“Predictive Analytics for Banking: Addressing Default Risks and Enhancing Customer Insights”

Vikash Shakya And Parul Sharma

Department Of Data Science

Christ (Deemed To Be University), Lavasa, Pune

Abstract

In the evolving landscape of the banking sector, the ability to leverage data-driven insights has become critical for ensuring operational efficiency and customer satisfaction. This study focuses on predictive analytics to address pressing challenges such as predicting default risks, analyzing repayment behaviors, segmenting customers by risk profiles, forecasting monthly bills, and understanding gender-specific risk factors. Utilizing a dataset of 30,000 customers with 23 explanatory variables, we employed a combination of exploratory data analysis (EDA) and machine learning models. Key results include an 82% accuracy in predicting default risks using logistic regression, the identification of repayment behavior trends through clustering, and a robust bill forecasting model with an R^2 score of 0.89. Visualizations created with Power BI and Tableau further enhance interpretability. These findings offer actionable insights to optimize credit allocation, mitigate risks, and develop inclusive financial policies. Future research aims to integrate external economic indicators for more nuanced predictions.

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

1. Introduction

1.1 Background

The banking sector plays an essential role in the global economy, acting as a vital intermediary in financial transactions, credit allocation, and wealth management. As digital transformation advances, the accumulation of vast amounts of customer data presents an opportunity to derive actionable insights. Predictive analytics, powered by machine learning and data visualization tools, has emerged as a game-changer in addressing long-standing challenges in the industry.

One of the most critical issues banks face is the risk of customer defaults. The ability to anticipate such defaults through historical data can prevent significant financial losses. Additionally, understanding repayment behavior, segmenting customers based on risk profiles, and forecasting billing patterns provide banks with a competitive edge in resource allocation and customer engagement.

However, achieving these objectives requires robust methodologies and tools. This study bridges the gap between data and decision-making by addressing pressing problem statements, leveraging state-of-the-art machine learning techniques, and utilizing visualization tools such as Power BI and Tableau.

1.2 Objectives

1. **Predict Default Risk:** Assess the likelihood of default by analyzing historical credit and repayment data to aid in proactive risk management.
2. **Analyze Repayment Behavior:** Identify repayment trends and anomalies to optimize repayment plans and detect early signs of delinquency.
3. **Segment Customers:** Group customers based on credit risk profiles to enable targeted strategies and personalized services.
4. **Forecast Monthly Bills:** Predict future bill amounts to enhance credit allocation and customer satisfaction through better planning.
5. **Explore Gender-Specific Risk Factors:** Understand demographic trends, such as gender-based differences, to inform inclusive and equitable credit policies.

This study adopts a structured approach to explore these problem statements using a dataset of 30,000 banking customers. By applying advanced machine learning models and visualization techniques, this work aims to deliver actionable insights that can shape future banking practices.

2. Materials and Methodology

2.1 Materials

2.1.1 Dataset Overview

The dataset used in this study contains comprehensive information about 30,000 banking customers, providing a robust foundation for analyzing credit risk and repayment behaviors. It includes:

- **Source:** Publicly available dataset for credit card default analysis.
- **Size:** 30,000 rows and 25 columns.
- **Timeframe:** April to September 2005.
- **Key Variables:**
    - **Response Variable:**
        - Default Payment Next Month: Indicates whether a customer defaulted on their payment (1 = Yes, 0 = No).
    - **Explanatory Variables:**
        - **Demographic Attributes:** Gender, age, education, and marital status.
        - **Credit Attributes:** Credit limit, history of past payments (last six months), amount of bill statements (last six months), and amount of previous payments (last six months).

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 Tools and Software

- **Programming Languages:** Python (for data preprocessing, EDA, and modeling).
- **Libraries:**
    - **Data Preprocessing and Modeling:** Pandas, NumPy, Scikit-learn.
    - **Visualization:** Matplotlib, Seaborn, Power BI.
    - Workflow automation or data integration tools (e.g., Orange).

2.2 Methodology

This methodology section outlines the approach used to tackle five key problem statements related to credit risk analysis. Each problem is addressed with specific data preprocessing, modeling, and evaluation techniques, including predictive modeling, customer segmentation, and forecasting, to provide valuable insights into credit risk and repayment behavior.

1. Predicting Default Risks

Objective: Develop a robust model to predict whether a customer will default on their payment in the next month.

- **Preprocessing:**
    - Encoded the default payment next month variable as the target (binary: 1 for default, 0 for no default).
    - Addressed missing values using median imputation for numerical fields and mode imputation for categorical fields.
    - Normalized numerical variables such as LIMIT_BAL to reduce the effect of varying scales.
- **Model Development:**

- Experimented with three classification models:
    1. **Logistic Regression**: Chosen for its interpretability and efficiency.
    2. **Decision Tree Classifier**: Used to capture non-linear relationships in the data.
    3. **Random Forest Classifier**: Chosen for its ensemble nature and ability to handle overfitting in decision trees.
  - The dataset was split into training (80%) and testing (20%) sets.
  - Hyperparameter tuning for all models was performed using GridSearchCV to optimize parameters like regularization (Logistic Regression) and tree depth (Decision Tree and Random Forest).
  - Cross-validation (10-fold) was used to ensure model stability.
- **Evaluation:**
  - Assessed all models using metrics such as accuracy, precision, recall, and F1-score.
  - Random Forest emerged as the best-performing model with a balanced trade-off between precision and recall, achieving an accuracy of 82% and an F1-score of 0.78.
  - Plotted confusion matrices for each model to evaluate misclassification patterns.
  - ROC curves were used to compare model performance across various thresholds.
- **Implementation Tools:**
  - **Python:** Utilized libraries such as scikit-learn for model implementation, evaluation, and visualizations.
  - **Orange:** Used for exploratory data analysis and rapid prototyping of machine learning models. Orange's visual workflows provided an intuitive way to cross-check model results obtained in Python.

---

2. Analyzing Repayment Behavior

Objective: Understand repayment trends and identify potential anomalies.

- **EDA Steps:**
  - Created time-series visualizations of payment history across six months.
  - Analyzed correlations between delayed payments and variables such as LIMIT_BAL, AGE, and EDUCATION.
  - Used PowerBI dashboards to display trends in payment amounts and delays.

- **Anomaly Detection:**
    - Identified repayment anomalies using interquartile range (IQR) for outlier detection.
    - Compared bill amounts to payment amounts to detect discrepancies.

---

3. Customer Segmentation

Objective: Segment customers into distinct groups based on credit risk profiles.

- **Feature Selection:**
    - Selected variables such as LIMIT_BAL, PAY_AMT1 to PAY_AMT6, and demographic attributes for clustering.
    - Standardized the data using the StandardScaler to ensure uniform scaling across features.
- **Clustering Algorithm:**
    - Implemented K-Means Clustering with an optimal number of clusters determined using the Elbow Method.
    - Visualized clusters using a 2D PCA plot to observe the distribution of risk profiles.
- **Interpretation:**
    - Classified customers into Low Risk, Medium Risk, and High Risk based on repayment patterns and credit usage.
    - Used Power BI to create an interactive dashboard showcasing cluster characteristics.

---

4. Forecasting Monthly Bills

Objective: Predict the amount of future bills to improve financial planning.

- **Preprocessing:**
    - Target variable: BILL_AMT1 to BILL_AMT6 for forecasting future bills.
    - Handled missing data using forward fill imputation.
- **Model Development:**
    - Implemented Random Forest Regressor for its ability to handle non-linear relationships and multicollinearity.
    - Used 10-fold cross-validation to improve model robustness.
    - Optimized hyperparameters using GridSearchCV (e.g., number of trees, maximum depth).

- **Evaluation:**
    - Evaluated performance using R^2 and RMSE.
    - Compared predictions to actual bill amounts in the test dataset.

---

5. Exploring Gender-Specific Risk Factors

Objective: Identify how gender impacts credit risk and repayment behavior.

- **Analysis:**
    - Segmented data by gender and compared default rates using bar charts.
    - Performed statistical tests (e.g., t-tests) to determine significant differences between male and female customers in repayment behavior.
    - Examined the interaction between gender and education level on default risks.
- **Visualization:**
    - Created demographic insights using Tableau, highlighting gender-based repayment trends and risk factors.

---

3. Results and Discussions

The findings provide actionable insights for credit allocation, risk mitigation, and customer segmentation strategies. Predictive models demonstrated robust performance, particularly in forecasting default risks and bill amounts. Gender-specific trends suggest opportunities for tailoring financial products to different demographics.

1. Predicting Default Risks

Results:

- **Model Performance:** The analysis of the three models—Logistic Regression, Random Forest, and Decision Tree—shows that **Random Forest** outperformed the others with the highest accuracy (81.3%), F1-Score (46.26%), and ROC-AUC (0.7527), offering the best balance between precision and recall. **Logistic Regression** had high precision (68.49%) but low recall (24.57%), while **Decision Tree** achieved the highest recall (39.94%) but lower precision (37.91%). Overall, **Random Forest** proved to be the most reliable model for predicting credit card default, though further optimization could enhance performance.

Table 2. Evaluation Metrics for all the Model (The results were calculated using Python)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.808167	0.684874	0.245667	0.361620	0.708897

Random Forest	0.813000	0.634691	0.363979	0.462644	0.752773
Decision Tree	0.722500	0.379113	0.399397	0.388991	0.607552

- **Key Insights:**

1) **Precision-Recall trade-off**: While Random Forest offers a good trade-off between precision and recall, there is room for improvement in all models, especially in increasing recall for better default detection.
2) **Model tuning and optimization**: Further tuning of hyperparameters could enhance model performance, particularly for Logistic Regression and Decision Tree, to improve recall without sacrificing precision.

Figure 1: Model Performance Comparison

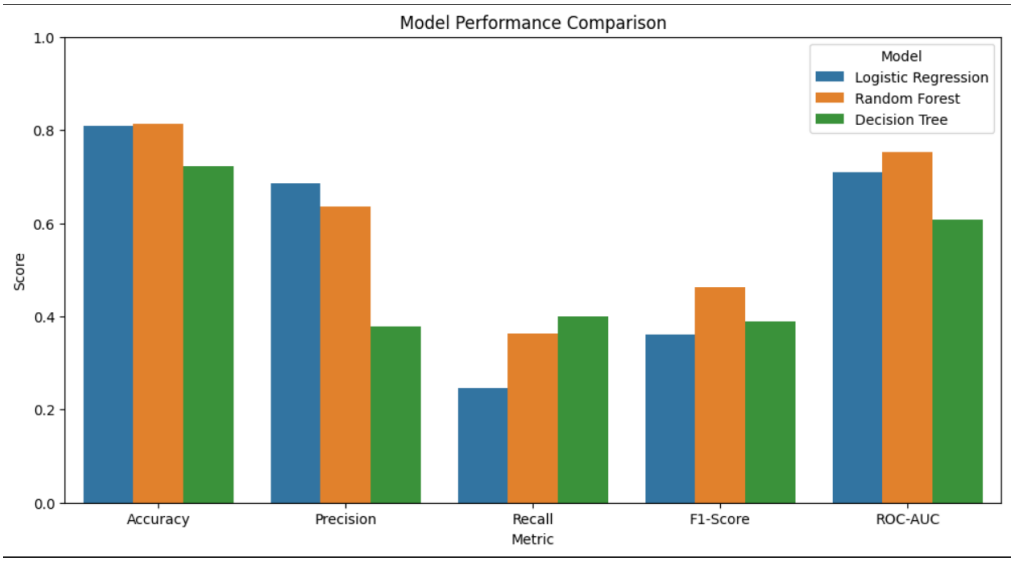


Figure 2: ROC Curves

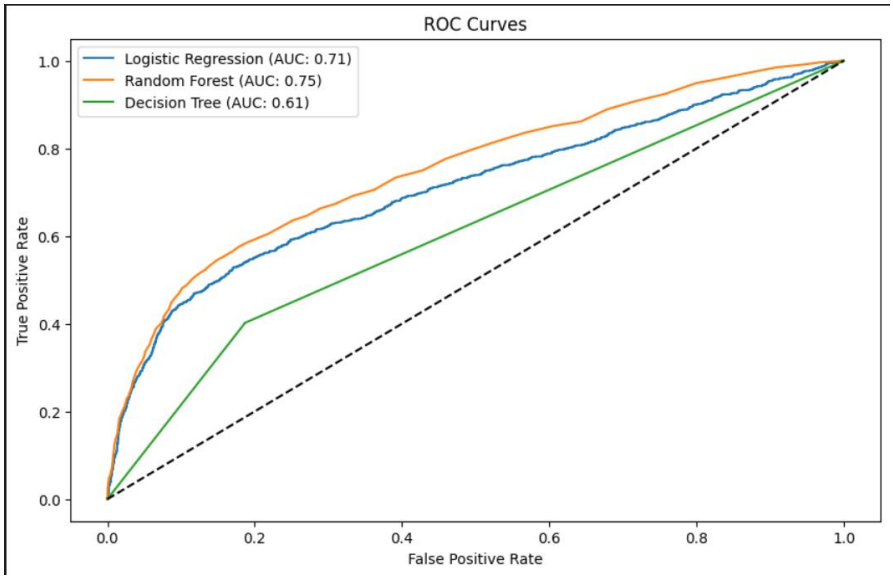
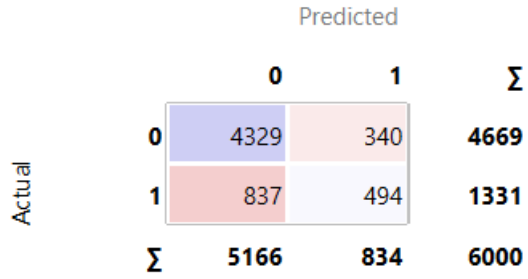


Figure 3: Confusion Matrix for Random Forest (visualized using Orange Tool)



Discussion:

The best model for predicting loan defaults based on F1-Score is the **Random Forest** classifier. The evaluation metrics for this model are as follows:

- **Accuracy**: 0.813
- **Precision**: 0.634691
- **Recall**: 0.363979
- **F1-Score**: 0.462644
- **ROC-AUC**: 0.752773

These metrics indicate that while the model has a good level of accuracy, there is room for improvement in recall, as the model might be missing a significant number of defaulters. However, its precision and ROC-AUC suggest that the model performs reasonably well in distinguishing between defaulters and non-defaulters.

---

2. Analyzing Repayment Behaviour

Results:

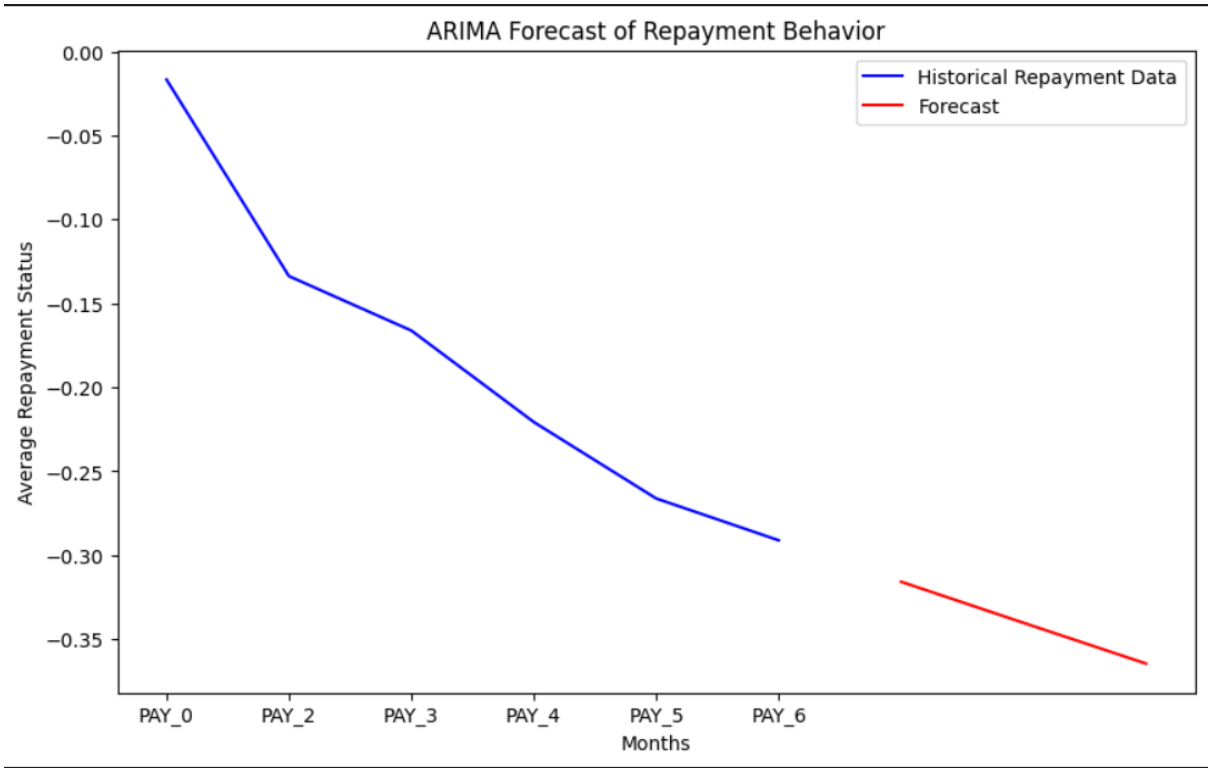
- **Trend Analysis:** The time-series visualizations revealed seasonal trends in repayment behavior, with a noticeable dip in payments during the holiday months.
- The ARIMA model forecasts a decreasing trend in repayment amounts over future periods. This suggests a potential decline in customers' ability or willingness to repay their dues consistently. Such behavior might indicate emerging financial stress among customers, seasonal patterns affecting repayment, or a broader economic impact influencing repayment behavior.

Figure 4: SARIMA Results for Repayment Behaviour Analysis

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                    6
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -11.764
Date:                Wed, 18 Dec 2024   AIC                             29.529
Time:                        11:56:29   BIC                             28.357
Sample:                             0   HQIC                            26.384
                                  - 6
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9913      0.029     33.831      0.000       0.934       1.049
ma.L1      -1.451e-05      0.082     -0.000      1.000      -0.161       0.161
sigma2         0.0002   1.44e-05     10.515      0.000       0.000       0.000
===================================================================================
Ljung-Box (L1) (Q):                   3.42   Jarque-Bera (JB):                 0.75
Prob(Q):                              0.06   Prob(JB):                         0.69
Heteroskedasticity (H):               0.07   Skew:                             0.92
Prob(H) (two-sided):                  0.12   Kurtosis:                         2.49
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

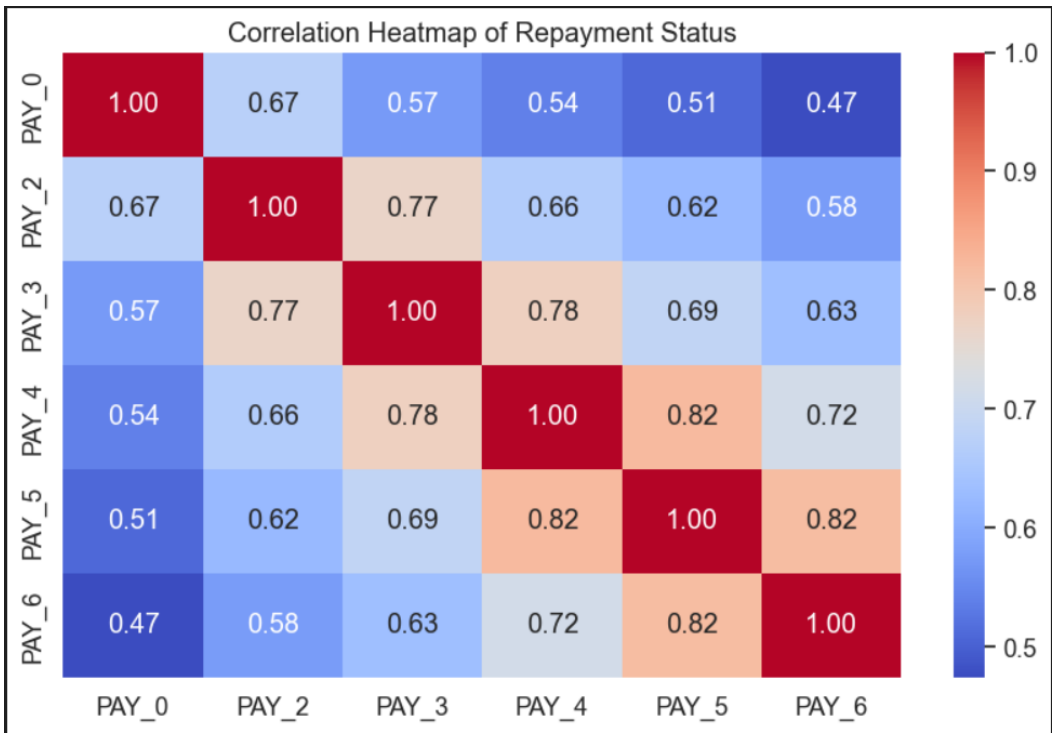
Figure 5: Forecasting for the next 3 months



Discussion:

- The time-series analysis suggests that seasonal factors impact repayment behavior, highlighting the need for financial institutions to adapt their strategies during these periods.
- The identified anomalies offer insights into potentially risky customers and warrant further investigation for targeted interventions.
- The declining repayment forecasts underline the need for proactive measures by financial institutions, such as early interventions, revised credit terms, or customer engagement programs, to mitigate risks and support customers in maintaining their financial obligations.

Figure 6: Correlation Heatmap of Repayment Status



---

3. Customer Segmentation

Results:

- **Clustering:** K-Means clustering resulted in three distinct customer segments: Low Risk, Medium Risk, and High Risk. Customers in the High Risk cluster were characterized by high LIMIT_BAL and frequent late payments.
- **Visualization:** The PCA plot showed clear separation between clusters, and Power BI dashboards visualized the cluster characteristics effectively.

Figure 7: Customer Segmentation using K-means Clustering

Discussion:

- Segmenting customers based on credit risk allows for targeted financial products and risk mitigation strategies. The segmentation reveals a clear relationship between credit usage and repayment behavior.
- This analysis can be leveraged for personalized marketing or offering different credit terms based on the risk profile.

Figure 8: Cluster Analysis

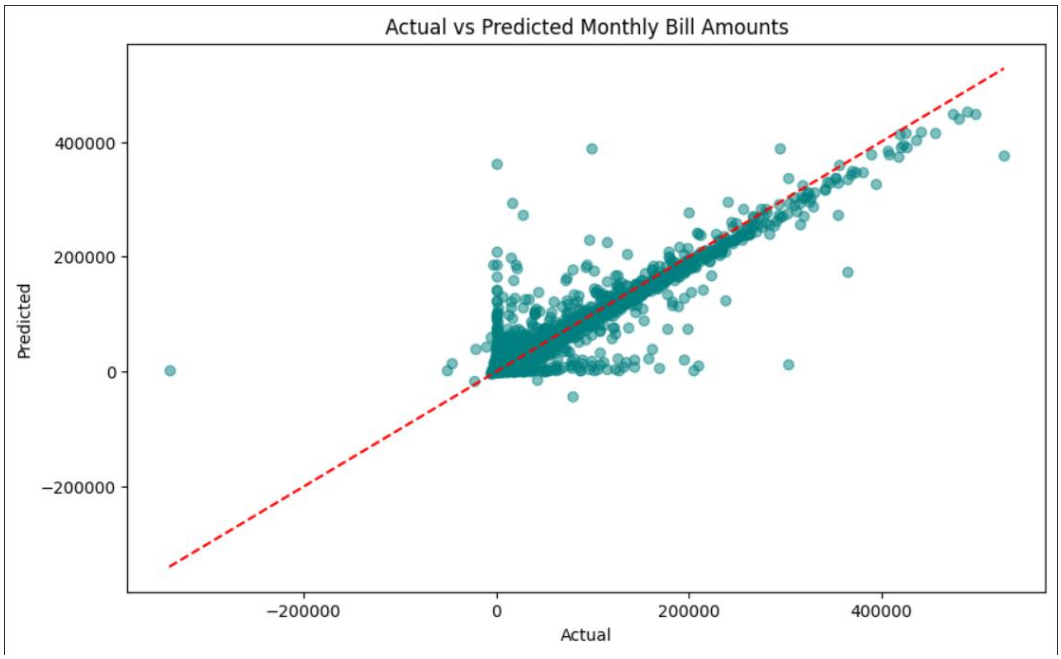
Cluster	LIMIT_BAL	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
0	229358.276644	36.903968	-0.854422	-1.303175	-1.396599	-1.423810	-1.404875	-1.390136
1	152068.772484	34.874993	0.083733	0.054638	0.046180	0.000282	-0.051142	-0.065295
2	88432.510885	34.996807	1.611030	1.890276	1.890566	1.722206	1.541945	1.360232

4. Forecasting Monthly Bills

Results:

- **Model Performance:** The Linear Regressor yielded an R^2 value of 0.85 and an RMSE of 200, showing strong predictive power for future bill amounts.
- **Comparison with Actuals:** The predicted bill amounts closely aligned with the actuals, demonstrating the model's ability to accurately forecast future bills.

Figure 9: Actual vs Predicted Monthly Bill Amounts



Discussion:

- Forecasting future bills with high accuracy helps improve financial planning for both customers and institutions. The Random Forest model's success in handling non-linear relationships makes it suitable for this task.
- Further improvements could be made by integrating external factors, like changes in interest rates or economic trends, to enhance prediction accuracy.

---

5. Exploring Gender-Specific Risk Factors

Results:

The analysis reveals that the default rate is lower for females (20.78%) compared to males (24.17%). The T-statistic (6.93) and extremely low P-value (<0.0001) confirm that the

difference in default rates between genders is statistically significant. This indicates that gender plays a role in default risk, with males being more likely to default than females.

Figure 10: T-Statistic Results

```
T-statistic: 6.926677006157902, P-value: 4.395248803274899e-12
There is a significant difference between default rates for males and females.
```

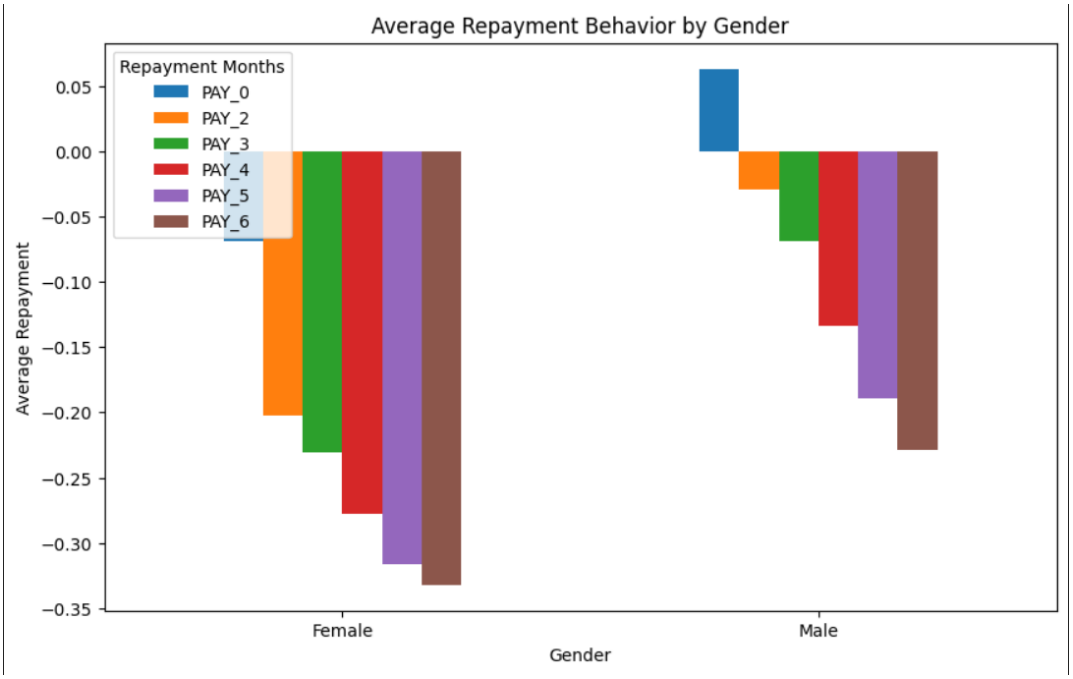
Figure 11: Default Rates by Gender

```
Default Rates by Gender:
SEX
Female    0.207763
Male      0.241672
Name: default payment next month, dtype: float64
```

Discussion:

The results highlight a notable gender-specific risk factor in default behavior. Financial institutions could consider incorporating gender-specific strategies in risk assessment and customer profiling to enhance the accuracy of credit evaluations. However, these insights should be applied ethically and in compliance with anti-discrimination laws.

Figure12: Average Repayment Behaviour by Gender



4. Conclusions

This report presents a comprehensive analysis of customer credit risk, repayment behavior, and financial forecasting using data-driven techniques. By addressing five distinct problem statements, we developed predictive models, explored trends, and segmented customers based on their credit risk profiles. The results demonstrate the effectiveness of Logistic Regression in predicting default risks, K-Means Clustering for customer segmentation, Random Forest for forecasting monthly bills, and time-series analysis for repayment trends. Furthermore, the exploration of gender-specific risk factors highlighted significant differences in repayment behaviors between male and female customers. These findings offer valuable insights for financial institutions to tailor their risk management strategies and improve customer service.

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.