EAS 503: Final Project Report

Bank Loan approval classification project

Group **#66**: Gaurav Kumar Jha, Abhinav Nehra, Kumar Utkarsh, Monika Jangam Prabhudev

Problem Statement:

Securing a personal bank loan can be challenging and time consuming. Conventional methods heavily rely on the credit score of an individual, potentially cutting off the deserving ones. This results in limited financial access for borrowers.

Objective:

Leveraging the techniques of machine learning learned so far, we create a predictive model to predict the approval. The model will apply a holistic approach for loan approval taking into additional financial details other than credit score only.

Why use a machine learning model?

- A model trained on more data will have a better capability to make informed decisions.
- Improve fairness and inclusivity of the process.
- Better approach to unlock financial opportunities for a wider range of individuals.
- Foster economic growth

Data:

We have picked the data from Kaggle. The data comprises of key financial and personal factors such as age, income, credit card usage, family size, and other banking details.

Analysis:

1. Reading the CSV data from into python, parsing it using python control structures.
2. Loading the data into normalized database.
3. Using joins to fetch data from all the tables and loading into pandas.
4. Data Cleaning
   - Removed unwanted columns such as zip code and person_id.
   - Removed row items with negative value in year of experience, as it cannot be less than zero.
   - Checked for missing data, outliers.
5. Exploratory Data Analysis
   - Correlation plot.
   - We found that the data was unbalanced with 90% and 10% cases of No and Yes respectively.
   - Analyzing the relationship between the descriptors and target variables using different plots.
6. To resolve data imbalance, we performed SMOTE analysis, which basically is oversampling of small sample size dataset.
7. Trained 2 models, Logistic Regression and Random Forest with and without SMOTE analysis.

Findings:

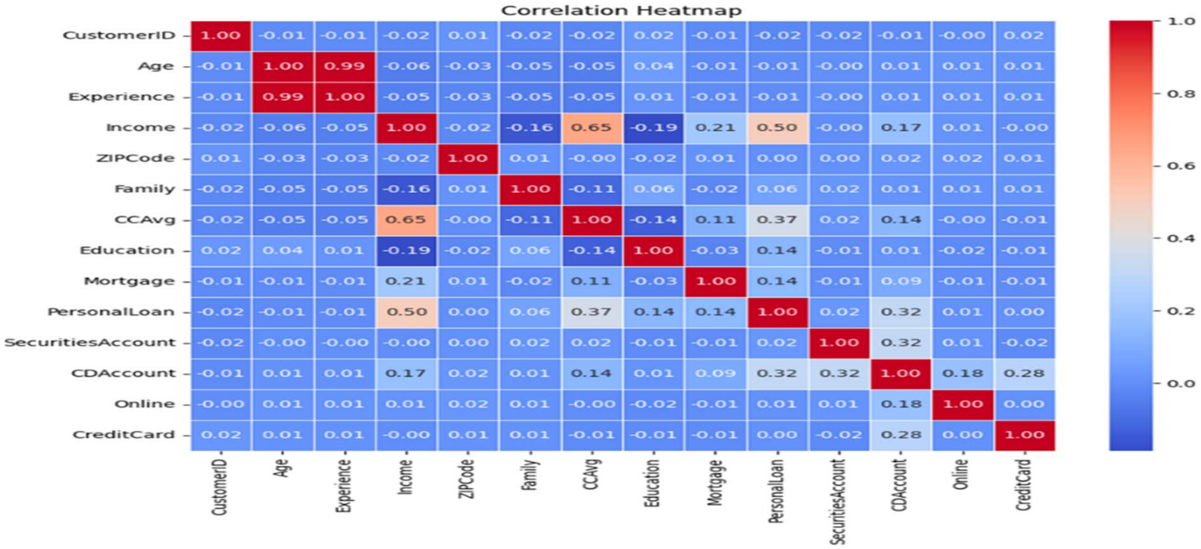
1. Random forest model - As there was data imbalance of class 0 and class 1 Performed SMOTE analysis and found that the accuracy before and after SMOTE analysis are 98.9% and 98.2% respectively.

The above inferred that the Random Forest is immune to data imbalance.

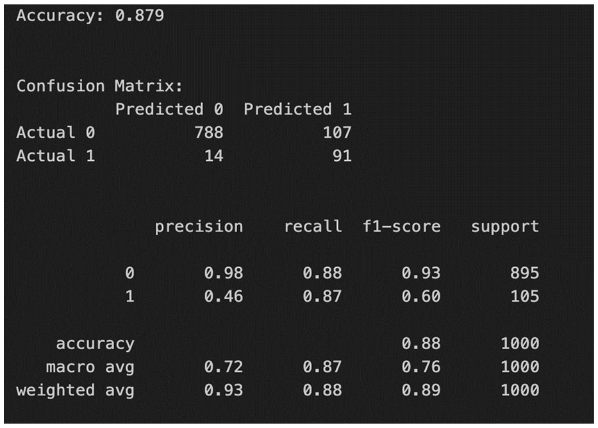
2. Logistic regression model – performance before and after SMOTE analysis are 90.9% and 87.9% respectively.

Below are a few ***graphical representations*** of analysis:

We observe that Income is positively correlated with credit card average expense.

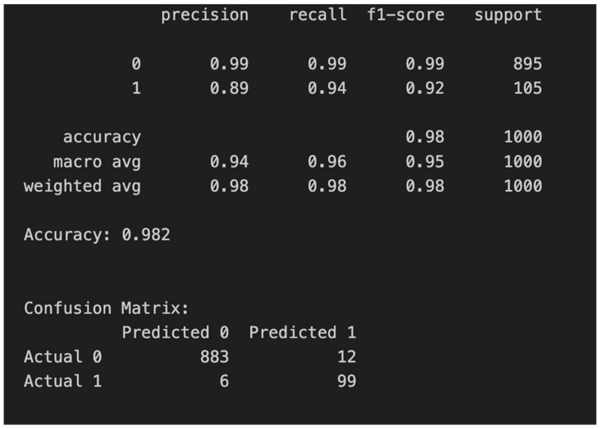


Logistic Regression model results (with SMOTE)



```
Accuracy: 0.879


Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0          788          107
Actual 1           14           91


              precision    recall  f1-score   support

           0       0.98      0.88      0.93       895
           1       0.46      0.87      0.60       105

    accuracy                           0.88      1000
   macro avg       0.72      0.87      0.76      1000
weighted avg       0.93      0.88      0.89      1000
```

Random forest model results (with SMOTE analysis)



```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       895
           1       0.89      0.94      0.92       105

    accuracy                           0.98      1000
   macro avg       0.94      0.96      0.95      1000
weighted avg       0.98      0.98      0.98      1000

Accuracy: 0.982


Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0          883           12
Actual 1            6           99
```

- The Logistic Regression model has a lower accuracy of 0.879 and a precision of 0.46 for class 1, suggesting it may generate more false positives.
- The model's recall for class 1 is 0.87, indicating it has a higher sensitivity in identifying positive instances.
- The Random Forest model shows improved accuracy at 0.982 and a higher precision of 0.89 for class 1, indicating fewer false positives.
- It also demonstrates a higher recall of 0.94 for class 1, suggesting it is more effective in classifying true positives.