**PREDICTIVE FRAMEWORK FOR CREDIT CARD DEFAULT RISK**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**Sravya Alapati (AP21110010651),**

**Jahnavi Tadikamalla (AP21110010655),**

**Niharika Juttuka (AP21110010681),**

**Yaswanth Yarasani (AP21110010751)**

**A picture containing text

Description automatically generated**

Under the Guidance of

**(Dr. M Krishna Siva Prasad)**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**April, 2024**

**Certificate**

Date: 30-April-24

This is to certify that the work present in this Project entitled “**Predictive Framework for Credit Card Default Risk**” has been carried out by **[Sravya Alapati, Jahnavi Tadikamalla, Niharika Juttuka, Yaswanth Yarasani]** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in the **School of Engineering and Sciences**.

Dr. M Krishna Siva Prasad

**Supervisor**

**Acknowledgment**

We extend our deepest gratitude to Dr Krishna Siva Prasad for his valuable guidance, unwavering support, and mentorship throughout the course of our Applied Data Science (ADS). Dr Siva Prasad's expertise, encouragement, and dedication have been instrumental in shaping the success of this project endeavour.

His insightful feedback, constructive criticism, and commitment to academic excellence have significantly contributed to the development of this project.

We express our sincere thanks to Dr Krishna Siva Prasad for his mentorship, which has been a guiding force in our academic journey.

Yours Sincerely,

Sravya Alapati-AP21110010651

Jahnavi Tadikamalla-AP21110010655

Niharika Juttuka-AP21110010681

Yaswanth Yarasani-AP21110010715

**Contributions:**

**Sravya - Data Preparation and Preprocessing**

Contribution: Sravya takes the lead on data cleaning, preprocessing, and feature engineering. She handles missing values, outliers, and normalization of data to ensure the dataset is ready for modeling. Sravya also conducts initial exploratory data analysis to identify patterns, correlations, and potentially useful features that could enhance model performance.

**Jahnavi - Model Development and Optimization**

Contribution: Jahnavi is responsible for building and tuning various machine learning models. Her experiments with different algorithms, including both traditional approaches like Logistic Regression and more complex ones like Random Forest and XGBOOST. Jahnavi focuses on adjusting hyperparameters, implementing cross-validation, and utilizing techniques like grid search to find the optimal settings for each model.

**Niharika - Analysis and Validation**

Contribution: Niharika takes charge of evaluating the models using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC curves. She analyses the outcomes to determine which models perform best and interprets the results to provide insights into the predictive power and reliability of each. Niharika also performs statistical tests and validations to ensure the robustness of the findings.

**Yaswanth- Reporting and Presentation**

Contribution: Yaswanth puts together a report and presentation for the project. He makes sure to organize the information in a way that makes sense and is easy to understand. He is great at using pictures and graphs to help explain the data and results. He also makes sure to explain everything in a way that everyone can understand, even if they aren't experts in the subject. Finally, Yaswanth takes care of any paperwork that needs to be submitted along with the project.

**Collective Contribution:** Together, the team works collaboratively to ensure that every aspect of the project from data handling to final presentation is thorough and polished. We regularly meet to discuss our progress, troubleshoot issues, and ensure that the project aligns with the course objectives and grading criteria. Our collective efforts lead to a comprehensive understanding of how different models can be applied to predict credit card defaults, showcasing our ability to apply theoretical knowledge to a practical problem, which is essential in data science education.

Novelty:

The novelty of this project in predicting credit card defaults in Taiwan lies in its comprehensive and integrated approach, utilizing a variety of finely tuned advanced machine learning models. Key innovative aspects include:

Diverse Model Deployment: Employing multiple algorithms such as Logistic Regression, SGD, SVM, Decision Trees, KNN, Random Forest, and XGBOOST, each optimized for the dataset, ensures a robust predictive framework.

Advanced Metrics Evaluation: The project goes beyond standard accuracy measures, using precision, recall, F1 score, and ROC-AUC to provide a nuanced assessment of model performance.

Focus on Tuned Performance: Each model is rigorously tuned to enhance its predictive power and ability to generalize, critical for applications in the dynamic financial sector.

Predictive Power of Socioeconomic Factors: The analysis incorporates demographic insights, enhancing the understanding of how different factors influence credit behavior and facilitating tailored risk management strategies.

Real-time Application Potential: Emphasis on real-time capable models like SGD aligns with the needs of financial institutions for immediate risk assessments.

Ethical and Regulatory Compliance: The project addresses potential biases and ensures compliance with regulatory standards, promoting ethical AI usage in financial practices.

Integration with Business Processes: Predictive insights are designed to integrate seamlessly into business workflows, enhancing decision-making in credit management.

**Contents**

[Abstract: 7](#_Toc165239053)

[Introduction: 8](#_Toc165239054)

[Background: 9](#_Toc165239055)

[Description: 10](#_Toc165239056)

[Proposed solution using Data Science Technique: 11](#_Toc165239057)

[Model Architecture: 12](#_Toc165239058)

[Experimentation Details: 13](#_Toc165239059)

[**Dataset:** 13](#_Toc165239060)

[**Model performances:** 13](#_Toc165239061)

[**Primary models with the best Accuracy:** 13](#_Toc165239062)

[**Other Models Used for Comparison:** 13](#_Toc165239063)

[**Comparative Accuracy Table:** 15](#_Toc165239064)

[**ROC Curve:** 15](#_Toc165239065)

[Conclusion: 16](#_Toc165239066)

[Future Recommendations: 16](#_Toc165239067)

[References: 16](#_Toc165239068)

# Abstract:

This project aims to develop a predictive model to determine whether credit card holders will clear their debts in the next month based on various attributes. With the increasing prevalence of credit card usage and the associated risks of default, accurate prediction of repayment behaviour is crucial for financial institutions to mitigate losses and make informed lending decisions. Leveraging a dataset containing historical transactional and demographic information, this study employs machine learning algorithms such as logistic regression, decision trees, random forests, and KNN to analyse. Feature engineering techniques and model optimization are employed to enhance predictive accuracy and generalizability. The proposed model not only assists financial institutions in identifying high-risk customers but also provides insights into the factors influencing repayment behaviour, thereby enabling targeted interventions and risk management strategies. The efficacy of the model is evaluated using metrics such as accuracy, precision and recall, demonstrating its potential to improve credit risk assessment and enhance financial stability.

In an era marked by economic volatility and evolving consumer behaviour’s, the effective management of credit risk remains a critical priority for financial institutions. This project proposes a novel approach to enhance credit risk assessment by leveraging predictive modelling techniques to forecast the likelihood of credit card holders clearing their debts in the upcoming month. Drawing upon a comprehensive dataset encompassing diverse attributes such as limit balance, gender, and default payment next month, this study employs a sophisticated ensemble of machine learning algorithms including logistic regression, support vector machines.

Through rigorous feature engineering and model optimization, the predictive model is fine-tuned to accurately classify customers into distinct risk categories: those expected to repay their debts and those at risk of default. Moreover, advanced techniques such as interpretability analysis and model explainability are employed to elucidate the underlying factors driving repayment behaviour, thereby empowering financial institutions to make informed lending decisions and devise targeted risk mitigation strategies. The performance of the proposed model is evaluated using robust evaluation metrics including accuracy, precision, recall, and F1-score, ensuring its reliability and effectiveness in real-world credit risk management scenarios. By facilitating proactive risk identification and strategic intervention, this predictive modelling framework offers a promising avenue for enhancing financial stability and safeguarding the interests of both lenders and borrowers in the dynamic landscape of credit markets.

# Introduction:

In today's financial landscape, credit card usage has become ubiquitous, facilitating convenient transactions and enabling consumers to meet their diverse financial needs. However, alongside the benefits, credit card usage also presents inherent risks, particularly for financial institutions tasked with managing credit risk effectively. One of the key challenges faced by these institutions is accurately assessing the likelihood of credit card holders repaying their debts, thereby minimizing potential losses and ensuring financial stability.

Traditional methods of credit risk assessment often rely on historical data and static scoring models, which may not adequately capture the dynamic nature of consumer behaviour and the evolving economic environment. As such, there is a growing imperative to leverage advanced analytical techniques and predictive modelling to enhance the accuracy and granularity of credit risk assessment processes.

This project endeavours to address this challenge by developing a predictive modelling framework that can forecast the probability of credit card holders clearing their debts in the next billing cycle. By analysing a diverse range of customer attributes, including transactional history, socio-demographic characteristics, and credit utilization patterns, our aim is to construct a predictive model capable of identifying customers at high risk of default and those likely to meet their repayment obligations.

The significance of this project lies in its potential to empower financial institutions with actionable insights for better risk management and decision-making. By accurately predicting repayment behaviour, lenders can proactively identify and mitigate credit risk, optimize loan portfolio management, and tailor financial products and services to meet the evolving needs of their customers.

Through this report, we will delve into the methodology employed, including data preprocessing, feature engineering, model selection, and evaluation techniques. We will also discuss the practical implications of our predictive modelling framework for credit risk assessment and highlight its potential to drive value for both financial institutions and consumers in an increasingly complex and dynamic financial landscape.

# Background:

In recent years, Taiwan's credit card industry has had big problems. Many people were not paying their credit card bills on time, and it was expected to be the worst in the third part of 2006. This happened because banks in Taiwan were giving out too many credit cards and cash cards to people who couldn't really afford them. They were doing this to get more customers. People were using credit cards too much, even if they couldn't really pay back the money they borrowed.

This caused a lot of trouble for people and made them worry about their money. Banks and the government had to make some changes to how they give out credit cards and how they check if people can really pay back what they borrow. The situation with credit cards in Taiwan got really bad because banks were giving out too many cards to people who couldn't handle them. People were spending too much money using these cards, even if they couldn't afford it.

This made it hard for them to pay back what they owed. Because of all this trouble, many people lost trust in using credit cards. Banks and the government had to step in to fix things. They made new rules to make sure that banks didn't give out too many cards to people who couldn't pay back what they borrowed. Our project wants to help fix this problem. We're using computer programs to look at past information about how people used their credit cards. This helps us see who might have trouble paying back what they owe. By doing this, banks can make better decisions about who gets credit cards and how to help people manage their money better. This could make things safer for everyone who uses credit cards in Taiwan.

# Description:

The credit card industry in Taiwan has been facing a severe cash and credit card debt crisis, with delinquency rates expected to reach their peak in the third quarter of 2006 (Chou, 2006). This crisis stems from several key factors:

1. Over-issuance of Credit Cards: In an effort to expand their market share, card-issuing banks in Taiwan have been issuing cash and credit cards to individuals who may not be financially qualified. This practice has led to a significant increase in the number of credit cards in circulation, potentially exacerbating the debt problem.

2. Excessive Credit Card Usage: Many cardholders in Taiwan, regardless of their ability to repay, have been using credit cards excessively for consumption purposes. This behaviour has resulted in the accumulation of heavy credit card debts and cash advances, further contributing to the crisis.

3. Impact on Consumer Confidence: The credit card debt crisis has eroded consumer confidence in the financial system, leading to widespread concerns about personal finances and financial stability. This lack of confidence poses challenges for both banks and cardholders, as they grapple with the repercussions of the crisis.

Our project focuses on developing a predictive model to forecast credit card holders' likelihood of defaulting on payments in the next month. By analysing historical transactional data and customer attributes like repayment status and bill statement amounts, we aim to pinpoint individuals at high risk of default. This model will provide banks and financial institutions with valuable insights, enabling proactive credit risk management and informed lending decisions. By accurately identifying at-risk individuals, targeted interventions such as financial counselling or repayment plan restructuring can be implemented, mitigating the crisis's impact on both borrowers and lenders.

Overall, our data-driven approach aims to stabilize Taiwan's credit card industry by addressing the challenge of credit card debt and delinquency. Through the predictive model's development, we seek to empower institutions with the necessary tools and insights to navigate the crisis effectively, fostering financial stability for all stakeholders involved.

# Proposed solution using Data Science Technique:

Based on the analysis of various machine learning models using ROC curve and accuracy metrics, we have identified decision tree and stochastic gradient descent as the top-performing models for our credit card project. Therefore, our proposed solution for addressing the challenges in the credit card industry in Taiwan is as follows:

1. Implementation of Decision Tree Model: We recommend deploying the decision tree model as a primary tool for predicting credit card holders' likelihood of defaulting on payments in the next month. Decision trees offer interpretability and ease of understanding, making them suitable for practical application by financial institutions. By utilizing decision tree-based predictions, banks can efficiently identify individuals at high risk of default and tailor interventions accordingly.

2. Utilization of Stochastic Gradient Descent (SGD) Model: In addition to decision trees, we propose integrating the stochastic gradient descent (SGD) model into our predictive framework. SGD offers the advantage of fast training times and scalability, making it suitable for handling large datasets and real-time prediction scenarios.

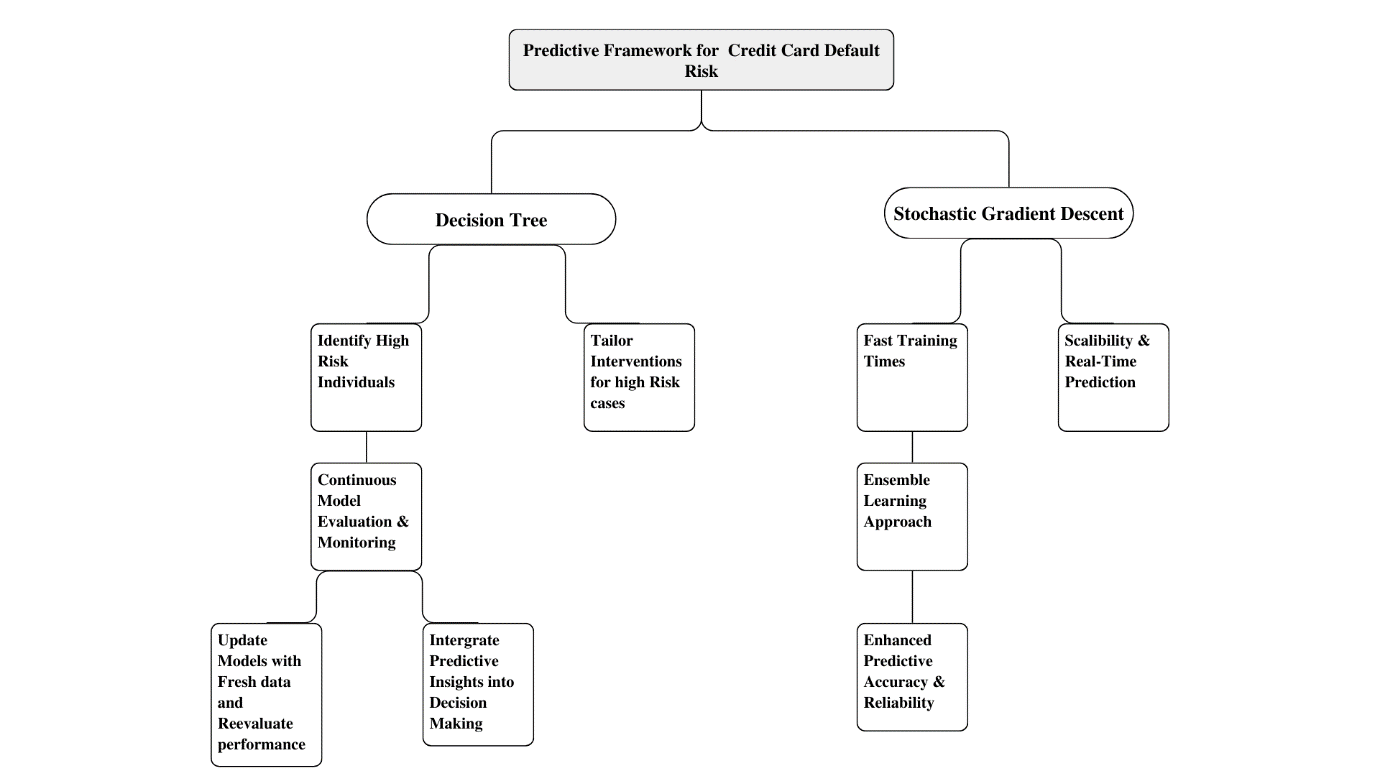
3. Ensemble Learning Approach: To further improve prediction performance, we suggest implementing an ensemble learning approach that combines the strengths of multiple models, including decision trees and SGD. Ensemble methods such as stacking or boosting can effectively leverage the diverse insights provided by individual models, resulting in enhanced predictive accuracy and reliability.

4. Continuous Model Evaluation: It is imperative to establish a framework for continuous model evaluation and monitoring to ensure the ongoing effectiveness and relevance of the predictive solution. Regularly updating the models with fresh data and re-evaluating their performance using appropriate metrics such as ROC curve analysis and accuracy assessment will enable banks to adapt to evolving market dynamics and maintain a competitive edge in credit risk management.

5. Integration of Predictive Insights into Decision-Making Processes: Finally, we recommend integrating the predictive insights generated by the decision tree and SGD models into banks' decision-making processes. By incorporating these insights into credit approval, risk assessment, and customer management workflows, financial institutions can make informed lending decisions, mitigate credit risk, and foster greater financial stability within the credit card industry in Taiwan.

By adopting these proposed solutions, banks and financial institutions can leverage the predictive power of decision tree and SGD models to effectively address the challenges posed by credit card debt and delinquency, ultimately contributing to the stabilization and resilience of Taiwan's credit card industry.

# Model Architecture:



* Tree Model: Primary tool for identifying high-risk individuals, aiding in tailoring interventions.
* SGD Model: Known for fast training times and scalability, suitable for real-time prediction scenarios.
* Ensemble Learning Approach: Combines decision trees and SGD to leverage their strengths, improving overall predictive accuracy and reliability.
* Continuous Model Evaluation: Essential for maintaining model performance through regular updates and re-assessment.
* Integration into Decision-Making: Ensures that predictive insights are actively used in the bank’s workflows, enhancing decision quality and reducing risks.

# Experimentation Details:

## **Dataset:**

The data in this dataset can be used to develop models that can predict whether a credit card client is likely to default on their payment in the next month. This prediction can help financial institutions to identify high-risk clients and take appropriate measures to minimize the risk of default and potential financial loss. The dataset can also be used to analyse the factors that contribute to credit card defaults, such as age, gender, education level, repayment status, and bill amount.

Moreover, this dataset is widely used in research, and many studies have been conducted using this data to explore various research questions such as the impact of social and economic factors on credit card defaults, the effectiveness of different machine learning algorithms in predicting defaults, and the role of financial education in reducing the risk of default. Overall, this dataset is an excellent resource for researchers, financial institutions, and anyone interested in credit risk analysis and prediction modelling.

## **Model performances:**

### **Primary models with best Accuracy:**

**1. Stochastic Gradient Descent (SGD):**

SGD is particularly well-suited for datasets with a large number of features and examples, like the credit card dataset from Taiwan which includes multiple attributes of cardholders.

The model's efficiency in handling large-scale data makes it ideal for real-time prediction, essential for financial institutions needing to make quick decisions based on the latest information.

Its robustness to feature scaling and ability to handle sparse data help in dealing with varied types of input features (like payment history, amounts paid, and demographic details).

**2. Decision Tree:**

Decision Trees provide clear and easy-to-understand models, which is crucial when financial institutions need to explain their decision-making processes to stakeholders, including customers and regulators.

This model can easily handle categorical variables without the need for dummy variables, which is useful since the dataset includes various categorical features such as education, marriage status, and repayment status.

The decision tree can incorporate non-linear relationships between features and the target variable, which are likely present in the complex dynamics of financial behaviours like defaulting.

### **Other Models Used for Comparison:**

**3. Logistic Regression:**

Logistic Regression is a go-to method for binary classification problems such as predicting default (yes/no).

It provides probabilities for outcomes which can be a valuable tool for banks to assess risk levels rather than just a binary outcome, allowing more nuanced decisions.

The logistic model can also utilize the rich demographic and historical transaction data in the dataset to weigh the impact of various factors on the probability of default.

**4. Support Vector Machine (SVM):**

SVM can be particularly effective in high-dimensional spaces, which is typical for financial datasets where many variables affect the outcome.

It is well-suited for classification problems with a clear margin of separation but can also efficiently perform in non-linear classification through the use of kernel tricks, potentially capturing more complex patterns in the data.

**5. K-Nearest Neighbour (KNN):**

KNN makes predictions based on the labels of nearest neighbours, reflecting the assumption that similar profiles tend to have similar default behaviours.

This model's reliance on locality and similarity can leverage the demographic and transactional similarities among clients in the dataset, offering personalized risk assessments.

**6. Random Forest:**

Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

The ensemble nature of Random Forest can effectively reduce the overfitting problem common to decision trees and is great for datasets with imbalanced classes like default predictions.

**7. XGBOOST:**

XGBOOST applies gradient boosting framework which is effective for both regression and classification problems.

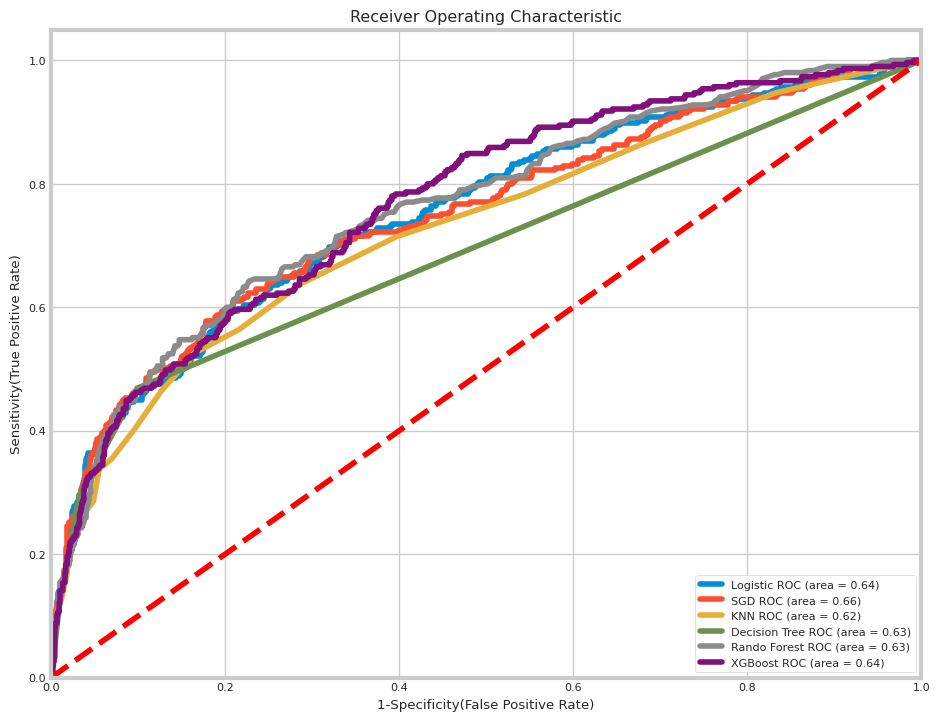
It is designed to optimize large-scale machine learning problems, making it a good fit for the dataset, which is likely complex and voluminous.

XGBOOST is also known for handling missing values and maintaining accuracy even when some data might be incomplete.

### **Comparative Accuracy Table:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | ROC |
| Logistic Regression | 0.825 | 0.683 | 0.311 | 0.427 | 0.637 |
| Stochastic Gradient Descent | 0.833 | 0.652 | 0.363 | 0.467 | 0.657 |
| SVM | 0.808 | 0.682 | 0.367 | 0.477 | 0.661 |
| KNN | 0.808 | 0.628 | 0.272 | 0.379 | 0.615 |
| Decision Tree | 0.828 | 0.697 | 0.295 | 0.414 | 0.631 |
| Random Forest | 0.810 | 0.627 | 0.298 | 0.404 | 0.626 |
| XG Boost | 0.821 | 0.624 | 0.337 | 0.438 | 0.642 |

### **ROC Curve:**



# Conclusion:

In this project, we explored several advanced machine learning models to predict credit card defaults, focusing on a dataset derived from Taiwan's credit card clients. The models tuned for this analysis included Logistic Regression, Stochastic Gradient Descent (SGD), Support Vector Machine (SVC), K-Nearest Neighbours (KNN), Decision Tree, Random Forest, and XGBOOST, each providing insights into their predictive accuracy, precision, recall, F1 score, and ROC values.

The SVC Model Tuned emerged as the top performer with the highest accuracy of 0.836667 and an ROC score of 0.661849, indicating its effectiveness in classifying the likelihood of defaults accurately. Other models also showed competitive performance, with the Stochastic Gradient Descent Tuned model and Decision Tree Tuned model providing valuable predictions but with slightly varying degrees of recall and precision.

# Future Recommendations:

**1. Model Improvement:**

Ensemble Techniques: Further explore ensemble methods that combine the predictions from multiple models to improve overall accuracy and robustness. Techniques such as stacking could be employed to harness the strengths of individual models effectively.  
Hyperparameter Optimization: Continue to refine the models by experimenting with more comprehensive grid and random search strategies for hyperparameter tuning to optimize their performance.  
Feature Engineering: Additional work on feature selection and engineering could reveal more subtle patterns and interactions in the data, potentially enhancing model performance.

**2.Deployment Strategy:**

Real-time Predictions: Develop an infrastructure for implementing these models in a real-time prediction environment. This would allow for dynamic risk assessment and more timely interventions.  
Integration with Existing Systems: Ensure that the deployment of these models is well integrated with the banks’ existing IT infrastructure for smooth operations and minimal disruption.

# References:

https://www.geeksforgeeks.org/machine-learning/

https://in.mathworks.com/discovery/machine-learning-models.html