Predicting Loan Approvals Using Neural Networks Report

Colin Buchheit

Introduction

The primary objective of this project was to develop a machine learning algorithm using a neural network to predict whether a customer would be approved for a personal loan. This binary classification task utilized a dataset containing demographic and financial features of 5,000 customers, with the target variable being Personal Loan (1 for approved, 0 for not approved). The project involved multiple stages, including data cleaning, visualization, neural network design, and model evaluation. This report details the steps taken to preprocess the data, the architecture of the neural network, and the results achieved, along with a discussion of challenges faced and areas for improvement.

Data Preparation

The dataset consisted of 14 columns, including demographic attributes such as Age, Experience, and Education, along with financial metrics like Income, CCAvg (average monthly credit card spending), and Mortgage. To prepare the data for the neural network, two columns—ID and ZIP Code—were removed as they were deemed irrelevant to the prediction task. The ID column was a unique identifier, and ZIP Code did not provide meaningful insight into loan approval decisions.

Next, the data was inspected for missing values, and no null entries were found. The dataset was then preprocessed using StandardScaler to normalize numerical features, including Age, Experience, Income, CCAvg, and Mortgage. This scaling ensured that the features contributed equally during model training. After preprocessing, the data was split into training and testing subsets, with 80% of the dataset allocated for training and 20% reserved for testing.

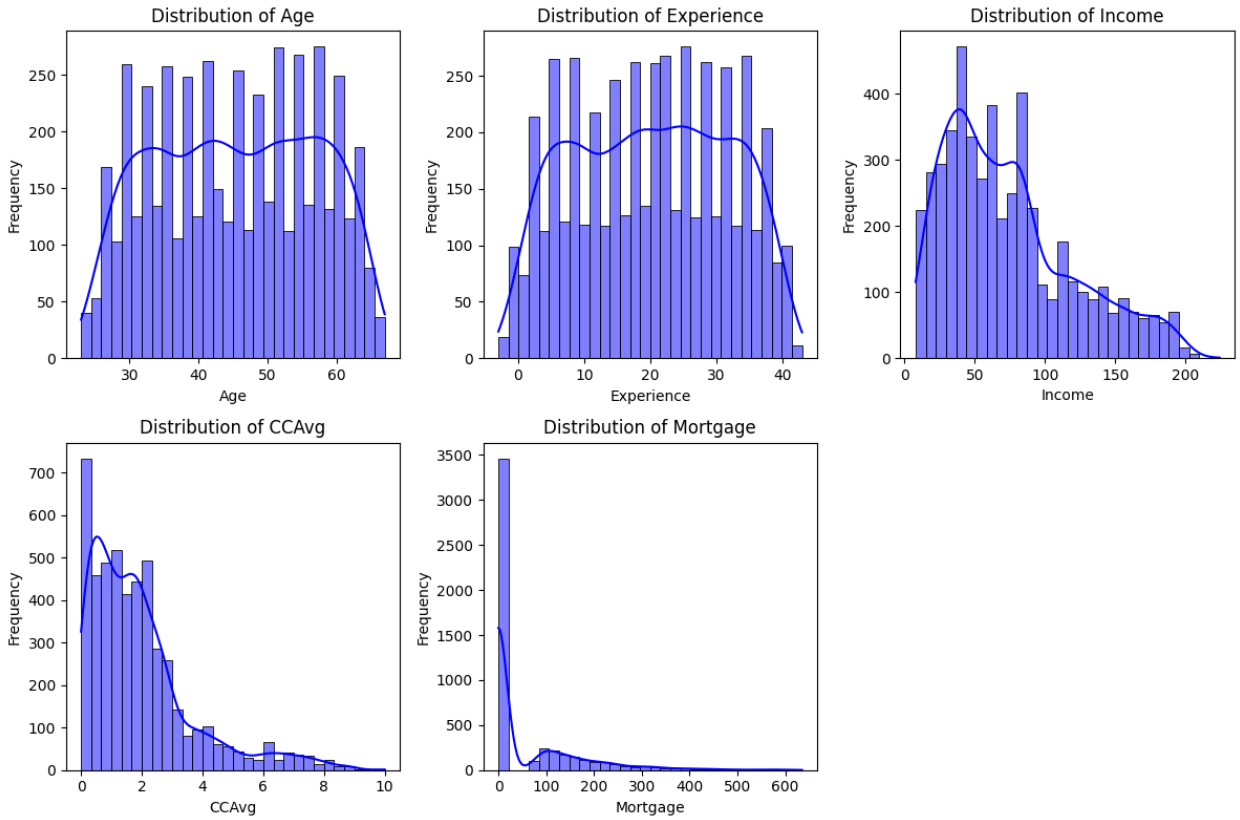
```
   ID  Age  Experience  Income  ZIP Code  Family  CCAvg  Education  Mortgage  \
0   1   25           1      49     91107       4    1.6          1         0
1   2   45          19      34     90089       3    1.5          1         0
2   3   39          15      11     94720       1    1.0          1         0
3   4   35           9     100     94112       1    2.7          2         0
4   5   35           8      45     91330       4    1.0          2         0

   Personal Loan  Securities Account  CD Account  Online  CreditCard
0              0                   1           0       0           0
1              0                   1           0       0           0
2              0                   0           0       0           0
3              0                   0           0       0           0
4              0                   0           0       0           1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  5000 non-null   int64
 1   Age                 5000 non-null   int64
 2   Experience          5000 non-null   int64
 3   Income              5000 non-null   int64
 4   ZIP Code            5000 non-null   int64
 5   Family              5000 non-null   int64
 6   CCAvg               5000 non-null   float64
...
CD Account           0
Online               0
CreditCard           0
dtype: int64
```

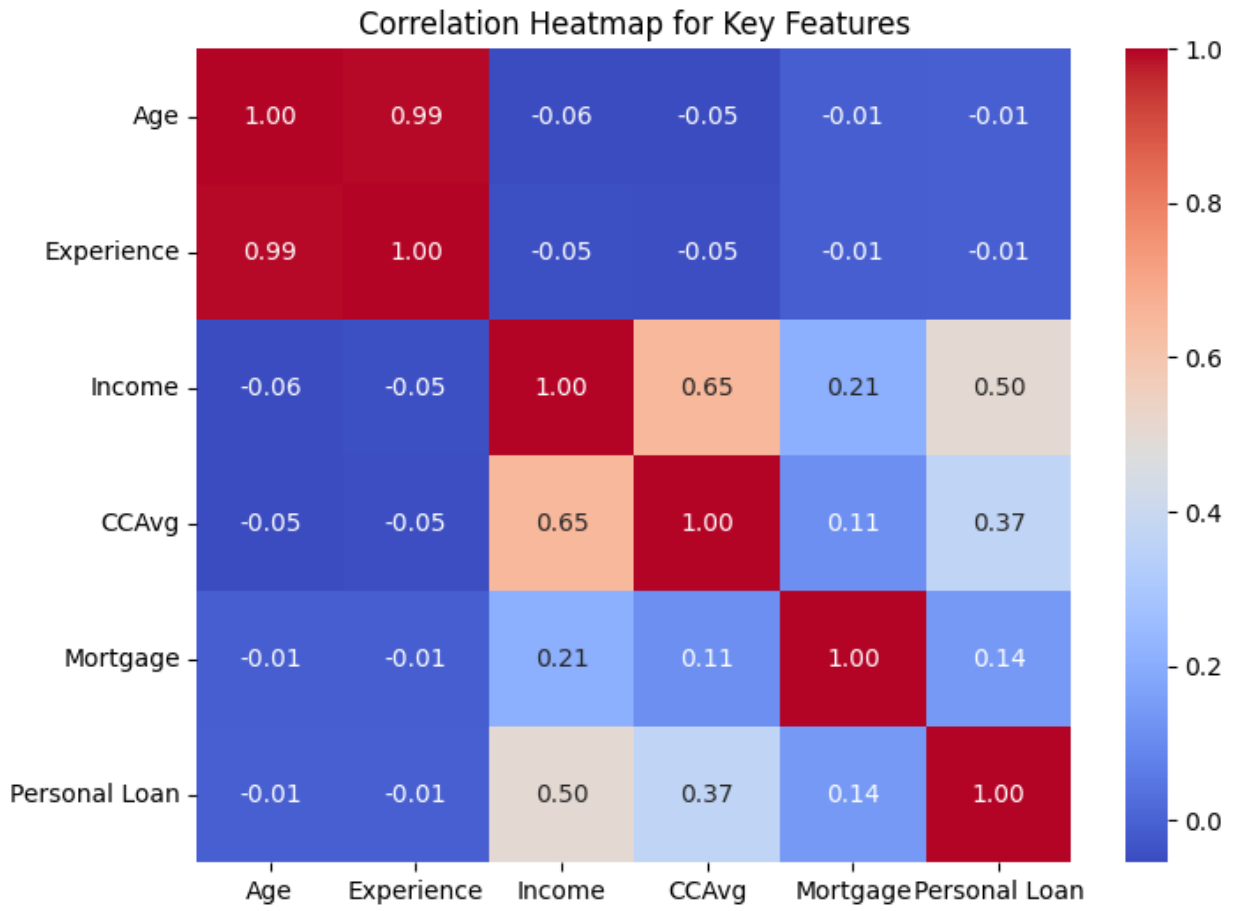
Data Visualization

To gain insights into the dataset, several visualizations were generated. The distributions of key numerical features, such as Age, Experience, Income, CCAvg, and Mortgage, were analyzed using histograms. These visualizations highlighted the variation in customer profiles. For instance, Income exhibited a right-skewed distribution, indicating that most customers earned a moderate income, while a smaller subset had significantly higher earnings. Similarly, CCAvg and Mortgage revealed that many customers had minimal monthly credit card spending or no mortgage, while a smaller segment exhibited higher

values.



A correlation heatmap was also generated to examine relationships between features and the target variable. The heatmap indicated that Income, CCAvg, and Education were positively correlated with Personal Loan, suggesting that higher income, higher credit card spending, and advanced education levels were associated with a higher likelihood of loan approval.



Neural Network Design

The neural network was designed to handle the binary classification task effectively. The architecture included an input layer corresponding to the 11 features of the preprocessed data. Two hidden layers were incorporated, with the first layer containing 64 neurons and the second containing 32 neurons, both using the ReLU activation function. To prevent overfitting, dropout layers were added after each hidden layer, with a dropout rate of 20%.

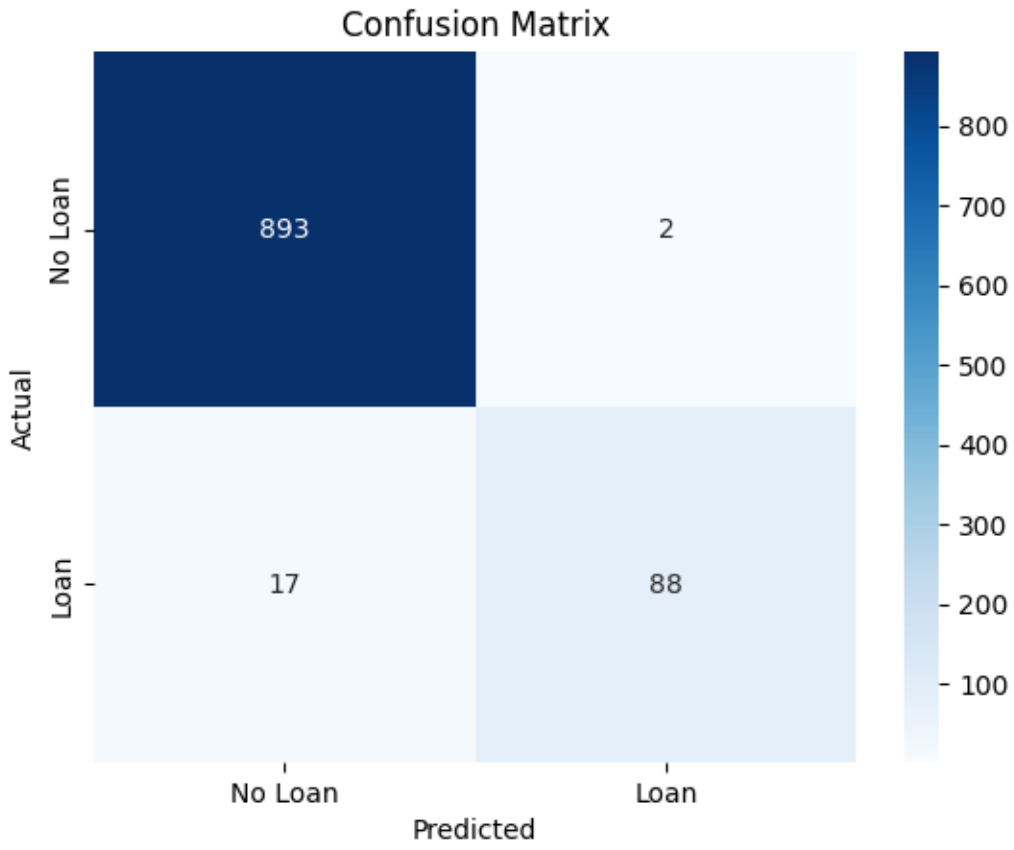
The output layer consisted of a single neuron with a sigmoid activation function to produce probabilities for binary classification. The model was compiled using the Adam optimizer, which provided efficient adaptive learning, and the binary cross-entropy loss function, which is suitable for binary classification tasks. Metrics such as accuracy were used to evaluate model performance during training and testing.

The model was trained for 50 epochs with a batch size of 32, and a validation split of 20% was applied to monitor performance on unseen data during training. Early stopping was implemented to halt training if the validation loss stopped improving, ensuring efficient resource usage and preventing overfitting.

Model Evaluation

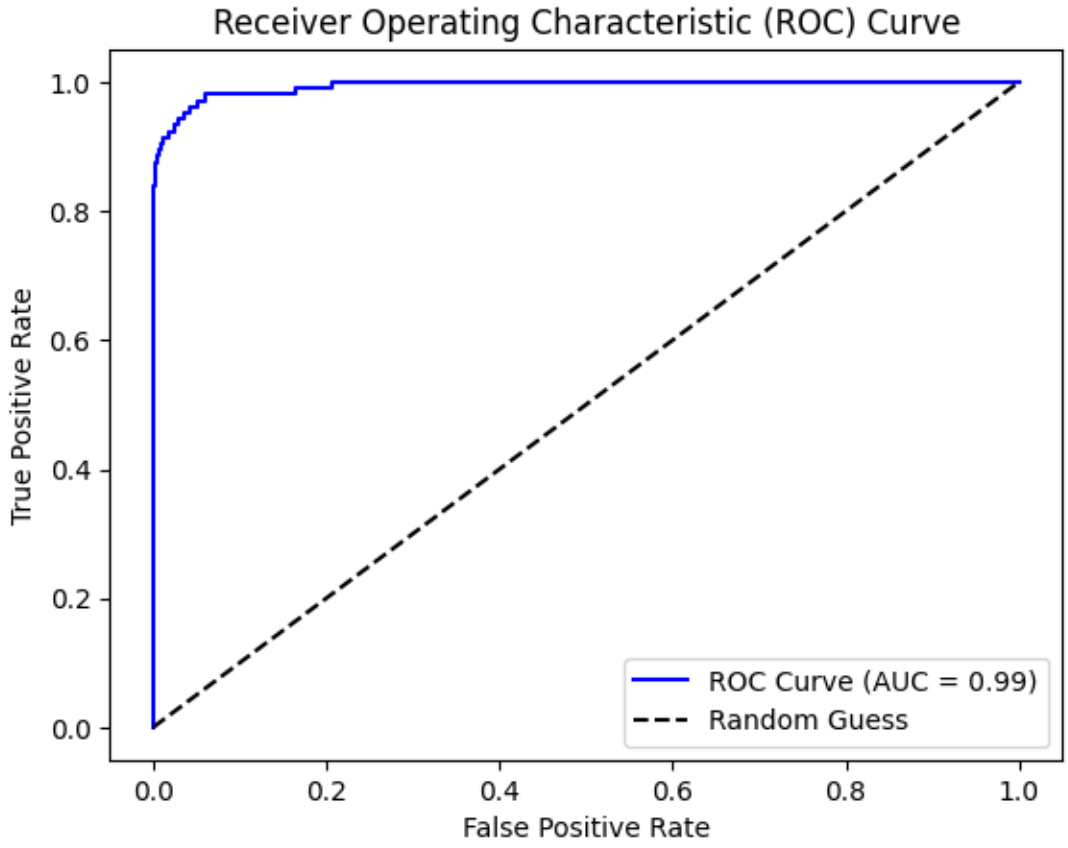
The neural network was trained using a maximum of 50 epochs; however, training stopped early after 39 epochs due to the EarlyStopping callback. The validation loss stabilized at 0.12, while the validation accuracy reached 97.3%. Early stopping prevented overfitting by halting training when further epochs would not yield significant improvements. This ensured efficient resource utilization and a well-generalized model.

The confusion matrix highlighted the model's strong predictive performance. Out of 1,000 test samples, the model correctly classified 88 loan approvals and 893 non-loan approvals, with only 19 misclassifications (17 false positives and 2 false negatives). These results translated to a high precision of 98%, indicating that most predicted loan approvals were correct. The recall of 84% showed that the model effectively identified most customers eligible for loans, while the F1 score of 90% balanced precision and recall effectively.



To further evaluate the model, a Receiver Operating Characteristic (ROC) curve was generated. Through additional research, this validation technique was implemented to assess the model's ability to rank predictions correctly. The ROC curve hugged the top-left corner of the plot, with an Area Under the Curve (AUC) score of 0.99. This high AUC score

validated the model's strong performance, indicating its ability to distinguish between approved and non-approved loans across various thresholds.



Challenges

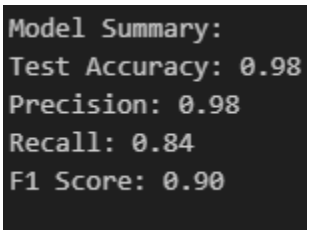
Several challenges were encountered during the project. Feature scaling was a critical step to ensure that all numerical features contributed equally during training. Without scaling, larger-valued features like Income would dominate the learning process, leading to suboptimal results. Additionally, while the dataset was slightly imbalanced, with more non-loan approvals than loan approvals, this did not significantly impact model performance. However, addressing this imbalance could further improve recall for the "Loan" class by ensuring that the model identifies more eligible loan approvals.

Improvements

While the model demonstrated excellent performance, there are still opportunities for further refinement. Hyperparameter tuning remains a key area for optimization, including adjusting the number of neurons, learning rates, or dropout rates to find the best balance

between model complexity and generalization. Although L2 regularization is already incorporated in the dense layers to mitigate overfitting, experimenting with different regularization strengths (e.g., adjusting the penalty parameter) could enhance robustness.

Lastly, refining the classification threshold from the default 0.5 offers another avenue for improvement. Depending on the business objectives, adjusting the threshold could improve recall (identifying more eligible customers for loans) or precision (reducing false positives), tailoring the model to specific organizational priorities.



Conclusion

The neural network successfully predicted loan approvals with high accuracy, precision, and recall. The project demonstrated the importance of thorough data preprocessing, careful neural network design, and performance evaluation using a variety of metrics. Additional research into the ROC curve and AUC validated the model's ability to generalize across varying thresholds, further supporting its robustness. With potential enhancements such as hyperparameter tuning and explainability techniques, the model could be refined further to provide even more robust predictions.