COMS 4995 Applied Machine Learning

Project Deliverable #1 - Project Proposal

Group members: Meilin Guo (mg4578), Chiara Wollner (crw2160), Angela Mu (aym2122), Hussain Doriwala (hd2551), Nagavasavi Jeepalyam (nj2506)

Background and Context

Over the years, a finance company has collected basic banking details along with extensive credit-related information. In an effort to streamline operations and reduce manual labor, the management is focused on creating a sophisticated system that can accurately classify and separate individuals into specific credit score brackets based on their financial history. With information about credit scores, companies can better assess the risk that is associated with lending money to certain individuals, make loan approval decisions, determine interest rates, etc.

This project intends to carry out a thorough comparative analysis of several machine learning models to assess their effectiveness in classifying credit scores. By conducting this empirical study, the aim is to uncover the most efficient techniques to improve the credit scoring process.

Datasets

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

The dataset consists of two files that contain information about each customer's credit-related information: Train.csv(100,001 rows, 27 columns) and Test.csv(50,001 rows, 27 columns). Given that the test dataset does not contain a labeled target variable, only the train dataset will be used, and its data will be split to form the train, validation, and test datasets.

The dataset contains financial and personal details of individuals, including numerical (Age, Income, Debt, Investments) and categorical (Occupation, Credit Score, Payment Behavior) data. It offers insights into customers' financial status, creditworthiness, and spending habits, using a mix of data types for a comprehensive analysis for assessing creditworthiness of an individual (Number of Loans, Type of Loan, Number of Bank Accounts, Number of Delayed Payments, Outstanding Debt, Credit Utilization Ratio, Credit History Age, etc).

Proposed ML Techniques

We propose several categories of ML techniques, including regression, classification, clustering, and artificial neural networks. For linear and logistic regression, we will encode categorical labels. Several classification techniques including Support Vector Machines, Decision trees, Random Forest & bootstrapping will also be applied. We will use clustering methods such as KNN and K-means, in which label columns will be dropped (worse performances are expected). We also aim to use Artificial Neural Networks, in which we will apply pre-trained and self-built Feed-Forward Neural Networks.