Project Deliverable #1 - Project Proposal

Sindhuja Rai , Joey Xia , Mansi Narendra Singh , Kevin Noble Taylor

Project - Predictive Analysis of Credit Card Application Outcomes

Background and Context

Financial institutions may use several factors to determine whether an applicant is accepted or rejected for a credit card. Regardless of how applications are processed by the issuing institutions or which attributes banks can use to determine the outcome of an application, we would like to know which factors, if any, can contribute to a particular application's outcome. Furthermore, we would like to know if features, such as gender, annual income, education status, or number of children, etc. have any correlation with an application's outcome, and the features that are most important in determining an application's outcome.

Dataset Description

Our dataset comes from Kaggle with the name "Credit card Details Binary Classification Problem" and contains two datasets. The first dataset includes all attributes of the applicant and a unique ID for the application, and the second dataset includes the same unique ID for the application and the outcome. The outcome, or label, of the record is a binary value, with 0 indicating the application is approved and 1 indicating the application is rejected. The dataset is imbalanced, containing 175 records of 1 (rejected) and 1,373 records marked 0 (accepted). The attributes of the credit applicants themselves include applicant gender (categorical), whether the applicant owns a car (categorical), number of children (numerical - discrete), annual income (numerical - continuous) and other attributes. There are 17 features excluding the unique ID and the application decision, and 1548 records.

Proposed Machine Learning Techniques

The problem we will be trying to solve in this project is binary classification - predicting the outcome of an application's status based on the features of the applicant. There are several techniques that can be used for binary classification, and based on our task and dataset, potential techniques we could use may include but are not limited to:

- K-Nearest Neighbors (KNN)
- Logistic Regression (Regularized / Non-Regularized)
- Support Vector Machines (Non-Kernel / Kernel)
- Decision Trees
- Random Forest
- Boosted Trees (AdaBoost / Gradient Boosting)
- Deep Learning / Neural Network Techniques.

Each of these models has a unique set of hyperparameters, which we will be tuning to optimize the evaluation metric. This project will be a good exercise in pre-processing data and features in multiple data types, tuning hyperparameters and evaluating the performance of several binary classification models. Additionally, because the dataset is imbalanced, this project will be an effective exercise in evaluating precision, recall, F1-Score, or other evaluation metrics, and potentially on calibrating the model(s).