CLARK UNIVERSITY
SCHOOL OF PROFESSIONAL STUDIES

FINAL PROJECT ON
PREDICTING CREDIT SCORES WITH ML TECHNIQUES

Course Name and Number: Applied Data Analytics MSCS3045-01-S25
Semester / Year: Spring 2025

By:

1. Charan Reddy Devapatla
2. Tarun Emmanuel Majhi
3. Bhanu Prakash Akepogu

Instructor:
Dr. Salem Othman

Teaching Assistant:
Shubhika Jain

Abstract — This project mainly aims at the development of predictive models in the classification of credit scoring that can help in identifying the good, standard, and bad credit profiles of a Customer. By using various machine learning techniques such as Random Forest, Logistic Regression, and Multilayer Perceptron, this paper will go through complex data processing; from PCA for dimensionality reduction and feature selection, to class balancing by the SMOTE algorithm. The data have a wide array of financial, including the timing of EMI payments and occupation-related salary data, allowing for so much more overall assessment in creditworthiness. Via good analysis and model evaluation, the project had to offer financial institutions tools far more powerful and precise in their assessment of credit risk—reducing default rates and promoting fair lending. Our results show the potential for machine learning to entirely transform the traditional credit-scoring system by providing insights into major driving calculations behind credit scores and their implications for the financial sector.

Index Terms —dimensionality reduction, feature selection, logistic regression, multilayer perceptron, predictive modeling, principal component analysis (PCA), random forest, SMOTE, supervised learning,credit risk evaluation, clustering analysis, K-means clustering, encoding, confusion matrix.

INTRODUCTION

Credit scoring has been one of the important tools within the financial industry in determining whether an individual can repay or not; it has an impact on decisions related to loan approval, interest rates, and financial service. In old days, traditional credit-scoring systems often use static models that are not able to make adjustments to complex patterns and relationships in the data of the lenders. Techniques in machine learning can, therefore, tranform this process by providing more dynamic, accurate, and fair credit assessments.

This project aims to create predictive models for improving the accuracy of Credit Score prediction using machine learning techniques. The models will be very helpful to financial institutions in making informed decisions, reducing default risks, and increasing access to credit for deserving individuals.

PURPOSE

In this project we mainly focused on usage of machine learning to improve credit score predictions. Accurate credit scoring allows lenders to manage risks more effectively; it helps people with limited credit histories have access to financial services. By using data exploratory analysis, machine learning can provide more transparency into the evaluation process for credit scoring.

The aim of this project will be the design and evaluation of machine learning models to predict credit scores according to borrower attributes. This involves improving the accuracy of predictions, identification of key columns that influence credit scores, and development of models that can cover to wide variety of borrower profiles.

In summary, this project demonstrates that machine learning can be used to enhance credit score prediction through increased accuracy and identification of significant influencing factors. The developed models are therefore very useful for financial institutions to practice better risk management and make informed decisions.

PROJECT DESCRIPTION

Objective:

In this project we developed machine learning models for credit score forecasting, enabling financial institutions to evaluate the creditworthiness of loan applicants more effectively. This would help improve decision accuracy while reducing default risks and ensuring greater financial inclusion.

Dataset:

The dataset we used utilized a very comprehensive dataset[1] which includes banking and credit information such as income, credit history, payment habits, outstanding debts, and patterns of loan repayment. These details enable a much-needed in-depth look into what really determines credit scores and will be used to develop a predictive model.

Methodology:

- Preprocessing: Preprocessing step includes data cleaning, handling missing values, normalizing numerical features, and encoding categorical variables.

- Dimensionality Reduction: Apply Principal Component Analysis (PCA) to reduce redundancy and improve computational efficiency.
- Feature Selection: Identifies the most relevant predictors, such as outstanding debt and payment behavior, to improve both model interpretability and accuracy.
- Class Balancing: Using SMOTE to handle class imbalances, ensuring that the models are trained fairly.
- Machine Learning Models: Implementation and Comparison in the Performance among Logistic Regression, Decision Trees, Random Forest, and Multilayer Perceptron.
- Evaluation Metrics: Assessing model performance using accuracy, F1-score, recall, and precision to determine the best-performing model. Random Forest achieved the highest accuracy of 82.2%, demonstrating its predictive power.

DATA DESCRIPTION

The dataset comprises 100,000 observations and extensive financial and behavioral information to predict credit scores. Some of the key numeric features include age, income, loan details, credit history, and monthly financial patterns. The data is so diverse that it guarantees robust predictive modeling applicable to a wide variety of customer profiles.

Overview of Dataset Attributes:

KEY FEATURES

- Age: Age of the candidate is 14 years to 72 years.
- Annual Income: tells the annual income, on average around 50,505 with a standard deviation of 38,299.
- Monthly In-Hand Salary: Shows the monthly take-home pay, which averages 4,198.
- Total EMI Per Month: Shows the average monthly loan repayment, which has a mean of 107.
- Number of Bank Accounts: Total active accounts per person.
- Number of Credit Cards: This calculates credit card ownership, spanning from 0 to 15 cards.
- Interest Rate: The interest rate on the loans; it has a mean of 14.5%.
- Number of Loans: The number of loans taken, with an average of 3.5 loans.
- Outstanding Debt: This includes unpaid loan balances, and the average is 1,426.

- Credit Utilization Ratio: Measures the percentage of credit in use, 32% on average.
- Credit History Age: This refers to how long credit history has been in months, averaging 221 months(18yrs)
- Monthly Balance: This is the average balance after expenses, at an average of 403.

DISCRETE FEATURES

- Month: Specific months in the dataset, bounded by a narrow range of unique values.
- credit_mix: Refers to the credit mix classification, e.g., Good, Standard, Bad.
- payment_behaviour: Describes the categories of payment behavior with limited unique values.

TABLE I DESCRIPTIVE STATISTICS OF NUMERIC FEATURES

Age	33.31	10.76	14.00	24.00	33.00	43.00	72.00
Annual Income	50,505.12	38,299.42	7,005.93	19,342.97	38,340.65	74,020.54	810,000.00
Monthly In-Hand Salary	4,198.77	3,187.49	303.64	1,626.76	3,268.23	6,014.88	30,000.00
EMI Per Month	107.04	130.03	0.00	29.20	76.00	139.00	1,200.00
Number of Bank Accounts	5.37	2.59	0.00	3.00	5.00	7.00	20.00
Number of Credit Cards	5.53	2.07	0.00	4.00	5.00	7.00	15.00
Interest Rate	14.53	8.74	1.00	7.00	12.00	20.00	40.00
Number of Loans	3.53	2.45	0.00	2.00	3.00	5.00	15.00
Outstanding Debt	1,426.22	1,155.13	0.23	566.07	1,034.18	1,789.08	10,000.00
Credit UtilizationRatio	32.29	5.12	20.00	28.05	31.00	36.00	60.00
Credit History Age	221.21	99.68	1.00	144.00	210.00	288.00	600.00
Monthly Balance	403.12	214.01	0.01	270.19	398.23	520.00	1,500.00

Some Exploratory Data analysis

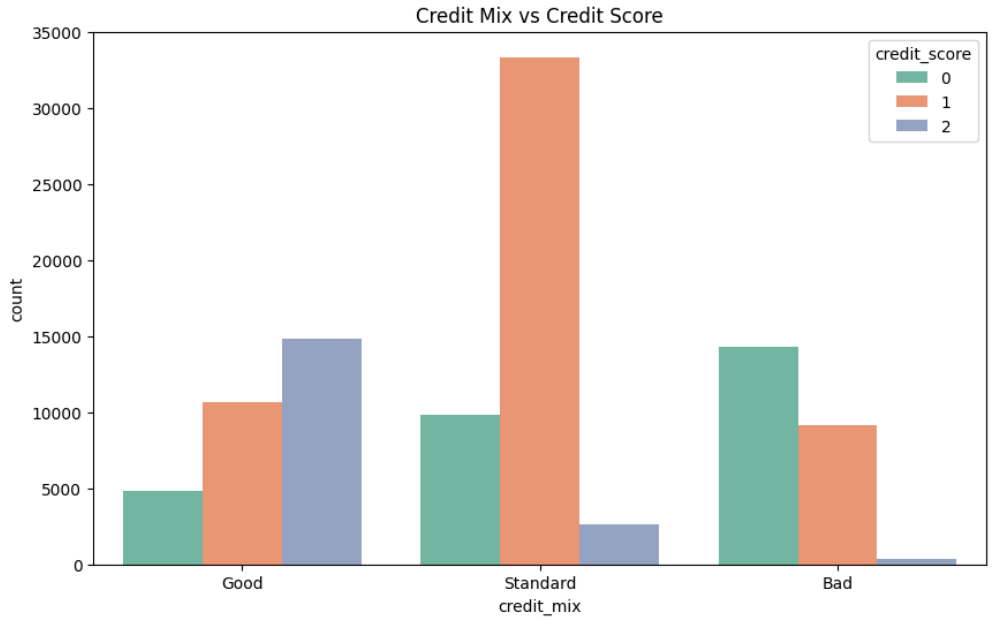

Fig.1 Bar graph for Credit Mix vs Credit Score

From the above figure 1,[2]we observed that Most individuals with a standard credit mix have a score of 1 (medium). In the case of the good credit mix, higher scores (2) dominate, while the bad credit mix has lower scores (0).

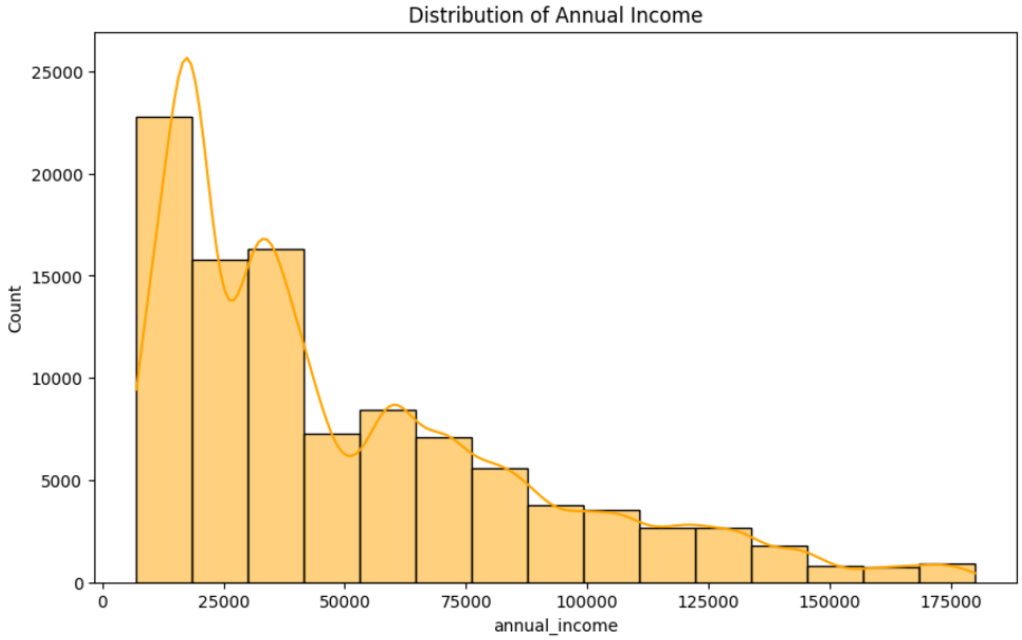

Fig 2. Histogram showing distribution of annual income

Above chart in Figure 2 represents the distribution of annual income. Most of the people in this group have yearly incomes between 10,000 and 40,000. The number slightly drops for incomes above 50,000, while few have an income of over 150,000.

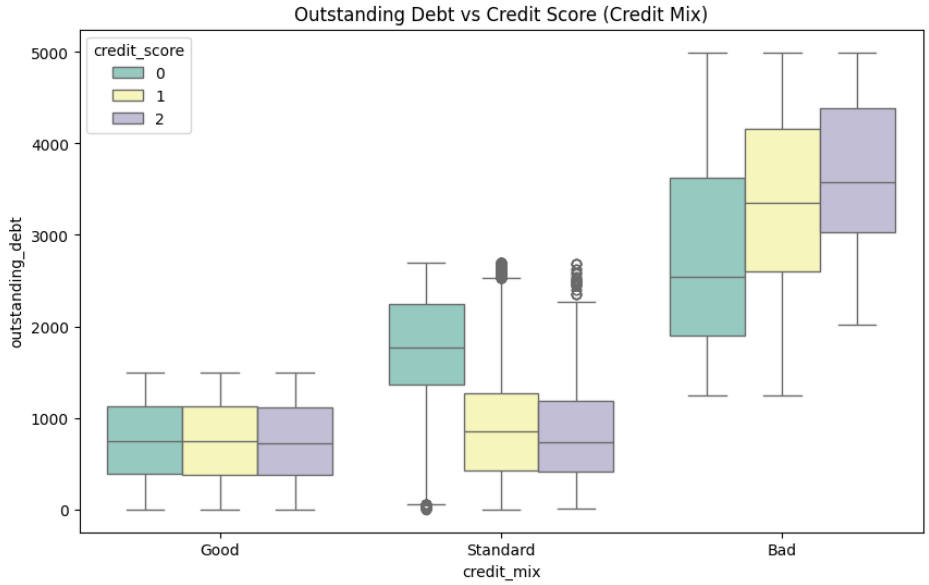

Fig 3. Box plot showing various credit score for debt

From Above Fig 3. We observed that those with a good credit mix have lower outstanding debts, generally below 1,000. The bad credit mix exhibits much higher outstanding debts, mostly over 3,000, showing how high debt impacts low credit scores.

DETAILS OF THE METHODOLOGY & DISCUSSION

Principal Component Analysis:

We first used Principal Component Analysis (PCA) for reducing dimensionality and thus improving model performance. Based on the explained variance ratio, we selected the optimal number of components to be used for model training. We set the explained variance ratio to 80%. This means that the selected principal components capture 80% of the variance in the data. The number of components required to achieve 80% variance was found to be 11.

With these 11 components, we trained several machine learning models and found the

accuracy to be highest of 72.7% with the Random Forest model. Increasing this number further improved performance by slight margin and in some cases reduced the accuracy as well due to redundant information in dataset.

Table II.PERFORMANCE METRICS WITH 80% VARIANCE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.640133	0.638133	0.640133	0.633754
LDA	0.636100	0.637923	0.636100	0.635295
Decision Tree	0.616767	0.617279	0.616767	0.617014
Random Forest	0.727267	0.726662	0.727267	0.725943
Multilayer Perceptron	0.692533	0.692469	0.692533	0.692103

For 90% variance: we set the explained variance ratio to 90%, which needed 16 components. Including these other components gave better accuracy in most models. We noted that, the highest accuracy of 75.4% was achieved using the Random Forest model. Similarly, the performance of other models, such as Logistic Regression and LDA, also improved slightly, though they did not outperform the Random Forest model.

Table III.PERFORMANCE METRICS WITH 90% VARIANCE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.645067	0.644007	0.645067	0.640413
LDA	0.644467	0.647340	0.644467	0.644575
Decision Tree	0.648333	0.648459	0.648333	0.648385
Random Forest	0.754133	0.753856	0.754133	0.753596
Multilayer Perceptron	0.714167	0.713803	0.714167	0.713287

Feature Importance:

To understand feature-level performance better, we looked at accuracy as features were added one by one in order of importance. A line plot (Fig. 5) was created where the x-axis represents the number of features and the y-axis represents the accuracy.

From below diagram we noted how accuracy changes when features are added one by one

in order of importance. Accuracy fluctuates at the beginning; between 5 and 10 features, accuracy takes a big dip, possibly due to some added features as they are irrelevant. From 10 to 20 features, accuracy improves constantly, showing these features contribute meaningfully to the model's performance.

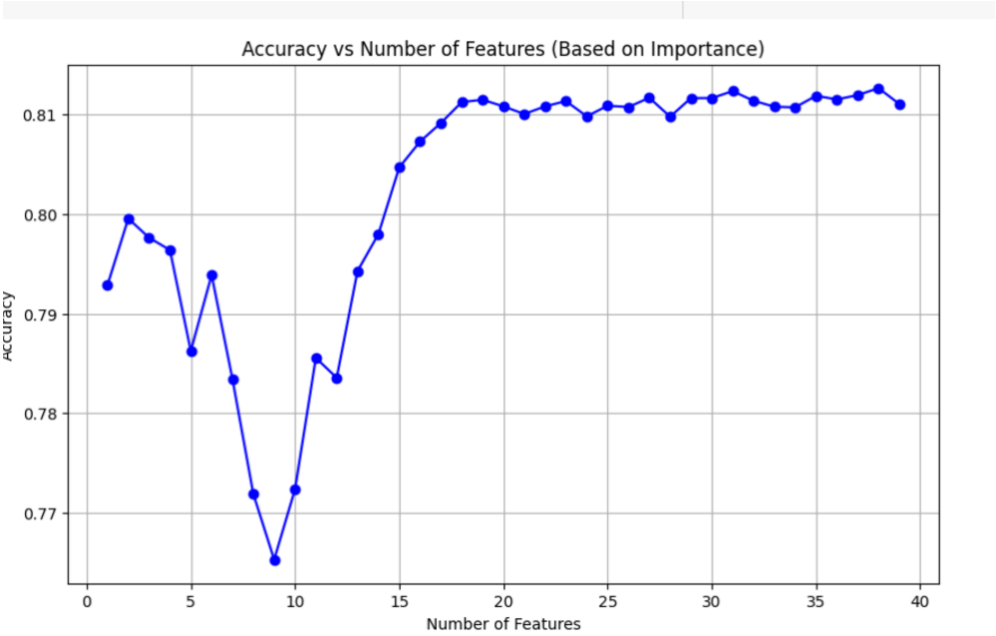

Fig5. Line chart for no. of features based on importance

Observations:

At the beginning, accuracy does change quite a lot for the first 5 features. This probably happens because the model had more influence by a few key features, but the lack of supporting features causes inconsistency in the predictions.

Feature selection:

Dip Between 5 and 10 Features:

Increasing the number of features to between 5 and 10, you can see a drop in accuracy. This could mean that some of the added features are either irrelevant which is not needed into the model.

Improvement Between 10 and 20 Features:

From 10 to 20 features, the accuracy increases at a constant phase, which means that such features bring useful information to help the model generalize better and improve its predicting power. Beyond 20 features, the accuracy flattens out, meaning that the additional features do not contribute to model performance. This means that feature selection is crucial in not including irrelevant data.

D. Evaluation metrics:

1.Accuracy: It shows how many predictions were correct out of all predictions made

Formula = TP+TN/TP+FP+FN+TN

2. Precision: It tells how many of the predicted positives were actually correct

Formula = TP/TP+FP

3. Recall: It shows how many actual positives were correctly identified

Formula =TP/TP+FN

4. F1-Score It balances precision and recall, calculated by Harmonic progression

Formula=2×Precision×Recall/Precision+Recall

Table IV. PERFORMANCE METRICS BEFORE CLASS BALANCING

Model	Accuracy	Precision	Recall	F1 - Score
Logistic Regression	0.637633	0.637307	0.637633	0.635724
LDA	0.639133	0.653649	0.639133	0.641580
Decision Tree	0.797533	0.797576	0.797533	0.797553
Random Forest	0.817933	0.818076	0.817933	0.817965
Multilayer Perceptron	0.699633	0.701599	0.699633	0.698979

DISCUSSION OF RESULTS

Logistic Regression:

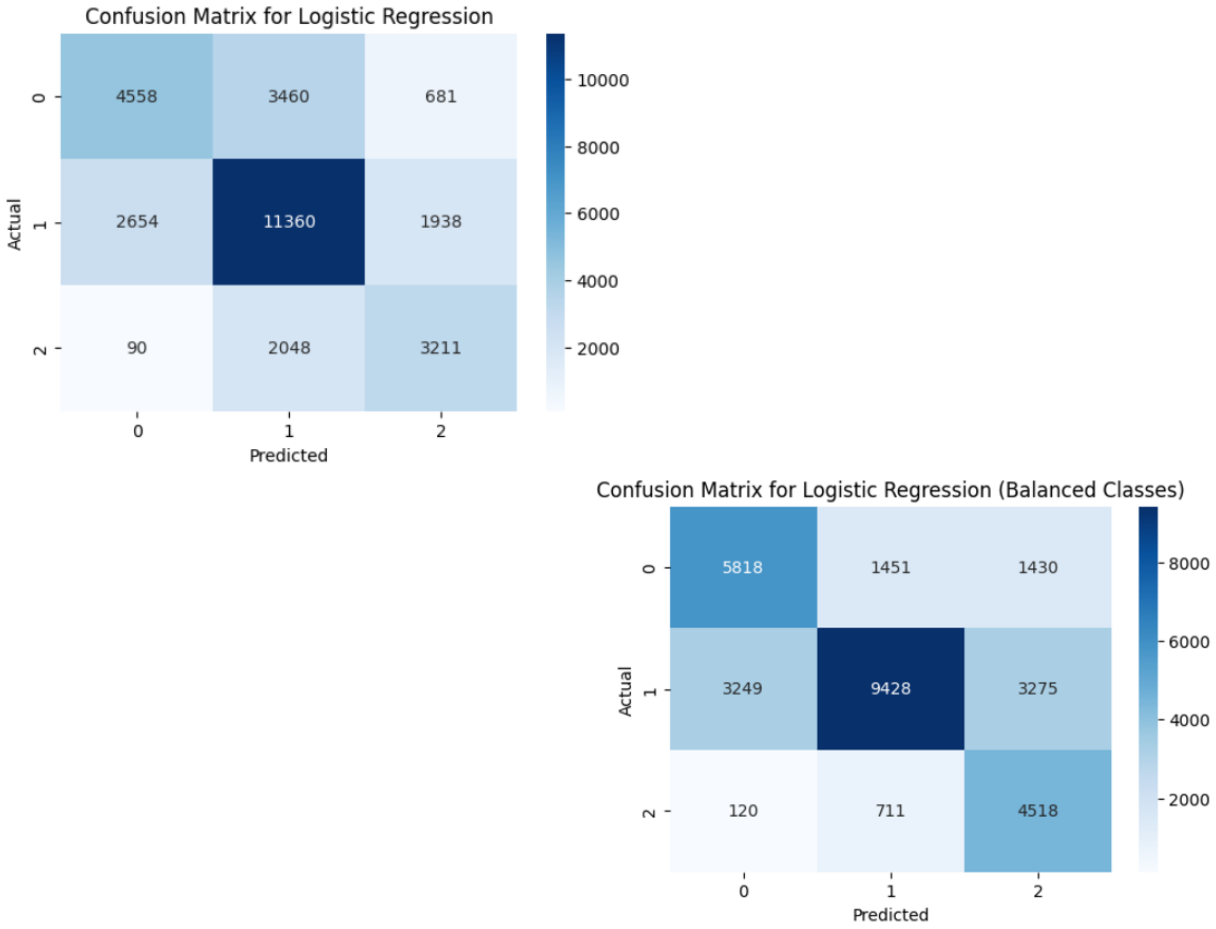

Fig6. Confusion matrix for Logistic Regression

For Logistic Regression with imbalanced classes, Class 1 has the highest false negative rate at 40.9%, which improved to 33.9% after balancing. The false positive rate for Class 2 was 51.0% and reduced to 43.6% with balanced classes. The greatest decrease in false negative rate occurred for Class 0 from 33.8% to 19.8%. Class 2 had the lowest false negative rate at 15.6%. This model realized an accuracy of 65.9%.Model performed relatively well in predicting class 1, with a precision of 70% and recall of 65%.

Linear discriminant analysis:

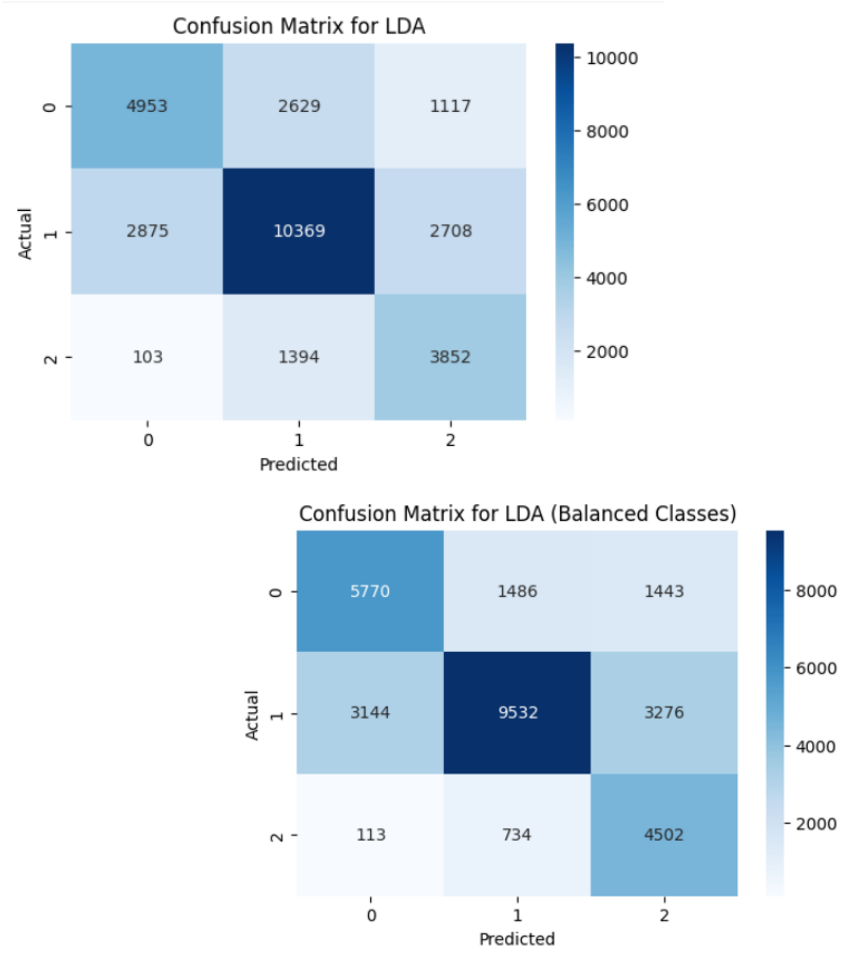

Fig7. Confusion matrix for LDA

For LDA, Class 1 also presented the worst false-negative rate of 37.3% with the original imbalanced matrix and moves to 33.0% after balancing. Again, the false-positive rate of Class 2 decreased, coming from the initial 50.2% to the balanced state of 41.2%. Class 0 also got improvement.

Its false negative went from 33.9 to20.4. LDA achieved 66%. It was better at predicting class 1, with precision and recall similar to Logistic Regression at about 70%, but misclassifications occurred for classes 0 and 2.

Decision tree:

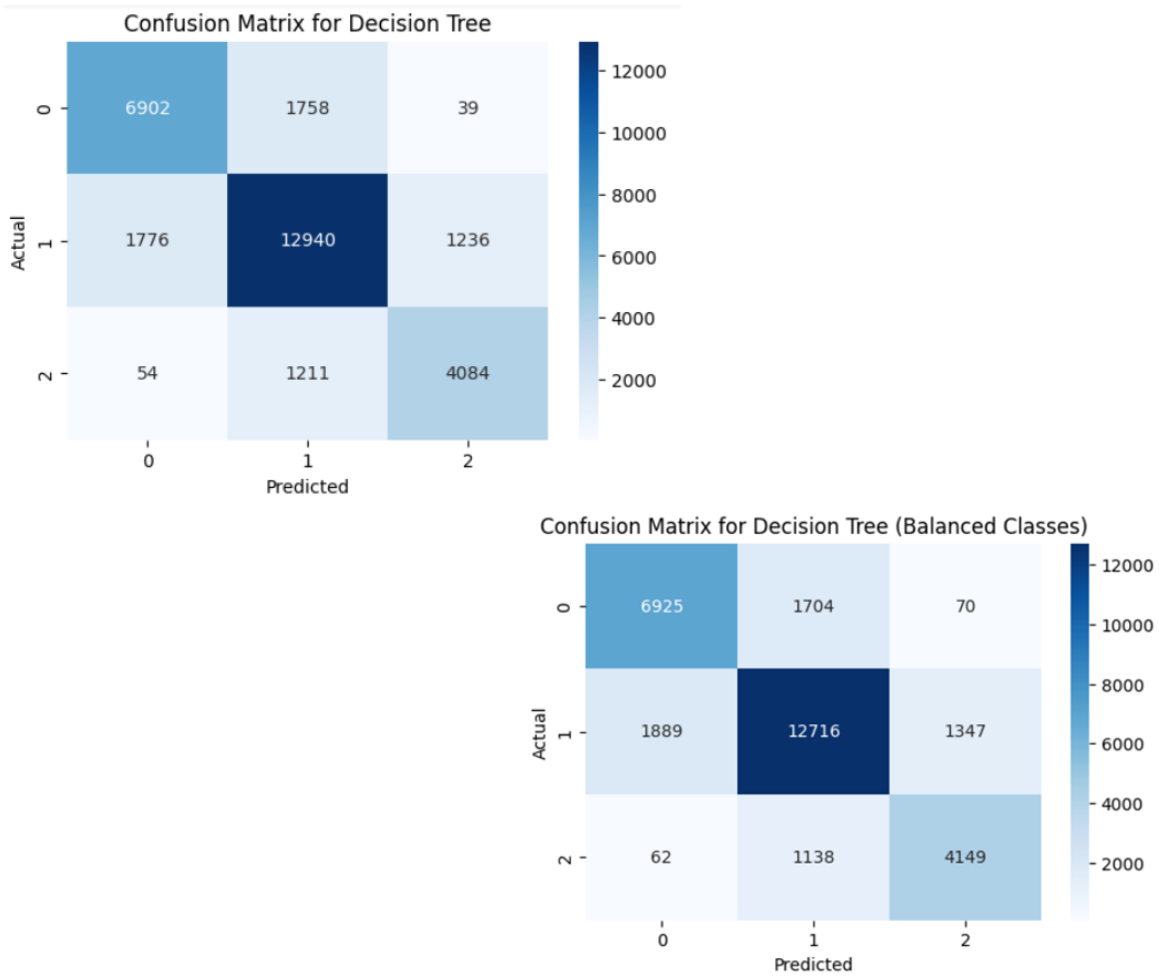

Fig8. Confusion matrix for LDA

Decision tree- For the imbalanced classes, Class 1 has the highest false negative rate at 23.0%, which decreased a little to 22.2% when classes are balanced. Class 2 had the false positive rate of 29.6%, and that also reduced to 27.4% after balancing. The false negative rate of 20.3% for Class 0 is actually the lowest and improved even further to 19.7% upon balancing. The lowest false negative rate is 22.9% for Class 2 for both cases.

This model got an accuracy of about 79.3%. It was quite good in class 1 prediction with precision and recall close to 79% and did fewer mistakes compared to LDA and Logistic

Regression

Random Forest:

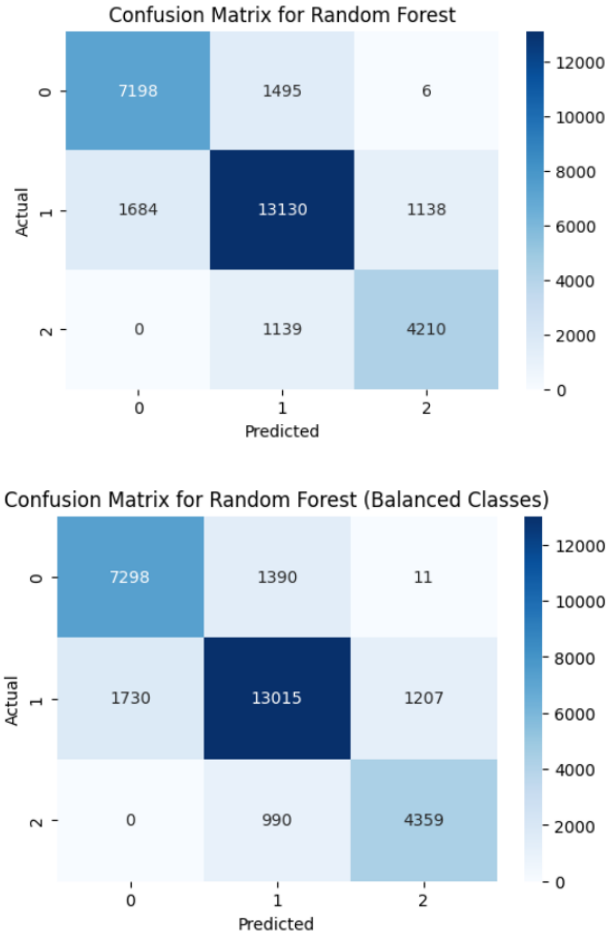

Fig9. Confusion matrix for Random Forest

[6]For Random Forest, Class 1 is the one with the highest false negative rate at 22.7% for the imbalanced classes; this slightly fell to 22.2% in balanced classes. The false positive rate for Class 2 was very high, being 21.3%, before improving a little after balancing at 18.5%. Class 0 has a false negative rate of 17.2%, and it shows a small reduction to 16.0% in the balanced model. Class 2 also held a rather low false negative rate that decreased from 21.3% to 18.5% after

balancing.

Random Forest handled the class imbalance well and kept a high overall accuracy of 82.2%, being outstanding in predicting class 1 & class 0 with precision and recall over 82%. while improving predictions for the minority class (Class 2).

Multi layer perceptron:

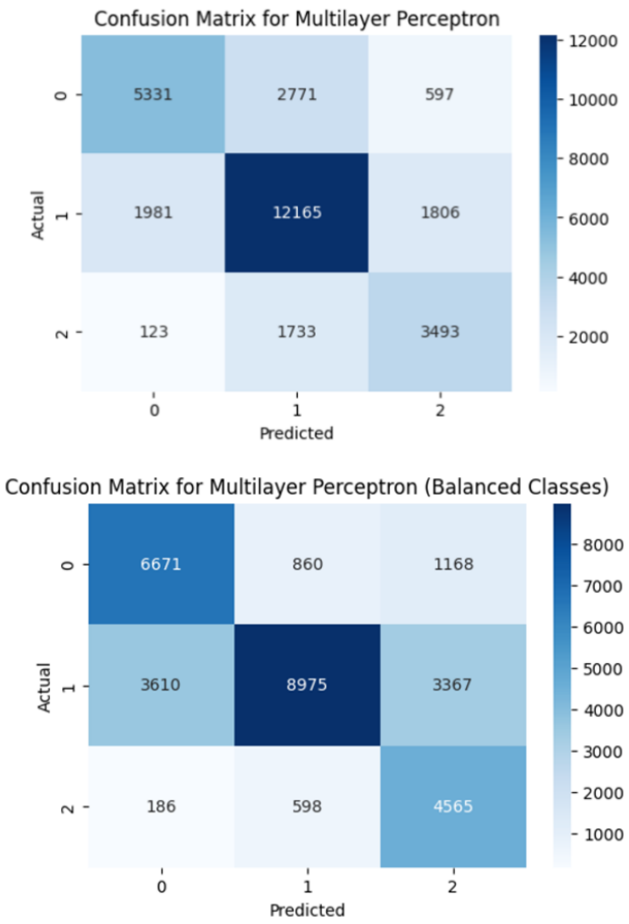

Fig10. Confusion matrix for Multilayer Perceptron

In the case of Multilayer Perceptron (MLP) and with imbalanced classes, Class 1 presented the highest false negative rate at 31.3%, which was highly improved to 28.7% after balancing. In addition, the false positive rate of Class 2 was 33.2%, reducing to 26.0% after balancing.

Moreover, Class 0 had a false negative rate of 34.2%, significantly improved to 21.4%. The false negative rate of Class 2 reduced from 33.2% to 22.0% with balanced classes.

This model gave an accuracy of around 67.3%. This performed reasonably well at class 1 predictions with precision at 73% but poorly for class 0 and 2.

TABLE IV PERFORMANCE METRICS ON BALANCED DATASET

Model	Accuracy	Precision	Recall	F1 – Score
Logistic Regression	0.6588	0.7035	0.6588	0.6632
LDA	0.6601	0.7036	0.6601	0.6648
Decision Tree	0.7930	0.7937	0.793	0.7932
Random Forest	0.8224	0.8232	0.8224	0.8225
Multilayer Perceptron	0.6737	0.7316	0.6737	0.6762

The **Random Forest** outperformed all others, with an accuracy of 82.2% and an F1-score of 82.3% after balancing; hence, strong in performance on all measures. Decision Tree is competitive with an accuracy of 79.3%, while Logistic Regression and LDA showed slight improvement from their imbalanced classes

STACKING-COMBINING PREDICTIONS:

TABLE V PERFORMANCE METRICS AFTER STACKING

	Accuracy	Precision	F1-Score	Recall
Stacking	0.823821	0.825897	0.823955	0.823821

We have implemented a stacking classifier, which combines machine learning models. It uses its base models: Logistic Regression, Decision Tree, Random Forest, and Multi-Layer Perceptron (MLP) to learn the pattern in the data, each independent of the other.
The stacking classifier is trained on a balanced version of the dataset to reduce the effect of class imbalance, which may show difference the predictions. Once trained, the model is tested

on a hold-out test set, and its performance is evaluated based on accuracy, precision, recall, and F1-score metrics. The results here are very good, as the stacking model has 82.4% accuracy and an F1-score of 82.4. This shows that stacking works well to catch difficult relationships in data by merging the strengths of multiple algorithms.

OPTIMAL NUMBER OF CLUSTERS USING SILHOUETTE SCORES

The clustering analysis identifies patterns and groups in the dataset as per K-Means clustering[4]. The Silhouette Score was calculated for the number of clusters ranging from 2 to 10 in order to determine the optimal number of clusters. From the below silhouette score graph, We observed that the highest score, about 0.28, was achieved with 2 clusters; thus, this is the most appropriate choice to group the data. Beyond 2 clusters, the Silhouette Score drops significantly to about 0.16 for configurations of 5 or more clusters, indicating less connection and separation among the clusters.

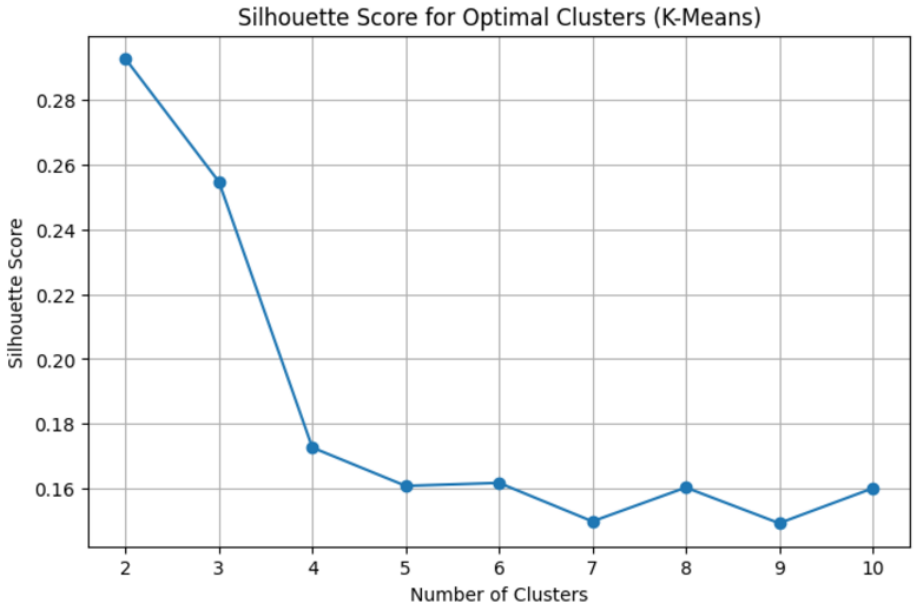

Fig6. Line Graph showing Silhouette score for clusters

K-MEANS CLUSTERING:

K-Means clustering was performed with n=2 clusters. The clusters obtained were visualized in two dimensions using Principal Component Analysis (PCA) for dimensionality reduction. The scatter plot below shows two clear clusters: Cluster 0 and Cluster 1. Points for Cluster 0 are clustered on the left side, ranging from about -4 to 1 along Principal Component 1, while points

for Cluster 1 have more in the right side, with values from about 0.5 to 6 along the same axis.

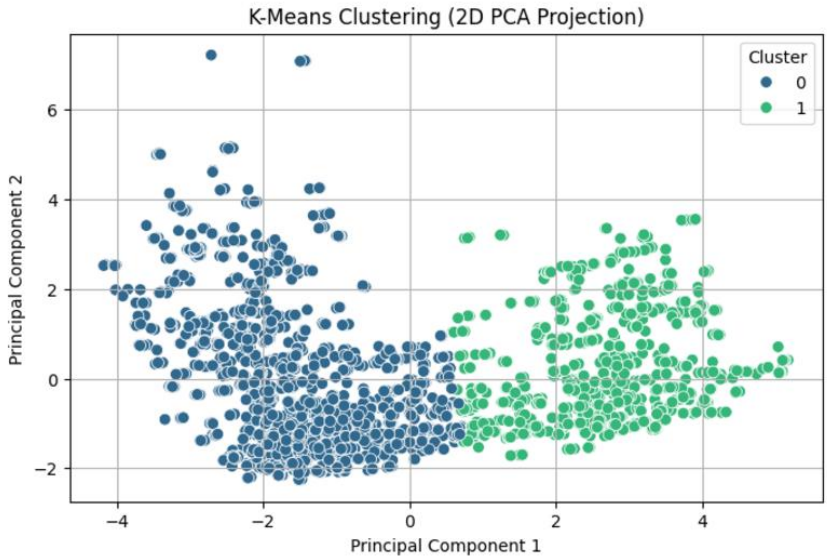

Fig7. Scatter Plot PC1 Vs PC2 for 2 clusters

Principal Component 1:

Credit Mix Standard: Differentiates individuals based on limited credit sources.

Amount Invested Monthly: Highlights monthly financial planning or income levels.

Total EMI Per Month: Reflects the loan burden based on installment basis.

Principal Component 2:

Payment of Minimum Amount (Yes): Indicates financial behavior based on minimum payment habits.

Number of Bank Accounts: Differentiates individuals by financial activity.

Number of Credit Cards: Captures credit usage and availability of credit lines.

OPTIMAL NUMBER OF CLUSTERS USING SCREE PLOT (Elbow Method)

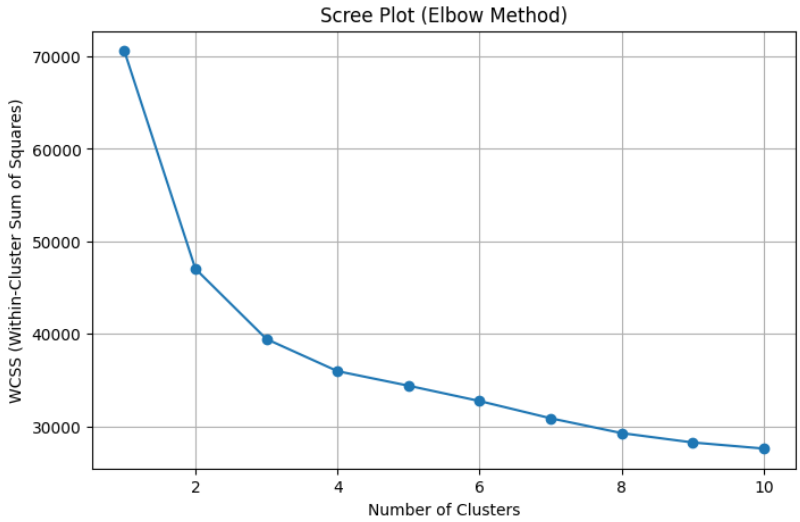

Fig 8. Scree Plot of number of clusters vs WCSS

The plot shows that the WCSS drops very steeply up to 1 to 4 clusters. From 4 onwards, the rate of decrease of WCSS really slows down, and the curve is almost flat. This would strongly suggest that the optimal number of clusters is probably 4, as beyond that, adding more clusters gives little return in terms of the reduction of WCSS.

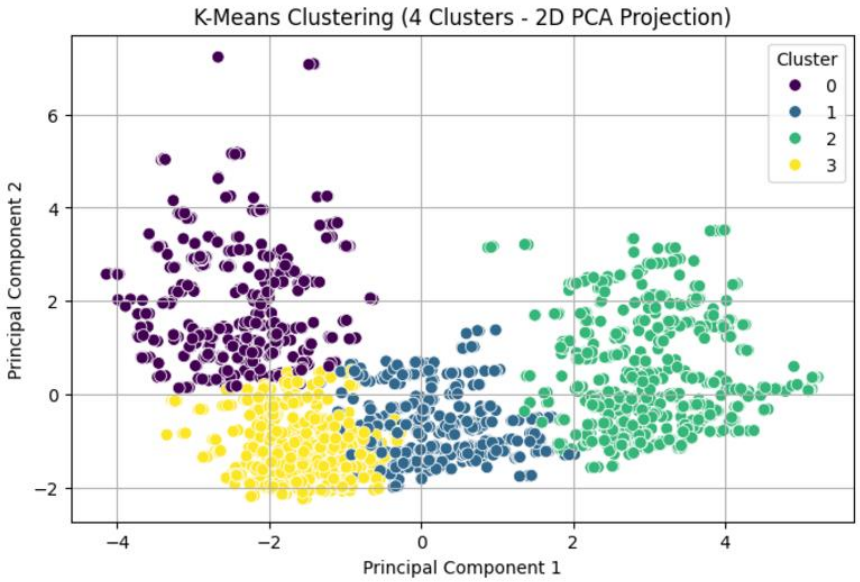

Fig 9. 2D- PCA Projection of 4clusters

From this PCA plot we observed well-separated clusters that show clear behavioral and financial patterns .While cluster 1 (Blue) characterizes the high financial activities of

participants, cluster 2 (Green) and cluster 3 (Yellow) point to some seperation like credit dependence versus debt pay-offs.

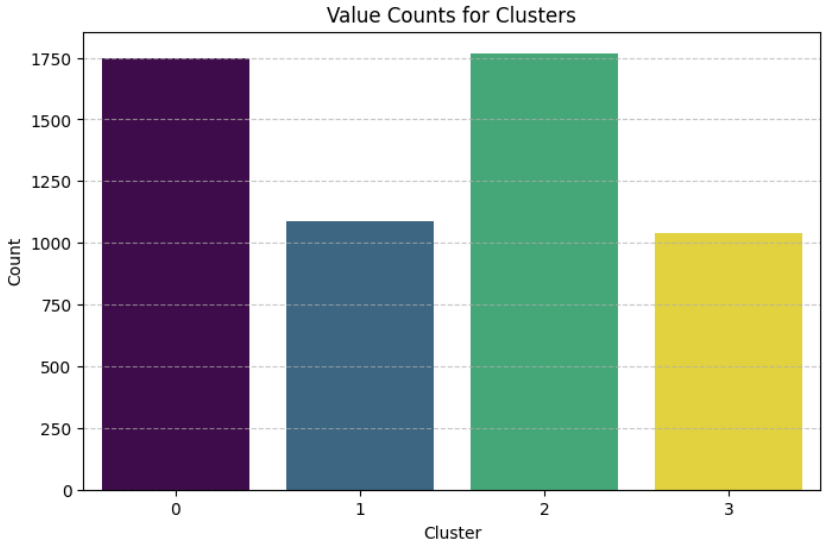

Fig 10. Bar graph showing value counts for clusters

Cluster 0 and Cluster 2 contain the largest number of data points, meaning that these clusters represent the most frequent patterns or groups in the dataset.

Cluster 1 and Cluster 3 contain fewer data points and so may represent niche or less frequent patterns in the data.

Balanced Distribution:

The distribution of clusters seems to be meaningful, as no cluster is too small or much larger than the others. This balance suggests that the partition of the data into 4 groups nicely captures the structure underlying the data without generating redundant or too scattered clusters.

CONCLUSION

This project is mainly based upon the loan approval process by applying advanced machine-learning techniques to predict credit scores effectively. With the implementation of robust data preprocessing techniques and models such as Logistic Regression, Decision Trees, Random Forest, and Multilayer Perceptron, this project assures a considerable increase in the accuracy of prediction. Class balancing techniques, such as SMOTE, reduced the bias in predictions and also improved the generalization aspect of the model toward the diverse profiles of customers.

Among the tested models, Random Forest is the most reliable, having the highest accuracy at 82.2% and the lowest false positive rate at 5.92%. It showed consistently good results in finding classes that are underrepresented and balanced false positives with false negatives. Other models, such as Decision Trees and Multilayer Perceptron, also had good results, which means machine learning could potentially lead a transformation in credit scoring by identifying important predictors like income and payment behavior.

Results underline the sensitivity of joining technical innovation with accurate considerations in credit risk evaluation. This study serves to compare the different models, hence it opens ways toward the creation of even more developed predictive analytics tools in the financial sector. In this context, machine learning techniques don't just give their help in operational efficiency but also to fair and informed decision-making in the banking industry.

References

[1] https://www.kaggle.com/datasets/parisrohan/credit-score-classification
[2] Shrawan Kumar Trivedi,A study on credit scoring modeling with different feature selection and machine learning approaches,Volume 63,2020,
https://doi.org/10.1016/j.techsoc.2020.101413.
[3] Bucker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2021). explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098
[4]"Credit Score Prediction Using Ensemble Model" by Dhillipkumar (2023)
This project demonstrates the application of ensemble models, including Random Forest, for credit score prediction.
https://github.com/Dhillipkumar/Credit-Score-Prediction-ML
[5]"Using Random Forest to Predict Credit Defaults Using Python" by Karina Kervin (2024)
https://developer.ibm.com/tutorials/awb-random-forest-predict-credit-defaults/
[6]"SMOTE for Imbalanced Classification with Python" by Jason Brownlee (2020)
This article offers implementing SMOTE in Python to handle imbalanced datasets.