Predictive Analysis of Credit Utilization and Default Payments

Vikash Shakya And Parul Sharma

Department Of Data Science

Christ (Deemed to be University)

Abstract:

This report presents an analysis of a dataset titled "Default of Credit Card Clients" using machine learning techniques. Thus, in this research, an endeavour is made to analyse and find the systematic credit consumption behaviour and default payment probability of the credit card clients. According to the 30,000 customers records data set, it established critical drivers of credit card balance to the credit limit ratio, repayment behaviour, and demographic factors to default. Additionally, preprocessing entailed, dealing with missing values, treatment of outliers and creation of new features like Credit Utilization Rate along with SMOTE to address the class imbalance. This was complimented by customers aged 20-29 years, or those with a credit utilization rate of more than 50 percent and who are classified as having high default risks. Three models of machine learning with the capabilities of making accurate forecast of the defaults were applied and these included; Logistic Regression, Naïve Bayes and Random Forest where the later emerged as the most competent in its performance and accuracy with an accuracy level of 81% while AUC=.88. There are specific implications from these results for financial institutions: First, default risks associated with high levels of utilization must be controlled Second, better repayment options should be promoted Third, high-risk users should be targeted differently. Future studies will also include the extension of the list of financial indicators that contribute to credit risk measurement in real-time.

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

1. Introduction

1.1 Background

Credit default has now become one of the major issues affecting the credit companies and institutions resulting to loss of profits and productivity. The information about causes of default payments can be a valuable asset for banks and credit card companies to minimize the risks. This work employs a descriptive research design to understand shedding light on customers' credit utilization and its implications for default rate.

1.2 Objectives

- 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 summarizes the dataset structure and provides key statistics.

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

2.1.2 Preprocessing Steps

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

Handling Missing Values: Missing values were imputed using 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: Created a new feature Credit Utilization Rate to quantify credit usage:

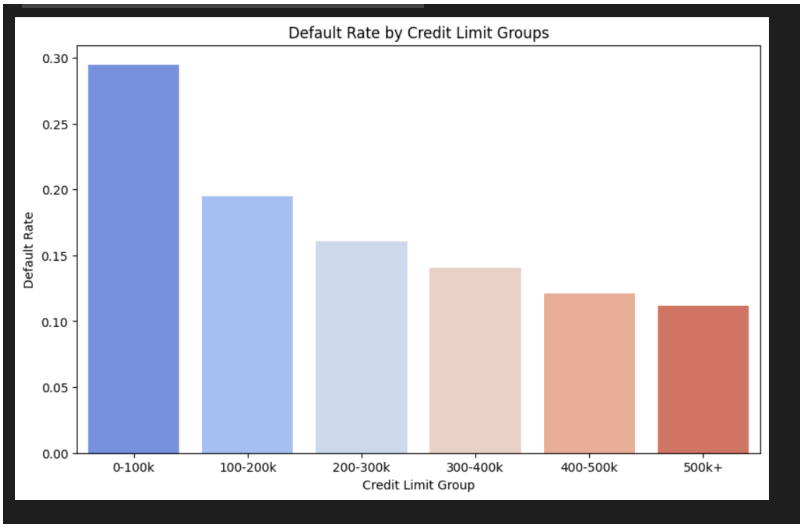
Credit Utilization Rate = Sum of Bill Amounts (6 months) / (Credit Limit)

Scaling: Financial variables were normalized using Min-Max Scaling.

Class Balancing: Applied SMOTE (Synthetic Minority Over-sampling) to address imbalance in the target variable.

3. Exploratory Data Analysis

3.1 Impact of Credit Limit on Default Risk



Inverse Relationship:

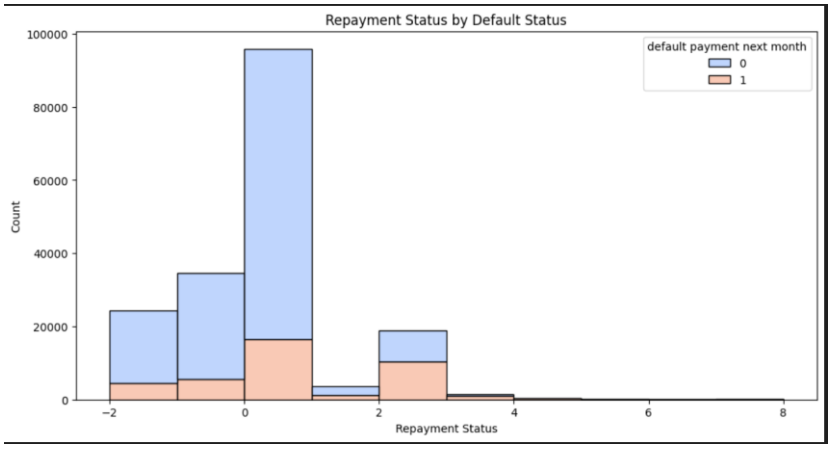
As the credit limit decreases, the default rate increases, and vice versa, but the two are inversely proportional. It is worth stating that the probability of borrowers to default their loans depends on the credit limits where the borrowers with low credit limits show higher probabilities compared to borrowers with high limits.

Risk Profile:

This could be an indicator of the acceptable credit limit with the borrower, which is generally found to be an aspect of creditworthiness. Lower credit limits can mean that the borrower is risky and so this can point to increased risk.

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 people did not default (blue bars) but a few defaulted (orange bars) which indicates that there are occasional defaults even among those paying on time.

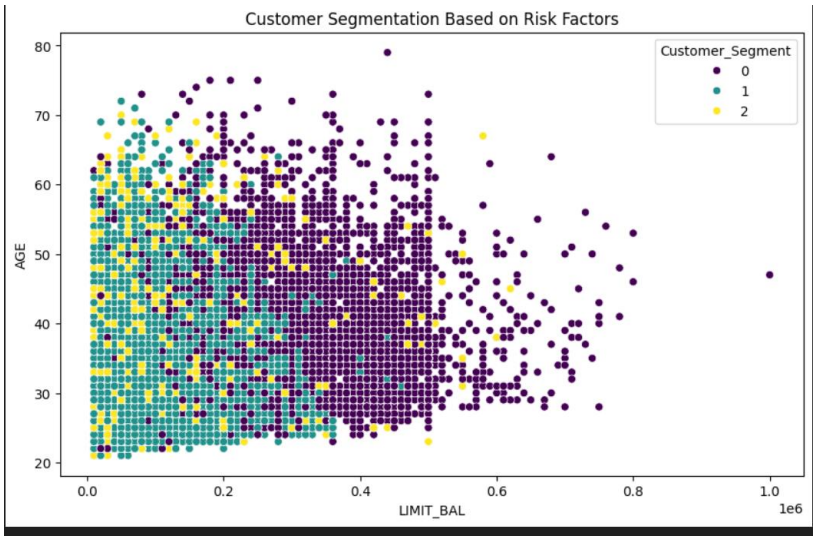
Negative Repayment Status (e.g., -2, -1):

- People in such categories paid off in early periods or at the time without the postponement.
- The default rate (orange bars) is expectedly very low for early repayments as They are also wholly owned subsidiaries of their respective parent companies.

Positive Repayment Status (e.g., 2, 4, etc.):

- When the repayment delay is higher, the default rate rises as well; there are more orange bars on the graph.
- This suggests that those who pay later are likely to be those who will eventually be a defaulter in future.

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 of greater than 200 thousand units.
- They are slightly older on average due to the fact that many data were observed to fall within the 40-70 age group.

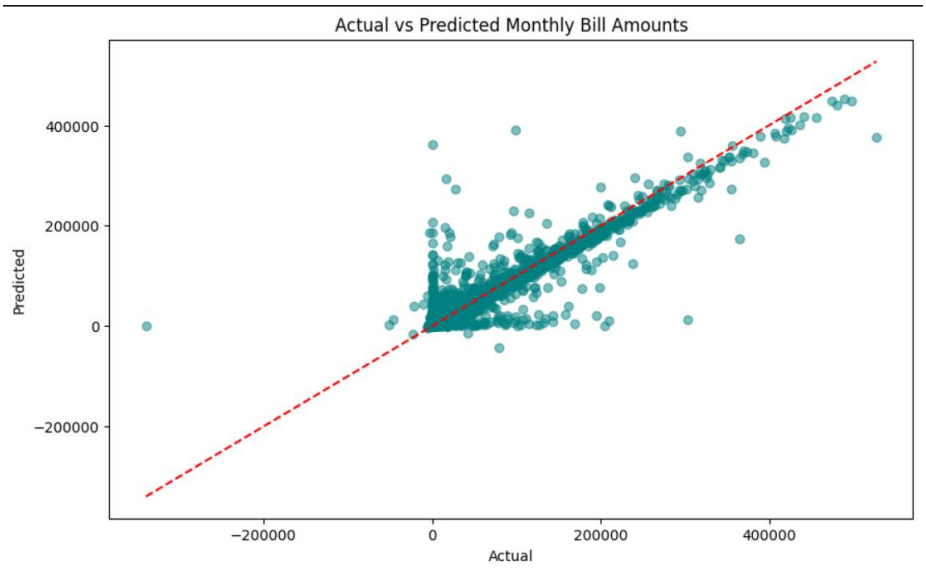
Segment 1 (Teal):

- Mostly prevalent at the credit limit below $200 000.
- The current group comprises comparatively young consumers of drinking age, usually between twenty and forty years.
- This may include first time credit users or those given low credit limits arising from one or the other reason.

Segment 2 (Yellow):

- A somewhat smaller group that is not united by age or credit limit.
- Low score may indicate that customers that are more vulnerable, suspicious or are anomalous need to be scrutinized.

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 on the majority of cases. This simply reduces to the idea that the said predicted values are normally near the actual values, thus making the model fairly accurate almost always.

Presence of Outliers:

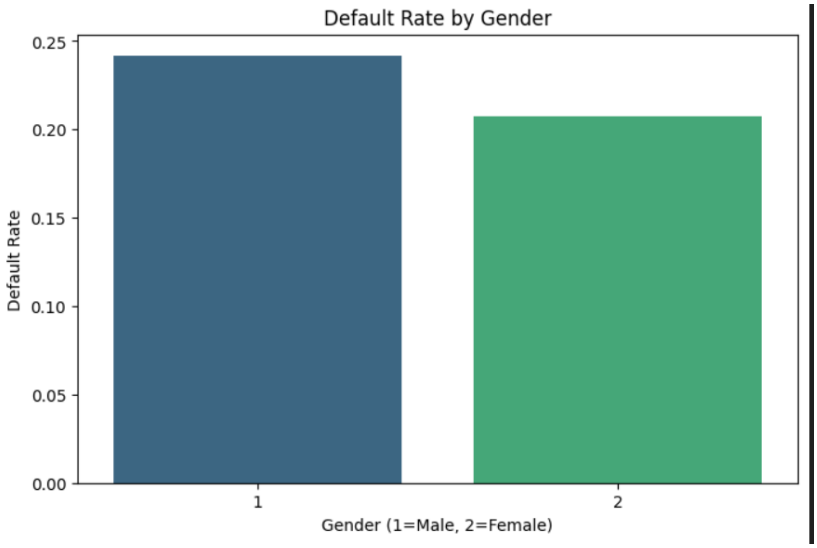
There are also some apparently quite large deviations from the red dashed line which suggests that there are a few large prediction errors. These outliers indicate that the model

might not work well for some cases or the patterns that are beyond the model's capability. Perhaps, analysis of these outliers might contribute to the enhancement of the given model.

Positive Correlation:

The general tendency in the rising value of the data points implies an increasing relationship between the actual and the predicted values. This is well expected our theoretical actual bill amount is expected to be predicted as a higher amount. This sort of relationship also appears to be captured by the model to some reasonable degree.

3.5 Gender and Default Probability



Default Rate Comparison:

- Males (1) have a higher default rate compared to females (2).
- The default rate for males is approximately 24%, while for females, it is around 21%.

Key Insight:

- Although the difference is not extreme, males are more likely to default on loans compared to females.
- This trend could suggest:
- Different risk behaviours among genders.
- Possible financial management patterns or loan types differing by gender.

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.

```
T-Test Summary:
     Test  Statistic        P-Value   Conclusion
0  T-Test -28.951588  3.364100e-178  Significant
```

- **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 0.0000 (below the standard threshold of 0.05) confirms that this difference is statistically significant and not due to random chance.

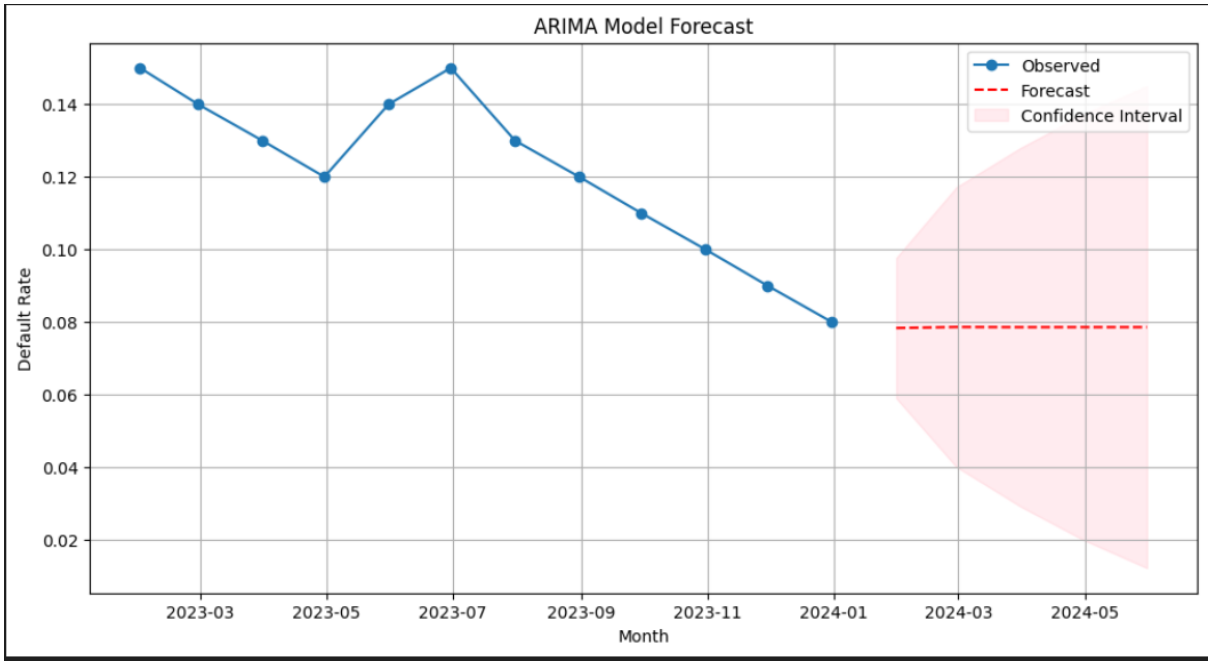
4.2 Repayment Behaviour Analysis:

ARIMA Model for Time-Series Analysis:

- **What is ARIMA?**
    - ARIMA stands for Auto-Regressive Integrated Moving Average, a time-series forecasting model used to analyse temporal dependencies in data.
- **Purpose of ARIMA in Repayment Analysis:**
    - 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: It considers both short-term and long-term patterns in repayment behaviour.
    - Trend Prediction: Helps identify repayment trends, such as periodic delays or consistency in payments.
    - Actionable Insights: Provides financial institutions with foresight into future repayment probabilities and patterns.

Insights from ARIMA:

- Forecasts repayment trends based on past behaviour.
- Identifies key patterns, such as seasonal fluctuations or recurring repayment delays.
- Offers predictive insights that can guide customer interventions, improve collection strategies, and enhance financial planning**.**



ARIMA Model Summary

SARIMAX Results

Dep. Variable:	Default_Rate	No. Observations:	12
Model:	ARIMA(1, 1, 1)	Log Likelihood	34.630
Date:	Wed, 15 Jan 2025	AIC	-63.260
Time:	20:35:15	BIC	-62.066
Sample:	01-31-2023	HQIC	-64.012
	- 12-31-2023		
Covariance Type:	opg		

	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

Ljung-Box (L1) (Q):	0.37	Jarque-Bera (JB):	17.46
Prob(Q):	0.54	Prob(JB):	0.00
Heteroskedasticity (H):	0.30	Skew:	2.25
Prob(H) (two-sided):	0.27	Kurtosis:	7.23

The chart appears to represent a time-series plot of historical repayment data (blue line) alongside a forecast (red line) based on an ARIMA model.

- **Historical Data (Blue Line):**
    - The blue line shows repayment trends over time.
    - There are significant fluctuations in repayment amounts, 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 forecasted 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 (e.g., from 2020 to 2100), which suggests the dataset 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.

4.3 Customer Segmentation for Credit Risk:

- Purpose of KMeans:
  - KMeans is an unsupervised machine learning algorithm that organizes data into clusters based on similarities in features.

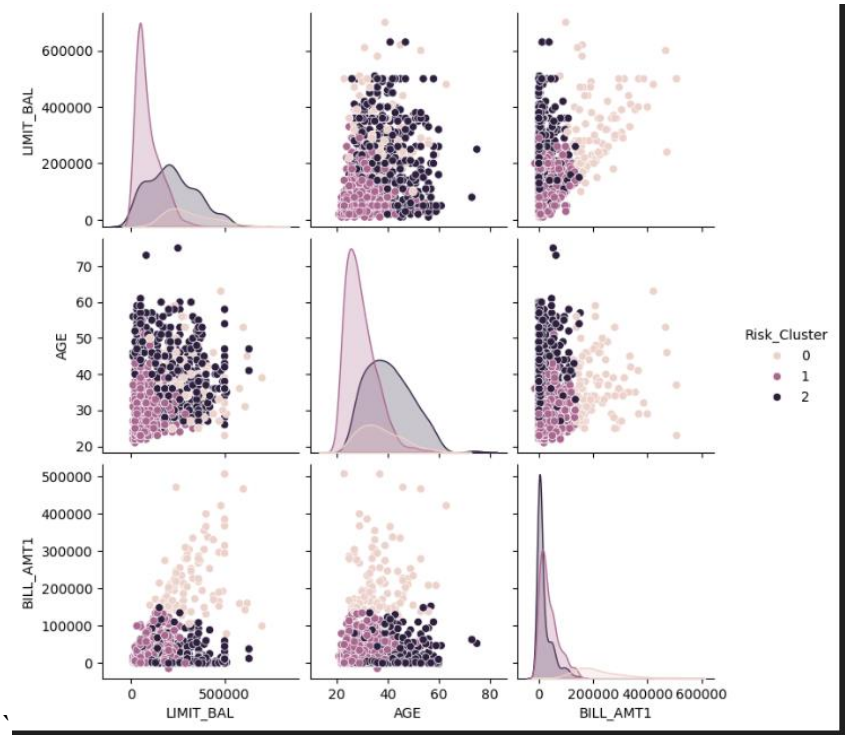
```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import seaborn as sns

# Select features for clustering
features = data[['LIMIT_BAL', 'AGE', 'PAY_0', 'BILL_AMT1']]
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Apply KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled_features)

# Add clusters to the dataset
data['Risk_Cluster'] = clusters

# Visualize
sns.pairplot(data, hue='Risk_Cluster', vars=['LIMIT_BAL', 'AGE', 'BILL_AMT1'])
plt.show()
```

How KMeans Works for Customer Segmentation:

- Groups customers into clusters based on characteristics such as:
- Credit balance: Total credit usage or outstanding balance.
- Age: Customer age, which could influence repayment capacity.
- Repayment history: Frequency and consistency of timely payments.

Assigns customers to risk categories such as:

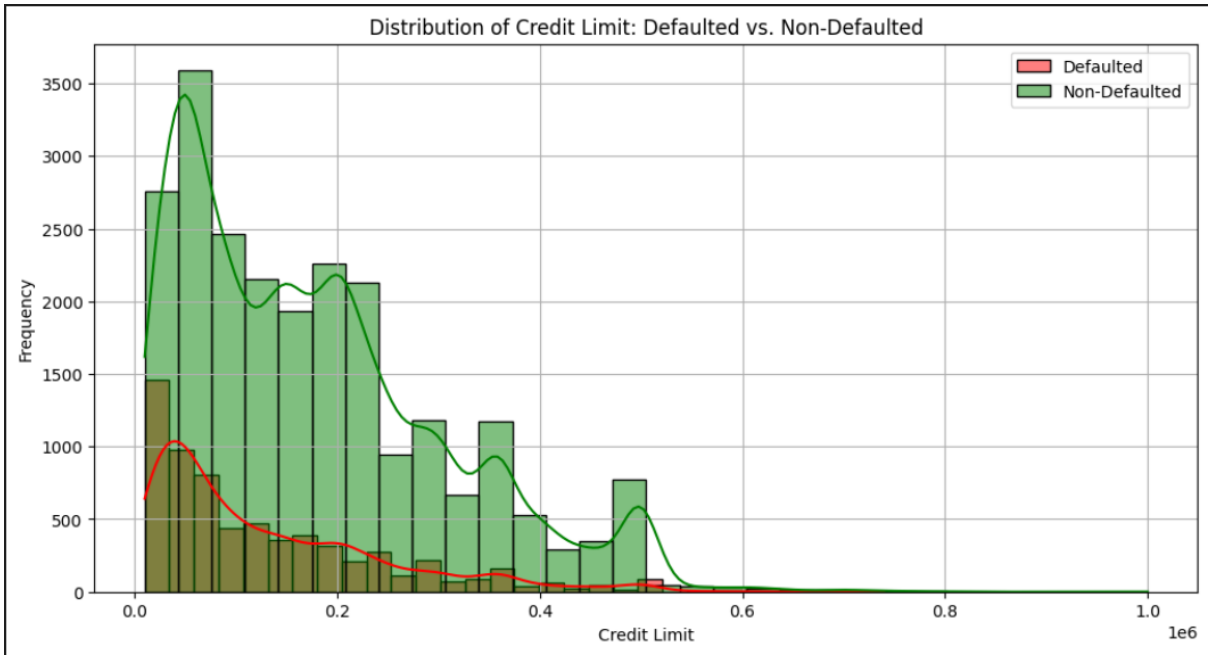
- Low risk: Regular repayment behavior and low balances.
- Medium risk: Occasional delays or moderately high balances.
- High risk: Frequent defaults Or significantly high balances.



This image represents a pair plot showing relationships between three variables—LIMIT_BAL, AGE, and BILL_AMT1—segmented by a Risk_Cluster categorical variable (with values 0, 1, and 2).

Insights:

- Risk and Financial Profile:
  - Cluster 0 appears to represent financially stable customers with higher credit limits and bill amounts.
  - Cluster 2 likely represents customers with limited financial activity and higher risk, as indicated by lower credit limits and bill amounts.
  - Cluster 1 could represent a transitional group with moderate financial behavior.
- Targeted Strategies:
  - For Cluster 0, marketing efforts can focus on premium credit products and loyalty rewards.
  - For Cluster 2, strategies should focus on financial education, budgeting tools, and products designed to build credit history.
  - For Cluster 1, products that encourage greater financial engagement (e.g., increased credit limits or spending rewards) might be effective.



	Group	Mean	Median	Std Dev	Min	Max
0	Defaulted	130109.656420	90000.0	115378.540571	10000	740000
1	Non-Defaulted	178099.726074	150000.0	131628.359660	10000	1000000

Insights:

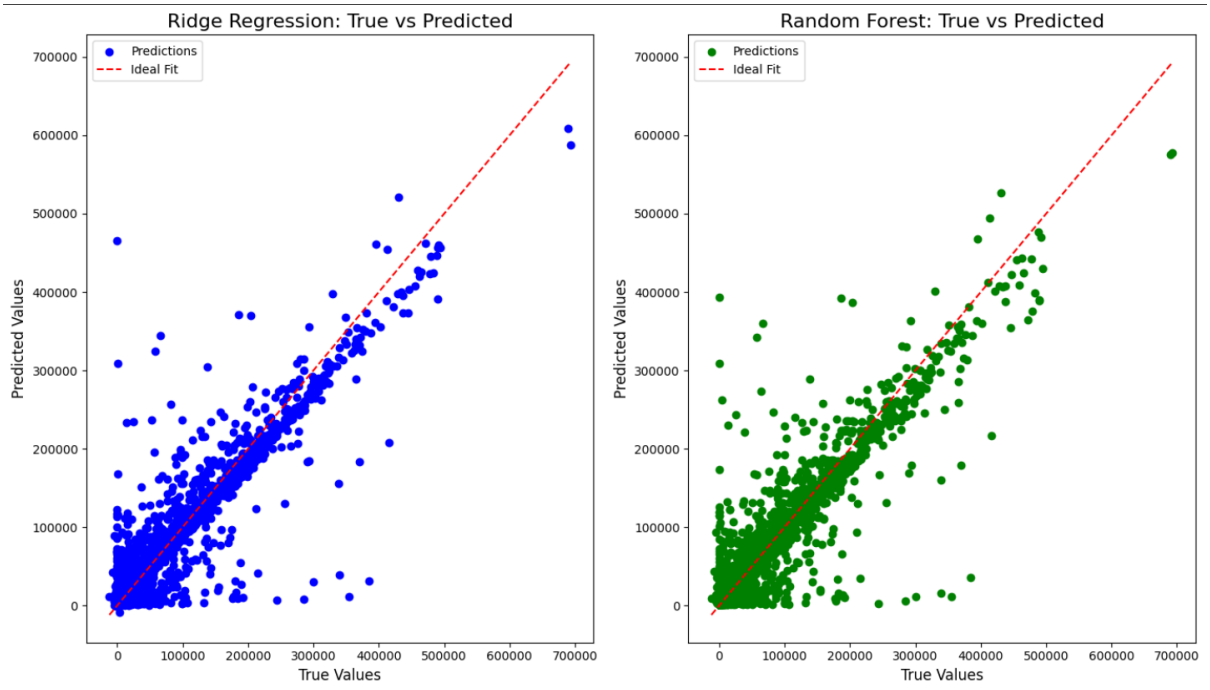
- Defaulted Customers: The distribution of credit limits is skewed towards lower values, indicating that customers with lower credit limits are more likely to default.
- Non-Defaulted Customers: The distribution is more spread out, with higher credit limits being common among non-defaulted customers.

4.4 Predicting Monthly Bill Amounts:

The choice of Linear Regression for predicting monthly bill amounts is reasonable given the nature of the problem and the simplicity of the model.

Why Use Linear Regression?

- **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.
  - In this case, features like LIMIT_BAL (credit limit) and BILL_AMT1 (previous bill amounts) likely have a direct and proportional impact on the monthly bill amount.




```
              Model           MSE
0  Ridge Regression  5.471846e+08
1     Random Forest  5.836739e+08
Ridge Regression is better with lower MSE.
```

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.

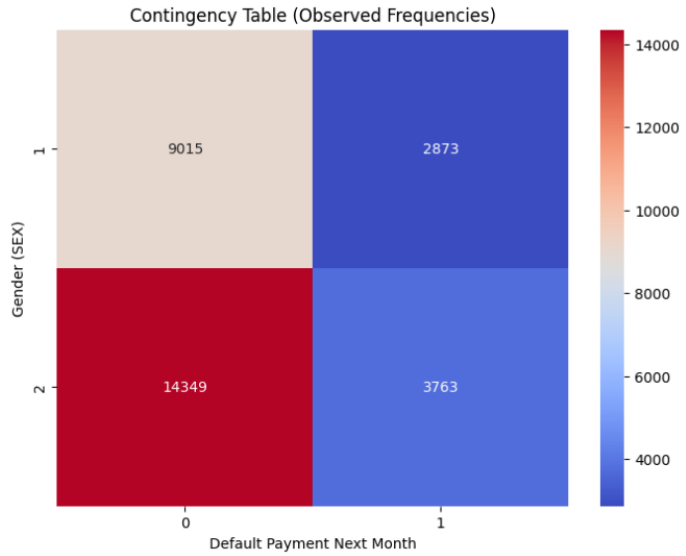
- High MSEs suggest that:
  - Some key features might be missing.
  - The model might not fully capture relationships between features and the target variable.
  - There could be noisy or irrelevant features affecting performance.\

4.5 Gender and Default Probability

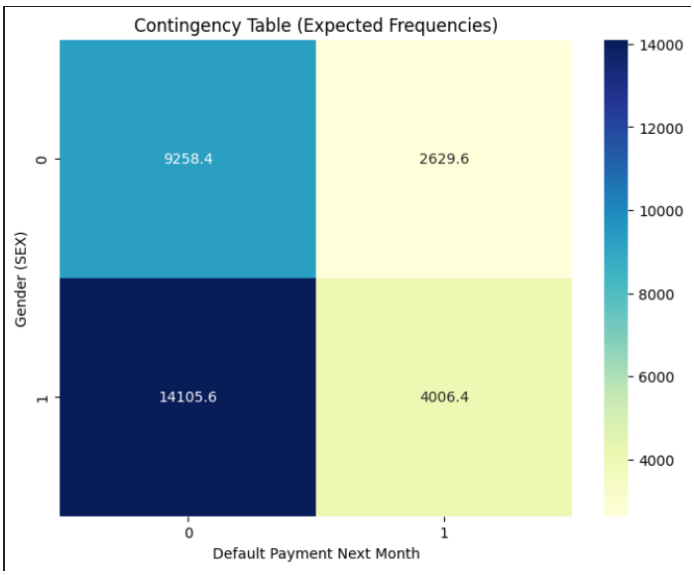
Using the Chi-Square test to examine the relationship between gender and default probability because it is designed to test associations between categorical variables.

```
Chi-Square Statistic: 1.980728826471716, p-value: 0.1593136474775643
No significant relationship between gender and default probability.
```

CONTINGENCY TABLE (Observed Frequencies)



CONTINGENCY TABLE (Expected Frequencies)



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

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

Interpretation

- **p-value > 0.05:**
  - Since the p-value is greater than the significance level (commonly set at 0.05), you fail to reject the null hypothesis. This means:
  - There is no statistically significant relationship between gender and default probability in your dataset.

- **Chi-Square Statistic (1.98):**
    - The low Chi-Square value indicates that the observed frequencies in the contingency table are not substantially different from the expected frequencies under the assumption of independence.

Insights

- Gender does not appear to be a significant factor in predicting whether someone defaults on payment, based on this test.
- While gender might not directly influence default probability, there could be interactions with other factors (e.g., income, age, or credit limit) that could indirectly affect default risk.

4.6 Conclusion:

The analysis provides valuable insights into customer behaviour and credit risk, and it highlights 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 Behaviour:** 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 behaviour 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 behaviour.

4.7 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.

5. 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.

5. Dataset Repository

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

Github Repository

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.