IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Credit Risk Analysis Of Personal Credit Card Default Using Machine Learning Algoritm

[1]Huzefa Chhatriwala, [2]Bhaskar Vishwakarma, [3]Devendra Kumar Pandey, [4]Nitin Kumar Choudhary

[1]Student, [2]Assistant Professor, [3]Professor, [4]Assistant Professor

[1]Management

[1]Medicaps University, Indore, India

Abstract: Credit loans are one of the most used loans. Credit loans earn a lot of money for the banking sector. If the loan defaults, it has impact on profit. The precise credit risk analysis is required to manage the defaults of the borrower. With a dataset containing different borrower attributes, the present research has different machine learning methods investigate a multi-model framework that integrates classification, Regression and Clustering methods. This research investigate a hybrid strategy by using a combination of different methods. Model of regression, classification are used in prediction of two target variable default_in_last_6_months and credit_card_default. This Model will aid in improving credit risk assessment and assist financial institution in streamlining loan approval procedures and risk management policies.

Keywords:- Credit Risk Analysis, Loan Default Prediction, Machine Learning, Classification Models, Predictive Modeling, Financial Risk Management, Supervised Learning, Unsupervised Learning, Logistic Regression, Polynomial Regression, K-means Clustering, Borrower Segmentation, Financial Decision-Making, Accuracy Optimization, Feature Engineering, Risk Profiling.

I. INTRODUCTION

1.1 Introduction

Credit risk assessment is one of the most important feature of the Financial industry, particularly in context of the personal loans. Credit card loan is one of the most common loan taken by individuals, it may not be the largest loan but it is the most common one cumulatively it becomes large in amount if we join all credit card loans. So, to reduce risk taken by the financial institution to give this loan credit risk assessment is necessary. Traditional methods of credit risk evaluation, such as credit scoring, have limitations in terms of flexibility, accuracy

This study aim to reduce risk taken by financial institution to provide loan by predicting default by borrowers in advanced. This can be tackled by evaluating a variety of models. Several models were applied in the analysis including Regression, Classification.

The prediction problem in a specific time horizon (e.g., default in the last six months and default of credit card payments), remains a significant challenge.

Credit risk prediction has so far relied on statistical models. These models are, however, prone to not capturing the complex, non-linear relationship in the financial data. To overcome this, advanced machine learning algorithms are used. This machine learning algorithm has opened up new avenues for improving the accuracy of credit risk default prediction. This research focuses on predicting loan defaults within the context of personal loans, with specific attention

to two distinct target variables: default_in_last_6months and credit_card_default. These target variables provide different valuable insight into credit risk analysis. While default_in_last_6_month focuses on default in the last 6 months, whereas credit_card_default the borrower has defaulted the loan or not. This dual-target approach of this research aim to provide the a more comprehensive understanding.

The study begins with exploratory data analysis (EDA) of the data to observe the distribution and interactions among features. Feature engineering techniques are used to identify the most significant predictors for each target variable. Then, the performance of each model is compared on the basis of metrics are used to cross-validate the performance of the regression, classification models to predict continuous outcomes, and the Silhouette score is used to cross-validate the quality of the clustering models. This comprehensive analysis provides a proper evaluation of the ability of these models to assist financial institutions to make data-driven decisions on loan approvals, risk management strategies, and customer segmentation.

II. LITERATURE REVIEW

Credit risk analysis is a critical field of research particularly in the case of financial institutions in a bid to effectively manage loans. Machine learning algorithms are becoming popular when it comes to forecasting and analyzing loan default since they can effectively handle complex relationships, and big data with ease in an attempt to minimize risk and maximize returns. There have been several studies that show the use of statistical techniques like Logistic regression and decision trees

- Suhadolnik et al. (2023) explain the application of models in credit risk analysis, focusing on ten algorithms for predicting loan default. The authors compare traditional statistical models with advanced machine learning techniques and conclude that ensemble techniques like XGBoost perform better than simple models like logistic regression in terms of predictive accuracy and reliability. The main conclusion is that XGBoost performs better than all machine learning models.
- Noriega et al. (2018) on credit risk forecasting is directed towards using machine learning (ML) for dealing with rising data complexities due to rising credit requirements. Methods such as SMOTE and Adaptive Synthetic Sampling (ADASYN) balance the data, whereas genetic algorithms and optimization methods (such as Ant Colony Optimization) improve feature selection as well as model performance. Models such as ensemble models such as Boosted Category models such as XGBoost, and neural networks have achieved improved predictive ability compared to classical methods, particularly dealing with c omplicated data.
- Emmanuel et al. (2024) explain stacking different machine learning classifier models with filter-based feature selection techniques increase accuracy. They use multiple dataset in research comparing stacked classifiers with individual models. For example, the stacked classifier approach is likely to enhance accuracy and F1-scores due to the ability of stacking to combine the strengths of different algorithms while compensating for their weaknesses. Feature selection techniques like information gain (IG), correlation, and mutual information have been used to reduce dimensionality and enhance model interpretability. Research shows that filter-based FS techniques, especially those based on IG, improve performance by picking the most informative features for the model.
- Wang et al. (2020) analyze credit scoring with several machine learning classifier algorithms a technique applied by financial institutions to determine an individual's creditworthiness. Five popular algorithms are compared to identify the most effective classifier to predict credit risk. Each algorithm has strengths unique to it The research proves that Random Forest performed better than other techniques with respect to precision, recall, AUC, and accuracy.

III. RESEARCH METHODOLOGY

This chapter outlines the systematic approach adopted to conduct the study, detailing the methods and procedures used to collect and analyze data. The core objective is to examine the extent to which Business Analysts (BAs) influence organizational development, focusing specifically on comparing performance indicators before and after their inclusion. Key elements of this chapter include the research design, sampling strategy, data collection methods, analytical tools, ethical practices, and limitations. The methodology is structured to ensure the study's reliability and validity while offering actionable insights for both scholarly and business audiences.

Data Collection

The data were collected from GitHub, which is an open-source repository that has a collection of datasets on credit risk available for free. Data were collected for the study from an open-access repository on GitHub, "Credit Risk Analysis Using Machine Learning" by Maix Bach (GitHub link). The repository contains a range of datasets relevant to credit risk analysis, including credit scores, past defaults, and other finance data. The available dataset was a good reference point for evaluating credit risk on the basis of machine learning, as per the study goals.

For analysis, 2000 random rows from the original data were extracted for research purposes. This was in order to achieve a dataset size that I could work with, but have a representative sample to estimate credit risk with machine learning techniques.

Data Source

Maix Bach, "Credit Risk Analysis Using Machine Learning," GitHub, available at https://github.com/maixbach/credit-risk-analysis-using-ML.

A. Data Preprocessing: -

This is an essential step in machine learning. It make sure that used dataset is clean and consistent. The dataset contained categorical as well as continuous features with some missing values and outliers to begin with.

1) Handling missing values: -

Missing values were imputed to preserve data integrity without causing any significant bias. For categorical variables, Mode function was employed to impute missing values. For numerical variables, median was employed so that it would not be influenced by outliers.

2) Encoding Categorical Variables: -

Categorical variables such as occupation_type, owns_car, and owns_house were transformed into numeric values by using label encoding and one-hot encoding where applicable. For example, occupation_type was label-encoded so that 0 to 18 were mappings of various occupation categories.

3) Outlier Detection and Removal: -

Outlier detection and removal is a important step as it ensures that data is not skewed, it did not distort model performance. Boxplot were initially used to visualize the outlier in the dataset. Once outlier was identified, then Interquartile range (IQR) method was employed to confirm and remove them.

Although Z-Score method was also used as an alternative for outlier detection and removal, it shoved very minimal impact on removing the outlier in dataset. Therefore it was removed in final methodology. Boxplots were useful for visual inspection, but the IQR method was the primary technique for effectively handling outliers in this case.

4) Feature Scaling: -

StandardScaler was used to scale continuous features so that all independent variables would be on the same scale which is beneficial during model training especially in distance-based models like K-Nearest Neighbors.

B. Exploratory Data Analysis (EDA): -

It was conducted to understand the data and dependency between independent variable and target variable. Exploratory data analysis guided me in
feature selection and model deployment by pointing out trends and associates that affect loan default risk. It included univariate, bivariate, and multivariate analysis, conducted independently for each target variable to acquire in-depth insights.

1) Univariate analysis: -

Univariate analysis is concerned with individual features.

a) Numerical variables: Histograms for all numerical columns were plotted to see the distribution of numerical features. It is used to identify the skewness, outlier, or data transformation needed for modeling
b) Categorical variables: The bar plots were drawn to determine the distribution of categories and detect any imbalances that might affect model results.

2) Bivariate analysis: -

Bivariate analysis investigates the interaction between every feature and target variable. Separate vis ualization is done for the two target variables.

a) Numerical variables: Box plots were utilized to evaluate how each numerical feature compares with the target variable.
b) Categorical variables: Count plots with hue of target variable were created for categorical column. This analysis assists to expose patterns like a greater percentage of loan defaults among specific occupations and asset-less individuals, offering insights into socio-economic factors that affect default risk.

3) Multivariate analysis: -

Multivariate analysis was performed to investigate relationships between numerical features and correlations between each target variable

a) Correlation Matrix: correlation matrix for numerical attributes was generated against every target variable in order to determine multicollinearity and see features which had high associations with each other. It makes accurate modeling.

C. Feature Engineering and Selection: -

Feature engineering and feature selection are also necessary in the pre-processing of a dataset for training a model. It is towards enhancing the accuracy of the model to correctly predict the
default. Different methods were employed to pick the most informative features:

1) Logistic Regression Coefficients:
It was used to know the relative importance of the features using the values of the coefficients. The coefficients are the weights of the features in determining the incidence of an event, i.e., loan default. The model estimates the probability of a binary outcome (i.e., default or otherwise) using the logistic function. The model identified the significant predictors of loan default.
2) Random Forest Feature Importance:
Random Forest orders feature importance ranks features by how many times a feature is employed to split data across all the trees. The higher a feature is used to reduce impurity in splits (e.g., Gini impurity or entropy), the more important it is.
3) Correlation Matrix Analysis:
The correlation matrix is a table of correlation coefficients for a variable set. The matrix informs us about how they correlate with one another. It also detects multicollinearity between independent variables that makes it difficult for some.

4) SelectKBest:
The SelectKBest was utilized for the selection of the top K features with the highest predictive power based on statistical tests to maximize the efficiency of dimension-reduced models.

D. Model Selection and Implementation: -

1) Regression Models:: A simple Linear Regression model was applied to predict the probability of loan default and explore linear relationships between independent variables and target variable The model was applied as an
exploratory effort to experiment with relationships within data, although it was poor in explaining bi nary outcomes.

- Linear Regression Model and Clustering Variants Linear
Regression was first employed to forecast the default of the loan and
to establish linear relationships.
- Then two models were developed one with Agglomerative Clustering and another with K-Means Clustering and Ridge Regression.
These cluster models helped in determining different segments of borrowers.

Formula:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

- Polynomial Regression: A Polynomial Regression model was added to account for non-linear associations between features. Further, one version of the model
was also optimized using GridSearchCV in order to reach the trade-off between complexity and precision

Formula:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n$$

2) Classification Models: Classification models were used to predict binary targets, e.g., whether a borrower would default or not.
These models were selected due to their strength and capability to address complex patterns within the dataset.

- Logistic Regression: Logistic Regression was applied in a straightforward
model for ease of interpretation and effectiveness in binary
classification. Moreover, one variation of the model was optimized using GridSearchCV for best hyperparameters.

Formula:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

- **Random Forest Classifier**:
It is employed to represent several decision trees to detect complex interactions among the features. Another model was constructed with GridSearchCV to enhance generalization and minimize overfitting.
Formula:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$$

- XGBoost Classifier:
  Extreme Gradient Boosting (XGBoost) enhances accuracy by progressively reducing errors from earlier trees. A second model was created with GridSearchCV to enhance generalization and prevent overfitting.
  Formula:

$$f(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

- **Support Vector Machine (SVM)**:
  It was utilized to maximize the margin between classes in order to identify the best hyperplane that discriminates loan default from non-default to achieve high accuracy in prediction. A GridSearchCV model was employed to optimize model parameters and enhance accuracy, stability, and generalizability, which is paramount in strong credit risk analysis.

  Formula:

$$f(x) = w^T x + b$$
$$y = \text{sign}(w^T x + b)$$

- **Linear Discriminant Analysis (LDA)**: It was utilized to convert high-dimensional data to lower dimensionality space to achieve maximum class separability. A GridSearchCV model was adopted to optimize model parameters to enhance accuracy, stability, and generalizability, which is essential in strong credit risk analysis.
  Formula:

The criterion for Maximizing Class Separation:

This equation is the objective of LDA, which is to determine the projection vector v that maximizes class separation.

$$J(v) = \frac{v^T S_b v}{v^T S_w v}$$

Posterior Probability Formula:

After training the model, LDA computes the posterior probability P(y=c|x) for every class ccc from the input features xxx. This is derived from Bayes' theorem, and it applies the likelihood and prior probabilities:

$$P(y = c|x) \propto P(x|y = c)P(y = c)$$

- **K**-Nearest Neighbors classifier: K-Nearest Classifier was used to predict loan default by finding the most similar borrowers on important features. Using GridSearchCV further,

the model was tuned to determine the best parameters, which enhanced classification accuracy and stability.

3) **Clustering Models:** Clustering models were used to cluster borrowers into segments in order to assist in understanding customer profiles and high-risk segment identification.
   - **K-Means Clustering**: an unsupervised clustering algorithm particularly tailored for clustering operations. It operates by partitioning data points into a fixed number (K) of clusters. By classifying borrowers into two clusters default or not default. Moreover, utilizing GridSearchCV, the model was tuned to determine the best parameters.
     Formula:

$$c^k = \frac{1}{N_{ck}} \sum_{x_i \epsilon c_k} x_i$$

E. Evaluation matrix:
In order to evaluate the model performance the following metrics were utilized, both for classification and clustering goals:

1) Confution matrix: Employed for the classification model, it yields comprehensive information regarding. The confusion matrix is also handy when comparing and assessing classification model errors, serving to aid in tuning.

2) Accuracy: Employed to classify a model,
The number of accurate decisions, giving the basic performance of a model.
Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

3) Mean Squared Error (MSE): It's employed for regression model to analyze the mean squared difference between actual and forecasted values, indicating precision in predictions.
Formula:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (yi - y\hat{}i)^2$$

4) R-squared (R^2): It estimates the regression model's performance by estimating the proportion of explained variance by the model, reflecting the explanatory ability of the model.

Formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - y\hat{}_1)^2}{\sum_{i=1}^{N} (y_i - y)\,2}$$

5) F1-Score: Used for classification models. It is a harmonic mean of precision and recall.
Formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

6) ROC-AUC: Mainly applicable to binnary classification models, A measure of the model's ability to discriminate between classes is offered by ROC-AUC.
Formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

7) Precision: Applied to classification models when False Positive cost is higher. Reflects the correctne ss of false positive.
Formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

8) Recall: Applied to classification models when the False Positive cost is higher.
Reflects the correctness of true positive.
Formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

9) Classification report: A detailed report for classification models. It includes different metrics.
10) Silhouette Score**:** Specific to cluster models such as KNN Clustering and K-means Clustering, The silhouette score estimates the similarity of data points in a given cluster with those of other clusters.

$$\text{Silhouette Score} = \frac{\text{b} - \text{a}}{\max(\text{a}, \text{b})}$$

IV. Result And Discussion

I assessed several algorithms with two target variables: default_in_last_6months and credit_card_default, following a dual-approach framework. The aim was to identify best-performing model in predicting the loan default risk and comparing the way these target variables influence the outcome and whether there will be impact on outcome

A. Target variable 1: default_in_last_6months

For this target variable, we utilized different classification models. Clustering methods like K-Means and Agglomerative Clustering were also utilized.

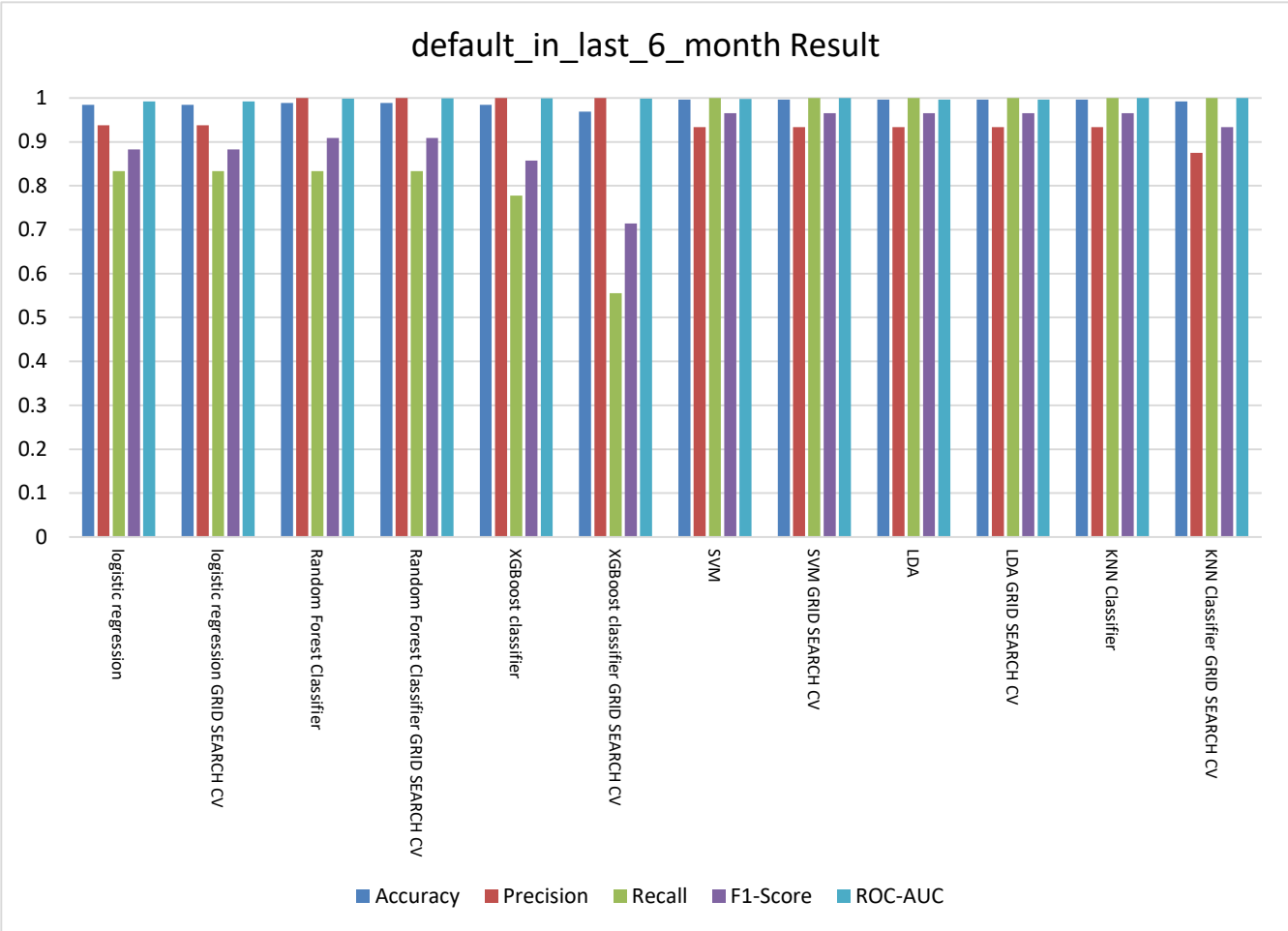
1) SVM model with Grid Search Cross-Validation (CV) and KNN Classifier is the best with an accuracy of 99.61%, recall of 100%, precision of 93.33%, F1-Score of 96.55, and ROC-AUC score of 99.97%, proving its strength in differentiating default and non-default cases.
2) Random Forest and LDA Classifier models also worked exceptionally well.

B. Target Variable 2: Credit_Card_Default

The model exhibits minimal variation in output in the evaluation matrix, this means that there is a variation in the nature of both target variable.

1) The Random Forest Classifier was the best-performing model, with an accuracy of 98.84%, precision of 100%, F1-Score of 0.9412, and an ROC-AUC of 98.84%.
2) XGBoost with Grid Search CV proved to be a powerful model with an accuracy of 98.84%, precision of 100%, and F1-Score of 0.9412.0.

Figure 1: default_in_last_6_month classification result



3) The Logistic Regression model performs well with an accuracy of 97.67% and balanced precision, recall, and F1-Score values.

Table 1: Clustering Result Table

Target variable	model	Silhouette Score
default_in_last_6_month	K-Mean Clustering	0.5715
default_in_last_6_month	K-Mean Clustering GRID SEARCH CV	0.5713
credit_card_default	K-Mean Clustering	0.5715
credit_card_default	K-Mean Clustering GRID SEARCH CV	0.5713

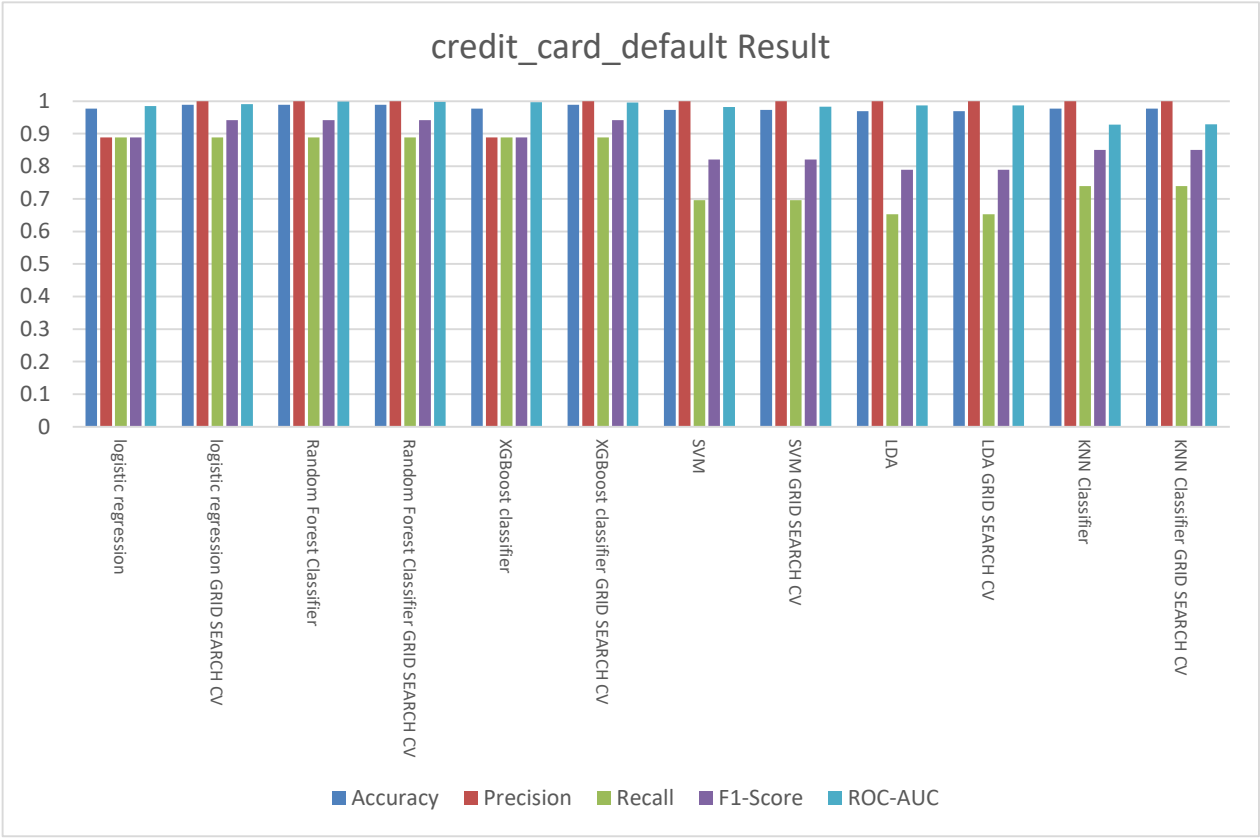
Table 2: Regression Result Table

Target variable	Model	mse	r^2
default_in_last_6_month	linear regression	0.00759	0.8519
default_in_last_6_month	linear regression standard scaler	0.76	0.8519
default_in_last_6_month	linear regression Agglomerative Clustering	0.00759	0.8519
default_in_last_6_month	linear regression K Means Clustering	0.0076	0.851
default_in_last_6_month	Polynomial regression	0.0096	0.8122
default_in_last_6_month	Polynomial regression GRID SEARCH CV with pipeline	0.0073	0.8578
credit_card_default	linear regression	0.03	0.629
credit_card_default	linear regression standard scaler	0.25	0.7335
credit_card_default	linear regression Agglomerative Clustering	0.03	0.629

credit_card_default	linear regression K Means Clustering	0.03	0.6308
credit_card_default	Polynomial regression	0.0231	0.716
credit_card_default	Polynomial regression GRID SEARCH CV with pipeline	0.0231	0.716

.

Figure 2: credit_card_default classification result



Dual strategy showed that model has excellent performance for the second, but discrepancies in recall and f1-score emphasize the significance of every feature set, For example:

1. Various models performed well for both target variables. There isn't a model that performs best for both target variable.
2. Random Forest classifier performed very well for both target variables.
3. XGBoost, though highly accurate, also indicated a recall trade-off on both variables

V. REFERENCES

1. Suhadolnik, Nicolas, Jo Ueyama, and Sergio Da Silva. "Machine learning for enhanced credit risk assessment: An empirical approach." Journal of Risk and Financial Management 16.12 (2023): 496.
2. Noriega, Jomark Pablo, Luis Antonio Rivera, and José Alfredo Herrera. "Machine Learning for Credit Risk Prediction: A Systematic Literature Review." Data 8.11 (2023): 169.
3. Emmanuel, Ileberi, Yanxia Sun, and Zenghui Wang. "A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method." Journal of Big Data 11.1 (2024): 23.
4. Wang, Yuelin, et al. "A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data." Procedia Computer Science 174 (2020): 141-149.
5. Maharjan, Menuka. "Comparative analysis of data mining methods to analyze personal loans using decision tree and naïve bayes classifier." International Journal of Education and Management Engineering 12.4 (2022): 33.
6. Sum, R. M., et al. "A New Efficient Credit Scoring Model for Personal Loan Using Data Mining Technique Toward for Sustainability Management." Journal of Sustainability Science and Management 17.5 (2022): 60-76.
7. Moradi, Saba, and Farimah Mokhatab Rafiei. "A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks." Financial Innovation 5.1 (2019): 1-27.