Predictive Analysis on Credit Card Defaults Based on Demographic Factors and Payment Behavior

Prepared for:
Supervisor's Name
CIND 820 XJH W2024

Prepared by:
Md Fahim Ferdous
ID: 501232653

Date of Submission:
January 22, 2024

Data Analytics, Big Data and Predictive Analytics
Toronto Metropolitan University

1.0 Introduction:

In today's world, the mode of transaction is taking a paradigm shift to keep up with the modern era progression. Credit card is taking the place of cash transactions and has installed the concept of contactless transaction, which opened a new door to the business world with lots of risk. With lots of promises, risk of default has been introduced as credit card is an arrangement where customer has to pay the due within a specific timeframe. It is noticed that the credit card default and due payment frequency is increasing, and it is prevalent amongst the people with specific demographic aspects. This is not only restricted to demographic factors but also to Limit allocation which is impacting the usage behavior. The study on The Relationship between Demographic Factors and Financial Behavior on Credit Card Usage in Surabaya (Memarista, Malelak and Anastasia, 2015) was done on specific demographic factors based on 105 respondents. This study was based on five demographic factors such as age, gender, education, income marital status and tried to show the impact of these demographic factors against financial factors using Chi-squared test and cross tabulation. And on the study The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients (Yeh and Lien, 2009) conducted the analysis to predict the card default based on six data mining techniques. Using different techniques, the authors tried to portray the comparative analysis of the best model output.

This project will focus on the importance and impact on credit card usage patterns and include demographic factors, limit and payment behavior; with higher data volume to conduct a predictive analysis with the impactful factors on credit card defaults.

2.0 Scope of the study:

This is a quantitative study to identify the combined factors (i.e: limit, demographic factors and payment factors) impact and predict the credit default. There were some previous works on credit default and associated factor impacts, however, there was an attempt to see the default event from as a whole point of view. There was no interaction between data dependency. And the study sample was also smaller. This project aims to predict the credit default based on two stages. On the first stage the limit will be justified based on demographic factors and payment behavior. The limits are fixed by credit analysts based on financial factors and past payment behavior. This study will focus on the impact of demographic factors on limit allocation. Limit is an important factor in credit default. The theme is, due to the limit amount, the usage amount and the bill amount go up. If the limit is allocated high while the demographic factors and payment behavior are indicating mismatch, that will lead to credit default. Upon finalizing the impacts and importance on limit, the study will further analyze how these factors are impacting credit default and the predictive analysis on credit default.

3.0 Objective:

The primary objective for this project is to:

- Identify the impact of Demographic factors and payment trend on Limit allocation
- Conduct effective predictive analytics on card default based on impactful demographic factors and payment behavior.

4.0 Methodology:

The study will be conducted with classification theme. The pattern of defaults will be identified with historical data with exploratory analysis. After that the predictive analysis will be conducted based on regression and decision tree. To find out the effectiveness of the model, confusion matrix will be developed, and accuracy, precision and recall will be calculated. Steps are provided below:

- Data collection and data preparation
- Dimensionality reduction
- Exploratory data analysis: Histogram and boxplot
- Experiment design: training set vs test set
- Data modelling:

  Stage 1: Decision tree analysis for limit

  Stage 2: Regression analysis for limit

  Limit=α+β*LIMIT_BAL+β*SEX+β*EDUCATION+β*MARRIAGE+β*AGE+β*PAY_0 +β*PAY_2+β*PAY_3+β*PAY_4+β*PAY_5+β*PAY_6+ε

  Stage 3: Decision tree for Credit default

  Stage 4: Regression analysis for Credit default

  default payment next month= α+ β*impactful demographic factors+ β* BILL_AMT1+ β* BILL_AMT2+ β* BILL_AMT3+ β* BILL_AMT4+ β* BILL_AMT5+ β* BILL_AMT6+ β*PAY_AMT1+ β* PAY_AMT2+β* PAY_AMT3+β* PAY_AMT4+β* PAY_AMT5+β* PAY_AMT6+ ε

- Model evaluation through confusion matrix and ratios

The Analysis will be conducted using python, R and for illustration tableu will be used.

5.0 The Dataset:

The dataset was collected from UCI Machine Learning Repository. The dataset includes a total of 23 variables and 30000 structured data points. The database presents factors such as:

X1: Amount of the given credit (NT dollar)

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: past payment history. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar September-April)

X18-X23: Amount of previous payment (NT dollar September-April)

Dataset link: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

References:

Memarista, G., Malelak, M., & Anastasia, N. (n.d.). *The Relationship between Demographic Factors and Financial Behavior on Credit Card Usage in Surabaya*. https://core.ac.uk/download/pdf/32453075.pdf

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Yeh,I-Cheng. (2016). Default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.