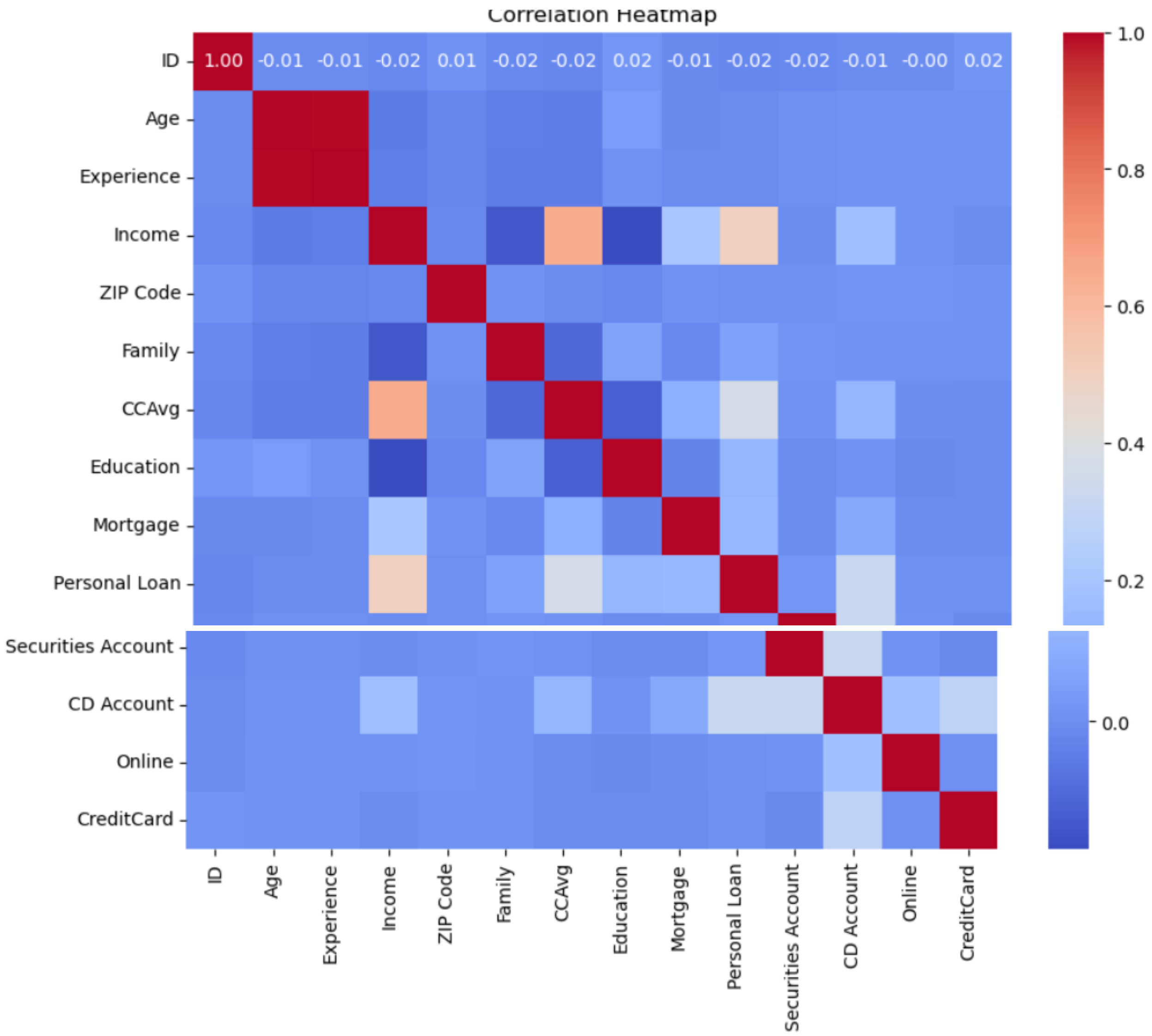
The logistic regression analysis provides valuable insights into the factors affecting loan approval. The attributes 'income,' 'CCAvg,' and 'mortgage' are the most significant variables influencing loan approval, while the 'age' attribute has the most negative impact. The model's high accuracy and the observed precision rates highlight its effectiveness and areas for improvement.

References

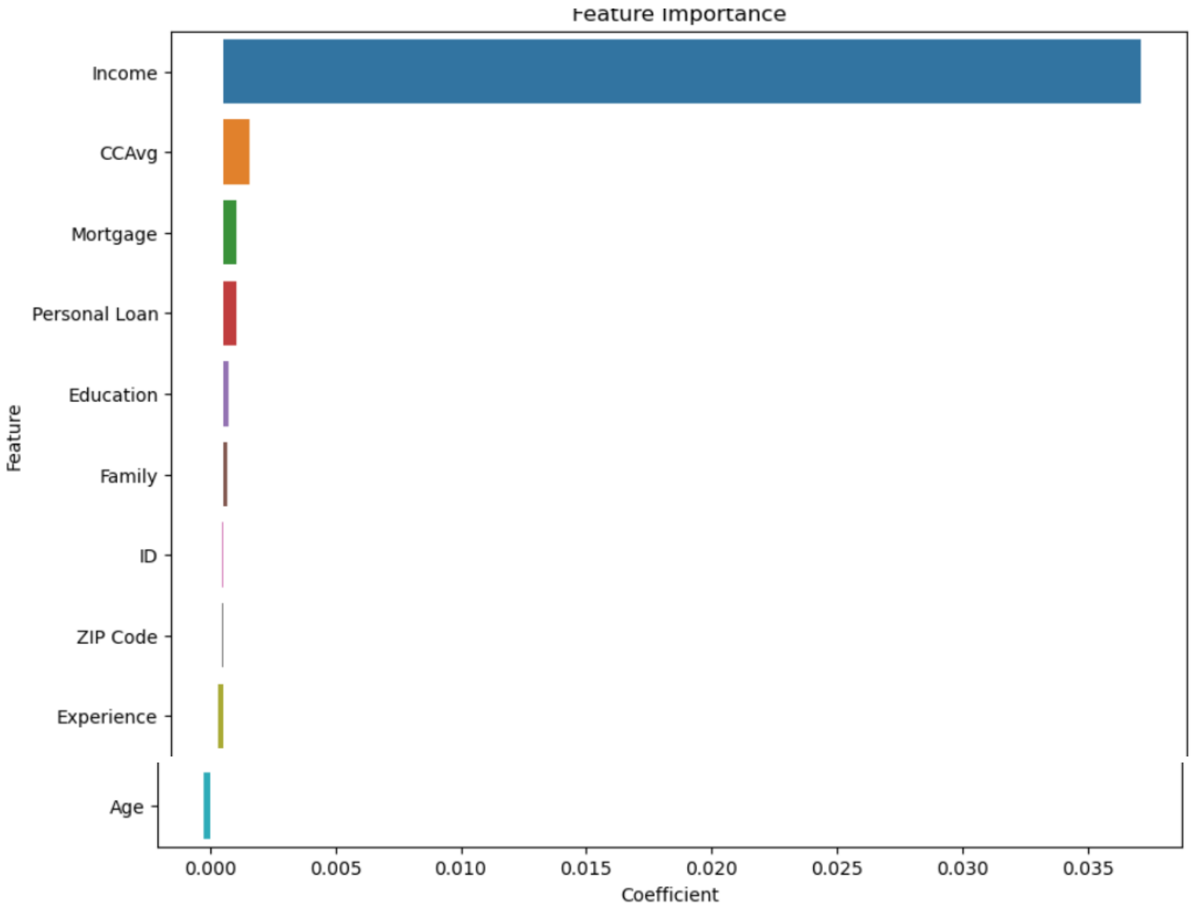
Anaconda Distribution. (2024). Anaconda (Version 4.12.3). Retrieved July 25, 2024, from https://www.anaconda.com

Appendix

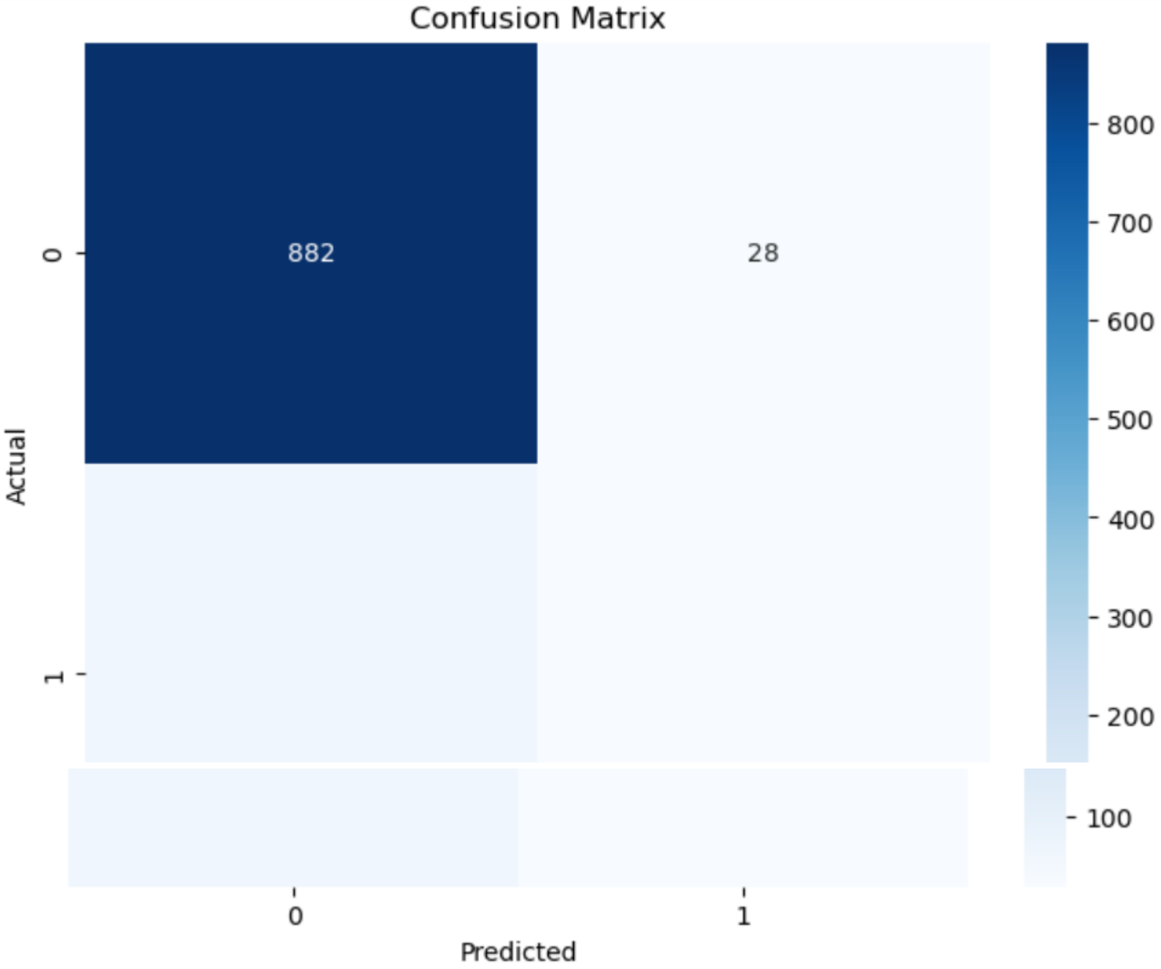
Appendix A



Appendix B


Feature Importance
Income
CCAvg
Mortgage
Personal Loan
Education
Family
ID
ZIP Code
Experience
Age
Feature
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Coefficient

Appendix C


Confusion Matrix
Actual
0
1
882
28
Predicted
0
1
800
700
600
500
400
300
200
100