TORONTO METROPOLITAN UNIVERSITY

CIND 820 - Big Data Analytics Project

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

PREDICT DEFAULT OF CREDIT CARD CLIENTS

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Author:**  Chandramathi Vitta

Student ID: 501144398

May 15, 2023

Credit card default prediction is one of the most common problems in the financial industry, as it can help banks identify clients that are most likely to default on their credit card products. The failure to make at least minimum payment on your credit card over an extended period of time results in default. When you are late paying your payment for a month, your credit card issuer will label your account as delinquent. When the due is more than 30 days late, then the payment issuer will report the late payment to credit bureaus, such as Equifax and TransUnion. Between 30 to 180 days, your card issuer will start contacting you in order to set up a repayment plan. If this is ignored and after 180 days your card issuer will close your account and label it as “Charge-Off’. This is also known as credit card default and further your card issuer will sell your case to a debt collection agency. Your credit score is badly affected and the longer you postpone the repayment the worse your credit score will become.

(Author, Steven Porrelllo (2021,October 8).What does a credit card default mean?, The Motley Fool, <https://www.fool.ca/personal-finance/credit-cards/what-does-a-credit-card-default-mean/>

The dataset that will be used for my research is "Default of Credit Card Clients" from Kaggle. I will be applying the data analytic and machine learning themes such as exploratory data analysis, classification and clustering to identify factors that lead to credit card defaults, patterns or trends in customer behavior and improve risk management in the credit card industry.

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset contains 25 variables and 30000 records.

Kaggle Datasets -<https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset> (2005).

I will be researching on the dataset using the predictive modeling techniques primarily on three main ideas :

1. Apply exploratory data analysis techniques to gain insights into the data to better understand the patterns within the data, detect outliers or find interesting relations among the variables.
2. Identifying the key factors that contribute to credit card defaults by applying classification algorithms.
3. Identify different groups of credit card users based on their payment behavior by conducting a clustering analysis, we can group users with similar payment behaviors and identify any patterns or trends that may be present.

The techniques and tools that will be used are classification and clustering techniques using python programming language. Various classification algorithms can be used, such as logistic regression, decision trees, and random forests. Next, clustering algorithms such as k-means clustering can be used to identify different groups of credit card users based on their payment behavior. Additionally, exploratory data analysis techniques can be used to gain insights into the data and identify relevant features. Python programming language will be mainly used as a tool for solving the problem and Tableau if needed. For exploratory data analysis I will be using data visualization libraries such as Matplotlib and Seaborn. For classification problems, algorithms such as logistic regression, decision trees, random forests and XG Boost and clustering algorithms such as k-means clustering. The project will also apply model evaluation and selection techniques such as k-fold and cross validation to identify the model with high accuracy.

Kaggle Datasets -<https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset> (2005).