CreditWatch: Predicting Credit Default

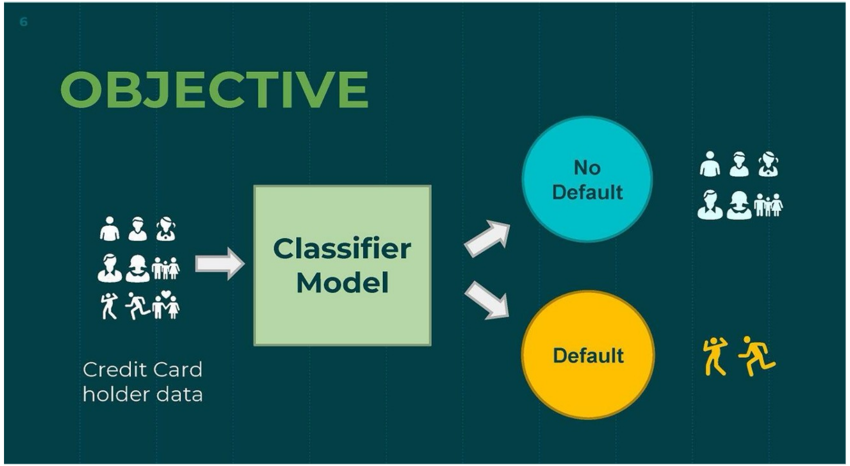
Team Members: Aarthi Padmanabhan, Madhurya Shankar, Sandhya Kilari

Introduction

- Credit default occurs when a borrower fails to meet the required debt payments on a loan or credit agreement

- Long-term: Impact on credit score, potential legal actions, and restricted access to future credit

- Risk reduction: Enhances decision-making processes by forecasting potential defaults



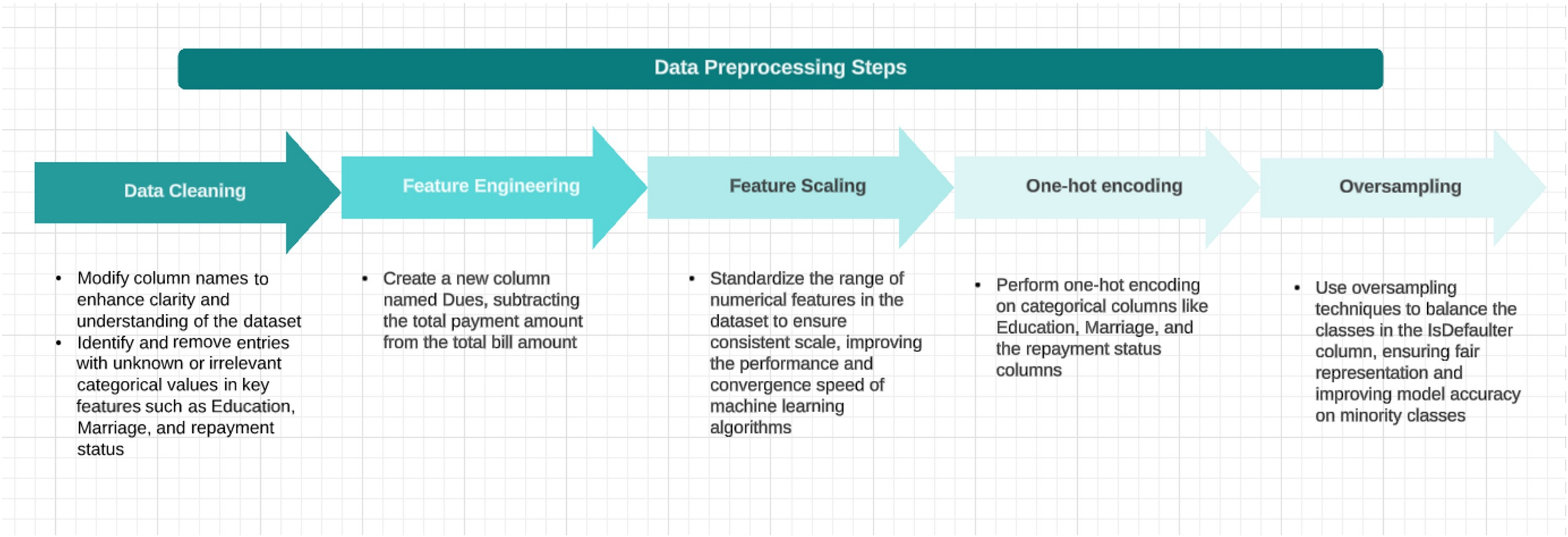
Data Description

- Scope of the Dataset
  - Comprehensive data comprising 30,000 customer records.
  - Aimed at predicting default payments among credit card users in Taiwan, from April 2005 to September 2005.
- Key features
  - Identity and Credit Info: Includes customer ID, credit limits, and demographics such as gender, education, marital status, and age.
  - Payment History: Tracks monthly payment status and bill amounts for the past six months.
  - Payment Outcomes: Documents payment amounts and flags potential defaults in the following month.

- One predictive binary label (Default: Yes = 1, No = 0)
- No missing values in the dataset

Dataset: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data

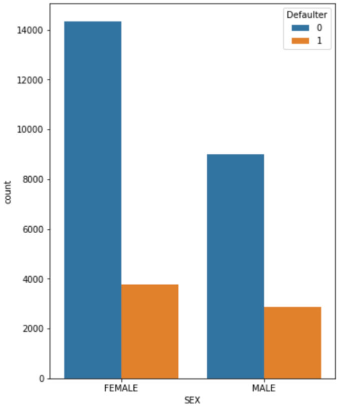
Data Preprocessing

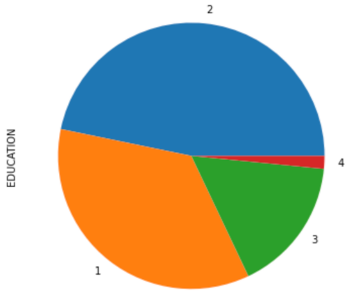


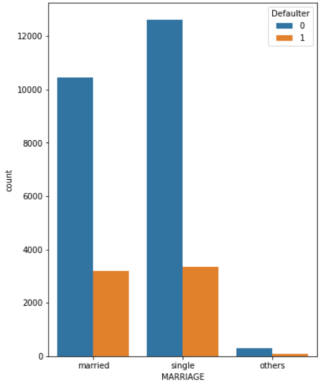
Exploration of univariate variables

Analysis on categorical variables:

- Dataset contains a higher number of female customers.
- Customers having university education are relatively higher followed by graduates.
- There are more single customers as compared to married.
- Relatively higher number of customers have duly paid bill amount each month.
- Approximately 30% customers are defaulters.

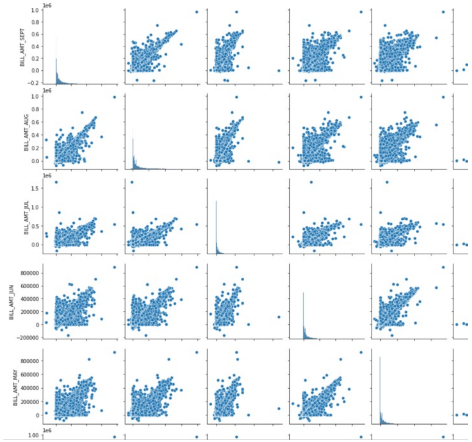


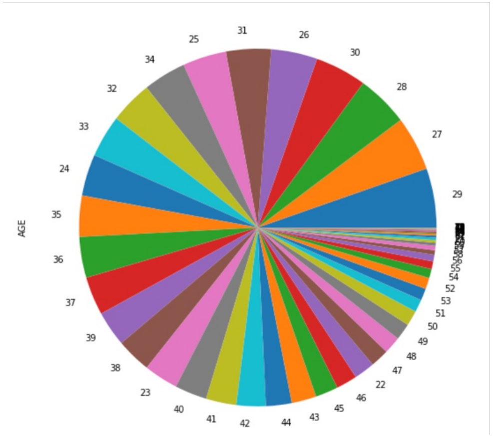


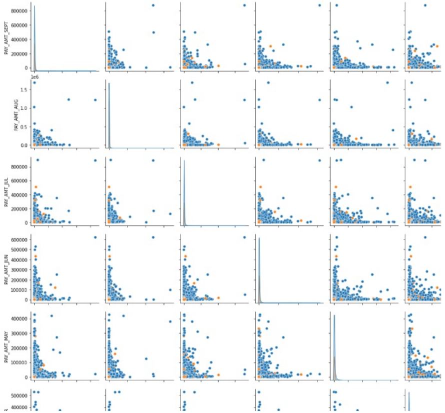


Analysis on continuous variables:

- Most customers are aged between 25 to 45.
- Balance limits mostly range from 50,000 to 250,000.
- Bill amounts usually span from -9,500 to 260,000, predominantly clustering between 1,000 to 55,000.
- Payments over the last six months typically range from 0 to 70,000, with most under 10,000.

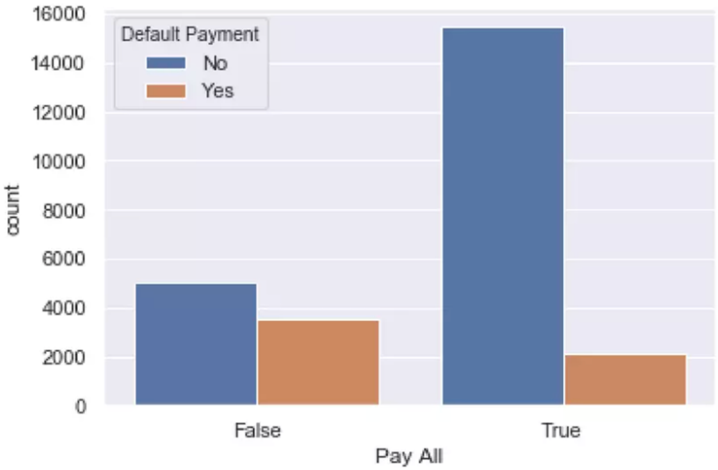




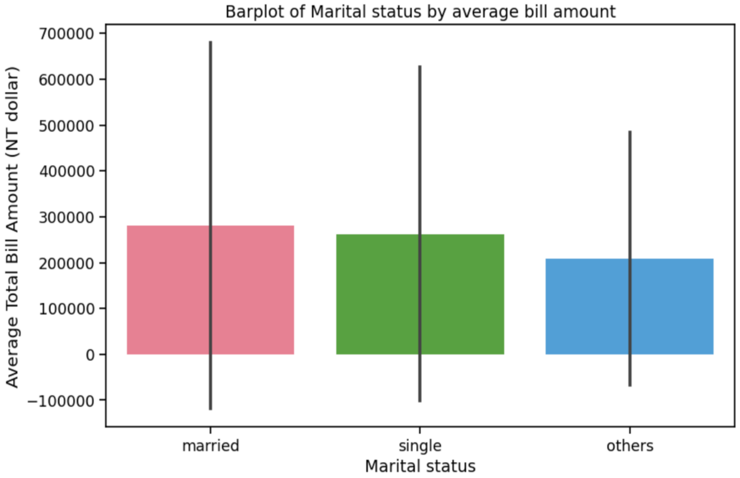


Exploration of pair of variables

Does having a delay in previous payment impact chances of default?



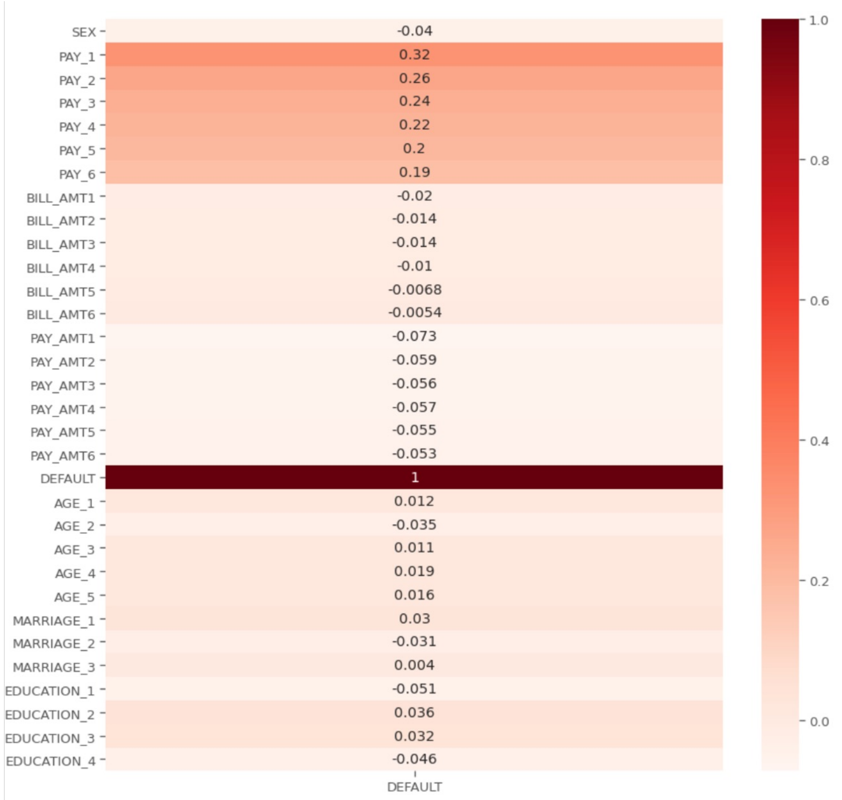
Do married people spend more than others?



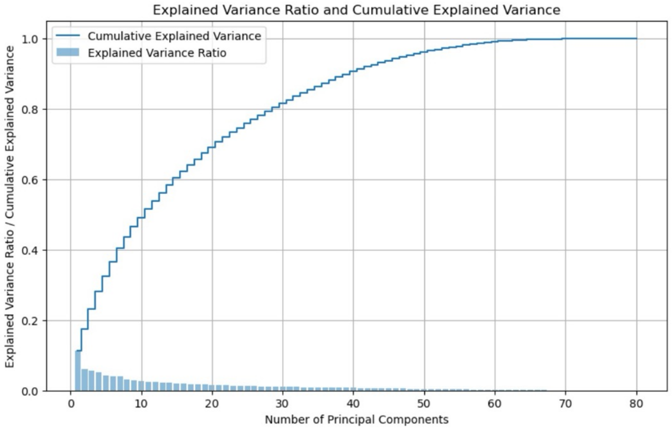
Pearson's Correlation

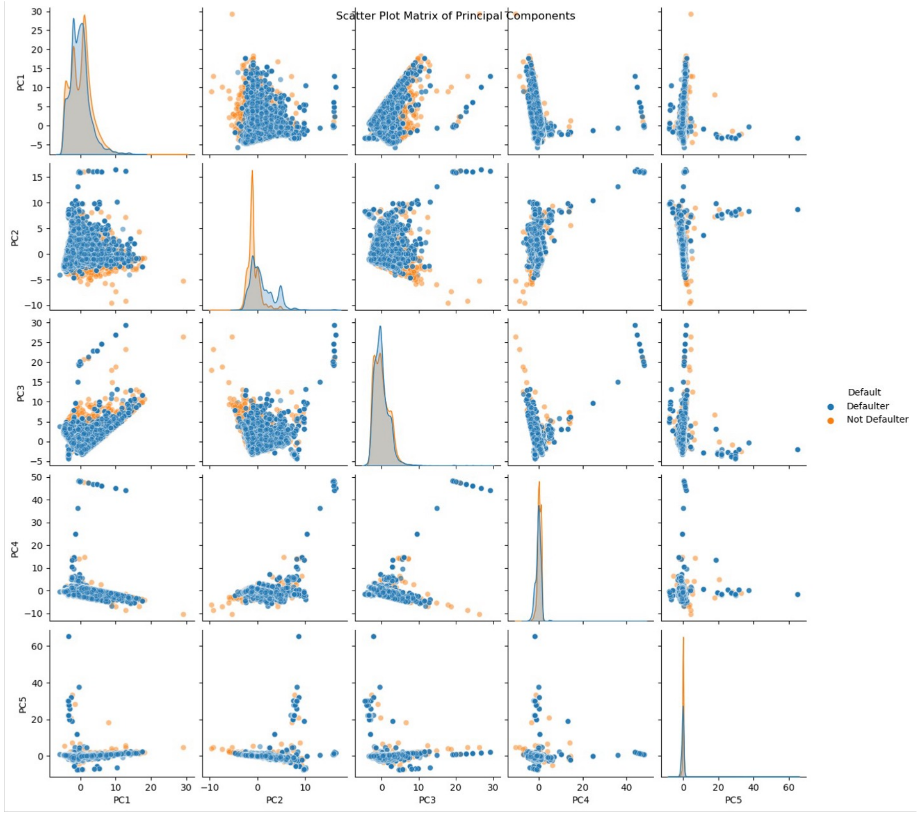
$$\rho = \frac{covariance(x_i, x_j)}{\sigma_i \times \sigma_j}$$

Repayment status of customers (PAY_1 - PAY_6) have the higher correlation towards the label.



Feature Selection/Dimensionality Reduction





Modeling Approaches Used

Linear Models

- Logistic Regression
- Gaussian Naive Bayes

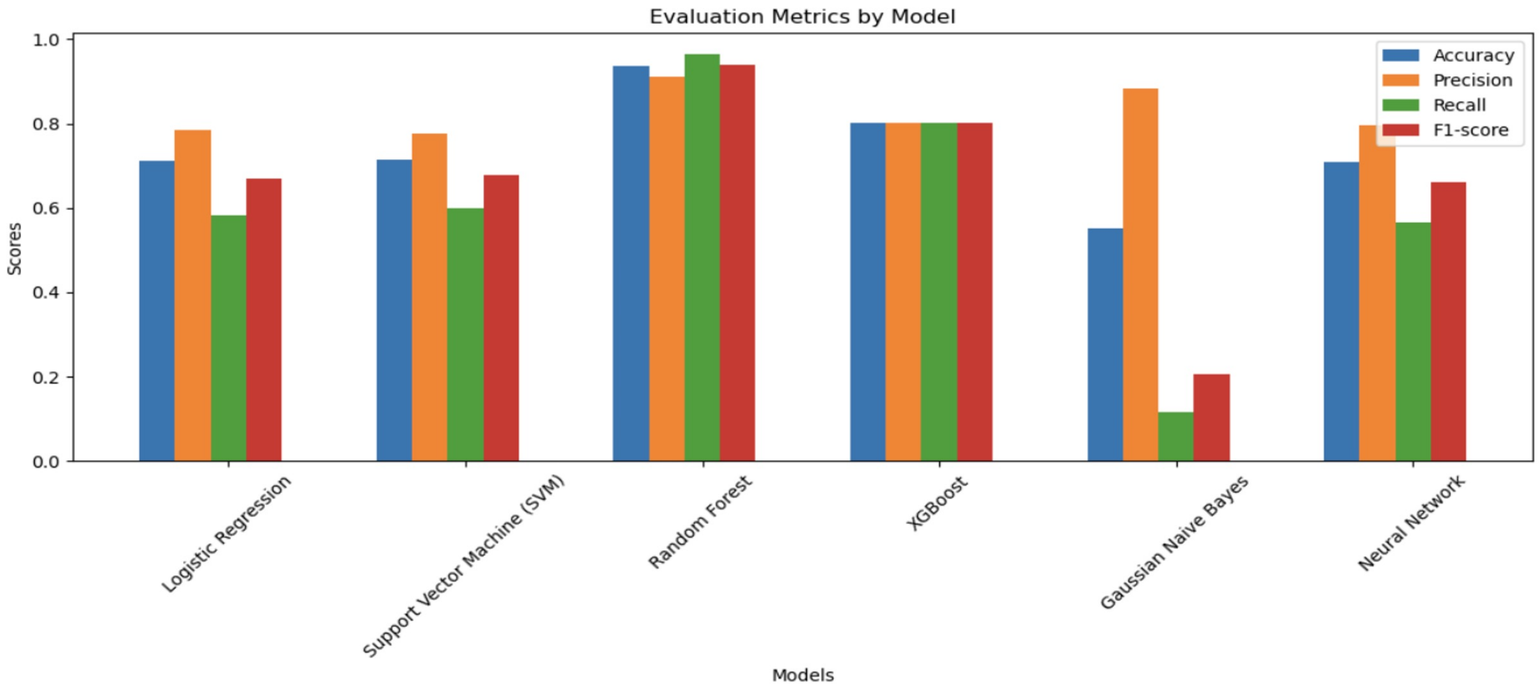
Non-linear Models

- Random Forest
- XGBoost
- Non-Linear Support Vector Machine
- Multilayer Perceptron (MLP) Neural Network

Exploring Classification Model Evaluation

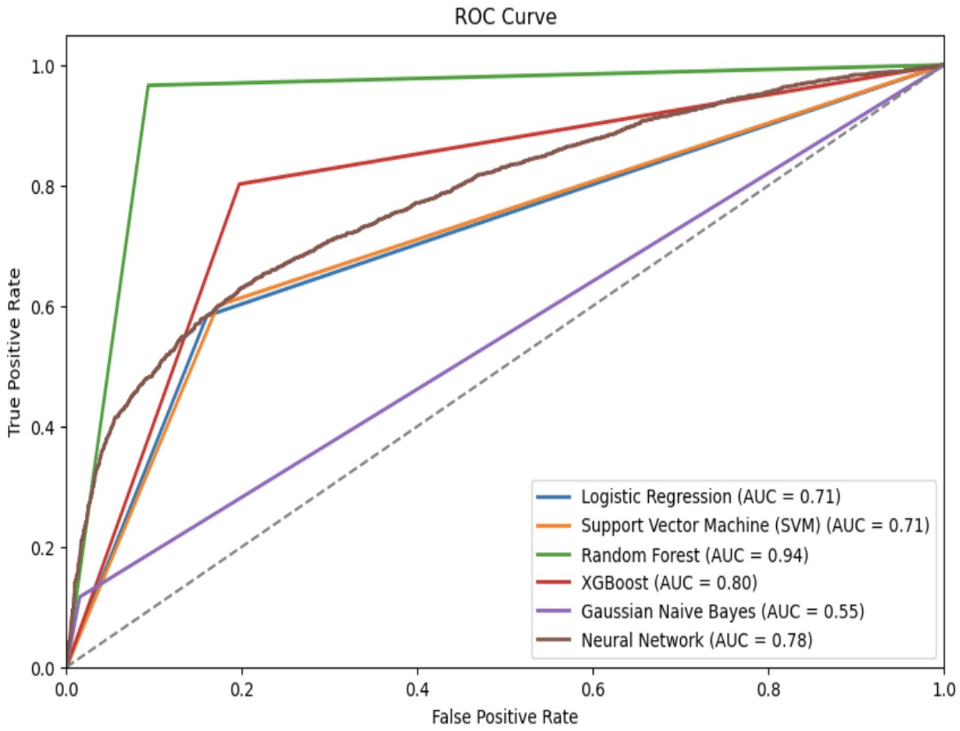
Models	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Logistic Regression	0.7115	0.7846	0.5832	0.669	[[3805 725] [1888 2642]]
Gaussian Naive Bayes	0.55	0.884	0.11	0.2	[[4461 69] [4002 528]]
Random Forest	0.93	0.91	0.966	0.93	[[4108 422] [154 4376]]
XGBoost	0.8025	0.8028	0.801	0.8024	[[3638 892] [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]]

Comparison of Evaluation Metrics



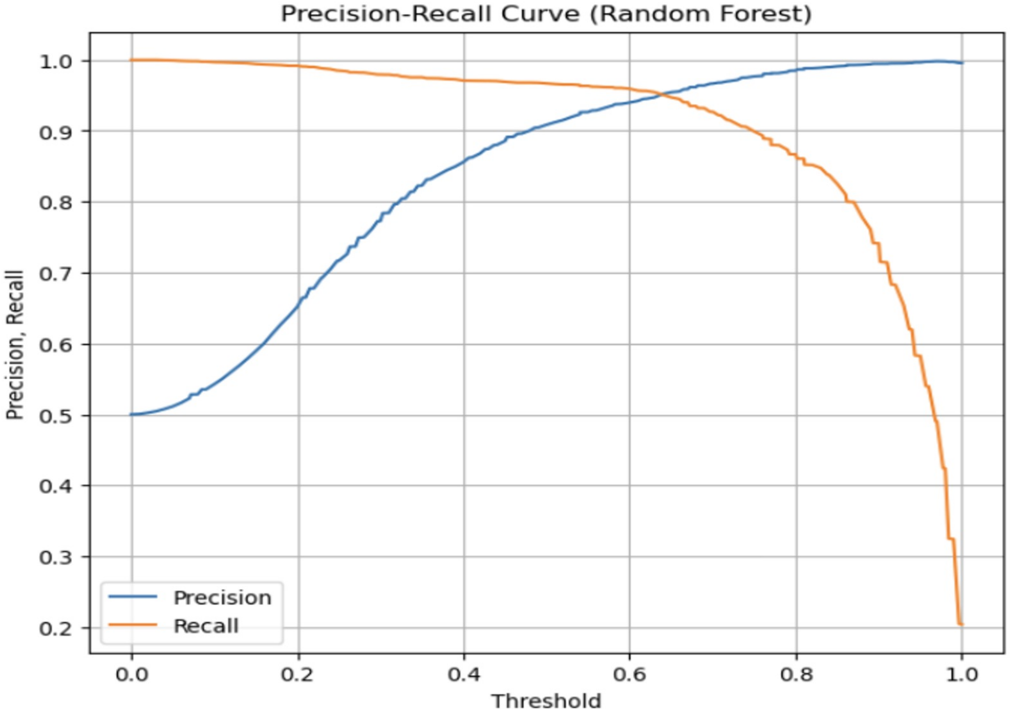
ROC Curve of Implemented Algorithms

- Random Forest outperforms with the highest AUC of 0.94
- Gaussian Naive Bayes shows the lowest performance with AUC of 0.55
- Logistic Regression and SVM have equal AUCs of 0.71, indicating similar performance



Precision-Recall Curve - Random Forest

The curve suggests that as the threshold for predicting a credit default increases, precision improves, but recall decreases.



Conclusion & Future Work

- The Random Forest algorithm yielded the most accurate predictions for credit default, suggesting it is well-suited for this application
- Investigate ensemble methods or advanced neural networks to further improve prediction accuracy
- Evaluation of model performance on a larger and more diverse dataset to ensure scalability and robustness
- Plan to integrate model predictions into a decision support system for financial institutions

Questions?