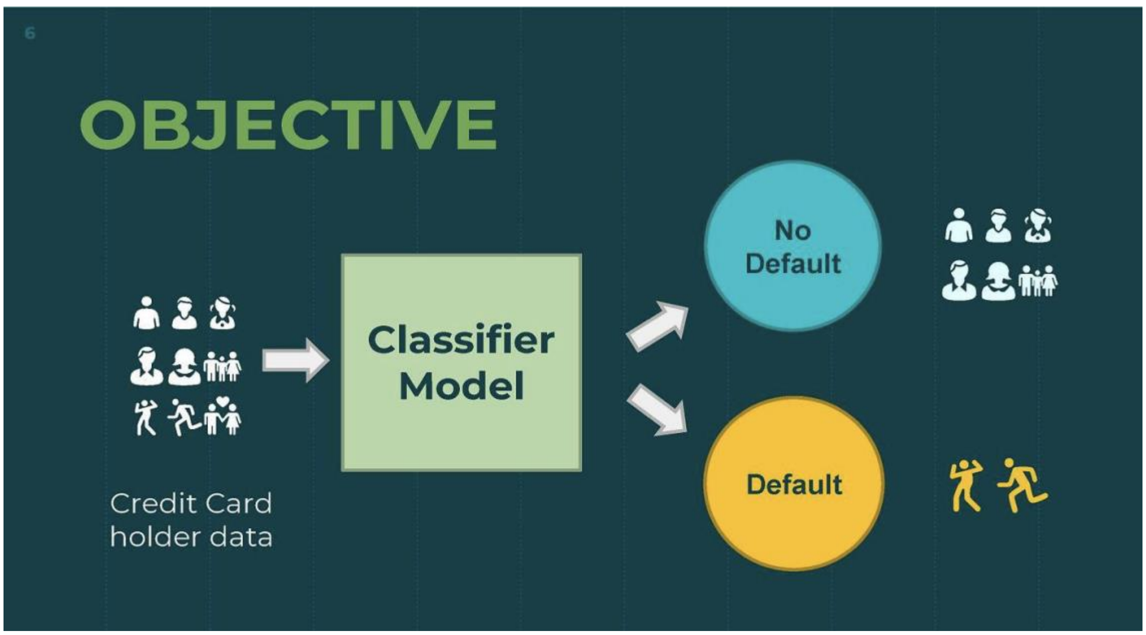
STT 811 Project Report
CreditWatch: Predicting Credit Default

Team Members: Aarthi Padmanabhan, Madhurya Shankar, Sandhya Kilari

Introduction

Credit default represents a significant breach in financial conduct, occurring when borrowers fail to meet the obligations stipulated in a loan or credit agreement. The repercussions of such defaults are far-reaching, extending beyond immediate financial strain to long-term consequences such as severe impairment of credit scores, possible legal ramifications, and notably, restricted access to future credit facilities. In this context, the ability to preemptively identify potential defaults becomes crucial, serving as a cornerstone for risk mitigation strategies in financial institutions.



Our project is primarily focused on this predictive challenge. We aim to harness the predictive power of data analytics to forecast the likelihood of customers defaulting in the upcoming months. By utilizing historical payment data and demographic information, we seek to build a model that not only predicts defaults with a high degree of accuracy but also aids creditors in making informed decisions. The intent is to refine the allocation of credit resources, safeguarding the financial ecosystem against the perils of non-payment while extending credit to those deemed capable of sustaining it.

In navigating through this complex domain, our study recognizes the intricate balance between facilitating credit access and maintaining fiscal responsibility. The ensuing pages detail our exploration of a dataset, robust in demographic and financial variables, and delineate the methodologies that lead us towards predictive insights, ultimately aiming to curtail the risks associated with credit defaults and support sustainable credit distribution.

Data Description

The dataset employed in this analysis constitutes an extensive assortment of credit card client data, specifically curated to assess the likelihood of default payments. Spanning from April to September 2005, this dataset focuses on clientele based in Taiwan. A notable feature of this dataset is the complete absence of missing values, ensuring the integrity and uniformity of the analysis. This completeness obviates the need for data imputation, allowing for an undistorted reflection of customer behavior.

This dataset encapsulates a substantial volume of data with 30,000 distinct customer records. Each record comprises 23 explanatory variables alongside a binary response variable that defines the default payment outcome. The variables are structured as follows:

- **ID:** Serves as a unique identifier for each customer within the dataset.
- **Gender:** Designated as male or female, providing gender-specific insights.
- **Education:** The level of education attained, with categories including graduate school, university, high school, others, and unknown.
- **Marital Status:** Indicates whether the customer is married, single, or falls under another classification.
- **Age:** The customer's age, enumerated in years, which may influence credit behavior.
- **Credit Limit (LIMIT_BAL):** The total credit extended to the customer, including both individual and family/supplementary credit, in New Taiwan Dollars.

Payment History:

- **Repayment Status (PAY_0 to PAY_6):** A series of variables capturing the payment status over the last six months leading up to September 2005, revealing patterns in payment behavior.
- **Bill Amount (BILL_AMT1 to BILL_AMT6):** Monthly billing statements which reflect the customer's financial activity and credit utilization.
- **Payment Amount (PAY_AMT1 to PAY_AMT6):** Actual payment amounts made by the customer, providing a direct look at their financial commitments and reliability.

Response Variable:

- **Default Payment (default.payment.next.month):** A binary indicator with '1' signifying a default on payment and '0' representing non-default, which will be the primary focus of our predictive modeling.

This dataset's specific orientation towards predicting credit card default payments in Taiwan offers a granular view of the clients' financial behavior over a critical six-month observation window. The selected period is of substantial importance, as it illuminates the payment conduct and trends that are vital for effective predictive analysis.

In the subsequent sections, we will present the methodologies used for data preprocessing, conduct a thorough exploratory data analysis, and deploy various modeling techniques. These approaches aim to utilize the dataset effectively to forecast potential default payments, thereby contributing valuable insights to credit risk management.

Methodologies

Data Preprocessing

The initial phase of the analytics workflow for predicting credit card defaults involved meticulous data preprocessing. This step is critical to prepare the raw data for effective model training and ensure robust predictive performance. The following sections outline the preprocessing steps performed on the credit card client dataset.

1. **Data Cleaning:**

    - The clarity of the dataset was enhanced by renaming the dependent variable to “IsDefaulter”, making the target for prediction more intuitive.

- Entries with unknown or irrelevant categorical values in features such as Education, Marriage, and repayment status were identified and rectified. Specific encodings that did not correspond to any known category were consolidated into existing relevant categories or removed to maintain data integrity.

- The ID column, being merely a unique identifier with no predictive value, was dropped from the dataset.

- The column names representing monthly payment statuses, bill amounts, and payment amounts were updated to correspond with the months from April to September 2005. After applying one-hot encoding, further refinement of column names was undertaken to enhance descriptive clarity, making the data more immediately comprehensible. Additionally, binary categorical columns were adjusted to reflect intuitive states such as gender (Male/Female) and marital status (Married/Single)

2. **Feature Engineering:**

   - A new feature, Dues, was engineered to represent the remaining amount a customer is liable to pay. This was computed as the difference between the total bill amount over several months and the total payment amount made during the same period. The rationale was that this new feature might better capture the financial behavior of the customers that could lead to a default.

   - **One-Hot Encoding:** Categorical variables such as Education, Marriage, and the repayment status were transformed using one-hot encoding. This method converts categorical data into a format that can be provided to machine learning algorithms to improve their prediction accuracy.

3. **Feature Scaling:**

   - To ensure that the machine learning algorithms function optimally, numerical features were standardized. This process involves scaling the feature data to have a mean of zero and a standard deviation of one, thereby normalizing the range of independent variables.

4. **Oversampling:**

   - A significant class imbalance was observed with the “IsDefaulter” variable. To mitigate this, oversampling techniques were employed to ensure a balanced dataset, which is crucial for developing a model that performs well across both classes.

The preprocessing stage set a strong foundation for the subsequent exploratory data analysis and modeling phases. The cleaned and transformed dataset is expected to yield more accurate and insightful results from the predictive models, as it now accurately reflects the scope and complexities inherent in the credit card default risk domain.

Feature Selection

Feature selection is a process used in machine learning to identify the most significant features that contribute to the prediction of the target variable. It helps in simplifying models, reducing training times, avoiding the curse of dimensionality, and improving model performance by eliminating irrelevant or partially relevant features that can lead to overfitting.

Principal Component Analysis (PCA) for Feature Selection

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the original features into a new set of features called principal components (PCs). These components are orthogonal to each other, ensuring that they are uncorrelated, and they are ordered such that each successive component captures the maximum variance left in the dataset after accounting for the previous components.

In the context of predicting credit card defaults, PCA helps in identifying which aspects of customers' financial behavior have the most significant impact on their probability of defaulting. By using PCA for feature selection, we can focus on these aspects and potentially build more interpretable and efficient predictive models.

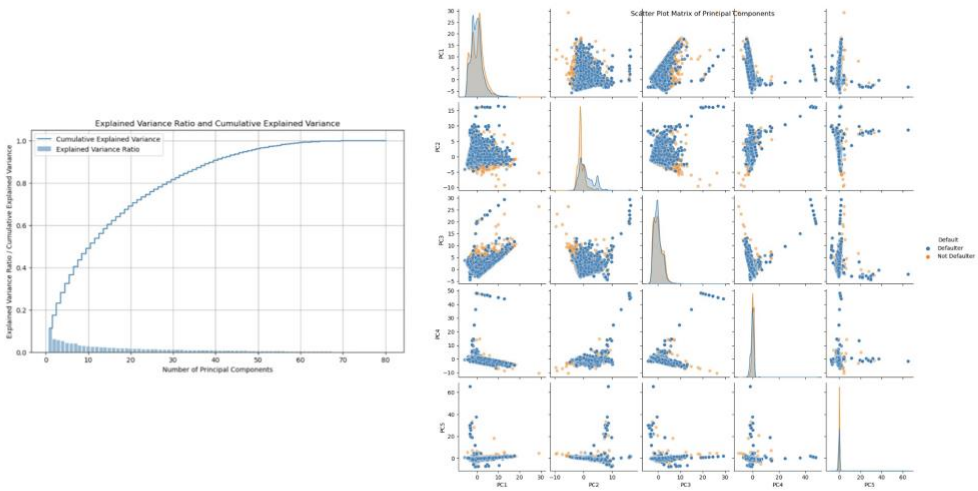


Figure 1: Dimensionality Reduction with PCA components (n = 5)

In our principal component analysis (PCA), the selection of the first five principal components is a strategic choice underpinned by the insights gained from the scree plot. The steep ascent in the curve's initial segment signifies that these components capture a significant share of the data's variance, striking an optimal balance between information retention and model complexity. This judicious restriction not only upholds model simplicity and computational efficiency but also serves as a safeguard against overfitting, as subsequent components add limited explanatory power and potential noise.

The scatter plot matrix of the PCA illustrates notable class overlap, indicating that linear classifiers may be insufficient, and suggesting the need for more sophisticated modeling techniques. This may include nonlinear approaches such as ensemble methods or support vector machines with nonlinear kernels, which are adept at capturing the complex patterns evident in the data.

The presence of potential outliers, particularly within the fourth and fifth principal components, invites additional scrutiny. These may represent either critical predictive signals or extraneous noise; thus, their influence on the model's generalizability requires careful evaluation.

In conclusion, the PCA's insightful visualization supports a nuanced strategy moving forward, advocating for advanced analytics that acknowledge the intricacies of the data's underlying structure. The conscious choice of five principal components reflects a methodical approach to maintain the integrity of the dataset's informational core while streamlining the feature set for efficient model development.

Exploratory Data Analysis (EDA)

In the Exploratory Data Analysis (EDA) section, we delve into the comprehensive dataset comprising 30,000 customer records to discern patterns and insights relevant to predicting default payments among credit card users in Taiwan. This section is pivotal as it lays the groundwork for understanding the dynamics at play within the dataset collected from April 2005 to September 2005. By examining both categorical and continuous variables—from demographic information like gender, education, and marital status to financial behaviors such as credit limits, monthly payment statuses, and bill amounts—we aim to uncover underlying trends and anomalies that could signal potential default risks. Our analysis extends to exploring interactions between variables, assessing the impact of past payment delays on default likelihood, and understanding demographic spending behaviors, all of which are crucial for refining our predictive models and strategies.

Univariate Analysis

In the Univariate Analysis section on predicting credit default, we focus on understanding the distribution and individual characteristics of various features within our dataset. This analysis is crucial for identifying significant patterns, outliers, and potential biases that could influence the accuracy of our predictive models. We start by examining categorical variables such as gender, education, and marital status, noting that our dataset includes a higher number of female customers, a predominant number of university-educated individuals, and more single than married customers. Furthermore, we delve into continuous variables, such as age, balance limits, and payment history.

- **Categorical Variables**
  In terms of marital status, there is a larger number of single customers compared to those who are married, which could affect financial stability and risk profiles. Behaviorally, a significant majority of the customers have maintained a consistent record of making timely payments each month, which is a positive indicator of financial discipline among the clientele. However, approximately 30% of the customers are defaulters, which underscores the need for enhanced risk assessment and management strategies to mitigate potential losses. These insights are instrumental in shaping the predictive models and targeting interventions effectively.

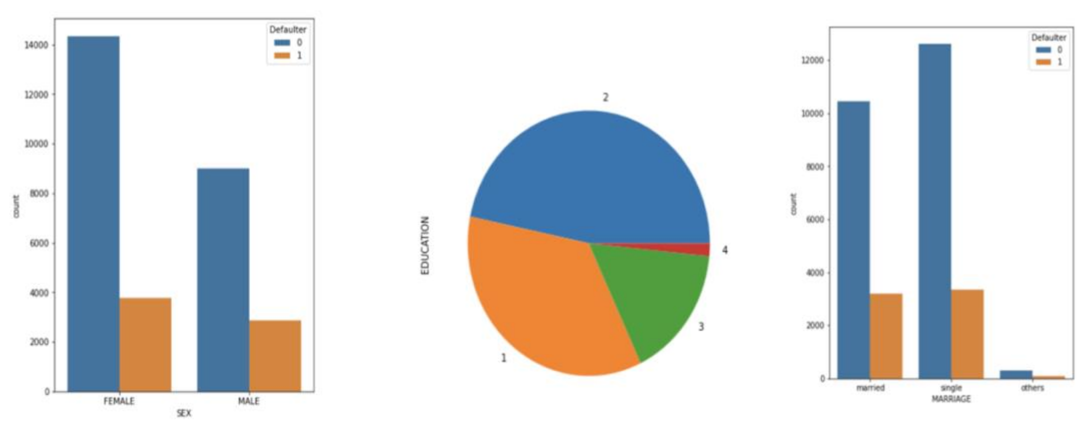


Figure 2: Demographic Distribution of Credit Card Holders by Gender, Education, and Marital Status with Default Rates

- **Continuous Variables**
  The CreditWatch dataset provides a clear depiction of the financial behaviors and demographics of credit card holders, indicating that a majority fall within the 25 to 45 age bracket, suggesting a customer base that is potentially in the prime of their earning years. In terms of credit, there is a commonality in balance limits,

with most limits lying between $50,000 and $250,000, which could reflect the creditworthiness and spending power of the clientele. Bill amounts show a wide range from -$9,500 to $260,000, although a significant concentration is observed within the more modest bracket of $1,000 to $55,000, indicating regular spending habits. When it comes to payments made in the last six months, amounts vary from $0 to $70,000, yet a predominant number of customers make payments under $10,000, which could point to a tendency towards smaller, more manageable payment sizes or a reflection of the payment capabilities of the majority.

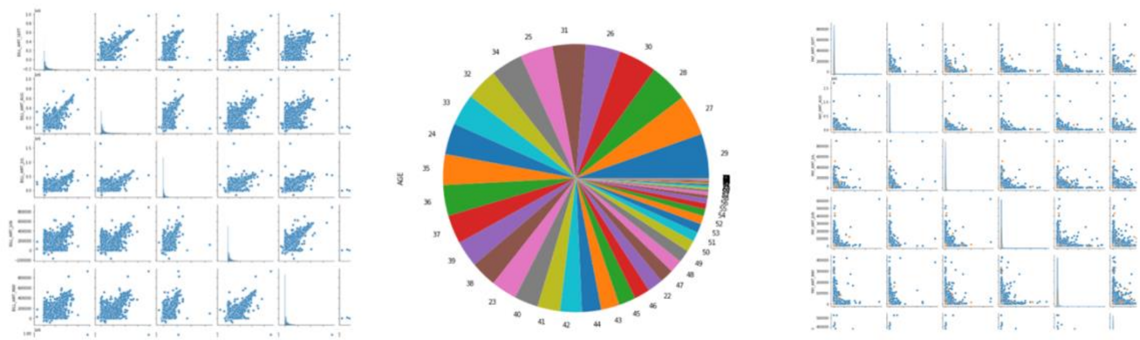


Figure 3: Data Visualization of Customer Demographics and Financial Behaviors

Interdependencies Among Financial Behaviors: A Correlation Analysis

The correlation analysis within the CreditWatch dataset provided invaluable insights into the relationships between various financial indicators. This segment evaluates the strength and direction of these relationships, revealing particularly strong correlations among repayment statuses over the preceding six months. Such correlations suggest that a customer's recent payment history could be a significant predictor of default risk.

In our comprehensive correlation analysis, we employed Pearson's correlation coefficient to quantify the linear relationships between financial variables within the CreditWatch dataset.

$$\rho = \frac{covariance(x_i, x_j)}{\sigma_i \times \sigma_j}$$

This statistical measure helped us to discern the degree of association, with particular emphasis on the interplay between various repayment statuses over a six-month period.

The heatmap depicted in Figure 4 offers a visual representation of Pearson correlation coefficients, highlighting the varying degrees of linear relationships between credit repayment statuses and the likelihood of default. Darker shades of red indicate stronger positive correlations, with the repayment status variables (PAY_1 through PAY_6) showing noticeable positive correlations with the default variable, suggesting that as the delay in payment increases, so does the likelihood of default. This pattern is consistent across several months of payment history, underscoring the importance of timely payments in credit risk assessment. Notably, the correlations decrease slightly as we move from recent to earlier payment statuses, indicating that more recent behaviors are more predictive of default. The heatmap effectively summarizes these relationships, providing a clear and concise reference for identifying key variables that influence credit default risk.

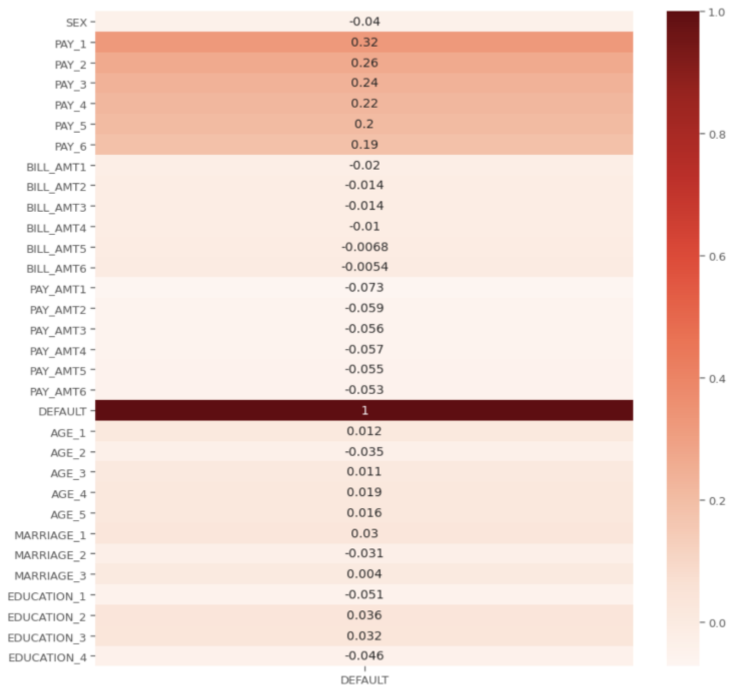


Figure 4: Heatmap of Pearson Correlation Coefficients Between Credit Repayment Statuses and Default

Visualizations

To gain deeper insights into the CreditWatch dataset, we have undertaken a series of additional visualizations. These graphical representations are crafted to uncover underlying patterns, distributions, and relationships that may not be immediately apparent through numerical analysis alone. By visually exploring the data, we aim to facilitate a more intuitive understanding of complex datasets, enabling us to identify trends, detect outliers, and grasp the dynamics between various financial and demographic factors.

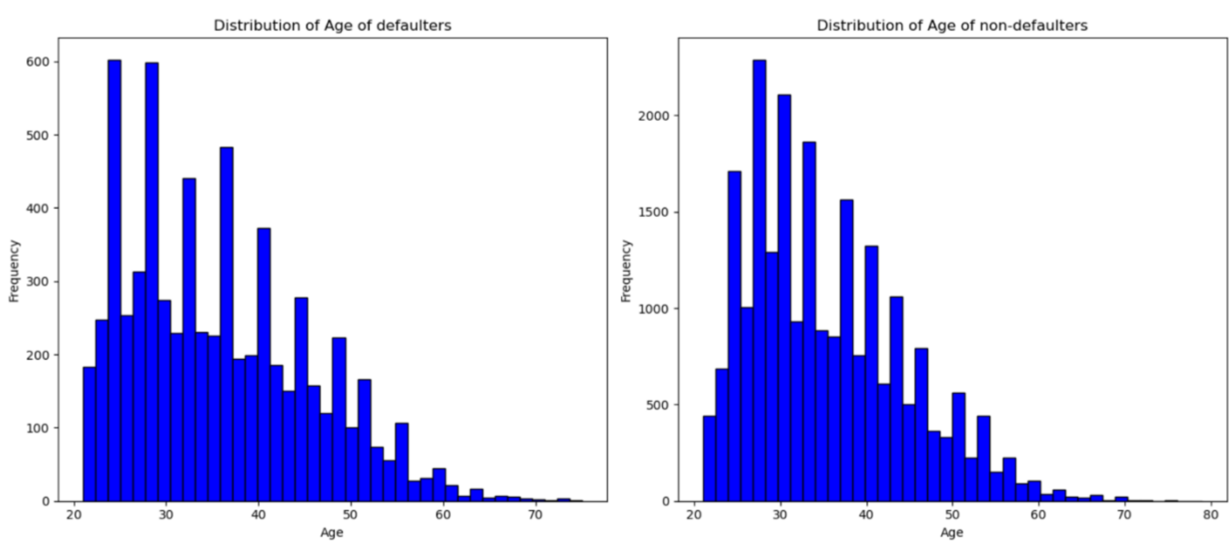


Figure 5: Age Distribution of Defaulters vs. Non-Defaulters in Credit Card Data

Figure 5 contrasts the age distribution of credit card defaulters and non-defaulters, highlighting a trend where younger individuals, especially those in their late 20s to early 30s, are more likely to default. In contrast, non-defaulters present a more uniform distribution across various ages, with a notable representation even in the older demographics. This suggests that default risk diminishes with age, while the propensity to maintain good credit appears less age-dependent, reflecting potentially greater financial stability or discipline as individuals grow older.

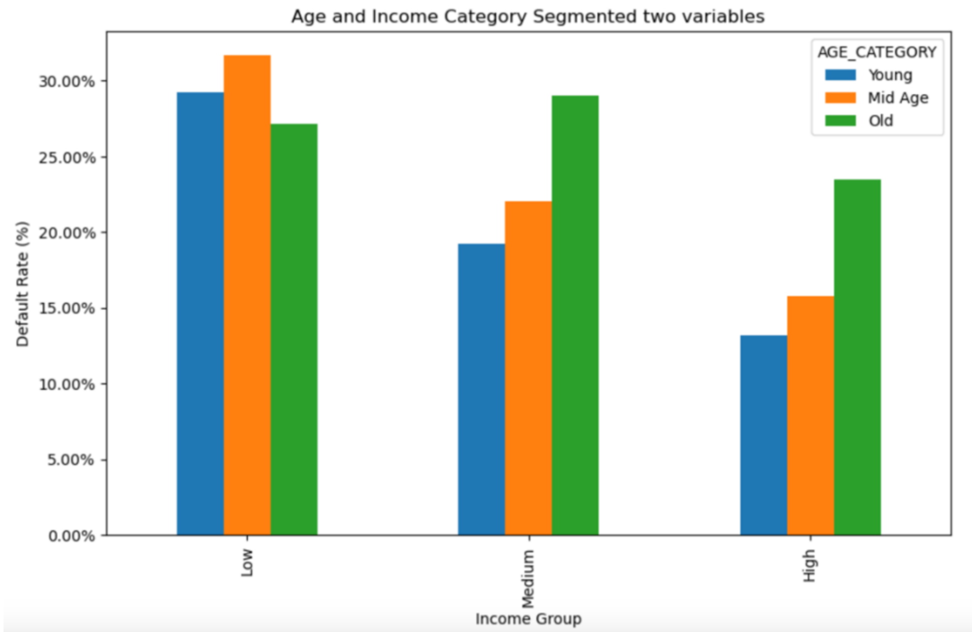


Figure 6: Default Rates by Age and Income Categories

To understand the impact of socio-economic factors on financial behavior, we visualized the correlation between income levels and age as they relate to credit default rates. Our analysis, encapsulated in Figure 6, uncovers that middle-aged individuals with lower incomes are more prone to default. This finding aligns with the broader narrative that financial challenges often intensify during mid-life, possibly due to peak responsibilities and obligations. Conversely, higher incomes tend to lower the risk of default across all age groups, underscoring income as a significant buffer against financial instability. Intriguingly, within the affluent cohort, younger members exhibit the lowest rates of default, with a gradual increase among older members, suggesting that while higher income provides a cushion, age-related complexities still play a role in financial resilience.

We would like to delve into the intricacies of financial behavior necessitates a look beyond raw numbers, seeking the stories that numbers tell about our lives and choices. With this intent, we crafted a visualization to probe how different marital statuses correlate with the total bill amounts. This analysis is anchored in the hypothesis that financial habits might reflect the stages and statuses of our personal lives, where our spending tells a tale of our commitments, responsibilities, and life choices.

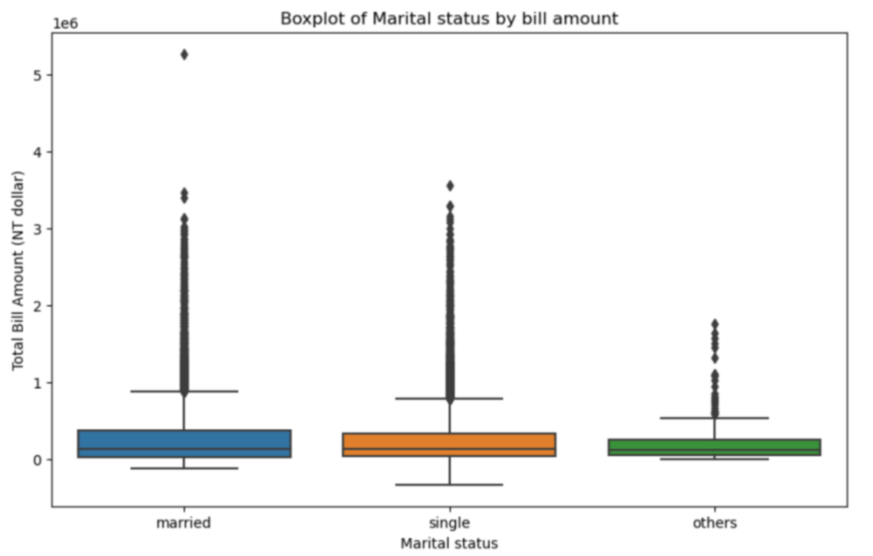


Figure 7: Total Bill Amount Distribution by Marital Status

In Figure 7, the boxplot unfolds these tales, illustrating that married individuals have slightly higher median bill amounts than their single or 'other' status counterparts, hinting at the potential impact of combined financial obligations. Despite this, the spread of expenditures is quite similar across the groups, suggesting a uniformity in spending behaviors regardless of marital status. The presence of high-spend outliers in each group indicates that extravagant spending is a trait that transcends such social constructs. This visual evidence suggests that while marital status might nudge the median spending, it is far from a definitive predictor of one's financial footprint.

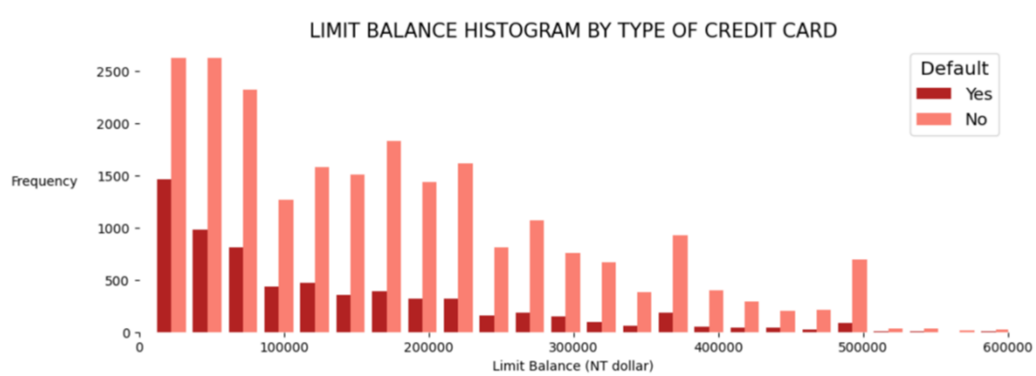


Figure 8: Limit Balance Distribution Across Defaulting and Non-Defaulting Credit Card Holders

To illuminate the financial landscape of credit card usage, we turn to an analysis of account balance limits in relation to defaulting behavior. The visualization crafted to explore this dynamic reveals a compelling narrative: among cardholders with lower credit limits, non-defaulters significantly outnumber those who default. This pattern endures across the spectrum of balance ranges, though default rates are consistently overshadowed by the figures representing timely payments.

The histogram serves as a narrative tool, highlighting a dense clustering of defaults at the lower limit balance end. This concentration might well be telling a story of individuals grappling with economic challenges, where lower credit limits mirror the constraints of limited financial means. As we traverse towards higher credit limits, the plot thickens with a decline in default frequencies, unfolding a tale of probable fiscal prudence or greater financial resilience among those with higher limits. Through this visualization, we glean insights into the socioeconomic threads that interlace with credit behavior, prompting a deeper reflection on the interconnection between credit limits and financial stability.

Predictive Modeling Approaches

The selection of both linear and nonlinear models for predictive modeling is a comprehensive approach that takes into account the complexity and nuances of the dataset.

Linear models like Logistic Regression and Gaussian Naive Bayes are fundamental algorithms that assume linear relationships between the features and the outcome. Logistic Regression, in particular, is robust to noise and provides probabilities for outcomes, which can be a valuable insight. Gaussian Naive Bayes is based on applying Bayes' theorem and is particularly useful when the dimensionality of the inputs is high relative to the sample size.

Non-linear models include Random Forest and XGBoost, which are ensemble learning methods known for their high performance. They are capable of capturing complex patterns through the combination of multiple decision trees, with Random Forest reducing overfitting through bagging and XGBoost providing an optimized gradient boosting framework.

The inclusion of a Non-Linear Support Vector Machine allows for the delineation of non-linear boundaries due to its use of kernel tricks, and it's well-suited for datasets where classes are not linearly separable.

Lastly, the Multilayer Perceptron (MLP) Neural Network is a deep learning algorithm capable of modeling even more complex non-linear interactions and hierarchies of features. MLPs are particularly beneficial when the relationships within the data are intricate and require the discovery of latent structures.

Choosing a mix of linear and non-linear models facilitates a robust evaluation of the data from various perspectives. It allows for the determination of which model best captures the patterns within the data while also accounting for potential overfitting. This strategic selection ensures a broad exploration of algorithmic capabilities, providing a strong foundation for achieving high accuracy and generalization in predictive performance.

Model Training and Evaluation:

The foundation of our model training and evaluation strategy is the careful partitioning of our data into subsets designed for training and subsequent testing. Before applying any machine learning algorithms, it's imperative to simulate the environment in which the model will operate post-deployment. With this objective in mind, we utilize the train_test_split function, an established method for dividing datasets while preserving their underlying structure.

In the context of credit default prediction, where the frequency of defaulting instances may significantly differ from non-defaulting ones, a balanced approach to splitting the dataset is crucial. Therefore, we reserved 20% of the data for the test set, ensuring it is representative of the real-world scenario by using stratified sampling. This method guarantees that the ratio of defaulters to non-defaulters remains consistent across both training and test datasets. By implementing such a stratified division, we aim to achieve a model evaluation that is as accurate and reliable as the model training itself.

Models were trained using the below algorithms:

1. **Logistic Regression:**
   Logistic Regression, as utilized in the CreditWatch project, serves as a fundamental statistical method for modeling the probability of a default event based on given predictors. This project implemented Logistic Regression through a refined approach, incorporating a grid search to optimize the model parameters. The

optimization process involved a parameter grid over different regularization penalties ('l1' and 'l2') and strength (C values ranging from 0.001 to 1000). These meticulous tuning aims to prevent overfitting and enhance the model's generalization to new data, thus making the predictions more reliable. Using cross-validation with three folds, the best parameters and the corresponding accuracy were systematically identified, demonstrating the effectiveness of Logistic Regression in predicting credit defaults.

2. **Gaussian Naïve Bayes**
   Gaussian Naive Bayes (GNB) is a probabilistic machine learning model that is particularly suited for classification tasks involving continuous data and where the features are assumed to follow a Gaussian distribution. In the context of the CreditWatch project, the GNB model was applied to predict credit defaults, capitalizing on its efficiency and straightforward implementation. The model operates under the naive assumption of independence between every pair of features, which simplifies the computation and often leads to surprisingly effective performance despite the simplicity of this assumption. GNB is notably fast and can handle large datasets well, making it a practical choice for initial exploratory analysis in predictive tasks where timeliness and computational efficiency are crucial.

3. **Random Forest**
   The Random Forest classifier, implemented in the CreditWatch project, represents a robust ensemble machine learning technique that combines multiple decision trees to produce a more accurate and stable prediction. By leveraging the principle of bagging, the Random Forest model increases its effectiveness through the aggregation of multiple decision trees trained on various sub-samples of the dataset, which helps in reducing the model's variance. This method effectively handles both binary and multiclass classification problems, making it highly applicable in predicting credit defaults where it's crucial to distinguish between defaulters and non-defaulters accurately.

4. **XGBoost**
   XGBoost, a cutting-edge implementation of gradient boosted trees, has proven to be an exceptionally powerful and versatile algorithm in the field of machine learning, particularly in the CreditWatch project for predicting credit defaults. This model distinguishes itself with its capability to handle a variety of data types, scalability, and improved accuracy through systematic hyperparameter tuning. In this project, the XGBoost model was finely tuned using a grid search approach to optimize parameters such as max depth and min child weight, enhancing predictive performance significantly. The model demonstrated high accuracy in both training and testing phases, highlighting its robustness and efficiency.

5. **Non-Linear Support Vector Machine**
   The Support Vector Machine (SVM) model, particularly with its radial basis function (RBF) kernel, provides a robust non-linear classification technique used in the CreditWatch project for predicting credit defaults. This model excels in finding the optimal hyperplane that maximizes the margin between different classes, making it highly effective for complex datasets with a clear margin of separation. By leveraging the RBF kernel, the SVM model can handle cases where the relationship between class labels and attributes is nonlinear, thus allowing for greater flexibility in capturing intricate patterns in the data.

6. **Multilayer Perceptron (MLP) Neural Network**
   The Multilayer Perceptron (MLP), a type of neural network utilized in the CreditWatch project, represents a sophisticated approach for predicting credit defaults by learning complex, non-linear interactions between variables. This model consists of multiple layers, including input, hidden, and output layers, with hidden layers featuring neurons that use ReLU activation functions for introducing non-linearity and Dropout layers for preventing overfitting by randomly ignoring certain units during training. The MLP was meticulously trained with a standard configuration of backpropagation using the Adam optimizer, which adjusts learning rates throughout training, and binary cross-entropy as the loss function, ideal for binary classification tasks.

We scrutinize the performance of six distinct machine learning models, each selected for its relevance and potential in tackling the complex challenge of credit default prediction. These models range from conventional logistic regression and decision trees to advanced techniques like the Multilayer Perceptron (MLP) and ensemble methods such as Random Forest and Gradient Boosting Machines. Each model has been thoroughly tested and evaluated using a series of key metrics that serve as indicators of their classification effectiveness.

The metrics we focus on include Accuracy, which measures the overall correctness of the model across all classes; Precision, which assesses the model's accuracy in predicting positive labels; Recall, or the sensitivity, that evaluates the model's ability to capture actual positive cases; the F1-score, which balances the precision and recall in a single metric; and the Confusion Matrix, providing a detailed breakdown of true positives, false positives, true negatives, and false negatives. These evaluation criteria are essential as they provide a holistic view of each model's strengths and limitations in predicting credit defaults. By analyzing these metrics, we can identify which models perform best under various scenarios and thus tailor our predictive strategies to enhance accuracy and reliability in real-world applications. This rigorous approach ensures that our predictions are not only statistically valid but also practically sound, offering actionable insights into managing credit risk effectively.

Models	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Logistic Regression	0.7115	0.7846	0.5832	0.669	[[3805 725] [1888 2642]]
Gaussian Naive Bayes	0.55	0.884	0.11	0.2	[[4461 69] [4002 528]]
Random Forest	0.93	0.91	0.966	0.93	[[4108 422] [154 4376]]
XGBoost	0.8025	0.8028	0.801	0.8024	[[3638 892] [897 3633]]
Non-Linear Support Vector Machine	0.714	0.777	0.6002	0.677	[[3750 780] [1811 2719]]
Multilayer Perceptron (MLP) Neural Network	0.71	0.79	0.56	0.66	[[3881 649] [1957 2573]]

Figure 9: Classification Models Evaluation Metrics

The summary table encapsulates the comparative performance of six different machine learning models applied in the CreditWatch project for the prediction of credit defaults. Logistic Regression shows respectable performance with an accuracy of approximately 71%, supported by decent precision but a relatively lower recall, suggesting it may miss a significant number of default cases. Gaussian Naive Bayes, despite its high precision, falls behind in recall and overall accuracy, indicating a cautious approach that is highly selective in predicting defaults, but often inaccurate.

The Random Forest model excels across all metrics, boasting an impressive 93% accuracy and balance between precision and recall, reflected in a high F1-score, which suggests it is highly reliable in distinguishing between defaulters and non-defaulters. XGBoost also performs well, with over 80% accuracy and a good balance between precision and recall. The Non-Linear Support Vector Machine shows a good balance between precision and recall as well, although with slightly lower accuracy than some other models.

The Multilayer Perceptron (MLP) Neural Network achieves an accuracy of 71%, with reasonable precision and recall, positioning it as a robust model, though slightly outperformed by the Random Forest and XGBoost models. The confusion matrices for each model provide deeper insight into the true and false positives and negatives, essential for understanding the context of each model's predictions. Overall, while some models demonstrate particular strengths, the Random Forest model stands out for its superior predictive power in this specific task.

Results and Discussion:

In this section, we delve into the detailed results of our study, analyzing the performance of six different machine learning models through a range of evaluation metrics: Accuracy, Precision, Recall, F1-score, and ROC Curve. Each metric sheds light on various facets of model performance and suitability for specific classification tasks. This comprehensive analysis aims to highlight the strengths and limitations of each model and provide insights into their practical implications in real-world scenarios.

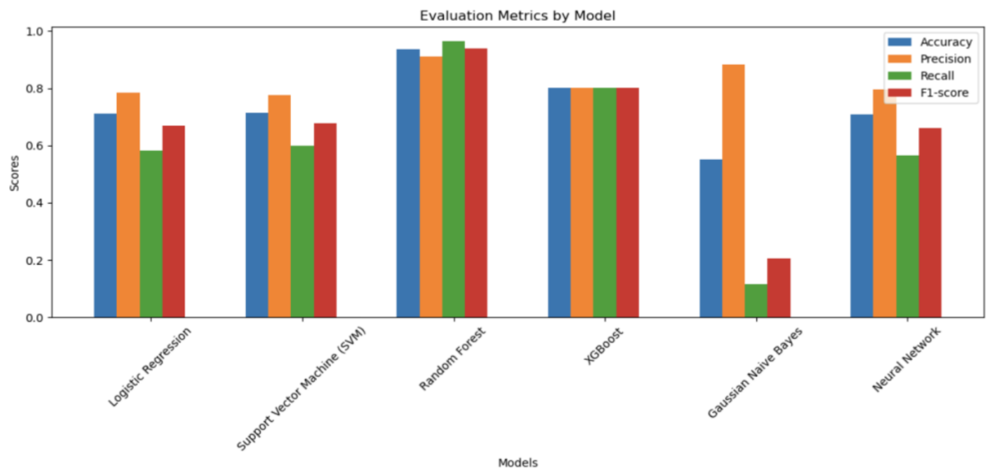


Figure 10: Comparison of Evaluation Metrics among all the used models

The provided bar chart serves as an evaluative comparison among six different machine learning models, gauging their performance based on four key metrics commonly used to assess classification algorithms: Accuracy, Precision, Recall, and F1-score. In this comparison, the Random Forest model emerges as the most competent, showcasing the highest scores across all metrics. This indicates a robust capacity to accurately classify data points, maintaining a strong balance between correctly predicting the positive class and minimizing the misclassification of the negative class as positive. Its performance suggests that it would be a reliable choice for a variety of classification tasks, particularly when both false positives and false negatives carry significant costs.

Conversely, the Gaussian Naive Bayes model registers the lowest Precision among the models, despite seemingly reasonable performance in other metrics. This particular aspect denotes a tendency to predict more false positives, potentially making it a less desirable option in scenarios where the precision of the positive prediction is critical.

Other models display varied performance profiles. The Support Vector Machine (SVM) and XGBoost both show commendable and balanced scores, signaling good overall performance, though not matching the superiority of the Random Forest in this dataset. Logistic Regression presents as a consistent, though not outstanding, contender, while the Neural Network boasts a high Recall, indicating a sensitivity towards identifying positive instances, but at a cost to its Precision and overall F1-score.

Each model's efficacy is context-dependent, with their suitability hinging on the particular demands of the application in question. Where one might prioritize avoiding false positives, Precision would be paramount, thus potentially discounting models like the Gaussian Naive Bayes and the Neural Network, despite the latter's high Recall. Ultimately, the decision on which model to deploy would balance these metrics against the specific objectives and costs associated with the classification task at hand.

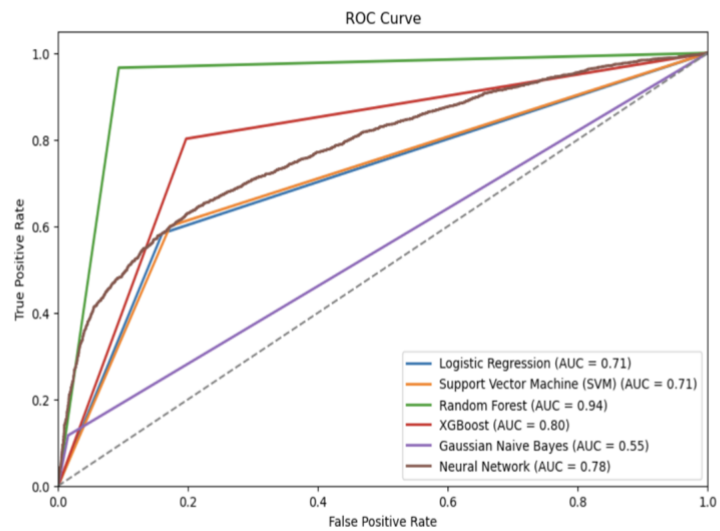


Figure 11: ROC Curve of Implemented Predictive Models

The Receiver Operating Characteristic (ROC) curve is an essential tool for evaluating the performance of binary classification models. It visualizes the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different thresholds. The TPR (sensitivity) measures the proportion of actual positives correctly identified by the model, while the FPR is the proportion of actual negatives incorrectly identified as positives. An effective classifier will maximize the TPR while minimizing the FPR, aiming for a top-left corner position on the ROC curve.

In the provided ROC curve chart, the performance of six different classification models is compared. The analysis indicates that the Random Forest model is the most effective classifier among those evaluated, with an AUC close to the ideal score of 1.0. The Gaussian Naive Bayes model underperforms significantly compared to its counterparts, with minimal separation between the classes. Other models like the SVM, Logistic Regression, and Neural Network show varying degrees of effectiveness, with performances that are above average but not outstanding. The AUC values provide a quantitative measure to compare the overall effectiveness of the models, with higher values representing better classification performance.

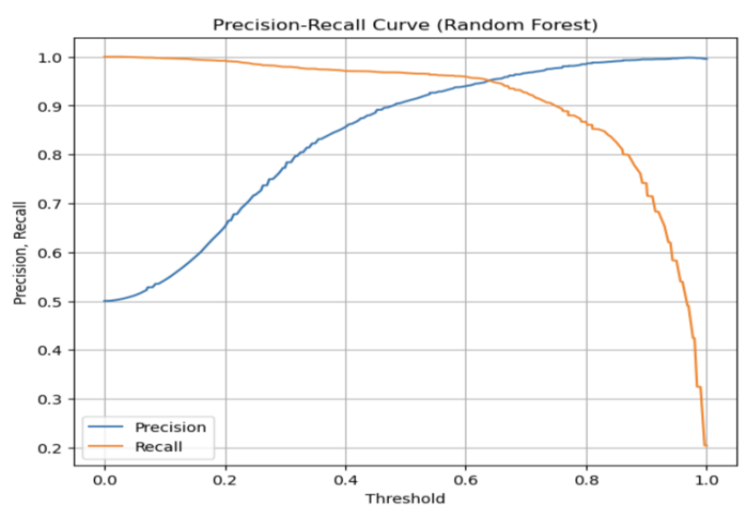


Figure 12: Precision-Recall Curve - Random Forest

The Precision-Recall Curve for the Random Forest model illustrates the balance between precision, the accuracy of positive predictions, and recall, the ability to capture all relevant instances. As the threshold for predicting a positive outcome increases, the model predicts fewer positives, leading to a higher precision but lower recall. This trade-off is crucial in scenarios where the consequences of false positives and false negatives differ significantly.

In regions of higher thresholds, the model exhibits increased precision at the expense of recall, suggesting a cautious approach that prioritizes the correctness of the positive predictions over the model's sensitivity to detecting positives. Conversely, at lower thresholds, the model captures most positives (high recall) but also makes more false positive errors (low precision). The curve indicates that there is a middle ground where precision and recall are more balanced, which might be considered the optimal operational point. However, the choice of the threshold should be guided by the specific application area—higher recall might be crucial in medical testing, while higher precision could be more important in fields like spam detection.

Overall, Random Forest model demonstrates flexibility in performance across the spectrum of thresholds, adaptable to different needs by altering the threshold.

Code

- https://github.com/SandhyaKilari/CreditWatch

Conclusion & Future work

The CreditWatch project, aimed at predicting credit defaults, revealed significant insights through its analytical approaches and data processing. The project employed various models, with the Random Forest algorithm emerging as the standout performer, demonstrating a remarkable accuracy in default prediction. This high level of precision underscores the algorithm's suitability for this kind of predictive task. Further, the project demonstrated a rigorous evaluation of model performance using various metrics like accuracy, precision, recall, and F1-score, confirming the robustness of the Random Forest model among the alternatives.

Looking ahead, there are several opportunities to enhance the predictive capabilities of the CreditWatch project. Exploring ensemble methods or advanced neural networks may offer avenues to further elevate the accuracy and reliability of the predictions. Moreover, evaluating these models on a larger and more diverse dataset would be instrumental in ensuring the scalability and robustness of the system, particularly for broader financial applications. Integrating these predictive models into a decision support system could provide significant benefits to financial institutions, assisting them in making more informed decisions regarding credit risks.

References

- https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data
- https://www.kaggle.com/code/zavidnikitamahbub/credit-card-default-eda-and-ensemble-methods?scriptVersionId=126256287&cellId=73
- https://www.kaggle.com/code/dnyaneshyeole/credit-card-default-a-comprehensive-eda-notebook
- https://medium.com/cuenex/credit-default-prediction-practical-tips-for-successful-execution-f62a92ab5df8
- https://towardsdatascience.com/catch-me-if-you-can-predicting-credit-card-default-using-classification-analysis-28b2955f7f7d