Report On

Personal Loan Approval Analysis and Prediction

To:

Treeleaf Technologies Pvt. Ltd.

By:

Ashutosh Chapagain
9861999836
ashutoshchapagain0@gmail.com

Date: 16 August, 2023

This report contains a brief summary on the approach, key findings and observations obtained from the analysis of the given dataset.

Approach:

1. **Data Preprocessing:** The project started with loading and understanding the dataset. Missing values were identified and handled appropriately. Categorical variables were encoded using one-hot encoding, and numeric features were scaled using the StandardScaler to prepare the data for modeling.
2. **Exploratory Data Analysis (EDA):** EDA involved visualizing the distribution of features, analyzing the relationship between features and the target variable, and exploring patterns and insights in the data.
3. **Model Selection and Training:** Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost models were trained on the preprocessed data. The training process involved feature scaling, fitting models to the training data, making predictions, and evaluating model performance.
4. **Model Evaluation and Tuning:** Model accuracy and classification reports were used to evaluate the performance of each model. Hyperparameter tuning was performed for Logistic Regression and SVM using GridSearchCV to optimize their performance.

Key Findings and Observations:

1. There were 31.92% missing values in Gender column, 1.34% in income, 23.78% in Home Ownership and 0.8% in Online column.
2. Missing values in Income and Online column were handled swiftly using the median and mode values respectively.
3. Gender and Home Ownership had large number of missing values. Since the distribution of gender was even between male and female. So, the missing values in Gender column was filled randomly with either male or female.
4. Home Ownership was checked to see if it had any correlation with income of an individual. Since no significant correlation was found, its null values were also filled randomly based on the percentage of each type of home ownership's frequency of occurrence.
5. Rows with age less than 18 and more than 97 were removed.

6. Negative experience values were converted into positive values assuming the negative signs were data entry mistakes.
7. Analyzing the target variable based on each feature, the following observations were made:

   i) Home Ownership does not affect the loan approval by much.

   ii) People with higher education tend to accept the loan more. People with bachelor's degree have lower rate of personal loan acceptance at around 4.34% while people with Master's degree and Professional Degree have higher rate of loan acceptance at around 12% and 13% .

   iii) Credit Card usage does not affect the loan approval by much.
   iv) Having a securities account with the bank slighly affects the acceptance of personal loan by about 2%.

   v) Having a Certificate of Deposit Account (CD Account) with the bank greatly affects the personal loan acceptance rate by around 38%.

   v) Usage of internet banking facilities does not affect the loan acceptance rate by much.

   vi) Age, Experience and ZIP Code do not affect the rate of personal loan approval much.

   vii) People with higher income have a higher rate of personal loan acceptance.

   viii) People with 1 and 2 family members have around 7-8% loan acceptance rate while 3 and 4 family members have around 10% and 12% of personal loan acceptance rate.

   ix) People with low CCAvg, upto around 3000 monthly, have low loan accpetance rate. But the acceptance rate from 3000 and above increases significantly.

   x) People with low mortgage have low loan acceptance rate and vice versa.
8. Standard Scaling, Robust Scaling and MinMax scaling all gave similar and slightly better result compared to Normalization in Logistic Regression.
9. Result of logistic Regression:

Accuracy: 0.9509

```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       919
           1       0.75      0.55      0.64        78

    accuracy                           0.95       997
   macro avg       0.86      0.77      0.81       997
weighted avg       0.95      0.95      0.95       997
```

10.Result of Random Forest:

Accuracy: 0.9880

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       919
           1       0.99      0.86      0.92        78

    accuracy                           0.99       997
   macro avg       0.99      0.93      0.96       997
weighted avg       0.99      0.99      0.99       997
```

11. Result of SVM:

SVM Model - Accuracy: 0.9719

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       919
           1       0.96      0.67      0.79        78

    accuracy                           0.97       997
   macro avg       0.97      0.83      0.89       997
weighted avg       0.97      0.97      0.97       997
```

12. Result of XGBoost:

XGBoost Model - Accuracy: 0.9920

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       919
           1       0.96      0.94      0.95        78

    accuracy                           0.99       997
   macro avg       0.98      0.97      0.97       997
weighted avg       0.99      0.99      0.99       997
```

The best metrics were given by XGBoost.

Conclusion:

The project successfully tackled the problem of predicting loan approval using machine learning. By employing data preprocessing, and exploring multiple models, valuable insights were gained into the dataset's patterns and predictive capabilities.