**Aim and Objectives**

This research aims to provide a comparative analysis of Logistic Regression, Random Forest, and XGBoost for credit default prediction, focusing on credit card consumers. The study will evaluate each model’s accuracy, interpretability, and robustness, examining how well they handle various feature interactions and respond to changes in data patterns. Key objectives include:

1. **Comparing Predictive Accuracy**: To measure and compare the accuracy of Logistic Regression, Random Forest, and XGBoost models in predicting credit defaults.
2. **Feature Impact Analysis**: To analyze how key features—such as financial history, income level, credit score, and spending behavior—impact model performance and prediction accuracy.
3. **Model Interpretability and Practicality**: To assess the interpretability and ease of use of each model in real-world banking scenarios, especially for decision-makers who may need to understand the reasons behind specific predictions.
4. **Recommendation of Best-Fit Model**: To identify and recommend the most effective model for credit default prediction based on the analysis.

### About the Dataset

The UCI Machine Learning Repository is the source of this dataset, which sheds light on the 2006 credit card debt problem that affected Taiwanese banks. Many banks gave credit cards to unqualified applicants in an effort to gain market share, which led to widespread debt buildup and higher cardholder delinquency rates. Customers' trust in personal finance was eroded by this practice, which presented serious problems for banks and consumers alike.

**Methodology**

This section explores the procedures and methods used to develop a model that uses a variety of machine learning approaches to forecast credit default. Data preparation, feature engineering, model selection, evaluation metrics, and model performance analysis are all covered in this part.

**Data Preparation**

The dataset was loaded and initially analysed using tools like df.head() and df.info() to explore its structure and identify key features, including demographics, credit history, and the target variable. Unnecessary columns, such as ID, were dropped to prevent overfitting. Categorical variables (e.g., SEX, MARRIAGE, EDUCATION) were encoded numerically using LabelEncoder. Exploratory Data Analysis (EDA) included statistical summaries (df.describe()), histograms for feature distribution, and a correlation heatmap to uncover relationships and multicollinearity. These steps ensured a clean, well-structured dataset for modeling.

**Feature Selection and Engineering**

Random Forest identified important features, highlighting variables with significant predictive power. Outliers were detected via box plots and managed using transformations or removals. Target variable analysis uncovered potential class imbalance, prompting strategies like resampling to enhance minority class predictions.

**Data Splitting and Scaling**

The dataset was split into training (70%) and testing (30%) sets using stratification to maintain class proportions. Features were scaled using StandardScaler for improved model performance.

**Modeling and Evaluation**

Three models were used: Logistic Regression, Random Forest Classifier, and XGBoost Classifier. Logistic Regression provided interpretability but struggled with minority class predictions. Random Forest offered insights into feature relevance and reduced overfitting through its ensemble approach. XGBoost excelled in handling class imbalance and provided robust accuracy. Evaluation metrics included confusion matrices, classification reports, and ROC-AUC scores, revealing trade-offs between sensitivity, specificity, and model robustness.

The analysis aids in selecting the optimal model for credit default prediction based on accuracy, interpretability, and feature importance.

**Results**

The models were evaluated on a dataset with two classes: class 0 (majority) and class 1 (minority). Key performance metrics, including precision, recall, and F1-score, were used to assess the ability of each model to classify the classes accurately.

Here’s a concise table summarising the performance metrics for each model:

| **Metric** | **Logistic Regression** | **Random Forest** | **XGBoost** |
| --- | --- | --- | --- |
| **Accuracy** | 68% | 81% | 81% |
| **Class 0 (Majority)** |  |  |  |
| Precision | 0.87 | 0.83 | 0.84 |
| Recall | 0.70 | 0.95 | 0.94 |
| F1-Score | 0.77 | 0.89 | 0.89 |
| **Class 1 (Minority)** |  |  |  |
| Precision | 0.37 | 0.64 | 0.63 |
| Recall | 0.63 | 0.34 | 0.36 |
| F1-Score | 0.47 | 0.44 | 0.46 |
| **Macro Avg F1-Score** | 0.62 | 0.67 | 0.67 |
| **Weighted Avg F1-Score** | 0.71 | 0.79 | 0.79 |

This table highlights the key metrics for all three models across both classes and their overall

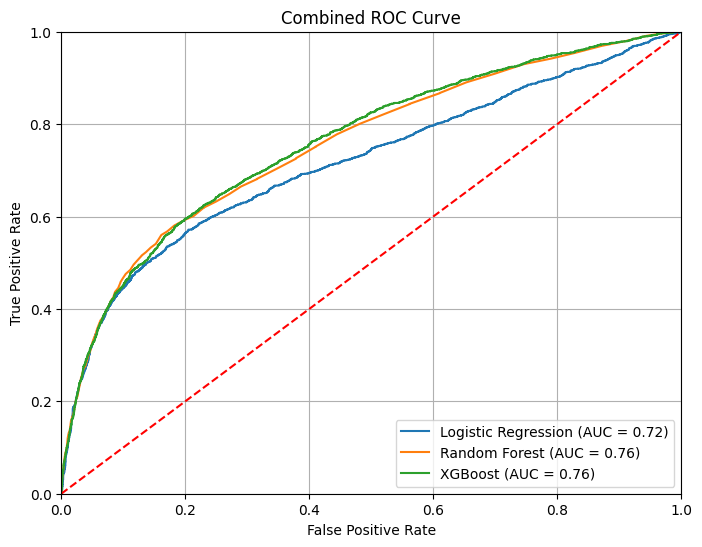
**Logistic Regression** achieved an overall accuracy of **68%**. Class 0 (majority) demonstrated strong performance with a precision of **0.87** and an F1-score of **0.77**, though its recall was relatively lower at **0.70**. Class 1 (minority) performance was weaker, achieving a precision of **0.37**, recall of **0.63**, and an F1-score of **0.47**. The model strongly favoured the majority class, as indicated by a weighted average F1-score of **0.71**.

**Random Forest Classifier** delivered a higher overall accuracy of **81%**. For class 0, it achieved a precision of **0.83**, recall of **0.95**, and an F1-score of **0.89**. However, its performance for class 1 was moderate, with a precision of **0.64**, recall of **0.34**, and an F1-score of **0.44**. The macro average F1-score of **0.67** highlights a trade-off between classes, though the weighted average metrics remained strong.

**XGBoost Classifier** matched Random Forest in accuracy at **81%**. Class 0 performance was similar, with precision of **0.84**, recall of **0.94**, and an F1-score of **0.89**. Class 1 metrics were comparable to Random Forest, with a precision of **0.63**, recall of **0.36**, and an F1-score of **0.46**. The macro average F1-score of **0.67** and weighted average of **0.79** reflect balanced overall performance.

**Key Insights**

Logistic Regression showed limited ability to balance minority class performance. Random Forest and XGBoost demonstrated better overall balance and robustness, with comparable accuracy but challenges in improving class 1 recall, resulting in moderate F1-scores for the minority class.



This Combined ROC Curve compares the performance of Logistic Regression, Random Forest, and XGBoost models in a classification task. The curves show the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across thresholds. Logistic Regression achieves an AUC of 0.72, indicating moderate performance, while Random Forest and XGBoost outperform it with AUC values of 0.76. The red diagonal represents random classification (AUC = 0.5), and all models exceed this baseline. Random Forest and XGBoost are the top-performing models, offering better TPR at various FPR levels. These models are preferable for deployment in this task, given their superior predictive power, provided computational complexity and interpretability are manageable concerns.