College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 3 Midweek Project

Submitted on:

Feb 1, 2024

Submitted to:

Prof: BEHZAD AHMADI

Submitted by:

NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The dataset is sourced from Kaggle and centers around personal loan modeling for a bank. It consists of various customer attributes, which are instrumental in understanding customer behavior and loan acceptance patterns.

Column Details
The dataset comprises the following columns:

- ID: Unique identifier for each customer.
- Age: Age of the customer.
- Experience: Number of years of professional experience.
- Income: The customer's yearly earnings, measured in thousands of dollars.
- ZIP Code: The ZIP code corresponding to the customer's home address.
- Family: The size of the customer's family or household.
- CCAvg: The average monthly expenditure on credit cards, in thousands of dollars.
- Education: The educational level of the customer.
- Mortgage: The amount of mortgage on the customer's house, measured in thousands of dollars.
- Personal Loan: Indicates whether the customer accepted a personal loan in the last campaign.
- Securities Account: This variable indicates whether the customer has a securities account with the bank.
- CD Account: This variable indicates whether the customer has a certificate of deposit (CD) account with the bank.
- Online: Indicates whether the customer uses online banking facilities.
- CreditCard: Indicates whether the customer uses a credit card issued by the bank.

The dataset consists of 5000 rows and 14 features. The dataset includes a mix of integer (int64) and floating-point (float64) data types.

The dataset provides a comprehensive statistical analysis of various attributes:

- Age: The average age is approximately 45 years, with a range of 23 to 67 years.
- Experience: Averages around 20 years, but includes some negative values, suggesting possible data entry errors.
- Income: The average monthly income is about $73.77K, showing a wide income range among individuals.
- Personal Loan: Only 9.6% of the individuals have a personal loan.

DATA CLEANING

- **Checking for the number of missing values in the dataset**
  The variables do not seem to have any null or NA values.

- **Removing duplicate rows**
  Duplicate rows in the dataset can lead to inconsistencies which may affect the accuracy of the analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis.
  There are no duplicate rows.

- **Dropping columns**
  Columns “ID”, and “Zipcode” were removed because these columns do not contribute to the analysis or the insights.

- **Handling negative values in Experience column**
  Initially, the dataset contained negative values for experience, indicating data errors. To rectify this, the absolute values of the 'Experience' were computed, effectively converting any negative numbers to positive.

- **Standardizing columns**
  'Income' column originally represented annual income, while the 'CCAvg' column denoted average monthly credit card spending. To standardize these measurements for consistent analysis, 'Income' was converted from annual to monthly figures.

EDA

Observations from the EDA:

Figure 1:

Here's an overview of the histograms for different variables -

- **Age**: The distribution of age is fairly consistent, with a slight right skew, indicating a larger number of younger customers.
- **Experience**: Similar to age, the experience histogram shows a wide range, with most values concentrated towards the middle range.
- **Income**: Income distribution is right-skewed, meaning a large number of customers fall into the lower income bracket, with fewer customers earning higher incomes.
- **CCAvg**: The average spending on credit cards is skewed to the right, indicating that most customers have lower average credit card spending.
- **Mortgage**: The majority of customers have low or no mortgage, with the distribution heavily skewed to the right.
- **Personal Loan**: This binary variable shows the count of customers who have accepted or not accepted a personal loan. The vast majority have not accepted a personal loan.
- **Online**: A majority of customers do use online banking services, as indicated by the tall bar at 1.
- **CreditCard**: Most customers don't have a credit card from the bank.

Figure 2:

- "Age" and "Experience" have a very high correlation (close to 1), which is logical since more age typically means more work experience.
- "Income" has a moderately positive correlation with "Personal Loan," suggesting that higher income levels might lead to a higher probability of accepting a personal loan.
- "CD Account" and "CCAvg" also show some positive correlation with "Personal Loan," indicating that customers with these accounts are more likely to accept a loan.

Figure 3:

- Income Density plot:
  Customers with higher income levels show a greater likelihood of accepting a personal loan. Those with lower incomes tend to decline personal loan offers. The distribution suggests that income could be a strong predictor of personal loan acceptance.
- CCAvg Density Plot:
  Customers who have higher avg credit card spending are more prone to accept personal loans. Lower average spending on credit cards is associated with declining personal loan offers.

LOGISTIC REGRESSION MODEL

The dataset has been split into two subsets: a training set, which comprises 70% of the data, and a testing set, which consists of the remaining 30%.

A **logistic regression model** is constructed. The model's accuracy is calculated, and a confusion matrix and classification report are generated to assess its performance in detail.

Accuracy: The model achieves a high accuracy of 94.9%, indicating that it correctly predicts the personal loan status (accepted or not accepted) for a large majority of the customers in the test set.

CONCLUSION

The model seems to perform reasonably well with a high accuracy and a good F1-score for class 0 (0.97). However, it has lower precision and recall for class 1 (0.85 and 0.63, respectively), indicating that it may struggle to correctly predict positive instances.

The three most significant variables based on their coefficients in the logistic regression model are CD Account, Education, and CreditCard.

The most negative variable, CreditCard, has a negative coefficient, suggesting that its presence negatively impacts the outcome.

REFERENCES

Munshi, A. N. (2022, March 22). *Logistic Regression using Python and Excel*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/02/logistic-regression-using-python-and-excel/

APPENDIX

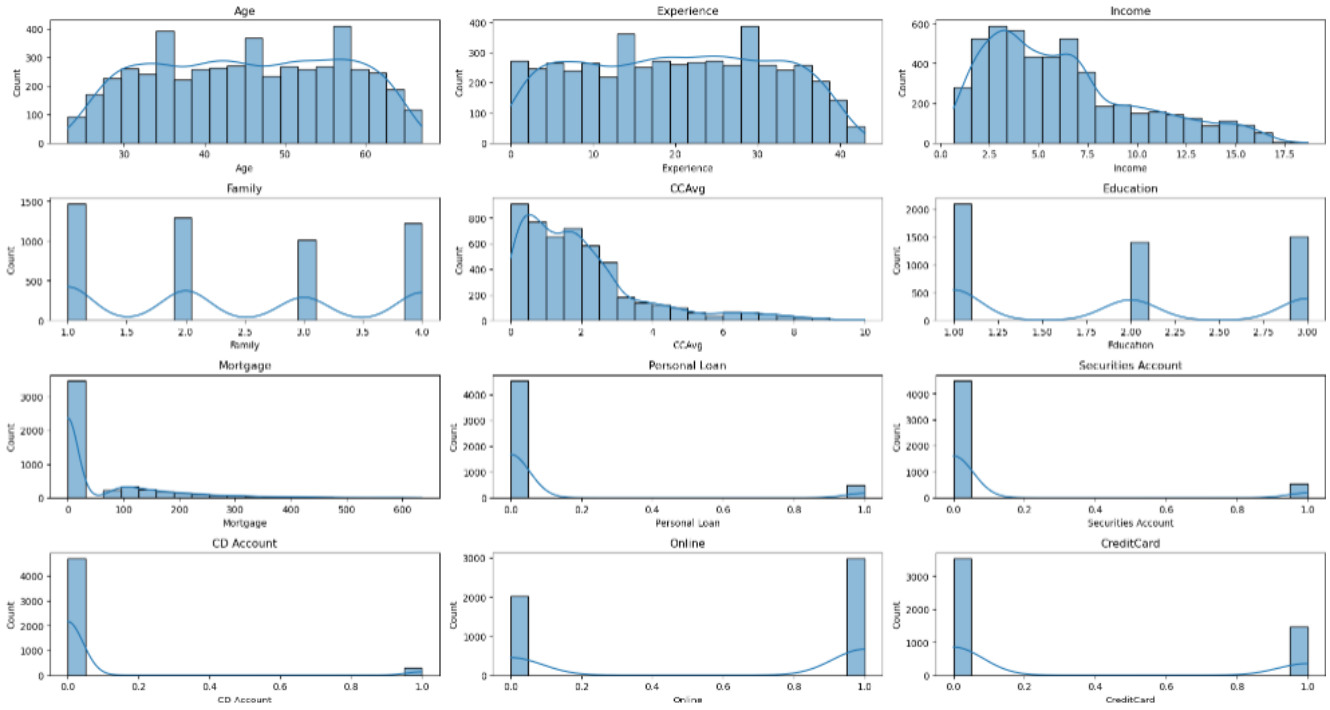

Figure 1 – Histograms for understanding distribution

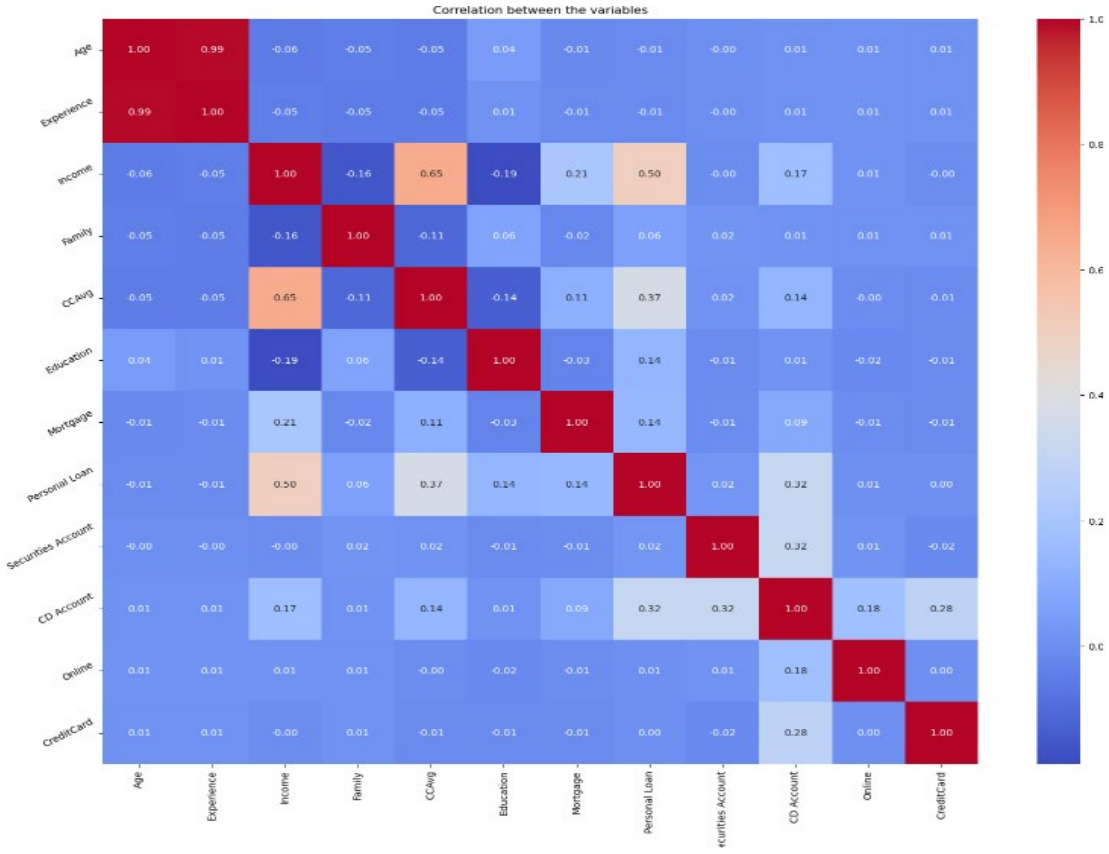

Figure 2 – Correlation matrix

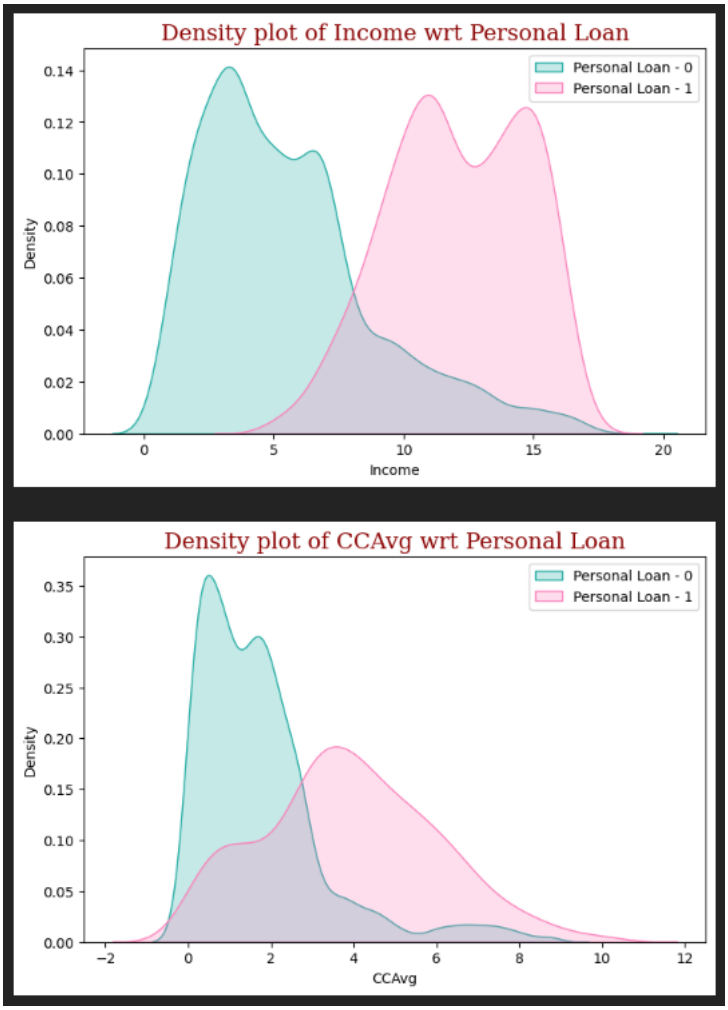

Figure 3 – Density plots

```
Accuracy: 0.949
Precision: 0.846
Confusion Matrix:
 [[1325   18]
 [  58   99]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97      1343
           1       0.85      0.63      0.72       157

    accuracy                           0.95      1500
   macro avg       0.90      0.81      0.85      1500
weighted avg       0.95      0.95      0.95      1500
```

Figure 4 – Logistic model

```
Top 3 significant variables:
             Coefficient
CD Account      3.116459
Education       1.676956
CreditCard      0.960295
Most negative variable: CreditCard
```

Figure 5 – 3 significant variables, negative influence