Bank Loan Approval using Classification

Bhavya Tummala
University of Missouri-St. Louis

December 3, 2023

Contents

1 Introduction **1**

2 Data Set **2**
2.1 Dataset Description 2
2.2 Input Data Visualization 2
2.3 Output Data Visualization 3

3 Data Processing and Analysis **4**
3.1 Data Splitting 4
3.2 Data Normalization 4

4 Model Selection and Evaluation **4**
4.1 Logistic Regression Model 4
4.2 Multi layer model 5
4.3 Model Evaluation 5

5 Feature importance and reduction **5**

6 Challenges Faced **6**

7 Access to code **6**

8 Conclusion **6**

1 Introduction

Bank Loans will be taken by people for their needs. When they take a bank loan they provide some information to the banks for verification, So that it makes easier for the loan providers whether to approve their loan or not. The "Bank Loan Approval using Classification" AI project aims to develop a predictive model for automating the decision-making process in granting loans by banks. The project involves thorough Exploratory Data Analysis of historical loan application data to identify patterns and insights that can aid in predicting the approval or

rejection of loan applications. The primary objective is to build a predictive model to assess the probability of loan approval based on various applicant features.

2 Data Set

The dataset was obtained from Kaggle website called the "Bankloanapproval". This dataset has been made publicly available. It contains the information about the customer who has applied for personal bank loan. The dataset has details of various attributes that are related to the person which shows his income, experience, mortgages etc.

2.1 Dataset Description

This dataset contains 5000 observations of 14 variables. Here, the dependent variable is the PersonalLoan. '0' indicates that the loan is not approved and '1' indicates that the loan is approved. The input features or the 14 variables are listed as:

- ID
- Age
- Experience
- Income
- ZIP.Code
- Family
- CCAvg
- Education
- Mortgage
- Personal.Loan
- Securities.Account
- CD.Account
- Online
- CreditCard

2.2 Input Data Visualization

The histogram plot of every input features showing their maximum and minimum value as well as how they are distributed can be seen in the images given below.

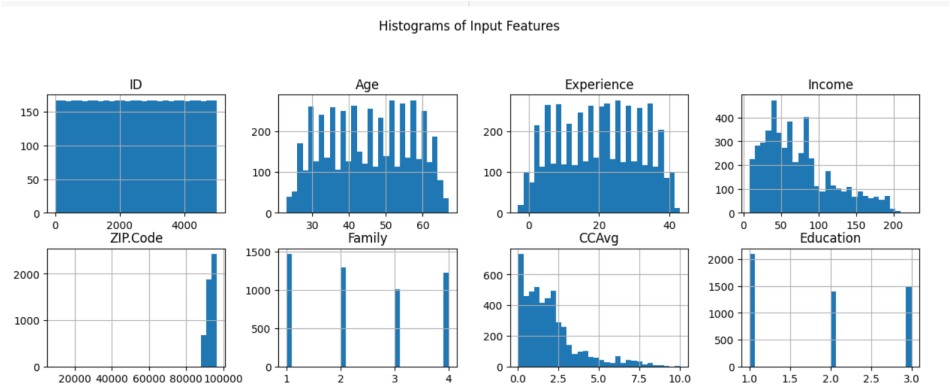

Figure 1: Input data distribution histograms(1).

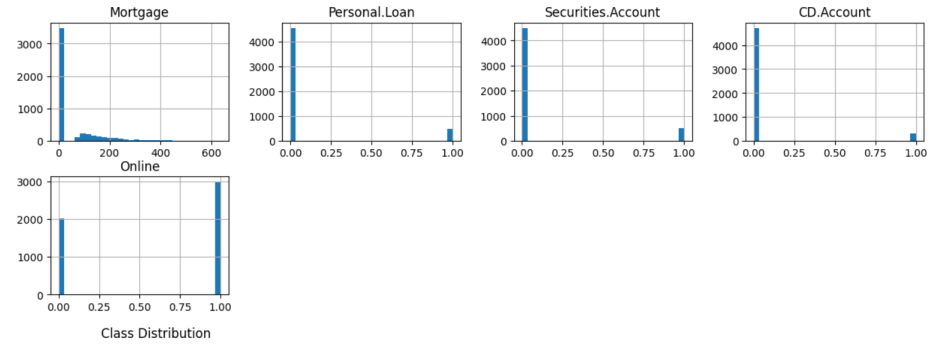

Figure 2: Input data distribution histograms(2).

2.3 Output Data Visualization

Here is the distribution of output labels.

Figure 3: Output data distribution histograms.

3 Data Processing and Analysis

3.1 Data Splitting

First of all, the data was randomly shuffled and then the dataset was split into training, testing. 80 percent of the dataset is allocated for training. And 20 percent of the dataset is allocated for testing.

3.2 Data Normalization

As we can see the data was not distributed uniformly. Therefore we need to pre-process the data with normalization technique. Min-Max normalization is used because it guarantees all features will have the exact same scale although it does not handle outliers well. The formula to calculate min-max value is as follow:

$$\frac{value - min}{min - max} \tag{1}$$

4 Model Selection and Evaluation

Logistic regression and multi layer model are used. Multi layer model is used for overfitting the model. In multi layer model, the neural network has 128-64-32 dense with relu as activation function in hidden layers, sigmoid in putput layer. The multi layer model is trained with 300 epochs. In Logistic regression model, we added 64 dense layers with relu as the activation function inn input layer and sigmoid as activation in output layer. This model is trained with 50 epochs. This model gives accuracy about 99.34 percent.

4.1 Logistic Regression Model

The performance of the logistic regression model is as shown. Training Accuracy : 83.00Validation Accuracy : 85.61Final Accuracy : 99.34

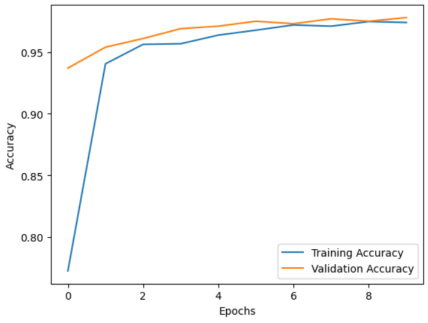

Figure 4: Plot for logistic Regression model.

4.2 Multi layer model

The performance of the multi layer model is as shown. Training Accuracy : 82.10Validation Accuracy : 84.78Final Accuracy : 99.98

4.3 Model Evaluation

Since, it is classification, I used three essential classification model metrics to evaluate:

- Precision
- Recall
- F1 score

The values for Evaluation metrics for our model is as shown below:

Table 1: Evaluation Metrics

Metric	Value
Precision	8.30%
Recall	6.70%
F1 score	7.41%

5 Feature importance and reduction

When we tried to plot the histogram of all the features with accuracy, it is a shown: From the

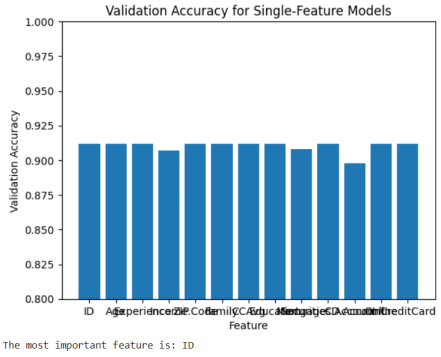

Figure 5: validation accuracy for single feature model.

above figure, we knew that ID is the most important feature. Now, we tried to remove features iteratively and calculated the accuracies. The graph indicates the accuracies for our model with iterative feature removal.

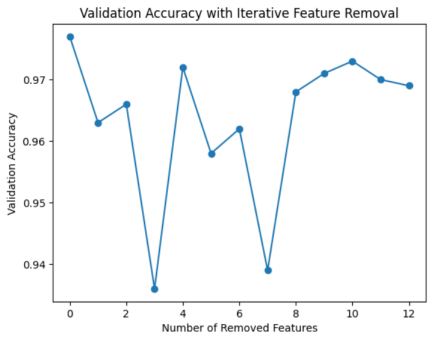

Figure 6: validation accuracy with iterative feature removal.

```
32/32 [==============================] - 0s 2ms/step - loss: 0.0828 - accuracy: 0.9690
Original Model Accuracy: 0.968999981880188
/usr/local/lib/python3.10/dist-packages/keras/src/engine/training.py:3079: UserWarning: You are saving your model as an HDF5
  saving_api.save_model(
32/32 [==============================] - 0s 2ms/step - loss: 0.0764 - accuracy: 0.9710
Reduced Model Accuracy: 0.9710000157356262
```

Figure 7: Accuracy before and after feature removal.

6 Challenges Faced

Loan approval statistics may be unbalanced if there is a notable difference between the amount of loans that are authorized and those that are denied. It is difficult to balance these datasets while maintaining their consistency.

7 Access to code

All the results here can be reproduced by using the google colab notebook. Here's the link to the notebook: Project Notebook.

8 Conclusion

In this project, I developed a neural network model to classify whether the bank loan is approved or not. Here I tested different activation functions(relu and sigmoid) and their effects on the performance of the model. I also tested the importance of each feature in the performance of the model by feature removal. This model made the task easier in classifying. While these models demonstrate promising performance in predicting loan approvals, they also highlight areas for improvement. Future enhancements could include incorporating additional data sources, exploring advanced algorithms, and addressing biases to further refine the models' accuracy and generalization capabilities.

References

Links for Overleaf:
Overleaf link to my Phase 1 : Phase 1 overleaf link.

Overleaf link to my Phase 2 : Phase 2 overleaf link.
Overleaf link to my Phase 3 : Phase 3 overleaf link.
Overleaf link to my Phase 4 : Phase 4 overleaf link.
Links for video recordings:
Video recording for Phase 1: Clip 1
Video recording for Phase 2: Clip 2
Video recording for Phase 3: Clip 3
Video recording for Phase 4: Clip 4