Credit Card Default Prediction

Nguyen Thi Minh Ngoc - 11219280

Abstract

This study presents a focused investigation into credit card default prediction leveraging Machine Learning models applied to the Taiwan dataset. Through extensive exploration encompassing diverse preprocessing techniques and model variations, our experimentation reveals that Catboost and Random Forest algorithms consistently outperform alternative methodologies in the identification of defaulting instances.

1 Introduction

In the realm of financial services, credit card default prediction stands as a critical challenge for both lenders and borrowers alike. As credit card usage continues to proliferate globally, understanding and accurately forecasting the likelihood of default becomes paramount for financial institutions to mitigate risks and maintain economic stability. Default prediction involves the utilization of various statistical and machine learning techniques to analyze historical transactional data, demographic information, and behavioral patterns of credit card users to forecast the probability of future defaults.

This predictive modeling task holds significant implications for both consumers and financial institutions. For consumers, accurate default prediction can help guide responsible financial decision-making, ensuring timely repayments to maintain a healthy credit profile and access to future credit opportunities. Conversely, financial institutions rely on robust default prediction models to assess creditworthiness, manage loan portfolios effectively, and minimize potential losses resulting from defaults.

Despite advancements in predictive analytics and machine learning algorithms, credit card default prediction remains a complex and multifaceted problem. Challenges arise from the dynamic nature of consumer behavior, evolving economic conditions, and the intricate interplay of various factors influencing repayment patterns. Moreover, the imbalanced nature of credit card default datasets, where defaults are relatively rare compared to non-default instances, adds another layer of complexity to model development and evaluation.

In this study, we delve into the intricacies of credit card default prediction, employing various preprocessing methodologies and leveraging sophisticated Machine Learning models to analyze the Taiwan dataset with the objective of maximizing the detection of defaulting individuals.

2 Literature Review

I-Cheng Yeh and Che-hui Lien [1], collectors of the dataset used in this study, have compares the predictive accuracy of probability of default among six data mining methods, including K-nearest neighbor (KNN), Logistic regression, Discriminant analysis, Naive Bayesian, Neural networks and Classification trees. After several experiments, they concluded that artificial neural network is the only one that can accurately estimate the real probability of default.

Using the same dataset, Sheikh Rabiul Islam et al [2] have used to approaches to tackle the defaulters prediction problem: machine learning approach and machine learning approach.

Additionally, with other datasets but in the same context of defaulters prediction, Zhaohong Wang et al [3] found that Random Forest and XGBoost perform better than Logistic Regression and KNN.

3 Methodology

3.1 Scaling Methods

The numerical features present in a dataset commonly display variations in their units and ranges, which can potentially influence model performance. To address this challenge, scaling techniques are employed to transform feature values into a similar scale. This standardization process aims to equalize the impact of all variables on model outcomes, thereby enhancing the overall performance of the algorithm. Two commonly used scaling methods to be discussed in this study include Standard Scaling and Min-Max Scaling.

3.1.1 Standard Scaling

Standard scaling is a technique where the values are adjusted to center around the mean with a unit standard deviation. Specifically, this method entails shifting the mean of the attribute to zero and adjusting the distribution's standard deviation to one. Formula of Standard scaling:

$$z = \frac{x - \mu}{\sigma}$$

where x is the feature values, μ is the mean and σ is the standard deviation of the feature values.

3.1.2 Min-Max Scaling

An alternative method of standardization is Min-Max scaling, aims to constrain features within a predetermined minimum and maximum range, typically ranging from zero to one. Formula of Min-Max scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where $\min(x)$ and $\max(x)$ are minimum and maximum values of the feature respectively.

3.2 Encoding Methods

Machine learning models are only able to operate on quantitative variables. Consequently, it is required for categorical features to be converted into numerical representations before fitting in the model. There are a huge range of techniques for encoding qualitative variables and in this study, I will explore three common methods of them, including ordinal encoding, one-hot encoding and weight of evidence.

3.2.1 Ordinal Encoding

Ordinal encoding involves the assignment of a sequence of consecutive integers, typically starting from zero, to distinct categorical values within a qualitative feature set. This encoding methodology is particularly beneficial when dealing with features exhibiting a clear ordinal relationship. The assignment of integers to each category is typically predicated on its hierarchical ranking within the feature set. *For instance, in an ordinal encoding scheme, the categories high, medium*, and *low* might be encoded as *2*, *1*, and *0*, respectively.

3.2.2 One-hot Encoding

One-hot encoding is a categorical variable transformation technique wherein each category is represented by a binary matrix column. Within this matrix, each instance is denoted by a *1* in the corresponding category column and *0* in all other columns. This method is particularly advantageous for nominal variables, characterized by a lack of inherent order among categories.

3.2.3 Weight of Evidence

The weight of evidence (WoE) serves as a measure of the predictive strength of an individual feature in relation to its independent feature, or the target. When a particular category or bin within a feature exhibits a disproportionately high proportion of events compared to proportion of non-events, the resulting WoE value is elevated. This elevated WoE value signifies that the corresponding category effectively separates the events from non-events.

The formula to calculate the weight of evidence of a category i of a feature is given by:

$$\text{WOE}_i = ln\left(\frac{\text{percentage of } y = 0_i}{\text{percentage of } y = 1_i}\right)$$

3.3 Resampling Methods

3.3.1 SMOTE

Imbalanced datasets can can indeed lead to models exhibiting bias toward the majority class. Indeed, the Synthetic Minority Oversampling Technique (SMOTE) offers a solution to address this imbalance issue by oversampling the minority class. Unlike simply duplicating records from the minority class, which may not add new information, SMOTE generates new instances synthetically from existing data. To put it plainly, SMOTE works by examining instances within the minority class and utilizing the k-nearest neighbor algorithm to select a random nearest neighbor. Subsequently, a synthetic instance is created randomly in the feature space, thereby augmenting the representation of the minority class in the dataset.

3.4 Feature Selection

3.4.1 Information Value

After introducing the concept of Weight of Evidence (WoE), it becomes evident that WoE provides insights into the predictive efficacy of individual bins within a feature. However, for the purpose of feature selection, a single value representing the predictive power of the entire feature would be invaluable and that is also the main idea of the Information Value. The formula to calculate the information of any feature is given by:

$$\text{IV} = \sum_{i=1}^{k} (\text{percentage of } y = 0_i - \text{percentage of } y = 1_i) * \text{WOE}_i$$

Feature selection based on information value:

Information Value	Meaning
< 0.02	Useless for prediction
$0.02 - 0.1$	Weak predictors
$0.1 - 0.3$	Medium predictors
$0.3 - 0.5$	Strong predictors
> 0.5	Suspicious or too good predictors

3.5 Machine Learning Models

3.5.1 Logistic Regression

Logistic regression serves as a method for estimating probabilities in binary classification tasks, where the outcome can take one of two possible values. It's an extension of the linear regression model to address classification problems. Specifically, instead of fitting a straight line or hyperplane to the data like, logistic regression employs the logistic function, or sigmoid function to constrain the output of a linear equation within the range of 0 to 1. The formula of sigmoid function:

$$\sigma = \frac{1}{1 + e^{-z}}$$

where $z = wX + b$ is a linear regression.

3.5.2 K-nearest Neighbours

K-Nearest Neighbors (KNN) for classification problem is an algorithm employed to predict the class of a new data point by examining the classes of its closest neighbors within the feature space. The method hinges on calculating the distances between the new data point and all existing points within the training dataset. Subsequently, it identifies the k nearest neighbors based on this distance computation. The class assigned to the new data point is determined by the majority class among its k nearest neighbors.

3.5.3 Random Forest

Random Forest is an ensemble learning technique that combines the output of multiple decision trees, each tree operates on a subset of the complete dataset, contributing its verdict to a final result. By amalgamating these diverse opinions, Random Forest mitigates overfitting tendencies and enhances generalization performance.

3.5.4 LightGBM

LightGBM is a software application that utilizes gradient boosted decision trees (GBDT), a technique where decision trees, considered as weak learners, are progressively combined in a serial manner, known as boosting. In LightGBM, each subsequent learner is trained to fit the residuals from the previous tree, resulting in iterative model improvements. The culmination of this process yields a final model that integrates the outcomes from each step, ultimately producing a robust learner. GBDT has earned a reputation for exceptional accuracy, consistently outperforming competitors in prominent machine learning competitions.

3.5.5 XGBoost

XGBoost, short for Extreme Gradient Boosting, represents a scalable and distributed gradient-boosted decision tree machine learning technique. Renowned for its capability in parallel tree boosting, XGBoost stands as the foremost choice for addressing regression, classification, and ranking challenges in the realm of machine learning.

3.5.6 Catboost

CatBoost stands as a powerful supervised machine learning method, employing decision trees for both classification and regression tasks. Its name aptly reflects its core strengths: handling categorical data (the *Cat*) and leveraging the prowess of gradient boosting (the *Boost*). In this methodology, numerous decision trees are sequentially crafted, with each subsequent tree refining the outcomes of its predecessor, culminating in superior results. Notably, CatBoost enhances the traditional gradient boosting approach by prioritizing efficiency, ensuring swift and optimized implementations.

4 Data

The data selected for this study is Default of Credit Card Clients from UCI Machine Learning, which contains payment data from April to October 2005 from a prominent financial institution, which serves as both a cash provider and credit card issuer, in Taiwan. The dataset contains 30,000 observations and 23 features, including demographic information (such as marital status, education, age, gender), historical repayment status and monetary amount (such as limit balance, historical bill statement amount and payment amount) with no missing values. The response variable of this data is the customer's default payment in October, which labeled as either 0 (non-default) or 1 (default). Within the dataset, the proportion of defaulters accounts for 22.1% (figure, which stands notably lower than the proportion of non-defaulters, indicating the presence of an imbalanced classification problem.

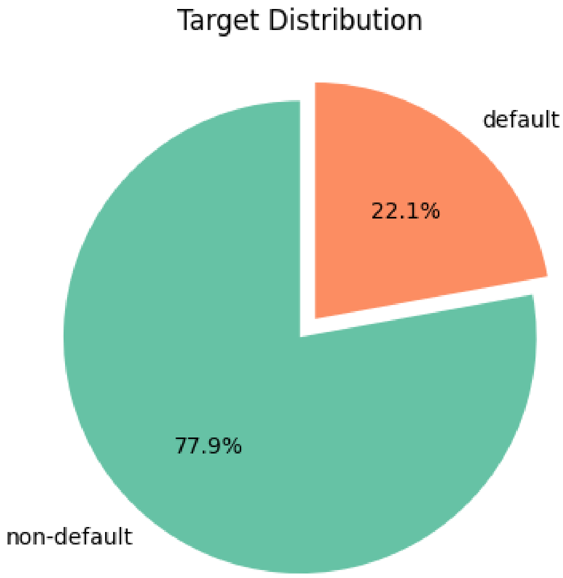


Figure 1: Target Distribution

All the categorical features of this data were previously converted into numeric manners. However, to make it easier to gain insights and find another way to encode categorical features, we will turn them back into their original meanings based on the information provided by I-Cheng Yeh [1], the collector of this data.

Distribution of some categorical features:

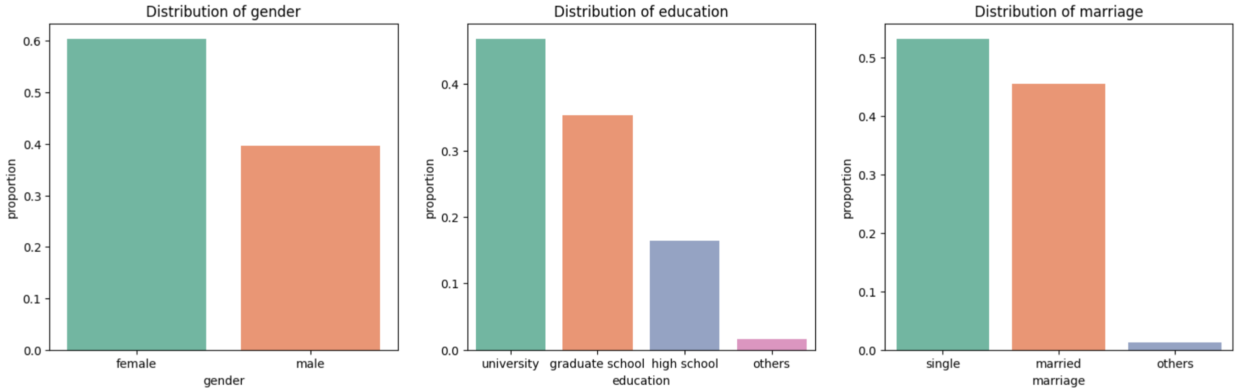


Figure 2: Distribution of Genders, Marital Status and Education

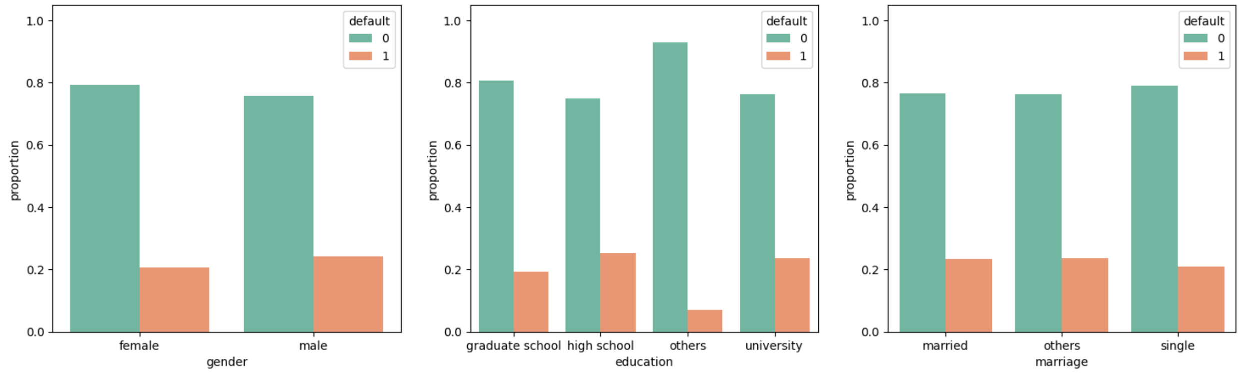


Figure 3: Target Distribution corresponding to each category

In repayment status features, there are two undocumented values, which are *0* and *-2* but we can not simply add them to an existing classes due to their large proportions. Hence, to preserve information, we will remain them as another particular status of repayment.

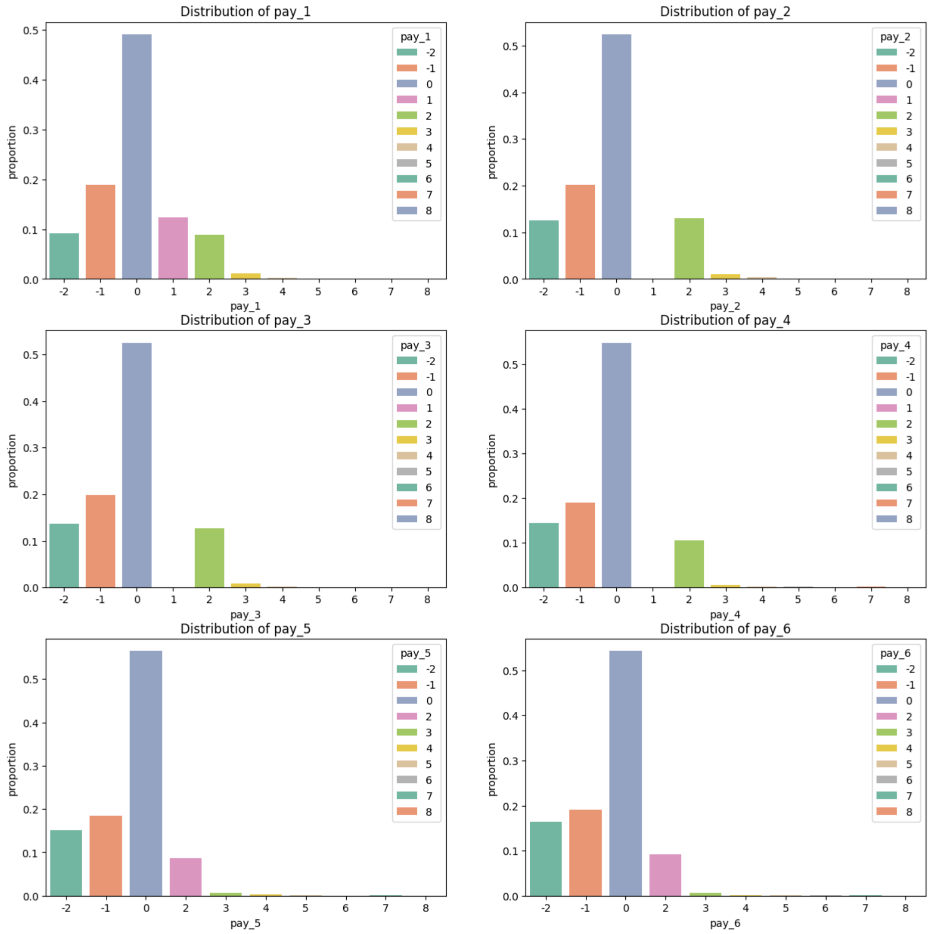


Figure 4: Distribution of Repayment Status

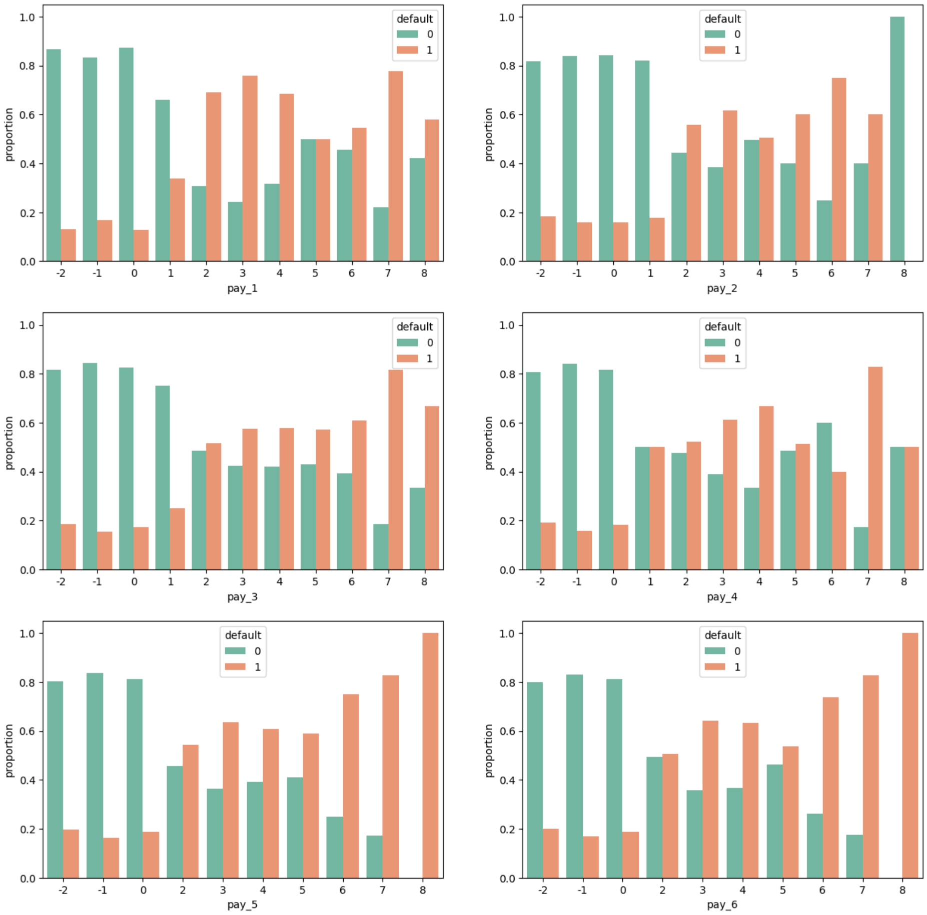


Figure 5: Target Distribution corresponding to each category of Repayment Status

When examining the percentage distribution of the target variable corresponding to each value in the repayment status features, it becomes evident that for statuses -2, -1, and 0, the proportion of default users is significantly lower than the rates of non-default. Conversely, for other statuses, an opposite trend emerges, notably with defaulters comprising a higher percentage compared to non-defaulters. Based on this observation, we can come to two conclusions:

- Status -2 and 0 might share a similar connotation with status -1, potentially indicating a duly paid status, to some degree.
- The information provided by the repayment status features appears to effectively discriminate between default and non-default users, owing to the distinct patterns exhibited by the two major types of repayment statuses.

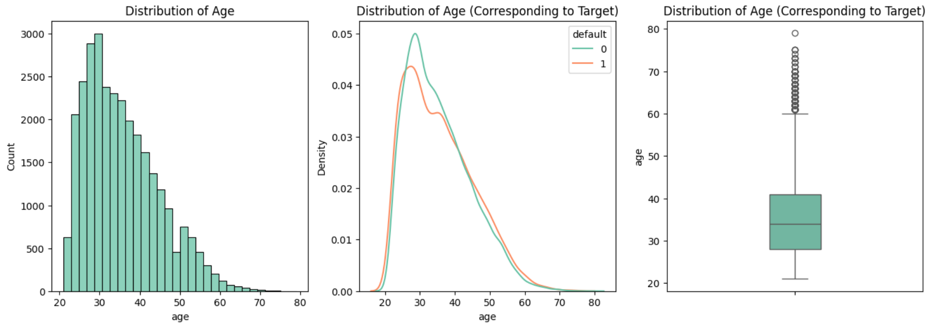


Figure 6: Age Distribution

The depicted age distribution, as illustrated in the figure 6, suggests that age alone may not serve as a robust discriminator between the two classes under consideration. A similar conclusion can be inferred for monetary amount features, indicating that they too may not adequately distinguish between the aforementioned classes.

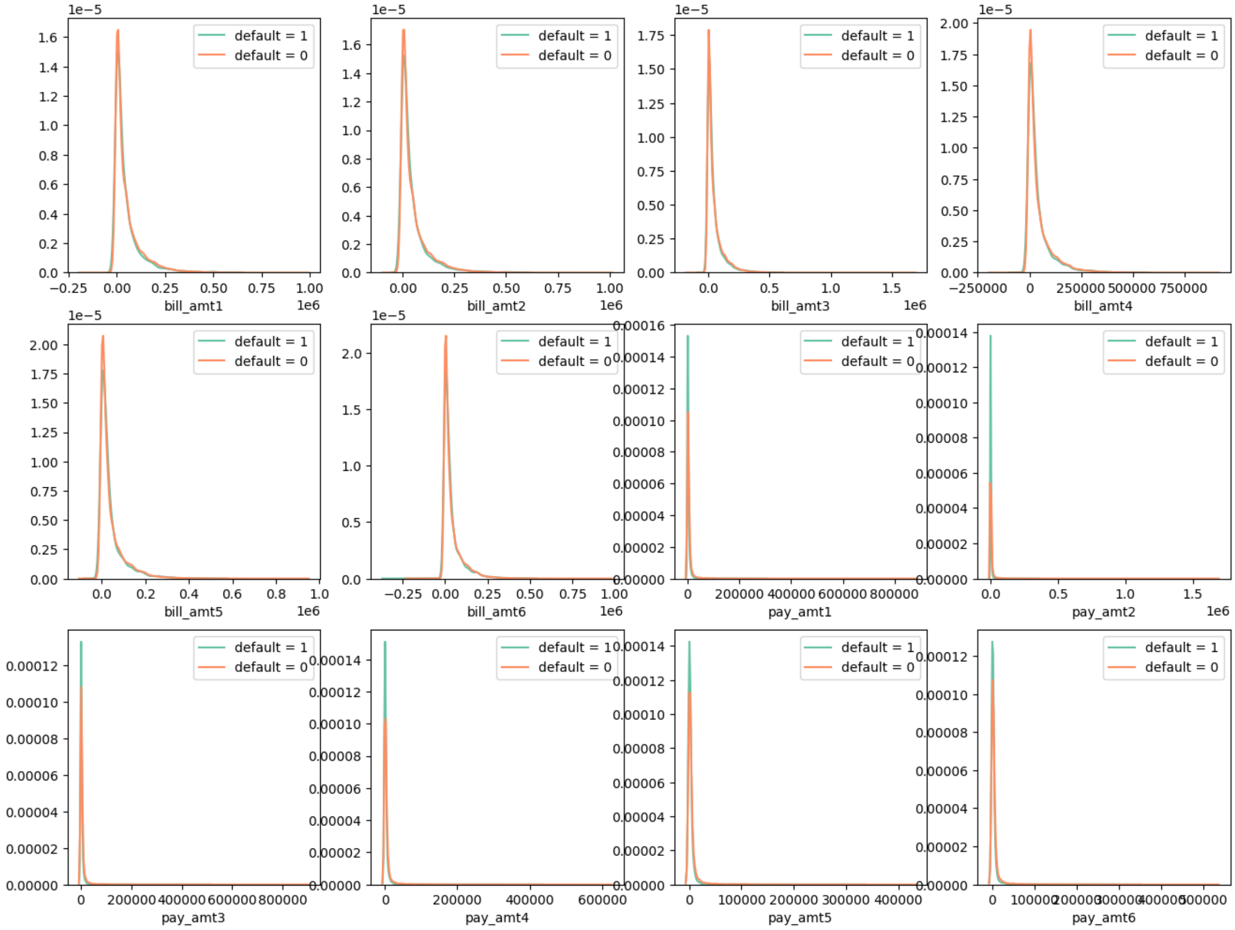


Figure 7: Bill Statement Amount an Payment Amount from April to September

5 Experimental Results

In order to enhance the comprehensibility of my research workflow, I shall partition it into two distinct phases (figure 8). Initially, I will employ Logistic Regression as the baseline model and systematically explore various preprocessing methodologies to ascertain the optimal approach when working with the selected dataset. Subsequently, transitioning into the second phase, I will apply the previously determined optimal set of preprocessing techniques alongside a variety of models, such as Logistic Regression, K-Nearest Neighbors (KNN), as well as ensemble learning algorithms including LightGBM, RandomForest, Catboost, and XGBoost, with the aim of identifying the most effective model for the task at hand.

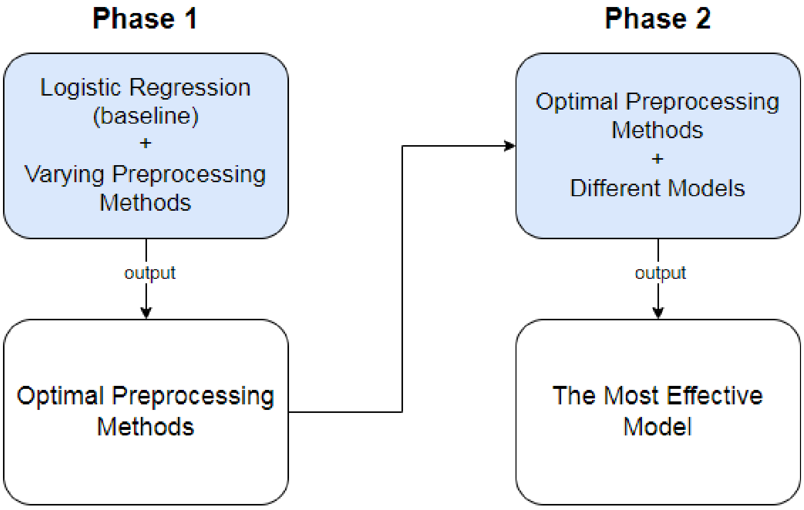

Phase 1
Logistic Regression (baseline) + Varying Preprocessing Methods
output
Optimal Preprocessing Methods
Phase 2
Optimal Preprocessing Methods + Different Models
output
The Most Effective Model

Figure 8: Flow of the study

5.1 Phase 1: Logistic Regression and Different Preprocessing Techniques

5.1.1 Varying Approaches to Dealing with Numerical Features

Treating as Categorical Features

The initial inquiry in this phase pertains to the diverse methodologies available for addressing numerical features. Should these features be converted into bins and subsequently treated as categorical variables? Alternatively, should they be retained as continuous numerical features and subjected to scaling before fitting the model?

Within the selected dataset, there exist 20 numerical features, which can be categorized into three primary groups: age, repayment status, and monetary amounts (comprising bill statement amounts and payment amounts from April to September). We intend to explore the consequences of treating each group respectively as categorical features, while keeping other components such as categorical encoder and model constant, in order to discern any resultant disparities (figure 9).

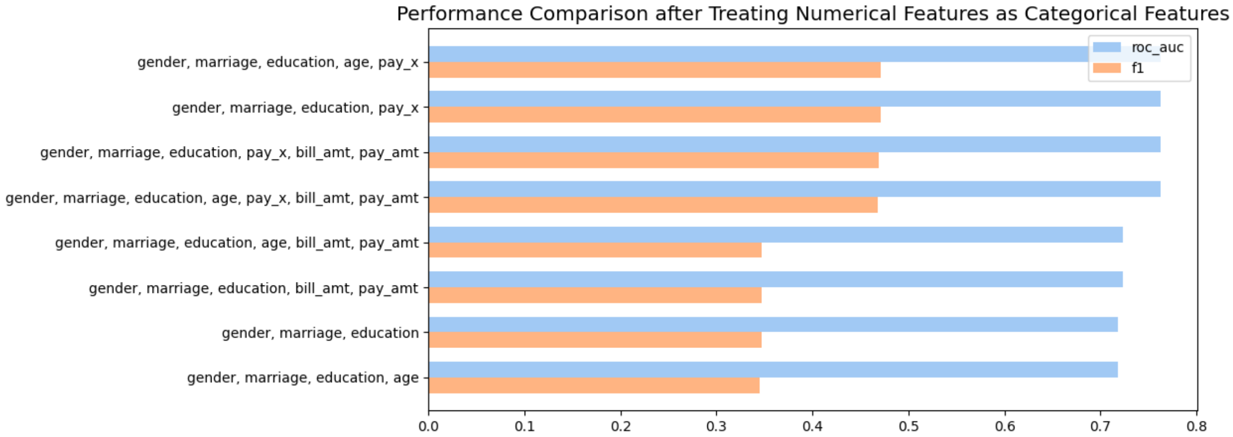


Figure 9: Treating as Categorical Features

The findings suggest a clear improvement in both f1 and roc auc scores, with the inclusion of categorical features related to repayment status. Besides, the combination of age and repayment status demonstrates the highest performance. As a result, we plan to classify these two features as categorical variables for subsequent testing phases.

Different Scaling Techniques

In addition to assessing whether numerical features should be treated as categorical variables, this study also investigates various scaling methods to identify the discrepancies in their outcomes (figure 10).

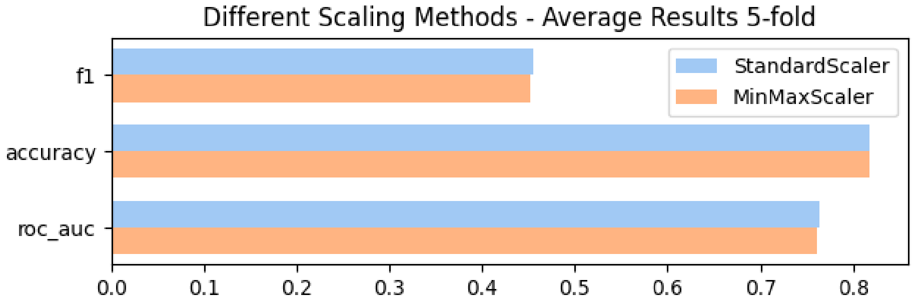


Figure 10: Different scaling techniques

Both the Min-Max and Standard scaling approaches demonstrate closely comparable outcomes, showing little difference between the two. However, during the implementation of 5-fold validation, the Standard scaling exhibits slightly superior performance over the Min-Max scaling across multiple folds. Therefore, we will use Standard scaling as the scaling method for testing in the second phase.

5.1.2 Different Encoding Methods

The performances of different encoding methods are given in figure 11.

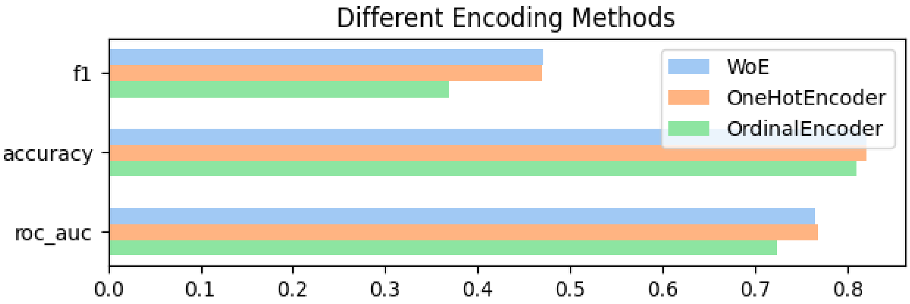


Figure 11: Performances of Different Encoding Techniques

Weight of Evidence (WoE) and One-Hot encoding significantly outperform Ordinal encoding. The result also shows that while both WoE and One-Hot encoding exhibit relatively comparable efficacy, it is noteworthy that WoE consistently yields superior results compared to One-Hot encoding across a majority of fold iterations during validation. Consequently, WoE emerges as the encoding method of choice for the forthcoming phase of experimentation.

5.1.3 Handling Imbalanced Dataset

The chosen dataset exhibits a pronounced imbalance in its target distribution. In this stage, we intend to address this issue through the application of the Synthetic Minority Over-sampling Technique (SMOTE) and subsequently assess the performance differential between employing SMOTE and not employing it. This comparative evaluation will inform our decision regarding the necessity of resampling the dataset using SMOTE.

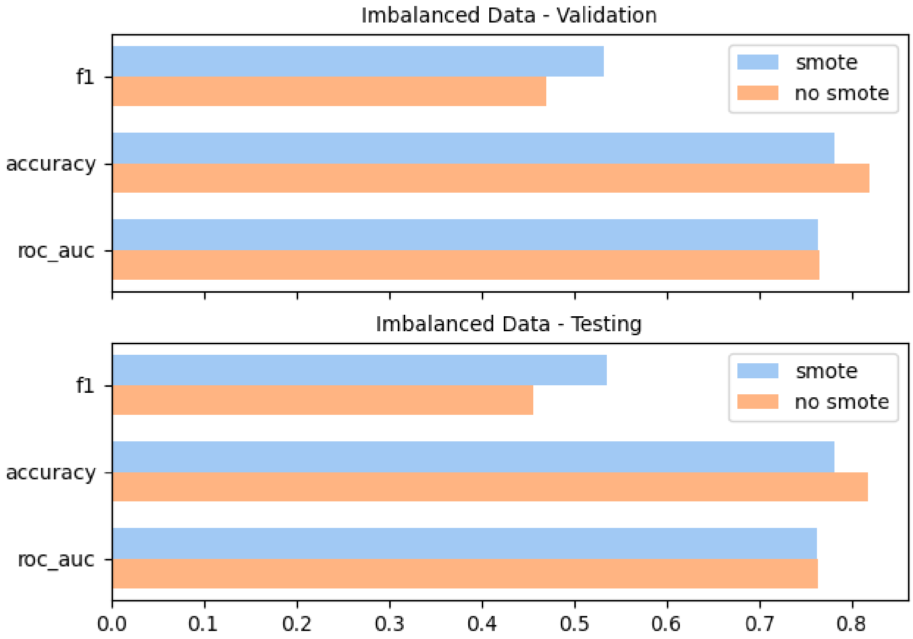


Figure 12: Using SMOTE to address imbalance issue

Although the utilization of the original dataset tends to result in higher accuracy, it is evident that the application of Synthetic Minority Over-sampling Technique (SMOTE) markedly enhances f1 scores. Given the inherent imbalance within the selected dataset, our assessment prioritizes f1 score and roc auc as primary evaluation metrics. Consequently, SMOTE will be employed in the forthcoming testing phase to ameliorate model performance and effectively address the imbalance issue.

5.1.4 Feature Selection using Information Value

The feature importance estimated by Information Value is given in figure 13

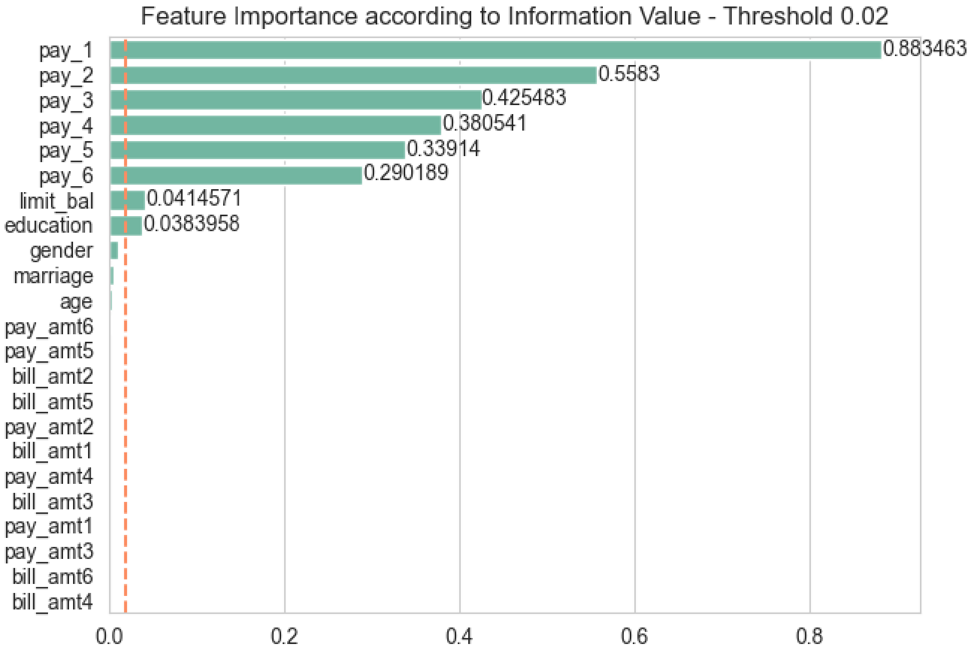


Figure 13: Feature Importance based on Information Value

Based on the assessment of information value, it is determined that monetary features, age, gender, and marital status are useless for prediction and therefore they will be eliminated after feature selection. The difference of performance between using all features and using only features selected by information value is given in figure 14.

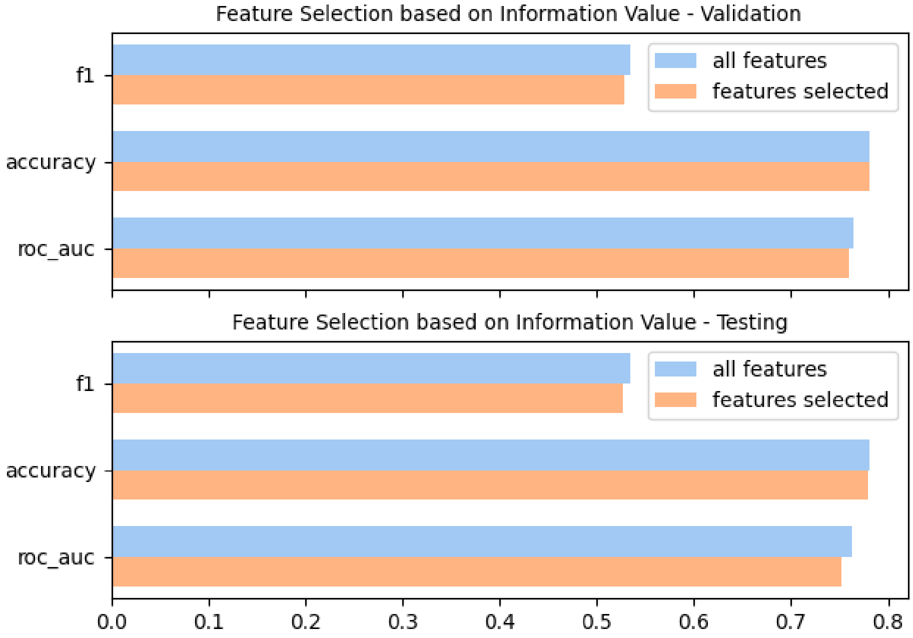
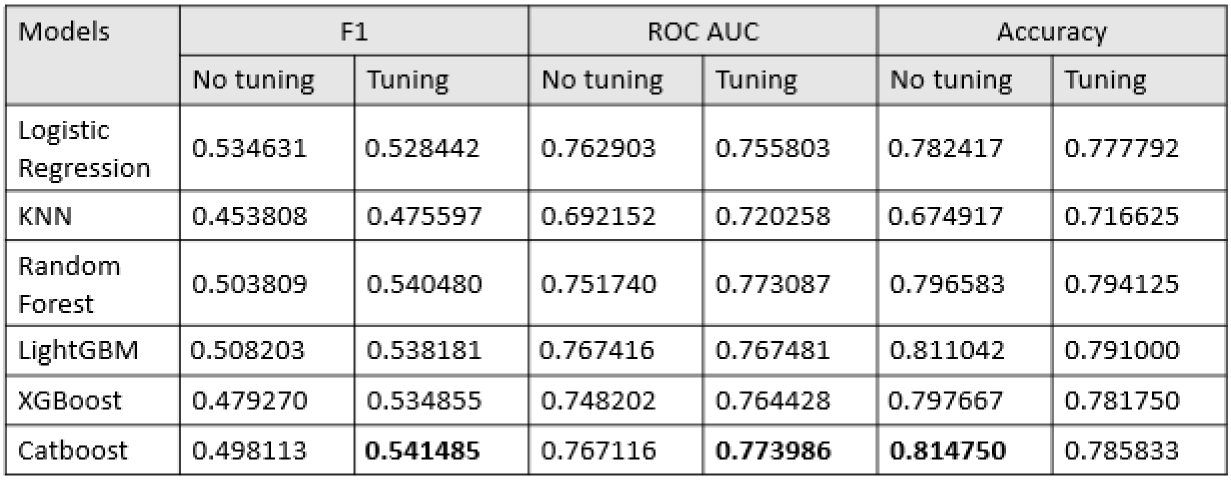


Figure 14: Performance Comparison between features used

The findings suggest that employing all available features yields better performance. This phenomenon could potentially be attributed to the limited size of the dataset and the relatively small number of features therein.

5.2 Phase 2: Models Comparison

The performance comparison between different model is given below.



Models	F1		ROC AUC		Accuracy	
	No tuning	Tuning	No tuning	Tuning	No tuning	Tuning
Logistic Regression	0.534631	0.528442	0.762903	0.755803	0.782417	0.777792
KNN	0.453808	0.475597	0.692152	0.720258	0.674917	0.716625
Random Forest	0.503809	0.540480	0.751740	0.773087	0.796583	0.794125
LightGBM	0.508203	0.538181	0.767416	0.767481	0.811042	0.791000
XGBoost	0.479270	0.534855	0.748202	0.764428	0.797667	0.781750
Catboost	0.498113	**0.541485**	0.767116	**0.773986**	**0.814750**	0.785833

Figure 15: Models Comparison Table Summary

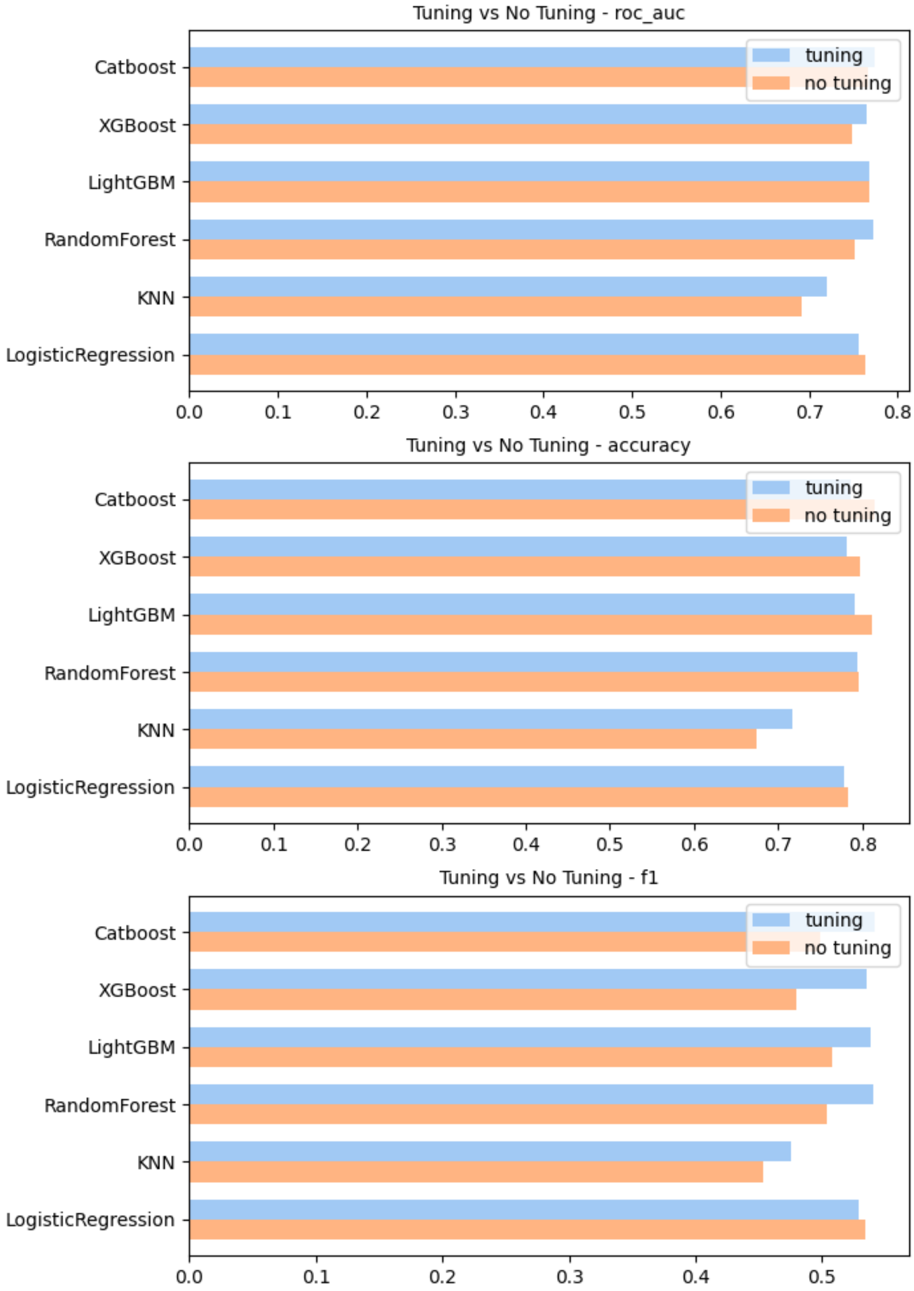


Figure 16: Models Comparison

After hyperparameters tuning, it is observed that despite a reduction in accuracy, a majority of the models exhibit improvements in both F1 scores and ROC AUC scores, indicating enhanced efficacy in addressing imbalanced class distributions, facilitating discrimination between disparate classes, and augmenting the identification of defaulters within the dataset.

Additionally, the findings also suggest that all ensemble models employed outperform traditional models, with notable excellence demonstrated by Catboost and Random Forest algorithms in particular.

6 Conclusion

Through the application of various preprocessing techniques and diverse model architectures, several key insights have emerged. Firstly, certain numerical features exhibit improved performance when treated as categorical variables. Secondly, both Min-Max and Standard scaling methodologies yield comparable results, suggesting flexibility in preprocessing choices. Thirdly, the utilization of one-hot encoding and Weights of Evidence surpasses the performance of Ordinal Encoding, with similar efficacy between the former two approaches. Moreover, employing Synthetic Minority Over-sampling Technique (SMOTE) effectively addresses imbalance issues inherent in the dataset. Lastly, our findings underscore the superiority of ensemble models over traditional counterparts, with notable excellence observed in the Catboost and Random Forest algorithms. These conclusions collectively contribute to enhancing the robustness of credit card default prediction models.

7 Limitations and Future Work

Given the constraints imposed by temporal and resource limitations, a considerable array of methodologies remains unexplored in the pursuit of optimizing model performance. Moreover, the selected dataset, characterized by its antiquity and relatively restricted dimensions in both size and feature richness, notably constrains predictive efficacy. Looking ahead, a prospective avenue involves dedicating additional time and resources towards sourcing a more contemporary dataset characterized by augmented dimensions, encompassing a broader spectrum of demographic and transactional attributes. This endeavor will enable a more comprehensive exploration of diverse techniques, thereby fostering the attainment of superior predictive outcomes.

References

[1] I-Cheng Yeh, Che-hui Lien (2009) The comparisons of data mining techniques for the predictive accuracy of the probability of default of credit card clients

[2] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor (2018) Credit Default Mining Using Combined Machine Learning and Heuristic Approach

[3] Zhaohong Wang, Cheng Han Wen, Wenda Zhou, Jun Zhang (2023) Credit Card Default Prediction with Data Modeling

Appendix

Link to Repository