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

Project Proposal: 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.

Proposed ML Techniques: Predicting the Credit Score of bank customers is a multiclass classification problem and we propose to employ the following ML techniques based on our preliminary analysis of the dataset:

- First, we need to preprocess and clean the data as it has a lot of missing/incorrect values. We also need to scale the numerical features and encode the categorical features.
- We will engineer new features like debt-to-income ratio and monthly balance to income ratio and assess their relevance in our models.
- Since we have an unbalanced dataset, we will use stratified splitting to get our training, validation and test datasets.
- The ML models we currently plan to experiment with are Logistic Regression, Support Vector Machine and Random Forest Classifier. We want to explore a diverse range of approaches, from linear to non-linear, to capture complex relationships within the data.
- We will use grid search to tune the models' hyperparameters and evaluate the models using cross validation.
- A use case for our classification would be to make loan approval decisions using the credit scores. In such cases, we would want to minimize the risk of loan defaults, so we prioritize identifying actual poor credit individuals (low false negatives) over incorrectly flagging good/standard credit individuals (low false positives) - so Recall will be an important metric that our models will try to optimize.