**Assignment 1 - Dataset**

Amal Joseph

Student ID: 4355187

Joel Jose

Student ID: 4338660

Business Analytics, St Lawrence College

ADMN5016-S23-101-2967-K: Applied Artificial Intelligence and Machine Learning

Professor Joseph Santarcangelo Ph.D

18 May 2023

**Description of dataset**

The data set is a part of a credit or loan approval dataset. Each row represents an individual client. Here is a brief description of each column:

* ID: Unique identifier for each client.
* CODE\_GENDER: Gender of the client (M for male, F for female).
* FLAG\_OWN\_CAR: Indicates whether the client owns a car (Y for yes, N for no).
* FLAG\_OWN\_REALTY: Indicates whether the client owns property (Y for yes, N for no).
* CNT\_CHILDREN: Number of children the client has.
* AMT\_INCOME\_TOTAL: Total income of the client.
* NAME\_INCOME\_TYPE: Client's income type (e.g., Working, State servant, Commercial associate, Pensioner, Unemployed).
* NAME\_EDUCATION\_TYPE: Level of highest education the client achieved.
* NAME\_FAMILY\_STATUS: Family status of the client (e.g., Single / not married, Married, Civil marriage, Widow, Separated, Unknown).
* NAME\_HOUSING\_TYPE: What is the housing situation of the client (e.g., House / apartment, Rented apartment, With parents, Municipal apartment, Office apartment, Co-op apartment).
* DAYS\_BIRTH: Client's age in days at the time of application. The negative value represents the number of days since birth.
* DAYS\_EMPLOYED: How long the client has been employed. The negative value represents the number of days since the client started working.
* FLAG\_MOBIL: Indicates whether the client has a mobile phone (1 for yes, 0 for no).
* FLAG\_WORK\_PHONE: Indicates whether the client has a work phone (1 for yes, 0 for no).
* FLAG\_PHONE: Indicates whether the client has a phone (1 for yes, 0 for no).
* FLAG\_EMAIL: Indicates whether the client has an email (1 for yes, 0 for no).
* OCCUPATION\_TYPE: What kind of occupation the client has.
* CNT\_FAM\_MEMBERS: The number of family members the client has.

These features can be used to predict whether a client's credit/loan application will be approved or not.

**Summary of the project**

1. Credit card applications: Gathering or loading the credit card application data to be used for the prediction model.
2. Inspecting the applications: Performing an initial examination of the data to understand its structure, features, and any apparent issues (like missing or inconsistent data).
3. Splitting the dataset into train and test sets: Dividing the dataset into a training set (to train the model) and a test set (to evaluate the model's performance on unseen data).
4. Handling the missing values: Identifying and starting to deal with missing or null values in the dataset, which may involve imputation or removing data points.
5. Preprocessing the data: Performing data preprocessing steps such as scaling numerical data, encoding categorical data, and feature engineering.
6. Fitting a logistic regression model to the train set: Training a logistic regression on the preprocessed training data.
7. Making predictions and evaluating performance: Using the trained model to make predictions on the test data and evaluating the model's performance using appropriate metrics like accuracy, recall, precision, or F1 score.
8. Grid searching and making the model perform better: Tuning the model's hyperparameters using techniques like grid search to improve the model's performance.
9. Finding the best performing model: Comparing various models or model configurations to find the one that performs best on the prediction task.

**Link to Dataset**

Github Link: <https://github.com/Amaljozef7/Datafile>

Kaggle Link: <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?resource=download>