CREDIT SCORE CLASSIFICATION

By Group 2

Tishani Wijekoon(S16379), Chami Sewwandi(S16028), W.K.Hiruni Hasara(S16210), S.Luxan(s16329)

ABSTRACT

This study explores the application of machine learning techniques to classify individual credit scores into three categories: Poor, Standard, and Good. Using a longitudinal dataset containing over 100,000 monthly credit records for 12,500 customers, we conducted thorough data preprocessing, feature engineering, and exploratory data analysis to understand key patterns and correlations.

Several classification models were tested, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, XGBoost, LightGBM, and CatBoost. Class imbalance was addressed using inbuilt class weighting. Among all models, **CatBoost with tuning and class weighting** achieved the highest performance, with a test accuracy of **85%**, followed by **Random Forest (79%)**, demonstrating the effectiveness of tree-based ensemble methods.

The results highlight the limitations of traditional static credit scoring approaches and show that machine learning can offer more adaptive and accurate credit risk assessments. This has significant implications for financial institutions seeking to improve loan decision-making and customer evaluation through data-driven methods.

INTRODUCTION

Credit scoring is a fundamental component of modern financial risk assessment, serving as a critical tool for lenders to evaluate the creditworthiness of potential borrowers. The ability to accurately classify individuals into appropriate credit risk categories directly impacts lending decisions, interest rates, and overall portfolio management strategies. Understanding the underlying factors that contribute to credit score classifications is essential for both financial institutions and consumers seeking to improve their financial standing. *(Arram, 2023)*

Traditional credit scoring models rely heavily on historical payment patterns and basic demographic information. However, with the increasing availability of comprehensive financial data, there is an opportunity to explore more nuanced relationships between various customer attributes and credit outcomes. This analysis aims to provide insights into these complex relationships through systematic exploratory data analysis.

The significance of this research extends beyond academic interest, as improved understanding of credit score determinants can lead to more fair and accurate risk assessment tools, better financial products, and enhanced consumer education initiatives. By identifying the most influential factors in credit score classification, financial institutions can develop more targeted strategies for risk management and customer acquisition.

THE QUESTION WE ARE GOING TO ANSWER

"What demographic, financial, and behavioral variables are most influential in differentiating customers with Good, Standard, and Poor credit scores?"

This research question addresses several key aspects:

- How do debt burdens and payment delays vary across score groups?
- Are certain occupations or age bands over-represented in Poor scores?
- Does the distribution of credit scores shift between January and August?

The question is designed to guide a comprehensive exploratory analysis that will identify the most significant predictors of credit score categories, ultimately providing insights for both theoretical understanding and practical application in credit risk assessment.

DATA SET

The dataset utilized in this analysis originates from Kaggle's "Credit Score Classification" dataset. This comprehensive dataset provides a unique longitudinal perspective on credit behavior, containing exactly 100,000 monthly observations representing 12,500 unique customers tracked across eight consecutive months from January to August.

Table 1- Data set description

Demographic Variables	
Variable	*Description*
Age	Customer age
Occupation	Professional category (16 distinct occupations)
Name	Customer name
SSN	Represents the social security number of a person
Customer_id	Represents a unique identification of a person

Credit Account Information	
Variable	*Description*
Total_EMI per month	The monthly EMI payments (in USD)
Num_Bank_Accounts	Number of bank accounts
Num of credit inquiries	Represents the number of credit card inquiries
Credit history age	The age of credit history of the person
Num_Credit_Card	Number of credit cards
Num_of_Loan	Number of active loans
Interest_Rate	Credit card interest rate
Types of loan	The types of loan taken by a person
Payment behavior	The payment behavior of the customer (in USD)

Financial Variables	
Variable	*Description*
Annual_Income	Yearly income
Monthly_Inhand_Salary:	Monthly disposable income
Outstanding_Debt	Current debt obligations
Credit_Utilization_Ratio:	Percentage of credit limit used

Behavioral Variables	
Delay_from_due_date:	Average payment delay in days
Payment_of_Min_Amount:	Whether minimum payments are made
Payment_Behaviour:	Spending and payment patterns
Credit_Mix:	Quality of credit portfolio
Changed_Credit_limit	The percentage change in credit card limit
Num of delayed payments	The average number of payments delayed by a person
Monthly balance	The monthly balance amount of the customer

Target Variable	
Variable	*Description*
Credit_Score	**Categorical outcome (Good, Standard, Poor)**

IMPORTANT RESULTS OF THE DESCRIPTIVE ANALYSIS

We began our analysis with rigorous data preprocessing. Missing values were handled using a structured imputation process—customer-wise forward-filling for stable variables, group-wise means for monthly-varying variables, and overall mean/mode imputation for any remaining gaps. Outliers were detected using the IQR method and retained, given the possibility that they represent valid financial extremes.

Feature selection involved removing irrelevant or problematic variables such as identifiers (Customer_ID, Name, etc.), Credit_History_Age (due to formatting issues), and Monthly_Inhand_Salary (due to multicollinearity with Annual_Income). We simplified high-cardinality categorical variables: occupations were grouped into three broader socioeconomic categories, and six payment behavior types were consolidated into three levels based on spending and payment amounts.

The target variable, Credit_Score, was imbalanced: 53.2% of observations were “Standard,” 29.0% “Poor,” and only 17.8% “Good.” This skew may affect model performance and suggests the need for resampling methods. Univariate analysis revealed skewed distributions and extreme values for variables such as Outstanding_Debt and Amount_Invested_Monthly. In bivariate analysis, key relationships were identified—lower credit scores were associated with a higher number of credit cards, loans, and payment delays. Importantly, customers who consistently paid only the minimum amount showed a significantly higher likelihood of having poor credit.

Although the dataset spans eight months (January–August) for 12,500 customers (100,000 records), we treated each record independently due to project constraints. A chi-square test showed that credit score distribution varies significantly across months ($p < 0.05$), with July having a slightly higher proportion of “Good” scores, suggesting mild seasonal effects.

Lastly, unsupervised clustering via K-means on FAMD-reduced data showed weak natural structure. The best silhouette score occurred at $k = 2$ (0.4416), but clusters overlapped considerably, indicating that credit behaviors don't separate cleanly into distinct groups. These EDA insights provided a strong foundation for informed model development.

IMPORTANT RESULTS OF THE ADVANCED ANALYSIS

1. MODEL PERFORMANCE COMPARISON

In the advanced analysis phase, we compared multiple classification models to evaluate their ability to predict credit score categories — Poor (0), Standard (1), and Good (2). All models were tested with **inbuilt class weighting** to address the class imbalance in our dataset.

Among all, the **CatBoost model with tuning and inbuilt class weights** performed the best overall on the test set, achieving **85% accuracy**. It maintained strong precision, recall, and F1-scores across all three classes, with a particularly high F1-score for the "Good" category (80%). This suggests CatBoost's effectiveness at capturing subtle relationships in structured, categorical-heavy data.

The **Random Forest model (tuned)** also performed competitively with **79% test accuracy** and highly balanced F1-scores (80%, 81%, and 74% for Poor, Standard, and Good respectively), demonstrating its robustness and interpretability.

Other models like **XGBoost** (73% accuracy) and **LightGBM** (70% accuracy) offered reasonable performance, though their precision and recall varied notably across classes. Ensemble methods like **AdaBoost** and **SVM** lagged slightly behind, with around **64–67% accuracy**, and lower scores for the Good class, likely due to overfitting or class confusion.

Logistic Regression, used as a baseline, performed the weakest with **64% accuracy**, further confirming the importance of non-linear models for this complex prediction task.

Model		Accuracy %	classes	Precision %	Recall %	F1_score %
Random Forest(tuned)	Inbuilt weights	79	0 1 2	77 83 74	83 79 74	80 81 74
XG Boost(tuned)	Inbuilt weights	73	0 1 2	72 87 54	81 64 56	76 74 67
Decision tree-pruning	With inbuilt class weights	71	0 1 2	60 68 80	74 78 67	67 72 73
SVM-with tuning	With inbuilt class weights	67	0 1 2	67 83 48	76 57 81	71 68 60
Adaboost	With inbuilt class weights	64	0 1 2	69 71 48	56 65 74	62 68 58
CatBoost	Inbuilt weights + Tuning	85	0 1 2	75 88 79	86 79 82	84 78 80
Light GBM	Class weights + Tuned	70	0 1 2	53 69 85	85 80 61	65 74 71
Logistic	original	64	0 1 2	57 67 64	43 59 70	48 59 70

Model		Accuracy %	classes	Precision %	Recall %	F1_score %
Random Forest(tuned)	Inbuilt weights	98	0 1 2	97 100 96	100 97 100	98 98 98
XG Boost(tuned)	Inbuilt weights	78	0 1 2	76 92 59	85 69 92	80 79 72
Decision tree-pruning	With inbuilt class weights	87	0 1 2	76 82 95	95 93 80	85 87 87
SVM-with tuning	With inbuilt class weights	73	0 1 2	72 90 53	82 61 91	77 73 67
Adaboost	With inbuilt class weights	73	0 1 2	72 90 53	82 61 91	77 73 67
catboost	Inbuilt weights + Tuning	87	0 1 2	88 87 90	85 90 89	87 88 89
LightGBM	Class weights + Tuned	74	0 1 2	56 73 89	90 83 64	69 77 74
Logistic	Original	64	0 1 2	57 67 64	46 52 77	51 58 70

2. MODEL-SPECIFIC ANALYSIS

CatBoost proved to be the most stable and accurate, with high recall and precision across all classes. Its gradient boosting on categorical data gave it a strong advantage, particularly in distinguishing the “Good” credit score class, which most models struggled with.

Random Forest, while slightly behind in overall accuracy, showed exceptional class balance and interpretability. Its performance was especially reliable for detecting “Poor” and “Standard” categories.

XGBoost and LightGBM demonstrated solid but slightly uneven performance, particularly struggling with the “Good” class, which may be due to overlapping feature distributions or underrepresentation in the training data.

Decision Trees (with pruning) achieved a moderate 71% accuracy but provided simple and interpretable outputs. However, they lacked the nuanced decision boundaries of ensemble methods.

SVM and AdaBoost failed to generalize well to the test set, achieving lower precision and recall in minority classes.

Logistic Regression, while useful for comparison, struggled due to its linear assumptions and limited ability to model class imbalance and feature interactions.

DISCUSSION AND CONCLUSIONS

we used machine learning to predict credit score categories based on customer financial and behavioral data. Through data cleaning, EDA, and feature selection, we identified strong associations between poor credit scores and higher debt, payment delays, and minimum-only payments.

We tested several models with class balancing techniques. **CatBoost (tuned)** gave the best performance with **85% accuracy**, followed by **Random Forest** at **79%**, both showing strong predictive power across all credit score categories. Traditional models like **Logistic Regression** and **SVM** performed poorly, highlighting the need for non-linear, robust methods.

Our results confirm that machine learning—especially tree-based ensemble models—can outperform traditional credit scoring approaches. For future work, more advanced temporal modeling and interpretability tools could provide even deeper insights

APPENDIX

BIBLIOGRAPHY

Arram, A. a. (2023). Credit card score prediction using machine learning models: A new dataset. *arXiv preprint arXiv:2310.02956*.

Zen, M. F. (2025, February 14). Building a credit score model: understanding the business context and dataset. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-understanding-the-business-context-and-dataset-426245d83e63

Zen, M. F. (2025, February 18). Building a credit score model: handling missing values and outliers. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-handling-missing-values-and-outliers-ceb501b3b7b7

Zen, M. F. (2025, February 20). Building a credit score model: feature engineering and encoding. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-feature-engineering-and-encoding-999373e0b9bb

Zen, M. F. (2025, February 20). Building a credit score model: Feature Selection - Muhammad Faizin Zen - Medium. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-feature-selection-516e5c54ed53