Loan Application Prediction

Introduction

This report outlines the approach, insights, and performance analysis for a binary classification task aimed at predicting the 'Application Status' (Approved or Declined) based on various features such as demographic information, transaction data, and other personal details.

Approach Taken

Data Preprocessing:

- **Handling Missing Values:**
    - Columns with less than 10% missing values had their rows removed.
    - Columns with more than 50% missing values were removed.
    - For ‘Cibil Score’ missing values were replaced with the median value.
    - For categorical columns null values were replaced with category “Unknown”.
    - For Social Media presence columns, null values were replaced with “0”.
- **Feature Engineering:**
    - **Dropping Irrelevant Columns:** Removed columns like 'name', ‘mobile’ that did not contribute to the classification task.
    - **Creating Useful Features:** Created ‘Pan Name Match’ and ‘days_since’ features from ‘name’ and ‘APPLICATION LOGIN DATE’ respectively.
    - **Encoding Categorical Features**: Utilised Label and Frequency Encoding for various Categorical Features.
    - **Scaling Numerical Features:** Applied StandardScaler to standardise numerical features such as ‘days_since’, 'Cibil Score' and 'APPLIED AMOUNT’ to improve model performance.

Model Selection

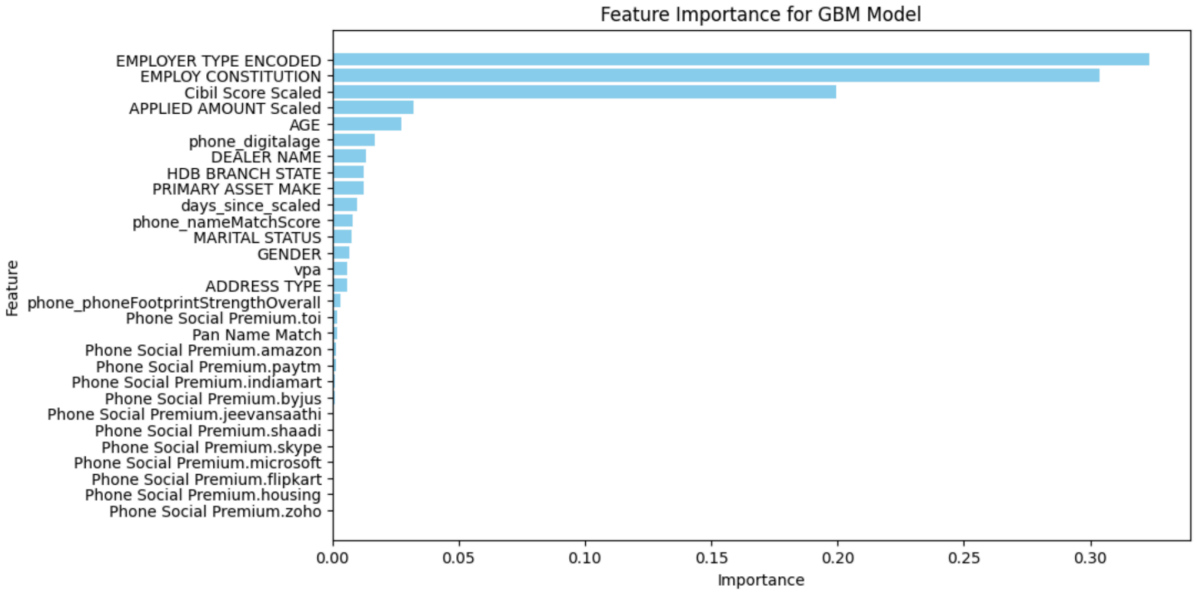
- Among the large number of models suitable for binary classification tasks, the commonly used ones are Random Forest Classifier, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Neural Network and Logistic Regression.
- Among them Gradient Boosting Machine (GBM) was chosen due to following reasons
    - Ability to fathom the complex relations between various features.
    - Effective Handling of Categorical and Numerical features.
    - Make better use of engineered features.
    - Quick access to importance of feature in classification

Insights and Conclusions from Data

Data Characteristics:

- The dataset had a mixture of numerical and categorical features, some with missing values and varying scales.
- Some features like mobile

Feature Importance:
Some Features turned out to be more influential for the classification like ‘EMPLOYER TYPE’ and ‘EMPLOY CONSTITUTION’ while other metrics like social media presence on many platforms had little influence, Influence of individual features given in the below graph.



Performance on train data

The GBM model managed to achieve an accuracy of 89% on the train data, the parameter results for the same are given below.

```
Accuracy: 0.89
Confusion Matrix:
[[521  82]
 [114 989]]
Classification Report:
              precision    recall  f1-score   support

       False       0.82      0.86      0.84       603
        True       0.92      0.90      0.91      1103

    accuracy                           0.89      1706
   macro avg       0.87      0.88      0.88      1706
weighted avg       0.89      0.89      0.89      1706
```

Conclusion

The Gradient Boosting Machine (GBM) model provided a robust solution for the loan application prediction task, achieving high accuracy and good generalisation capabilities on the training data.