Team 2 - Gokul Natesan, Madhulika Balakumar, Shumail Sajjad, Yash Dange

Project Report: Credit Score Classification

Introduction: In today's increasingly digital and 'cashless' world, credit cards have become vital tools for regular purchases and emergencies alike. With the help of cutting-edge data analytics and machine learning models, the project will focus on developing a robust credit score classification model which will classify the credit score of an individual into one of three categories: Poor, Standard and Good.

Data Source: https://www.kaggle.com/datasets/parisrohan/credit-score-classification

Data Description: The dataset chosen for our project delineates information about credit scores attributed to 12,500 clients. This information illustrates key details, including their names, age, id, occupation, annual income, number of bank accounts, and an extensive set of metrics to measure their credit history, such as number of loans, delayed payments, payment behavior, outstanding debt, etc. This dataset is robust for our analysis, offering a comprehensive sample size of 28 columns and 100,000 rows. Notably, there are eight entries for each customer corresponding to the months from January to August, highlighting the monthly relevance of these credit score checks. The distribution of the target variable, representing credit score categories - 'Poor' 'Standard,' 'Good', indicates class imbalance, with a majority in the 'Standard' category and fewer in 'Poor' and 'Good' categories.

Methodology: Data Preprocessing

As discussed in the presentation submitted in part ii of the assignment, by looking at some basic summary information about the data, we identified various issues that needed to be addressed before it is ready to be trained using a model. Some issues we noticed were:

a. Many columns had certain missing values. The dataset has 1,00,000 total rows but the number of non-null counts in columns like Name, Monthly salary, Loan Type, etc. are less than that.
b. The datatype for numerical columns like 'Age', 'Annual Income' should not be *object*. We needed to convert it to a numerical datatype like float or int, whichever is more suitable.

The required processing techniques were applied to all columns like Number of Credit cards, Interest rates, Number of loans, etc. As per standard protocol, features like ID', 'Customer_ID', 'Name', 'SSN', 'Type_of_Loan' were excluded from the model building stages. Further, encoding and scaling were performed on the categorical and numerical features in the cleaned version of the dataset, completing our data-preprocessing stage. We used one-hot encoding for other training categorical variables, and Ordinal Encoding for the target variable, given the inherent order possessed by its classes.

Methodology: Model Building, Tuning & Comparisons

As described in the initial project proposal, we identified various models that were suited for the classification task at hand and built instances of the same for deriving insights and comparisons.

a) Logistic Regression (One-vs-Rest):

Accuracy: 62.32%

- Precision:
  - Class 0: 62%, Class 1: 64%, Class 2: 55%
- Recall:
  - Class 0: 47%, Class 1: 74%, Class 2: 53%
- F1-score:
  - Class 0: 54%, Class 1: 69%, Class 2: 54%

The model performs relatively well in predicting Class 1 (with the highest precision, recall, and F1-score). However, it struggles more with Class 0 and Class 2, especially in terms of recall and F1-score, indicating it's not as effective at correctly identifying instances of these classes.The weighted average of precision, recall, and F1-score indicates an overall performance of the model, giving more weight to the classes with larger support.

While the model achieves an accuracy of around 62.32%, its performance varies across different classes. The model's ability to predict Class 0 and Class 2 have lower recall rates.

b) Support Vector Machines (SVM):

Accuracy: 53.00%

- Precision:
  - Class 0: 76%, Class 1: 53%, Class 2: 23%
- Recall:
  - Class 0: 3%, Class 1: 98%, Class 2: 2%
- F1-score:
  - Class 0: 5%, Class 1: 69%, Class 2: 4%
- Analysis:
  - Consistent pattern with other sets—high precision for Class 1, extremely low recall for Class 0 and Class 2.

The SVM model shows high precision for Class 1 across all sets. However, it performs poorly in recalling instances of Class 0 and Class 2, indicating a significant misclassification of these classes.There is a considerable imbalance in class prediction with a severe skew towards Class 1. Improvement in recall for minority classes (Class 0 and Class 2) is essential for a more balanced and accurate model.

c) Decision Trees (Tuned):

Accuracy: 70.005%

The model demonstrates overall decent performance across all classes, with an accuracy of 70.005% on the test set. Class 1 has the highest precision, recall, and F1-score, indicating the model's strength in predicting this class. Class 0 and Class 2 also show reasonable precision and recall scores, though slightly lower than Class 1.The model seems relatively balanced in predicting all three classes without a significant bias towards one class.

With Random Undersampling – After undersampling, the model's performance on the test set reveals a varied proficiency across different classes. It demonstrates a balanced precision and recall for Class 0, resulting in the highest F1-score among the classes. However, while exhibiting high precision for Class 1, the model struggles with lower recall. Conversely, for Class 2, it showcases high recall but lower precision. The overall accuracy stands at 67%, indicating a need for improvement in precision for Class 2 and recall for Class 1 to achieve a more balanced performance across all classes.

With Random Oversampling - The model's evaluation on the test set indicates diverse performance across distinct classes. Class 0 exhibits a balanced precision and recall, resulting in an F1-score of 0.70, which is the highest among the classes. However, for Class 1, while precision is high, there's a challenge with lower recall. Conversely, for Class 2, there's a higher recall but lower precision. The overall accuracy of the model stands at 67.17%, suggesting the need to enhance precision in predicting Class 2 and improve recall for Class 1 to achieve a more uniform and balanced performance across all classes.

d) Random Forest (Ensemble Method)

Accuracy: 68.22%

The model's assessment on the test set demonstrates varying performance across individual classes. Class 0 showcases a balanced precision and recall, resulting in the highest F1-score among the classes at 0.71. Conversely, for Class 1, the model exhibits higher precision but struggles with lower recall. In contrast, Class 2 displays a higher recall but lower precision.

Conclusion & Result:

After assessing the tested models, it's evident that the **Decision Tree** model stands out as the most effective among the evaluated models. This model exhibits a commendable overall accuracy of 70.005% on the test set and demonstrates reasonably balanced precision, recall, and F1-scores across all classes. Class 1, in particular, displays the highest precision, recall, and F1-score, indicating the model's strength in predicting this class without a significant bias towards any single class. In contrast, other models, such as Logistic Regression, Support Vector Machines (SVM), Random Forest (Ensemble Method), and the instances of Decision Tree models employing Random Undersampling and Random Oversampling techniques, face challenges with imbalanced performance across classes. These models encounter issues with precision or recall in specific classes, impacting their overall effectiveness. Therefore, based on consistent and balanced performance across all classes, the Decision Trees model emerges as the most reliable and effective model for this dataset. Hence, we have achieved the goal of this project to build a Machine Learning model using the techniques learned in class that can reliably classify the person’s credit worthiness into one of the three given categories.