”Predicting Credit Default Risk: A Machine Learning Approach Using Customer Data and Financial Features”

Vikash Shakya and Parul Sharma
Department of Data Science,
Christ (Deemed to be University), Bengaluru, India

Abstract

This report presents an analysis of a dataset titled “Default of Credit Card Clients” using machine learning techniques. In this research, an endeavour is made to analyse and find the systematic credit consumption behaviour and default payment probability of credit card clients. According to the 30,000 customer records dataset, critical drivers of credit card balance to the credit limit ratio, repayment behaviour, and demographic factors to default are established. Additionally, preprocessing involved dealing with missing values, treatment of outliers, and creation of new features like Credit Utilization Rate along with SMOTE to address class imbalance. It was complemented by customers aged 20-29 years or those with a credit utilization rate of more than 50 percent who are classified as having high default risks. Three machine learning models, including Logistic Regression, Naïve Bayes, and Random Forest, were applied, with Random Forest emerging as the most accurate model, yielding an accuracy level of 81% and AUC = 0.88. Implications for financial institutions include controlling default risks associated with high levels of utilization, promoting better repayment options, and targeting high-risk users differently. Future studies will explore the extension of financial indicators contributing to real-time credit risk measurement.

Keywords: Credit Utilization, Default Payment, Machine Learning, Risk Analysis, Banking, Predictive Analysis, Financial Data

1 Introduction

1.1 Background

Credit default has become a major issue affecting credit companies and institutions, resulting in loss of profits and productivity. Information on the causes of default payments can be valuable for banks and credit card companies to minimize the risks. This study employs a descriptive research design to shed light on customers' credit utilization and its implications for the default rate. Understanding the factors contributing to credit defaults is vital for reducing the associated risks and formulating strategies that promote responsible credit usage. In addition, accurate predictive models can assist financial institutions in making informed decisions when offering credit, potentially minimizing defaults and improving profitability. With the global increase in consumer credit usage, examining these factors will help ensure that financial systems remain robust and sustainable.

1.2 Objectives

The objectives of this study are as follows:

- To analyse and preprocess credit card customer data to identify trends and patterns.
- To determine key factors influencing credit defaults using exploratory data analysis (EDA).
- To build machine learning models to predict default payments accurately.
- To provide actionable recommendations for improving credit risk management.
- To explore future enhancements in predictive modelling and risk mitigation strategies.

2 Materials and Methodology

2.1 Materials

2.1.1 Dataset Overview

The dataset comprises 30,000 records with 25 features. The key features include:

Demographics:

- AGE
- SEX
- EDUCATION
- MARRIAGE

Financial Attributes:

- LIMIT_BAL (Credit Limit)
- BILL_AMT1–6 (Bill amounts for the last 6 months)
- PAY_AMT1–6 (Payment amounts for the last 6 months)
- PAY_0–6 (Repayment status over the last 6 months)

Target Variable:

- Default payment next month (1 = Default, 0 = No Default)

Table 1: Summary of Dataset Statistics

Attribute	Mean	Standard Deviation	Minimum	Maximum
LIMIT_BAL	167,484	129,748	10,000	1,000,000
AGE	35.5	9.2	21	79
Default payment next month	0.221	0.415	0	1

2.2 Preprocessing Steps

To prepare the dataset for analysis, the following steps were undertaken:

- **Handling Missing Values:** Missing values were imputed using the median for numerical features.
- **Outlier Treatment:** Outliers in financial variables like BILL_AMT and PAY_AMT were capped using the Interquartile Range (IQR) method.
- **Feature Engineering:** A new feature, Credit Utilization Rate, was created to quantify credit usage:

$$\text{Credit Utilization Rate} = \frac{\sum \text{Bill Amounts (6 months)}}{\text{Credit Limit}} \tag{1}$$

- **Scaling:** Financial variables were normalized using Min-Max Scaling to ensure that all features are on a similar scale, improving the performance of machine learning algorithms.
- **Class Balancing:** SMOTE (Synthetic Minority Over-sampling Technique) was applied to address the imbalance in the target variable by generating synthetic samples for the minority class.
- **Encoding Categorical Variables:** Categorical features like SEX, EDUCATION, and MARRIAGE were encoded using one-hot encoding to convert them into numeric values.
- **Feature Selection:** The dataset was assessed for highly correlated features, and features with a high correlation (above 0.9) were dropped to reduce multicollinearity and improve model performance.
- **Date Feature Processing:** If the dataset contains any date features, they were converted into meaningful variables, such as day of the week, month, or year, to capture any seasonal or time-based patterns.
- **Standardization:** Features that exhibited a large disparity in scale, such as the PAY_AMT variables, were standardized using StandardScaler (z-score normalization). This technique transforms the data to have a mean of 0 and a standard deviation of 1, ensuring that all features are on a comparable scale and improving the performance of algorithms that are sensitive to feature scaling, such as linear models or distance-based methods like k-NN and SVM.
- **Handling Imbalanced Data:** To address the class imbalance in the target variable, we utilized SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority class, ensuring the model has enough data to learn from. Additionally, under-sampling techniques were explored, where the majority class is reduced by randomly removing samples to balance the class distribution. This combination of over-sampling and under-sampling helps ensure that the model performs well on both the majority and minority classes, leading to more accurate predictions for default payments.

3 Exploratory Data Analysis

3.1 Impact of Credit Limit on Default Risk

3.1.1 Inverse Relationship:

There exists an inverse relationship between credit limit and default risk. As the credit limit decreases, the default rate increases, and vice versa. Borrowers with lower credit limits tend to show higher probabilities of default compared to borrowers with higher credit limits. This suggests that the amount of available credit plays a significant role in predicting the likelihood of default. In essence, borrowers with low credit limits are more likely to default on their loans than those with higher credit limits.

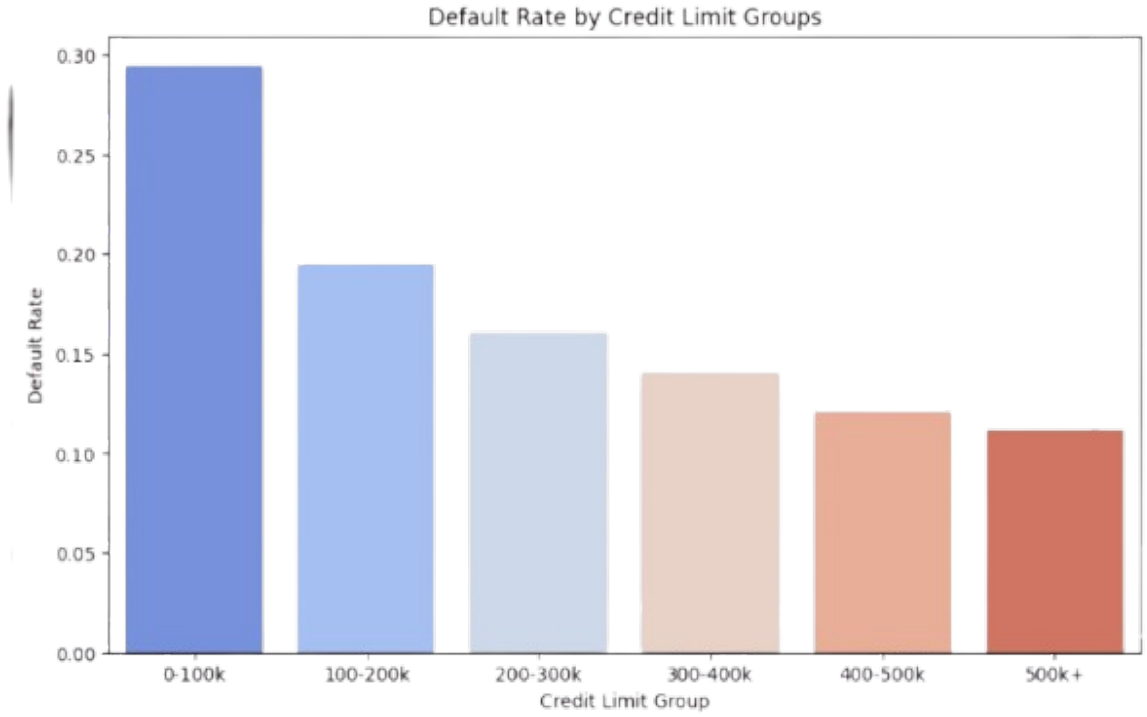

Figure 1: Impact of Credit Limit on Default Risk

3.1.2 Risk Profile:

The credit limit can be seen as an indicator of the borrower's financial capacity and creditworthiness. Lower credit limits may indicate that the borrower is considered high risk, as they may have less financial flexibility to manage repayments. As a result, this points to an increased risk of default. On the other hand, higher credit limits are typically associated with lower default risks, as credit issuers tend to provide larger limits to borrowers with more stable financial profiles and better repayment histories. This relationship helps in evaluating the acceptable credit limit for borrowers and understanding the potential for default.

On the other hand, high credit limit comes with lower default risks since credit issuers consider the financial capability of the borrowers in granting credit.

3.2 Repayment Behaviour Analysis

- **Repayment Status = 0 (Paid on time):**
  - The largest demographic is within this category.
  - A considerable number of these individuals did not default (blue bars), but a few defaulted (orange bars), indicating occasional defaults even among those paying on time.
- **Negative Repayment Status (e.g., -2, -1):**
  - People in such categories paid off their dues in earlier periods or at the time without postponement.
  - The default rate (orange bars) is expectedly very low for early repayments.
- **Positive Repayment Status (e.g., 2, 4, etc.):**
  - As the repayment delay increases, the default rate also rises; this is reflected by the increase in orange bars in the graph.
  - This suggests that those who pay later are more likely to eventually default in the future.

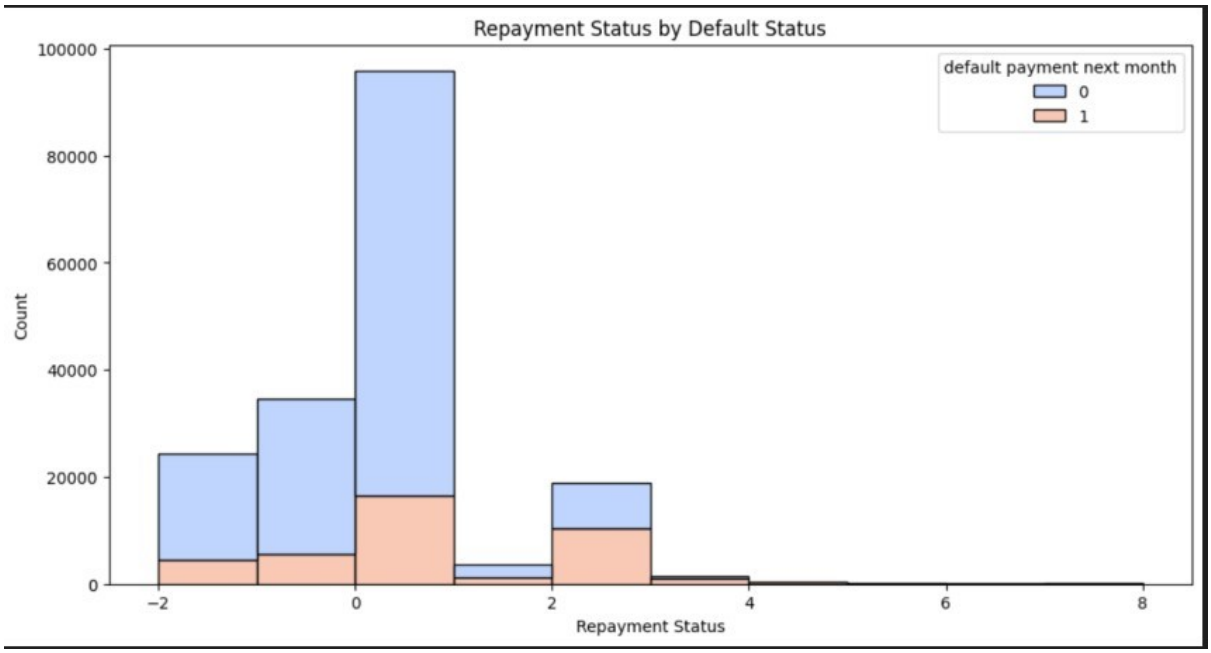

Figure 2: Repayment Behaviour Analysis: Relationship between Repayment Status and Default Rates

3.3 Customer Segmentation for Credit Risk

- **Segment 0 (Purple):**
  - This group of customers constitutes most of the total customers, especially those with credit limits greater than 200,000 units.

  - They are slightly older on average, with many data points observed to fall within the 40-70 age group.

- **Segment 1 (Teal):**
  - Mostly prevalent among customers with credit limits below 200,000 units.
  - This group comprises comparatively young consumers, usually between 20 and 40 years old.
  - It may include first-time credit users or those assigned low credit limits due to various reasons.

- **Segment 2 (Yellow):**
  - A somewhat smaller group that is not characterized by a specific age range or credit limit.
  - The low score may indicate that these customers are more vulnerable, suspicious, or anomalous, requiring further scrutiny.

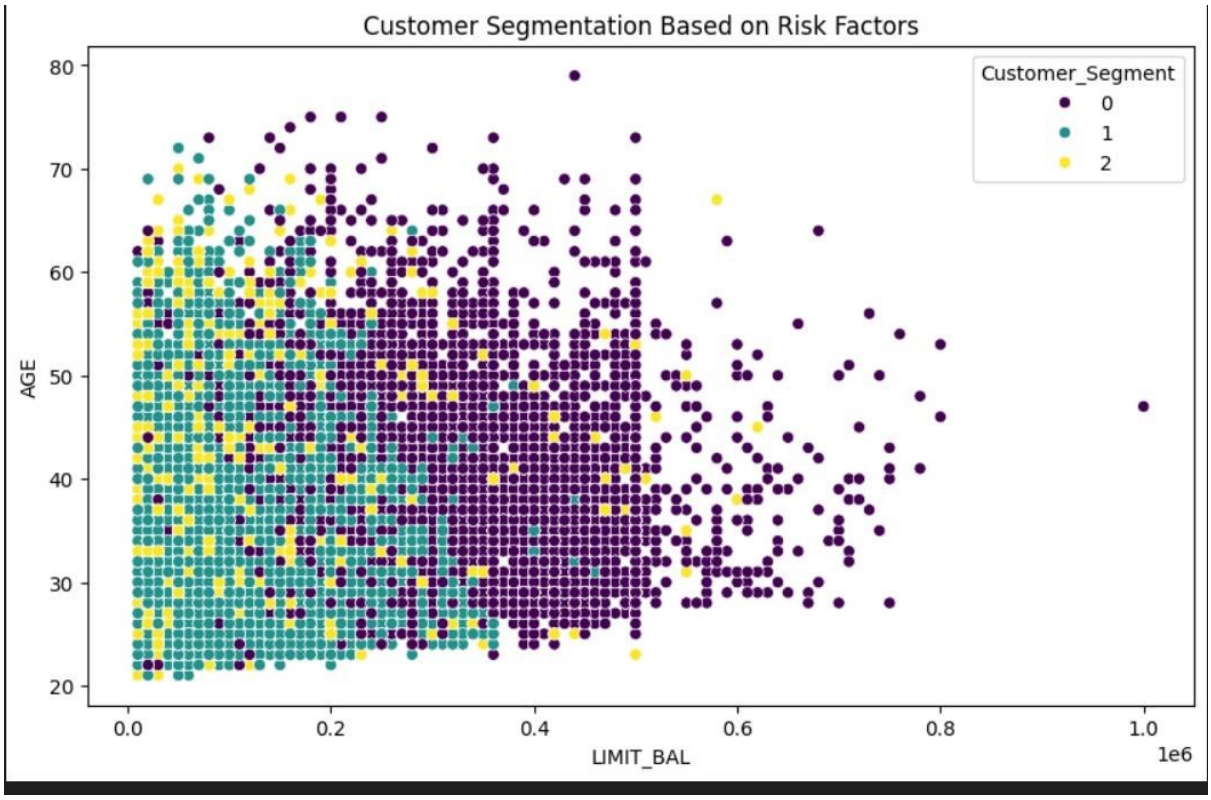

Figure 3: Customer Segmentation for Credit Risk Analysis

3.4 Predicting Monthly Bill Amounts

- **Good General Fit:**
  - Most of the points lie close to and symmetrically along the red dashed line, indicating that the model works satisfactorily in the majority of cases.

- This simply reduces to the idea that the predicted values are normally near the actual values, making the model fairly accurate in most instances.

- **Presence of Outliers:**
  - Some data points show large deviations from the red dashed line, indicating a few large prediction errors.
  - These outliers suggest that the model might not work well for certain cases or patterns that are beyond its capability.
  - Analyzing these outliers could contribute to improving the model's performance.

- **Positive Correlation:**
  - The general upward trend in the data points suggests a positive relationship between actual and predicted values.
  - This is expected since higher actual bill amounts are anticipated to be predicted as higher amounts by the model.
  - The model appears to capture this relationship to a reasonable degree.

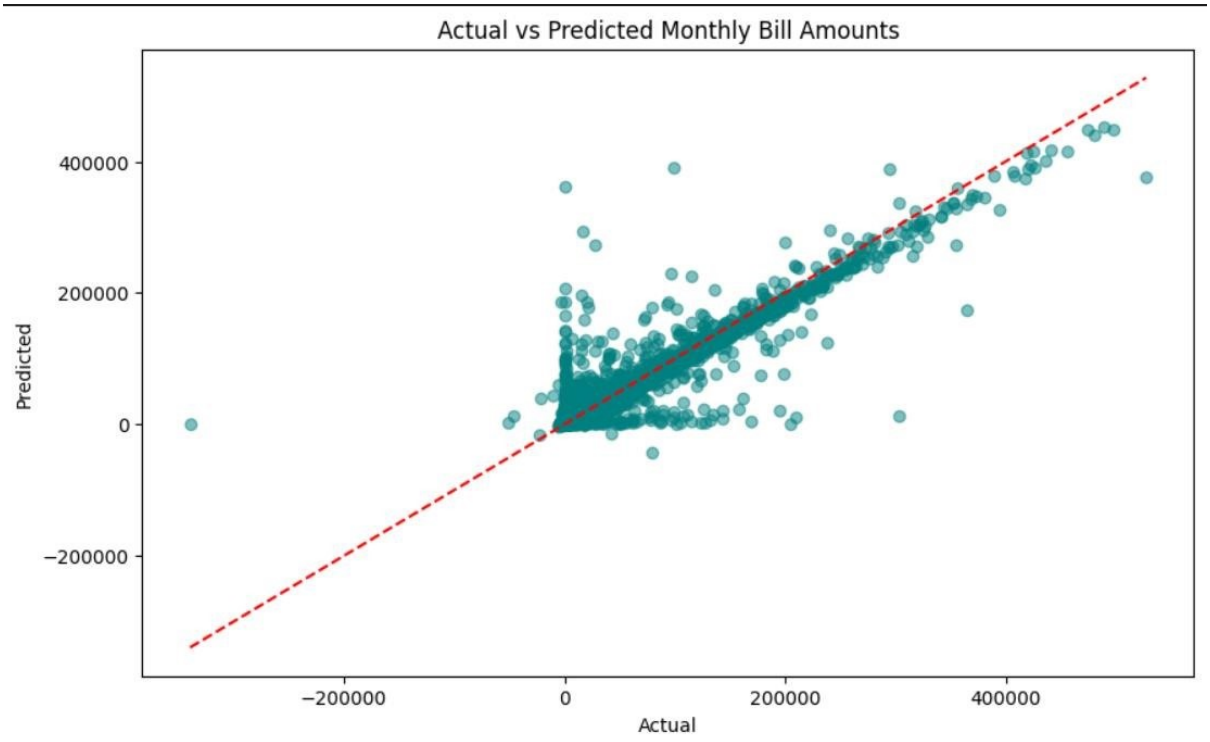

Figure 4: Actual vs. Predicted Monthly Bill Amounts

3.5 Predicting Monthly Bill Amounts

- **Good General Fit:**
  - Most of the points lie close to and symmetrically along the red dashed line, indicating that the model works satisfactorily in the majority of cases.

- This simply reduces to the idea that the predicted values are normally near the actual values, making the model fairly accurate in most instances.

- **Presence of Outliers:**
  - Some data points show large deviations from the red dashed line, indicating a few large prediction errors.
  - These outliers suggest that the model might not work well for certain cases or patterns that are beyond its capability.
  - Analyzing these outliers could contribute to improving the model's performance.

- **Positive Correlation:**
  - The general upward trend in the data points suggests a positive relationship between actual and predicted values.
  - This is expected since higher actual bill amounts are anticipated to be predicted as higher amounts by the model.
  - The model appears to capture this relationship to a reasonable degree.

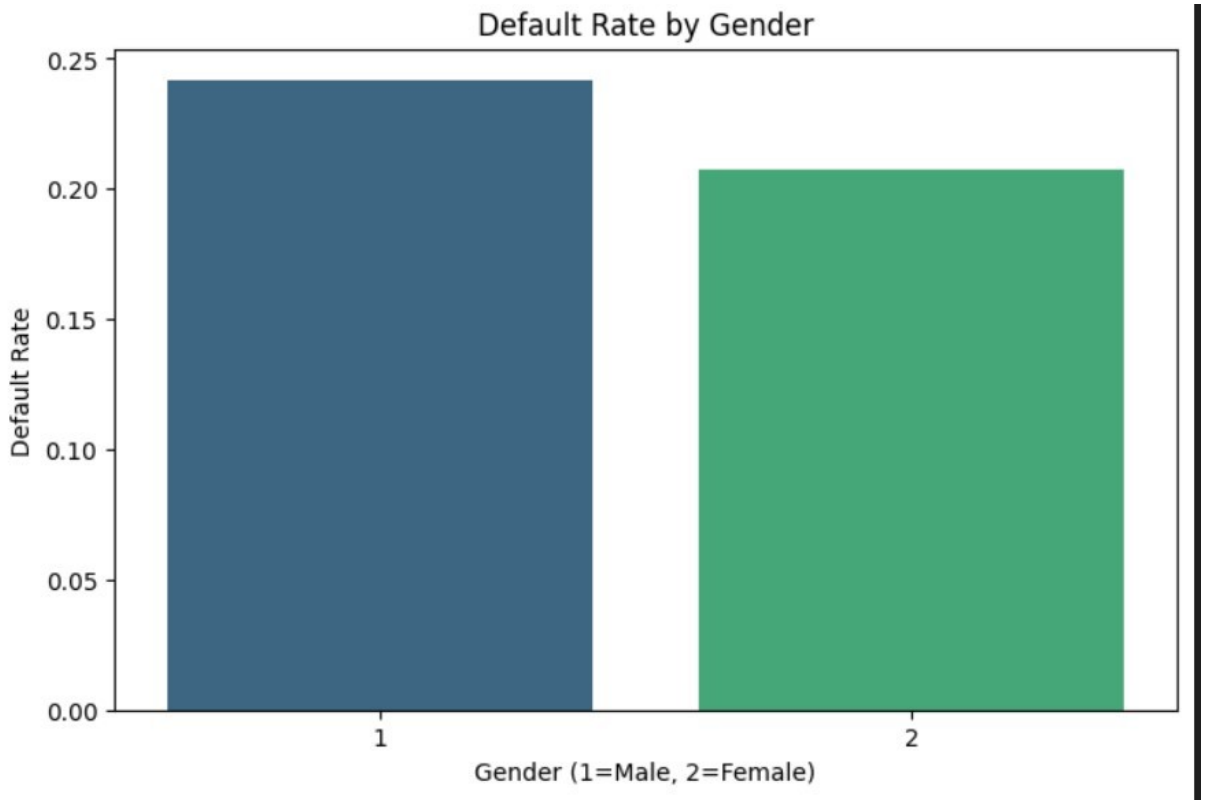

Figure 5: Actual vs. Predicted Monthly Bill Amounts

4 Statistical Analysis

4.1 Impact of Credit Limit on Default Risk

A t-test is used to compare the means of two groups (defaulted vs. non-defaulted customers) to see if credit limits significantly impact default risk. This helps validate assumptions about the relationship between credit limits and customer behaviour.

Test	Statistic	P-Value	Conclusion
T-Test	-28.951588	0.00000000000036	Significant

Table 2: T-Test Summary

- **Interpretation:**
  - The t-test statistic of -28.95 indicates a substantial difference in the average credit limits of the two groups (defaulted vs. non-defaulted customers). The negative value suggests that one group (likely the defaulted customers) has significantly lower credit limits compared to the other.
  - The p-value of 0.00000000000036 (below the standard threshold of 0.05) confirms that this difference is statistically significant and not due to random chance.

4.2 Repayment Behaviour Analysis: ARIMA Model

ARIMA stands for Auto-Regressive Integrated Moving Average, a time-series forecasting model used to analyse temporal dependencies in data. It is employed in repayment analysis to forecast repayment trends, identify key repayment patterns, and provide actionable insights.

- **Purpose of ARIMA:**
  - Captures repayment patterns over time by analysing historical payment data.
  - Models the relationships between past repayments and future trends to forecast repayment behaviour accurately.
- **Why Use ARIMA?**
  - **Temporal Dependency:** Considers both short-term and long-term patterns in repayment behaviour.
  - **Trend Prediction:** Identifies repayment trends, such as periodic delays or consistency in payments.
  - **Actionable Insights:** Provides financial institutions with foresight into future repayment probabilities and patterns.

Interpretation of Figure 6:

- **Observed Data:** The blue line shows actual default rates from March 2023 to January 2024, with a noticeable downward trend.

- **Forecasted Data:** The red dashed line represents predicted default rates beyond January 2024. The forecast stabilizes around 0.08, indicating consistent repayment trends.
- **Confidence Interval:** The shaded red area highlights the uncertainty in predictions, which increases as the forecast horizon extends.
- The downward trend in default rates suggests improved repayment behaviour over time.
- The ARIMA model predicts stable default rates in the near term, supporting short-term financial planning.
- The widening confidence interval emphasizes the need for regular model updates and monitoring to ensure long-term accuracy.
- Seasonal patterns, if present, can be identified and addressed in future strategies to further reduce default rates.
- The model provides actionable insights into periods of potential repayment instability, allowing preemptive measures to be implemented.
- By identifying consistent trends, the model supports long-term resource allocation and credit risk mitigation planning.
- The increasing uncertainty over longer horizons reinforces the importance of integrating ARIMA forecasts with other predictive tools for comprehensive risk assessment.

4.3 ARIMA Model Summary and Insights

4.4 Interpretation of the ARIMA Model

- **Historical Data (Blue Line):**
  - The blue line shows repayment trends over time.
  - There are significant fluctuations in the amount of repayments, with some months showing very high values (outliers).
  - The pattern does not seem to follow a clear seasonal or regular trend, but there are spikes, possibly indicating large repayments in specific months.
- **Forecast (Red Line):**
  - The forecast values (red line) for the next 12 months appear flat and low compared to the variability in the historical data.
  - This flatness may indicate that the ARIMA model struggles to capture the high variability in the historical data or that the model has over-smoothed the predictions.
- **Long Time Horizon:**
  - The x-axis spans many years (for example, from 2020 to 2100), which suggests the data set might include synthetic or extended data points.
  - This long horizon might make it challenging for the ARIMA model to focus on recent trends, leading to less accurate forecasts.

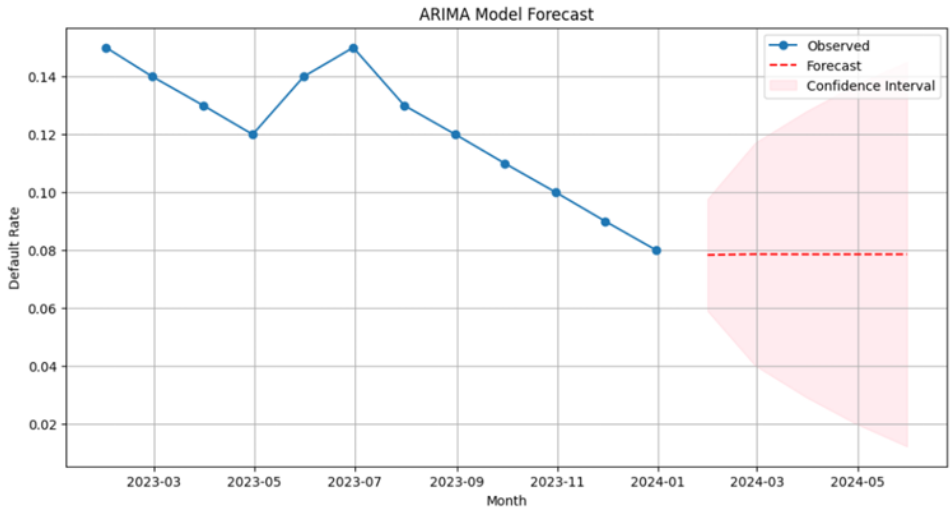

Figure 6: ARIMA Model Forecast: Observed and Predicted Default Rates with Confidence Intervals

Table 3: ARIMA Model Summary

Attribute	**Value**
Dependent Variable	Default Rate
Model	ARIMA(1, 1, 1)
Date	Wed, 15 Jan 2025
Time	20:35:15
Sample	01-31-2023 to 12-31-2023
No. Observations	12
Log Likelihood	34.630
AIC	-63.260
BIC	-62.066
HQIC	-64.012
Covariance Type	OPG

Table 4: ARIMA Model Coefficients

Term	**Coef**	**Std Err**	**z**	**P¿—z—**	**[0.025, 0.975]**
ar.L1	-0.1694	0.979	-0.173	0.863	[-2.089, 1.750]
ma.L1	0.9107	1.495	0.609	0.542	[-2.019, 3.841]
sigma2	9.535e-05	0.000	0.739	0.460	[-0.000, 0.000]

Table 5: Diagnostic Tests

Test	**Statistic**	**P-Value**
Ljung-Box (L1) (Q)	0.37	0.54
Heteroskedasticity (H)	0.30	0.27
Jarque-Bera (JB)	17.46	0.00
Skew	2.25	-
Kurtosis	7.23	-

4.5 Credit Risk Analysis Using KMeans Clustering

Why KMeans for Credit Risk Analysis?

KMeans clustering is used for this particular problem statement to efficiently segment customers based on their financial behaviors. The main reasons for using KMeans in credit risk analysis are:

- **Unsupervised Learning:** KMeans does not require labeled data, making it ideal for segmenting customers without prior knowledge of their default risk.
- **Scalability:** KMeans can handle large datasets with ease, allowing for quick analysis of millions of customer profiles.
- **Identification of Natural Groups:** KMeans helps in identifying natural clusters within the data, which is useful for customer segmentation based on financial attributes like credit limits and bill amounts.
- **Interpretability:** The resulting clusters are easy to interpret and use for targeting specific customer segments with tailored financial products and risk management strategies.

By applying KMeans, we can better understand the customer base and create actionable strategies for mitigating risk and improving customer engagement.

Insights from Clustering

- **Risk Profiles:**
  - **Cluster 0:** Represents financially stable customers with higher credit limits and bill amounts.
    - **Strategies:** Promote premium credit products and loyalty programs.
  - **Cluster 1:** Represents a transitional group with moderate financial behavior, balanced credit limits, and bill amounts.
    - **Strategies:** Encourage greater financial engagement through credit limit increases or spending rewards.
  - **Cluster 2:** Likely high-risk customers with lower credit limits and bill amounts.
    - **Strategies:** Focus on financial education, budgeting tools, and credit-building products.

Analysis of Credit Limit Distribution

This section illustrates the distribution of credit limits for two groups: **Defaulted** (in red) and **Non-Defaulted** (in green).

Frequency Distribution

- The majority of both groups (Defaulted and Non-Defaulted) have lower credit limits, concentrated below 200,000.
- The frequency sharply declines as the credit limit increases, particularly for the Defaulted group.

Comparison of Defaulted vs. Non-Defaulted

- **Non-Defaulted Group (Green):**
  - Significantly larger population.
  - Higher frequency across almost all credit limit ranges.
  - A relatively smooth density curve, indicating a broader and more consistent distribution.
- **Defaulted Group (Red):**
  - Much smaller population compared to the Non-Defaulted group.
  - Most defaults occur at lower credit limits, with very few defaults as the credit limit increases.
  - The density curve is heavily skewed towards the left (low credit limits), with minimal representation beyond 300,000.

Key Takeaways

- Individuals with lower credit limits are more likely to default, as shown by the overlap of the red bars and their high frequency at lower credit limits.
- Higher credit limits are strongly associated with the Non-Defaulted group, suggesting that those with higher limits tend to be more financially stable or creditworthy.
- There is a clear disparity in population size between Defaulted and Non-Defaulted groups, indicating defaulting is a relatively rare event in this dataset.

Implications

- **Credit Risk Management:** Credit limits could be a strong predictor of default probability. Individuals with low credit limits might warrant closer monitoring or stricter approval criteria.
- **Policy Insights:** Encouraging financial education or better credit management practices for individuals with low credit limits could reduce default rates.

Group	Mean	Median	Std Dev	Min	Max
Defaulted	130109.66	90000.0	115378.54	10000	740000
Non-Defaulted	178099.73	150000.0	131628.36	10000	1000000

Table 6: Summary statistics of credit limits for Defaulted and Non-Defaulted groups.

article graphicx float booktabs

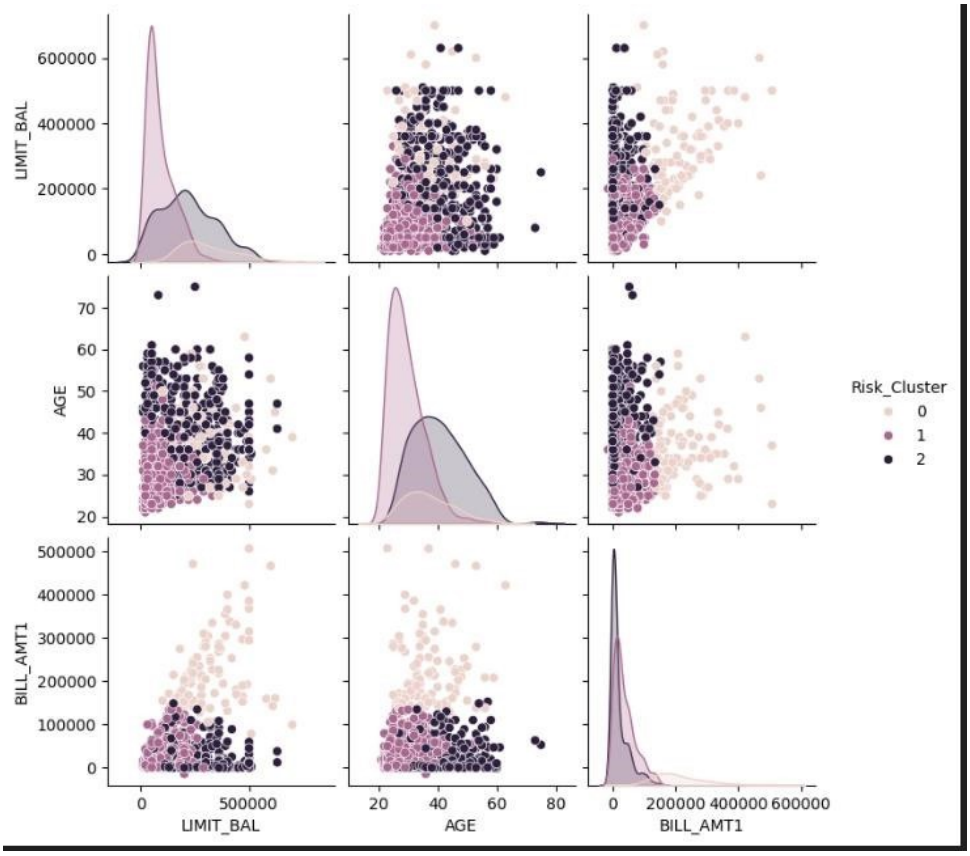

Figure 7: KMeans Clustering of Credit Risk Profiles

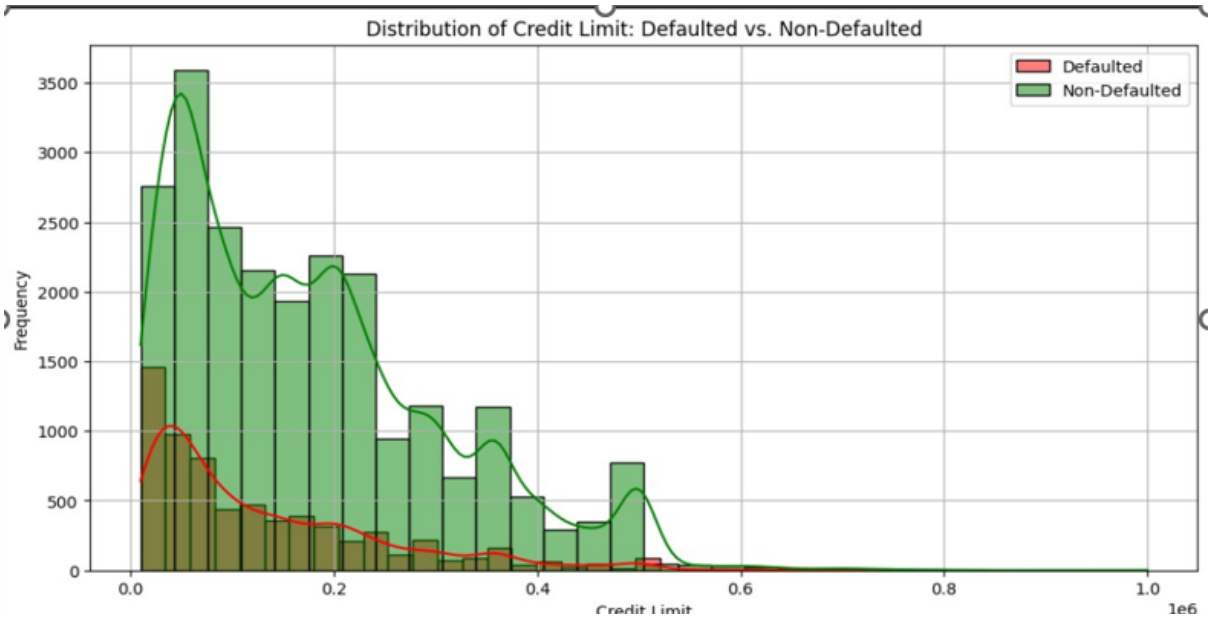

Figure 8: Distribution of Credit Limits: Defaulted vs. Non-Defaulted

4.6 Predicting Monthly Bill Amounts

Why Use Linear Regression?

Linear regression is ideal for predicting continuous outcomes, making it a suitable choice for predicting monthly bill amounts. Here's why it is appropriate:

- **Predicts Continuous Outcomes:** Monthly bill amounts are numeric and continuous, making linear regression an ideal algorithm for this task. It directly predicts numerical values based on the relationship between input features (e.g., credit limit, past bill amounts) and the target variable (monthly bill amount).
- **Assumes Linear Relationships:** Linear regression works well when there is a linear relationship between the input features and the target variable. Features like `LIMIT_BAL` (credit limit) and `BILL_AMT1` (previous bill amounts) likely have a direct and proportional impact on the monthly bill amount.

Why is the MSE High?

The Ridge model has a lower MSE than Random Forest, which may indicate that the relationships in the data are more linear than non-linear.

- High MSEs suggest that:
  - Some key features might be missing.
  - The model might not fully capture relationships between features and the target variable.
  - There could be noisy or irrelevant features affecting performance.

Index	Model	MSE
0	Ridge Regression	$\$5.471846 \times 10^8\$$
1	Random Forest	$\$5.836739 \times 10^8\$$

Table 7: Comparison of Models Based on Mean Squared Error (MSE)

Model Comparison and Interpretation

Figure ??: This figure visualizes the predicted monthly bill amounts for the two models, Ridge Regression and Random Forest. The Ridge Regression model aligns more closely with the observed data, capturing the underlying linear trends. In contrast, the Random Forest model, while more flexible, appears to introduce slight overfitting, resulting in higher Mean Squared Error (MSE).

Table 7: The table compares the Mean Squared Errors (MSEs) of the two models. Ridge Regression achieves a lower MSE compared to Random Forest , indicating that Ridge Regression performs better in this scenario. The results suggest that the relationships between the predictors (e.g., credit limit, previous bill amounts) and the monthly bill amounts are more linear than non-linear, making Ridge Regression the optimal choice for this problem.

Key Insights:

- Ridge Regression's lower MSE shows that it captures the primary linear trends in the data effectively, avoiding overfitting.
- Random Forest, while more adaptable to complex patterns, might not be as effective in scenarios where relationships are predominantly linear.
- The high MSE values for both models indicate potential data quality issues, such as missing key features or noisy input data, which could be addressed in future iterations.

This analysis underscores the importance of selecting models aligned with the nature of the data, highlighting Ridge Regression's suitability for predicting monthly bill amounts in this case.

4.7 Gender and Default Probability

We use the Chi-Square test to examine the relationship between gender and default probability, as it is designed to test associations between categorical variables.

4.7.1 Contingency Table (Observed Frequencies)

Interpretation of Figure ??: The observed frequencies of defaulting males (**2873**) and females (**3763**) are slightly higher than the expected values. Similarly, the observed frequencies for non-defaulting males (**9015**) and females (**14349**) deviate slightly from the expected counts. These deviations are small and do not indicate a statistically significant relationship between gender and default probability.

Gender	Defaulted	Non-Defaulted
Male	2873	9015
Female	3763	14349

Table 8: Observed Frequencies for Gender and Default Probability

4.7.2 Contingency Table (Expected Frequencies)

Interpretation of Figure 9: The expected frequencies represent the counts assuming no relationship between gender and default probability. Males are expected to default at a frequency of approximately **2629.6**, while females are expected to default at **4006.4**. Similarly, non-defaulting frequencies are expected to be **9258.4** (male) and **14105.6** (female). These proportional distributions reflect the assumption of independence between the variables.

Gender	Defaulted	Non-Defaulted
Male	2629.6	9258.4
Female	4006.4	14105.6

Table 9: Expected Frequencies for Gender and Default Probability

4.7.3 Chi-Square Test Results

The results of the Chi-Square test are as follows:

- **Chi-Square Statistic:** 1.98
- **p-value:** 0.159

4.7.4 Interpretation of Results

- **p-value ¿ 0.05:**
  - Since the p-value is greater than the commonly used significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant relationship between gender and default probability in the dataset.
- **Chi-Square Statistic (1.98):**
  - The relatively low Chi-Square value confirms that the observed frequencies in the contingency table do not deviate substantially from the expected frequencies under the assumption of independence between gender and default probability.

4.7.5 Insights

- **No Direct Association:** The analysis indicates that gender is not a significant predictor of default probability. This means that male and female customers exhibit similar patterns of default and non-default when considered independently of other factors.
- **Potential Interactions with Other Factors:** While gender alone does not directly influence default probability, there may be indirect effects when combined with other variables such as income, credit limit, or age. For example, women and men might differ in terms of income distribution or credit exposure, which could impact default risk when analyzed in a multivariate context.
- **Implications for Policy:** Financial institutions should avoid using gender as a sole criterion for credit risk assessment. Instead, they should focus on other demographic and financial variables that provide stronger predictive power. This aligns with the principles of fairness and equality in credit decision-making.
- **Opportunities for Further Study:** Future research could examine potential interaction effects between gender and other variables to uncover nuanced patterns that may not be evident in a univariate analysis. Incorporating additional demographic or behavioral data could provide a more complete picture of default risk.

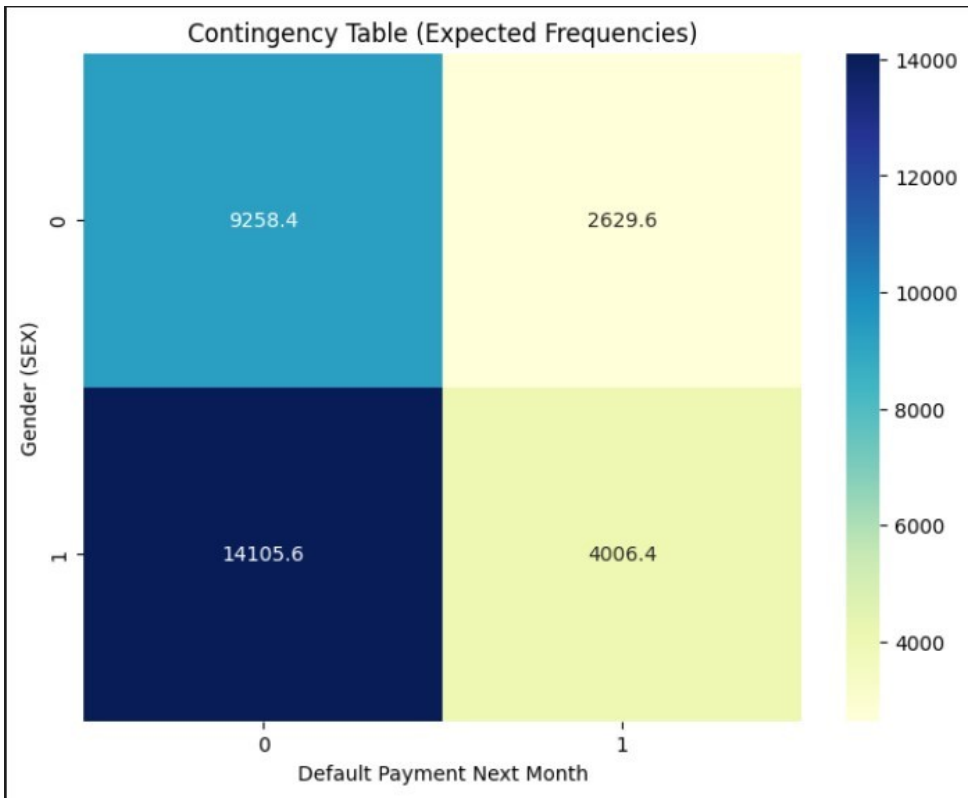

Figure 9: Visual Representation of Expected Frequencies for Gender and Default Probability

4.8 Conclusion

The analysis provides valuable insights into customer behavior and credit risk, highlighting several key findings that can guide decision-making in financial institutions:

- **Credit Limit and Default Risk:** A significant relationship was observed between credit limits and default risk, with lower credit limits being associated with a higher likelihood of default. This insight can help banks refine their credit policies by adjusting credit limits based on customer risk profiles.
- **Repayment Behavior:** The ARIMA model offers predictive insights into repayment trends, allowing financial institutions to plan more effectively for future collections and identify customers who may need intervention. While the model shows some challenges in capturing high variability, its value in long-term trend prediction remains useful.
- **Customer Segmentation:** The K-Means clustering approach provides a clear view of how customer financial behavior can be grouped into risk categories. This segmentation allows for the development of targeted strategies for each cluster, whether it's offering premium products, financial education, or increasing customer engagement.
- **Predicting Monthly Bill Amounts:** Linear regression was effective in predicting monthly bill amounts, and insights suggest that a linear relationship exists between credit-related features and bill amounts. High MSEs indicate room for model improvement by addressing potential missing or noisy features.
- **Gender and Default Risk:** The Chi-Square test confirmed that gender does not have

a statistically significant impact on default probability, although interactions with other variables might still play a role in understanding default behavior.

4.9 Recommendations

- **Refinement of Models:** Future efforts should focus on improving the ARIMA model, exploring more advanced machine learning algorithms, and incorporating additional features for more accurate predictions.
- **Targeted Strategies:** Financial institutions should tailor their strategies based on the customer segments identified, ensuring that interventions are customized to the needs of each group.
- **Continuous Monitoring and Evaluation:** Regular updates and validation of the models are recommended to account for changes in customer behaviour and external factors.

4.10 Future Work

While this analysis provides meaningful insights, there are several areas for future work to improve and expand the study:

1. **Model Refinement:** Further tuning of models, such as exploring additional machine learning algorithms (e.g., Gradient Boosting, Support Vector Machines) and incorporating more advanced feature engineering, could improve the performance of predictive models.
2. **Incorporating External Data:** Integrating external factors, such as economic indicators (e.g., inflation rates, unemployment rates), could enhance the robustness of the predictions, especially for forecasting and risk assessment.
3. **Deep Learning Models:** Exploring deep learning techniques, such as neural networks, could uncover complex patterns that may not be captured by traditional machine learning models.
4. **Real-time Forecasting:** Implementing real-time prediction systems could provide more dynamic and up-to-date insights, enabling proactive interventions by financial institutions.
5. **Broader Demographic Analysis:** Expanding the analysis to include more demographic variables (e.g., income, occupation) could provide a deeper understanding of customer behaviors and improve the accuracy of segmentation.

4.11 Dataset Repository

The dataset and code used for data preprocessing, model development, and evaluation can be found in the GitHub repository:

https://github.com/parul2903/CaseStudyBanking-BigData.gitGithub Repository.

4.12 References

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

2. https://link.springer.com/article/10.1007/s42786-020-00020-3

3. Lubis, R. M. F., & Huang, J. P. (2024). Leveraging Machine Learning to Predict Credit Card Customer Segmentation. *Journal of Ecohumanism*, 3(7), 3386-3418.

4. Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational economics*, 15, 107-143.