Credit Card Default Prediction Using Customer Behaviour Data

1. Introduction

Credit risk assessment is an important process in the financial industry, particularly for credit card issuers, as it helps identify customers who are likely to default on their payments. Accurate prediction of credit card default enables financial institutions to minimize financial losses and improve decision-making related to credit approval and risk management.

In this project, a secondary dataset containing real customer demographic and behavioural information was downloaded and analysed to perform credit card default prediction. The study applies descriptive analysis, statistical techniques, and predictive modelling approaches to understand customer behaviour and estimate default risk using real-world data.

2. Objectives of the Study

1. To explore and summarize the demographic characteristics, credit limits, billing amounts, and repayment patterns of credit card clients using descriptive statistics and data visualizations.
2. To examine the relationship between historical repayment behaviour and credit card default status through statistical analysis and correlation techniques.
3. To develop predictive models to estimate the probability of credit card default using traditional statistical methods such as logistic regression and machine learning classifiers including Decision Trees and Random Forests.
4. To evaluate and compare model performance using metrics such as accuracy, precision, recall, F1-score, and ROC–AUC in order to perform real credit risk prediction based on customer behaviour data.

3. Dataset Description

The dataset used in this study is a secondary dataset downloaded from a publicly available source. It contains detailed information on credit card clients, including demographic attributes, credit limits, billing amounts across multiple months, repayment history, and default status.

The dataset represents real customer behaviour and is suitable for analysing repayment patterns and predicting credit card default risk. The target variable indicates whether a customer has defaulted on credit card payments, while the remaining variables describe customer characteristics and historical financial behaviour.

4. Methodology

A quantitative analytical approach was adopted in this study using the downloaded secondary dataset. The analysis was conducted in accordance with the defined objectives and involved the following steps:

- ❖ Exploratory data analysis was performed to understand the structure of the dataset and summarize demographic, billing, and repayment characteristics using descriptive statistics and visualizations.
- ❖ Statistical and correlation analyses were conducted to examine the relationship between historical repayment behaviour and credit card default status.
- ❖ Predictive models, including logistic regression and machine learning classifiers such as Decision Trees and Random Forests, were developed to estimate the probability of credit card default.
- ❖ The performance of the developed models was evaluated and compared using classification metrics such as accuracy, precision, recall, F1-score, and ROC–AUC to assess their effectiveness in real credit risk prediction.

5. Findings of the study

- **Exploratory Data Analysis Findings**
    1. Exploratory analysis reveals that the dataset predominantly consists of working-age and well-educated clients, with female customers forming the majority of credit card holders.
    2. Credit limits, billing amounts, and repayment values exhibit wide variability and strong right-skewness, indicating heterogeneous financial behaviour among clients.
    3. While most clients maintain moderate credit limits and billing balances, a small proportion of customers demonstrate very high credit usage and repayment levels, suggesting unequal credit consumption patterns.
    4. Analysis of repayment behaviour shows that non-defaulters consistently exhibit higher mean repayment amounts than defaulters across all six observed months.
    5. On average, non-defaulters repaid approximately ₹5,200 to ₹6,600 per month, whereas defaulters repaid only about ₹3,100 to ₹3,400.
    6. These findings confirm that lower repayment amounts are strongly associated with higher default risk, validating repayment behaviour as a key indicator of credit card default.

- **Predictive Modelling and Evaluation Findings**

7. Model performance comparison indicates that the Random Forest model outperforms the Decision Tree and Logistic Regression models, followed by Decision Tree and Logistic Regression respectively.
8. The Logistic Regression model achieved an AUC of 0.715, indicating reasonable baseline discriminatory power and good interpretability.
9. The Decision Tree model improved predictive performance with an AUC of 0.749 by capturing nonlinear relationships between customer behaviour and default risk.
10. The Random Forest model demonstrated the best performance with an AUC of 0.775, highlighting the effectiveness of ensemble learning techniques in credit risk prediction.
11. These results illustrate a clear trade-off between model interpretability and predictive accuracy, with simpler models offering transparency and ensemble models providing superior performance.

- **Credit Risk Estimation and Model Inference**

12. The estimated overall probability of default in the dataset is approximately 22%, reflecting the average level of credit risk among customers, while individual predicted probabilities enable identification of higher-risk clients.
13. Based on evaluation metrics such as accuracy, precision, recall, F1-score, and ROC–AUC, the Random Forest model demonstrates effective performance in predicting credit risk using customer behaviour data.
14. The model exhibits higher accuracy in identifying low-risk (non-default) customers compared to high-risk (default) customers.
15. In the analysed case, the model predicts that the customer has a low credit risk, indicating a high likelihood of timely credit card bill repayment.

6. Conclusion

- ✓ Predictive models, including Logistic Regression, Decision Tree, and Random Forest, were developed and evaluated to estimate the probability of credit card default. Model performance comparison showed that the Random Forest model outperformed the other approaches, achieving the highest ROC–AUC score and demonstrating superior predictive accuracy due to its ability to capture complex, non-linear relationships in customer behaviour data. However, simpler models such as Logistic Regression provided greater interpretability, emphasizing the trade-off between model transparency and predictive performance.
- ✓ Overall, the findings indicate that customer demographic and behavioural data can be effectively leveraged for real-world credit risk prediction. The study highlights the potential of machine learning models, particularly ensemble techniques, in supporting credit risk management and decision-making processes. These insights can assist

financial institutions in identifying high-risk customers, improving credit assessment strategies, and reducing potential default-related losses.

✓ This study successfully demonstrated the application of descriptive analysis, statistical techniques, and predictive modelling to assess and predict credit card default risk using customer behaviour data from a secondary dataset. Exploratory data analysis revealed significant variability in credit limits, billing amounts, and repayment behaviour, highlighting the heterogeneous financial characteristics of credit card clients. The analysis confirmed that repayment behaviour is a critical indicator of default risk, with non-defaulters consistently exhibiting higher repayment amounts than defaulters.