International Journal of Statistics and Applied Mathematics

**ISSN:** 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(8): 196-205
© 2025 Stats & Maths
https://www.mathsjournal.com
Received: 21-07-2025
Accepted: 23-08-2025

**Talawar Basavaraj**
Department of Statistics,
Karnatak University, Dharwad,
Karnataka, India

**Talawar AS**
Department of Statistics,
Karnatak University, Dharwad,
Karnataka, India

**Corresponding Author:**
**Talawar Basavaraj**
Department of Statistics,
Karnatak University, Dharwad,
Karnataka, India

# A statistical evaluation of machine learning classifiers for credit risk prediction

**Talawar Basavaraj and Talawar AS**

**DOI:** https://www.doi.org/10.22271/maths.2025.v10.i8c.2140

## Abstract

In the rapidly evolving financial landscape, the precision of credit scoring models has become a cornerstone of risk management and profitability for lending institutions. This study embarks on a pioneering exploration into the predictive capabilities of different classifier models within the realm of credit scoring. By delving into the intricate dynamics of behavioral and collection scoring, we reveal how these models not only assess creditworthiness but also preemptively identify potential defaulters with remarkable accuracy. Our analysis transcends traditional methodologies, integrating advanced machine learning techniques to enhance predictive power and decision-making efficacy. The findings illuminate the nuanced interplay between borrower characteristics and default probabilities, offering unprecedented insights into the optimization of credit risk strategies. This work not only augments the toolkit of financial institutions but also sets a new benchmark in the scientific discourse on credit scoring. Through this endeavor, we aim to elevate the standards of credit risk assessment, ensuring that the allocation of credit is both judicious and equitable. Our unique approach promises to transform how financial institutions navigate the complexities of creditworthiness, ultimately fostering a more robust and resilient financial ecosystem. Based on the analysis, we conclude that, the loan percent income is most significant variable and home ownership, the other is the least significant. Among statistical models, linear discriminant analysis (LDA), logistic regression (LR) and Gaussian Naïve Bayes (GNB), GNB has higher receiver operating characteristic curve (ROC) and area under the curve (AUC). The support vector classifier (SVC) is considered as the best classifier with consistent evaluation metrics across train and test split.

**Keywords:** Credit scoring, machine learning, risk assessment, credit risk management, predictive modeling

## Introduction

The significance of assets in banking is reflected in the profitability derived from issuing credit cards to customers. Credit scoring endeavors focus on identifying the impact of various applicants' characteristics associated with illicit behavior and payment defaults. The central aim of any banking system is to distinguish valuable stakeholders from whom they can achieve optimal returns on investments in assets (Kumar, 1998) [1]. The banking sector significantly influences credit card holders through its myriad services. Financial institutions issue credit cards only after thorough verification and validation process. However, this rigorous scrutiny does not always guarantee that the most deserving candidates receive these credit cards (Bhattacharya, 2021) [2].

Credit scoring model predictions have become integral part of the commercial sector. Credit scoring involves a suite of decision models and underlying techniques that aid lenders in issuing consumer credit (Thomas *et al*., 2002) [3]. This method, primarily utilized in consumer credit, assists credit granters in making informed lending decisions (Abdou and Pointon, 2011) [4]. It functions as a crucial decision-making tool for lenders in allocating consumer credit (Halima and Humira, 2012) [5]. Credit scoring, a traditional decision-making model, assesses the risk associated with credit products, such as credit cards and loans, by analyzing the historical data of applicants. This process aids credit lenders in making well-informed decisions regarding the issuance of credit products (Altman and Saunders, 1998) [6].

In contemporary practice, financial institutions employ a variety of risk assessment tools and techniques to mitigate risk. The analysis of customer credit data is crucial in identifying potential defaulters. The effectiveness of a credit scoring model is paramount for managing and establishing credit risk, which is essential for the institution's profitability. Research indicates that ensemble-based approaches are particularly effective in credit score assessment (Dastile *et al.*, 2020) [7].

Credit scoring systems are predominantly used to estimate the probability of loan default. The primary objective of this paper is to assess and compare the predictive power of different classifier models in credit scoring. This involves analyzing the strengths and weaknesses of these models. We also provide recommendations for financial institutions regarding model selection and the automatic identification of defaulters using advanced machine learning methods. Thus, the purpose of this paper is to identify key characteristics of customers to support credit decisions, compare various variable selection methods, make credit decisions using the generalized linear model (GLM), specifically Logistic Regression model, Linear discriminant analysis & support vector classification and compare the performance of GLM with machine learning methods (Hand and Henley, 1997) [8].

## Methodology

The data collection process varies depending on the type of data. Datasets can be sourced from files, databases, sensors, and other sources. For this study, we sourced our data from the Kaggle website: Kaggle Credit Risk Dataset. Data pre-processing is a crucial step in machine learning, significantly impacting the accuracy of model building. According to the 80/20 rule in machine learning, 80% of the time is spent on data pre-processing, and 20% on analysis. Data pre-processing involves cleaning raw data to produce clean data suitable for training models. Effective data pre-processing is essential for achieving good results from machine learning models.

**Table 1:** Credit Data Description.

| Sr. No. | Feature Name | Description |
|---|---|---|
| 1 | loan_status | Loan status (0 is non default 1 is default) |
| 2 | person_age | Age of Customer |
| 3 | person_income | Annual Income of Customer |
| 4 | person_home_ownership | Home ownership (Categorical: RENT /OWNED/etc) |
| 5 | person_emp_length | Employment length (in years) |
| 6 | loan_intent | Loan intent (Categorical: Home Improvement/ Medical/ Personal/Venture) |
| 7 | loan_grade | Loan grade (Categorical: A to E) |
| 8 | loan_amnt | Loan amount |
| 9 | loan_intrate | Interest rate |
| 10 | loan_percent_income | Ratio of Loan to income |
| 11 | cb_person_default_on_file | Historical default (Categorical: Y/N) |
| 12 | cb_preson_cred_hist_length | Credit history length |

To train the best-performing model using the pre-processed data, we utilized supervised learning techniques (Tsai and Wu, 2008) [9].

### 2.1 Supervised Learning

Supervised learning involves training an AI system with labeled data, where each data point is tagged with the correct label. Supervised learning can be divided into two primary categories: "Classification" and "Regression."

Classification: A classification problem occurs when the target variable is categorical, meaning the output can be classified into distinct classes, such as "default" or "non-default." The commonly used classification algorithms are:

### 2.1.1 Gaussian (Normal) Naive Bayes

The Gaussian (Normal) Naive Bayes can be adapted to handle real-valued attributes by assuming a Gaussian (Normal) distribution. This classification method simplifies the model by requiring only the estimation of the mean and standard deviation from the training data (Saunders and Allen 2010) [10]. Let $X$ represents the values for an input variable in the training data. Probabilities of new $x$ values are estimated using the Gaussian probability density function (pdf). The parameters values can be plugged into the Gaussian pdf with a new input for the variable, providing an estimate of the probability of that new input value for that class.

$$f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \times e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$

Where

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\sigma^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$, n is the number of instances

### 2.1.2 Logistic Regression

Logistic regression is similar to multiple linear regression, with the primary difference being that the outcome is binary and a structured model approach. It is a popular method due to its fast computational speed and the ability to produce a model that allows for rapid scoring of new data.

**Data and Pre-processing:** The dataset used is the Credit dataset from the machine learning data archive, containing 12 covariates (8 numerical, 4 categorical) and 32,581 observations. Each observation represents an individual customer, with the response variable indicating their classification (1 = "Bad" or 0 = "Good"), and the covariates providing various attributes related to the customer's personal or financial information.

A binary logistic model is fitted to the data using the logit link function. This models the classification of the $i^{th}$ customer as either good or bad using a Bernoulli random variable.

$$Y_i = \begin{cases} 1, if\ the\ customer\ is\ creditworthy \\ 0, if\ otherwise \end{cases}$$

Conditional Probability is

$$P(Y_i = 1|x_i) = \pi_i\ and\ P(Y_i = 0|x_i) = 1 - \pi_i$$

Where $x_i$ is a vector of covariates associated with this customer.

The conditional expectation is then given by:

$$E[Y_i|X_i] = \pi_i$$

and this is associated to a linear predictor via the logit function,

$$i.e., logit\ \pi_i = lo\ g\left(\frac{\pi i}{1-\pi i}\right) = (x_i)'\beta = \eta_i$$

where $\beta$ denotes the vector of parameters to be estimated. This estimation process utilizes iterative weighted least squares (IWLS), a method thoroughly explained in Davison (2003). The conditional joint probability of $Y_1, Y_2 \dots\dots . Y_n$, assuming conditional independence, can be expressed as:

$$\prod_{i=1}^{n} \pi_i^{yi} (1 - \pi i)^{1-yi} = exp\left[\sum_{i=1}^{n} \log\left(\frac{\pi i}{1-\pi i}\right) + \sum_{i=1}^{n} \log(1 - \pi i)\right]$$

This indicates that the probability distribution belongs to the exponential family. The choice of link function is primarily driven by the desire for straightforward interpretation of model parameters.

The linear predictor model, we are fitting is given by:

$$\eta = \beta_0 + \beta_1 . personel + \beta_2 . edcation + \beta_3 . home\ loan + \beta_4 .\ medical + \beta_5 . venture + \beta_{6.} debit\ consolidation$$

To assess which variables influence creditworthiness and their impact, we start by evaluating the significance of variable groups. This involves performing a likelihood ratio test by comparing the deviance of a full model with a reduced model where one group of variables is omitted. The test statistic is then compared to a $\chi^2$distribution, where $k$ is the number of excluded parameters [4].

### 2.1.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA), a traditional technique for classification based on assumptions that, the data for each variable is normally distributed and all attributes have the same variance. LDA estimates the mean and variance for each class from the data. In a univariate scenario with two classes, the mean for each class is computed as the average of values in that class.

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i$$

Here, $\overline{X_k}$ represents the mean value of $x$ for class $k$, and $n_k$denotes the number of instances in class $k$. Variance is computed across all classes as the average of the squared deviations of each value from the mean.

$$\sigma^2 = \frac{1}{n-k} \sum_{i=1}^{n} (X_i - \overline{X_k})^2$$

### 2.1.4 Support Vector Machine (SVM)

The core concept of a Support Vector Machine (SVM) is to identify the optimal hyperplane that maximizes the margin between two classes. In a two-dimensional space, this hyperplane is a line; in three dimensions, it is a plane; and in higher dimensions, it is a hyperplane. The objective is to find the best boundary that separates two classes.

### 2.1.5 Extreme Gradient Boost (XGBoost)

The XGBoost, or Extreme Gradient Boosting, typically uses decision trees as base learners. These trees are constructed iteratively until a stopping criterion is met. XGBoost is an ensemble method that employs Classification and Regression Trees (CART), where each tree contains real-valued scores at the leaves, which can be used for classification if necessary.

### 2.1.6 K-Nearest Neighbor's (KNN)

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning method used for classification and regression. It classifies or predicts based on the proximity of data points. For classification, a class label is assigned based on the majority vote among the nearest neighbors. While "majority vote" is commonly used, it does not strictly require more than 50% of the votes, especially in multi-class problems.

### 2.1.7 Random Forest

Random Forest is a powerful ensemble learning algorithm known as Bootstrap Aggregation or bagging. It builds multiple decision trees from different subsets of the training data and combines their predictions to enhance model performance.

- **Bootstrap Aggregation (Bagging):** Involves creating multiple models from a single training dataset.
- **Random Forest:** A modification of bagging where decision trees are built using a random subset of features at each split, reducing correlation among trees and improving classification performance.

### 2.1.8 Decision Trees

Decision Trees, specifically the Classification and Regression Trees (CART) algorithm, are fundamental in predictive modelling. The CART model is represented as a binary tree where each node splits based on an input variable, and leaf nodes contain the prediction outcome.

### Training and Testing the Model

To train a model, the data is first divided into two distinct sets: the Training Set and the Testing Set. The training set is used to teach the model how to process information by fitting it to the data. During this phase, only the training data is utilized; the testing data remains unseen to ensure the model's performance is evaluated on truly new data.

**Training Set:** This subset is used to fit the model's parameters. It represents the information from which the model learns and adapts.

**Testing Set:** This subset consists of previously unseen data used to assess the performance of the fully-trained model. It is crucial that this set is not used during the training phase to avoid any bias in performance evaluation.

Once the data is split, the training process begins with one of the selected models. After training, the model is applied to the testing set to predict outcomes. The performance is then evaluated using a confusion matrix, which provides a detailed breakdown of the model's accuracy.

**Model Evaluation:** This evaluation process helps in identifying the most accurate and effective model, guiding further refinement and tuning, including adjusting hyper parameters to improve accuracy and optimize performance based on the confusion matrix results.

**Table 2:** The Model of Confusion Matrix

| ------- | **Positive** | **Negative** | --------------- |
|---|---|---|---|
| Positive | True Positive (TP) | False Negative (FN) | Sensitivity $\frac{TN}{(TN+FN)}$ |
| Negative | False Positive (FP) | True Negative (TN) | Specificity $\frac{TN}{(TN+FN)}$ |
| -------- | Precision $\frac{TP}{(TP+FP)}$ | Negative Predicted value $\frac{TN}{(TN+FN)}$ | Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$ |

Key Terms for Evaluating Classification Models

***Accuracy:*** The percent (or proportion) of cases classified correctly.

$$Accuracy = \frac{\sum_{i=1}^{n} True\ Positive + \sum_{i=1}^{n} True\ Negative}{SampleSize}$$

**Sensitivity /Recall:** The percent (or proportion) of all 1s that are correctly classified as 1's.

$$Sensitivity = \frac{\sum_{i=1}^{n} True\ Positive}{\sum_{i=1}^{n} True\ Positive + \sum_{i=1}^{n} False\ Negative}$$

**Specificity:** The percent (or proportion) of all 0s that are correctly classified as 0's.

$$Specificity = \frac{\sum_{i=1}^{n} True\ Negative}{\sum_{i=1}^{n} True\ Negative + \sum_{i=1}^{n} False\ Positive}$$

**Precision:** The percent (proportion) of predicted 1s that are actually 1's.

$$Precision = \frac{\sum_{i=1}^{n} True\ Positive}{\sum_{i=1}^{n} True\ Positive + \sum_{i=1}^{n} False\ Positive}$$

### 2.2 F1- Score / F1- Measure

Classification accuracy is a popular metric because it provides a straightforward summary of a model's performance in a single value. However, to get a more nuanced view of a model's effectiveness, especially in the context of imbalanced datasets, the F-Measure (or F1 Score) is used. The F-Measure combines both precision and recall into a single metric. While precision measures the accuracy of positive predictions, and recall measures how well all positive instances are identified, neither metric alone provides a complete picture. A model might show high precision but poor recall, or vice versa.

In summary, the F-Measure integrates both precision and recall into a comprehensive metric, providing a more holistic evaluation of a model's performance in scenarios where both metrics are important.

$$F1 = \frac{2}{Recall^{-1} + Precision^{-1}} = 2\ \frac{Precision.Recall}{Precision + Recall}$$

### 2.3 ROC Curve

The ROC curve is a graphical representation of a classifier's performance, showing the trade-off between recall (sensitivity) and specificity. The ROC curve illustrates how the balance between recall and specificity shifts as the classification threshold changes. A good classifier will achieve high recall while maintaining high specificity. In other words, it should correctly identify as many positive cases as possible without misclassifying too many negatives as positives.

The AUC is a summary metric derived from the ROC curve. AUC measures the overall performance of the classifier:

- An AUC of 1 indicates a perfect classifier that correctly classifies all positive and negative instances.
- An AUC of 0.5 signifies a classifier with no discriminative power, equivalent to random guessing.

### 2.4 Variable Importance and the Dalex Package

The Dalex package is a valuable tool for interpreting machine learning models. It helps in exploring and explaining model behavior by creating an "Explainer" object, which wraps around a predictive model. This allows users to:

Variable Importance: Assessing the importance of variables in a model is crucial for:

- Model Simplification**:** Identifying and excluding non-influential variables.
- Model Exploration**:** Comparing variable importance across different models.
- Model Validation**:** Validating model predictions against domain knowledge.
- Knowledge Generation**:** Discovering new factors that influence the outcome.

**Variable importance methods fall into two categories**

- Model-Specific Methods**:** Tailored to the structure of the model, such as using regression coefficients for linear models or feature importance from tree-based models like random forests.
- Model-Agnostic Methods**:** Applicable across various models, often based on permutation or perturbation of variables to observe changes in model performance.

**Permutation-Based Importance**: This approach measures how the performance of the model degrades when the values of a specific variable are permuted. If performance drops significantly, the variable is considered important. Overall, the permutation-based method, inspired by Leo Breiman's work with random forests, is a robust, model-agnostic tool for evaluating variable importance and exploring model performance. Let $\underline{\hat{y}} = \left(f(x_1), \ldots\ldots, f(x_n)\right)'$ denote the corresponding vector of predictions for $\underline{y}$ for model $f()$. Let $L\ (\underline{\hat{y}},\ \underline{x}, \underline{y})$ be a loss function that quantifies goodness-of-fit of model $f()$. For instance, $L\ (\underline{\hat{y}},\ \underline{x}, \underline{y})$ may be the value of log-likelihood or any other model performance measure. Consider the following algorithm:

a) Compute $L^0 = \ L\ (\underline{\hat{y}},\ \underline{x}, \underline{y})$i.e., the value of the loss function for the original data. Then, for each explanatory variable $X^j$ included in the model, do steps 2-5.

b) Create matrix $X^j$ by permuting the $j^{th}$ column of, i.e., by permuting the vector of observed values of $X^j$.

c) Compute model predictions $\underline{\hat{y}}^j$ based on the modified data$x^j$.

d) Compute the value of the loss function for the modified data:

$$L^j \ = \ L\ \left(\underline{\hat{y}}^j,\ x^j, \underline{y}\right)$$

e) Quantify the importance of $X^j$ by calculating

$$vip^j_{Diff} = L^j - L^0 \text{ or } vij^j_{Ratio} = \frac{L^j}{L^0}$$

**3. Results and Discussion:** We consider Kaggle's credit risk dataset for the analysis. We focused on two types of analysis (univariate and bi-variate analysis).

**3.1 Univariate Analysis:** Univariate analysis is the simplest form of data analysis, focusing on a single variable. Its primary purpose is to describe the data and uncover patterns within it, without examining relationships between variables. In this analysis, we utilized libraries such as Pandas and NumPy for data manipulation. We started by checking the shape, size, and information about the data, followed by obtaining summary statistics.

We visualized the target variable using a pie chart, which illustrates the distribution of the two classes (0 and 1) in terms of percentage. Additionally, we analyzed the frequency of unique values in each categorical variable to better understand the dataset.

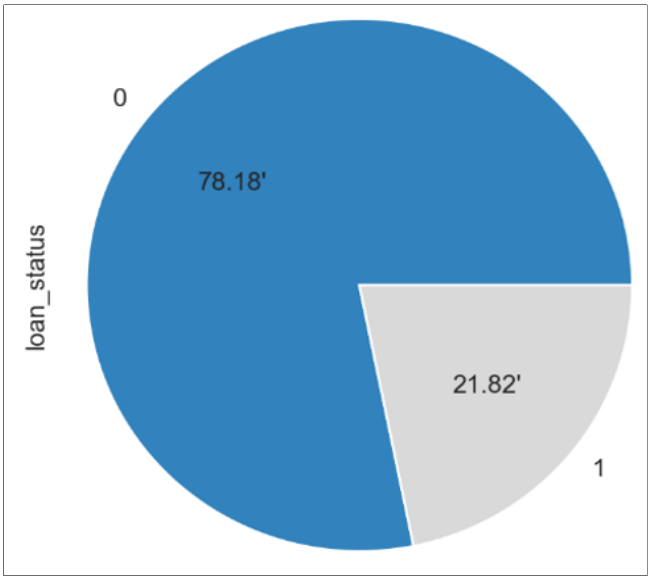

**Fig 1:** The Dataset Splitting into Train Data and Test Data.

### 3.2 Bivariate Analysis

Bivariate analysis investigates the relationship between two variables to understand their interaction and the nature of their association. This analysis aims to uncover whether there are any significant relationships, discrepancies, or causes of differences between the two variables.

### Key Objectives

- Explore Relationships: Determine how two variables are related.
- Identify Discrepancies: Spot any inconsistencies or variations between the variables.
- Assess Relationship Depth: Gauge the strength and significance of the relationship.

In our analysis, we used histograms to evaluate the distribution of the data. This visualization helps in understanding how values are spread across different ranges, providing insights into the data's overall distribution.

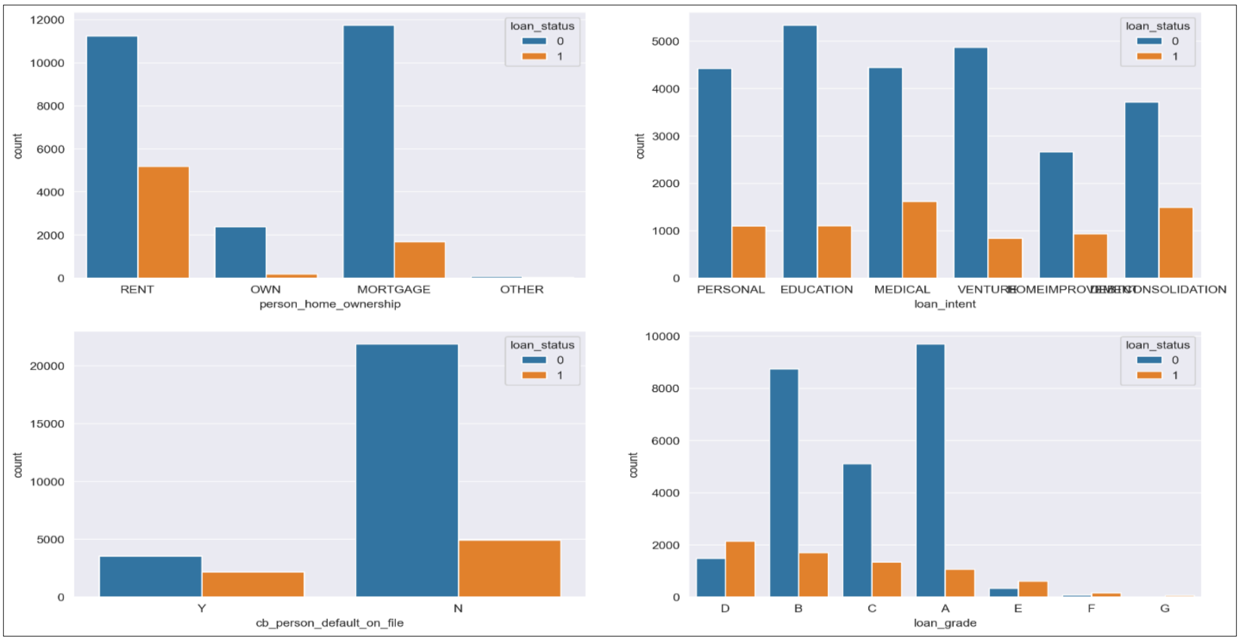

**Fig 2:** To Finding Missing Values by using Histogram.

From the Figure 3 we observed that neither feature of the graphs follows a normal distribution. Consequently, we decided to fill the missing values (NaN's) with the median values for both the loan interest rate and employment length features. Additionally, we identified some unrelated data entries, which were cleaned using cross-tabulations and median values. Notably, clients who had not defaulted on loans had a median interest rate that was 4% lower than those who had defaulted. Given that issuing loans to high-risk clients can have adverse outcomes for both the lender and the client, our goal is to enhance credit risk modeling using machine learning algorithms to mitigate these risks.

### Chi-Square test for independence

**Table 3:** Chi-Square test for independence

| Variables | $\chi^2_{cal}$ | df | p-value |
|---|---|---|---|
| person home ownership | 1908 | 3 | $p - value < 2.2e - 16$ |
| loan intent | 520.51 | 5 | $p - value < 2.2e - 16$ |
| loan grade | 5609.2 | 5 | $p - value < 2.2e - 16$ |
| cb person default on file | 1044.4 | 1 | $p - value < 2.2e - 16$ |

**Interpretation:** There is no association between variable loan status and all other categorical variables.

### 3.4 Student's t-test for significant difference

**Population 1:** Defaulted vs Population 2: Non defaulted

**Table 4:** Student's t-test for significant difference

| Variables | $t_{cal}$ | df | p-value |
|---|---|---|---|
| person_age | 3.9421 | 11533 | $8.1 \times 10^{-05}$ |
| person_income | 35.853 | 10867 | $p - value < 2.2e - 16$ |
| person_emp_length | 14.737 | 10867 | $p - value < 2.2e - 16$ |
| loan_amnt | 17.389 | 10099 | $p - value < 2.2e - 16$ |
| loan_int_rate | −57.757 | 9641 | $p - value < 2.2e - 16$ |
| loan_percent_income | −59.086 | 8904 | $p - value < 2.2e - 16$ |
| cb_preson_cred_hist_length | 2.7793 | 11244 | $p - value < 2.2e - 16$ |

**Interpretation:** There is no significant difference between loan status and all other variables.

**Table 5:** Confusion Matrix

| Variables | $\hat{Y} = 1$ | $\hat{Y} = 0$ |
|---|---|---|
| $Y = 1$ | 0.8859 | 0.9485 |
| $Y = 0$ | 0.5623 | 0.7516 |

**Model evaluation:** The following table gives model evaluation of classifiers based on different variables.

**Table 6:** Model Evaluation based on all variables.

| Classifier | Train Accuracy | Test Accuracy | Train F-1 Accuracy | Test F-1 Accuracy | Train ROC-AUC | Test ROC-AUC | AUC DROP |
|---|---|---|---|---|---|---|---|
| GNB | 0.8173 | 0.8132 | 0.3886 | 0.3858 | 0.6116 | 0.6177 | 0.0009 |
| LR | 0.8658 | 0.8673 | 0.6444 | 0.6410 | 0.7540 | 0.7519 | 0.0021 |
| LDA | 0.8626 | 0.8649 | 0.6486 | 0.6466 | 0.7605 | 0.7563 | 0.0041 |
| SVC | 0.9181 | 0.9189 | 0.7794 | 0.7781 | 0.8254 | 0.8200 | 0.0053 |
| XGB | 0.9584 | 0.9365 | 0.8958 | 0.8345 | 0.9073 | 0.8661 | 0.0412 |
| KNN | 0.9291 | 0.8818 | 0.8199 | 0.6914 | 0.8598 | 0.7845 | 0.0753 |
| RF | 1.0000 | 0.9345 | 0.9999 | 0.8263 | 0.9999 | 0.8573 | 0.1426 |
| DT | 1.0000 | 0.8926 | 1.0000 | 0.7565 | 1.0000 | 0.8492 | 0.1508 |

The difference between train accuracy and test accuracy is less for LR, LDA and SVC classifier methods as compared to all other classifier methods. But SVC showed better performance on both the training and test datasets. Among these classifier methods, SVC is considered as the best classifier due to its consistent performance across all evaluation metrics.

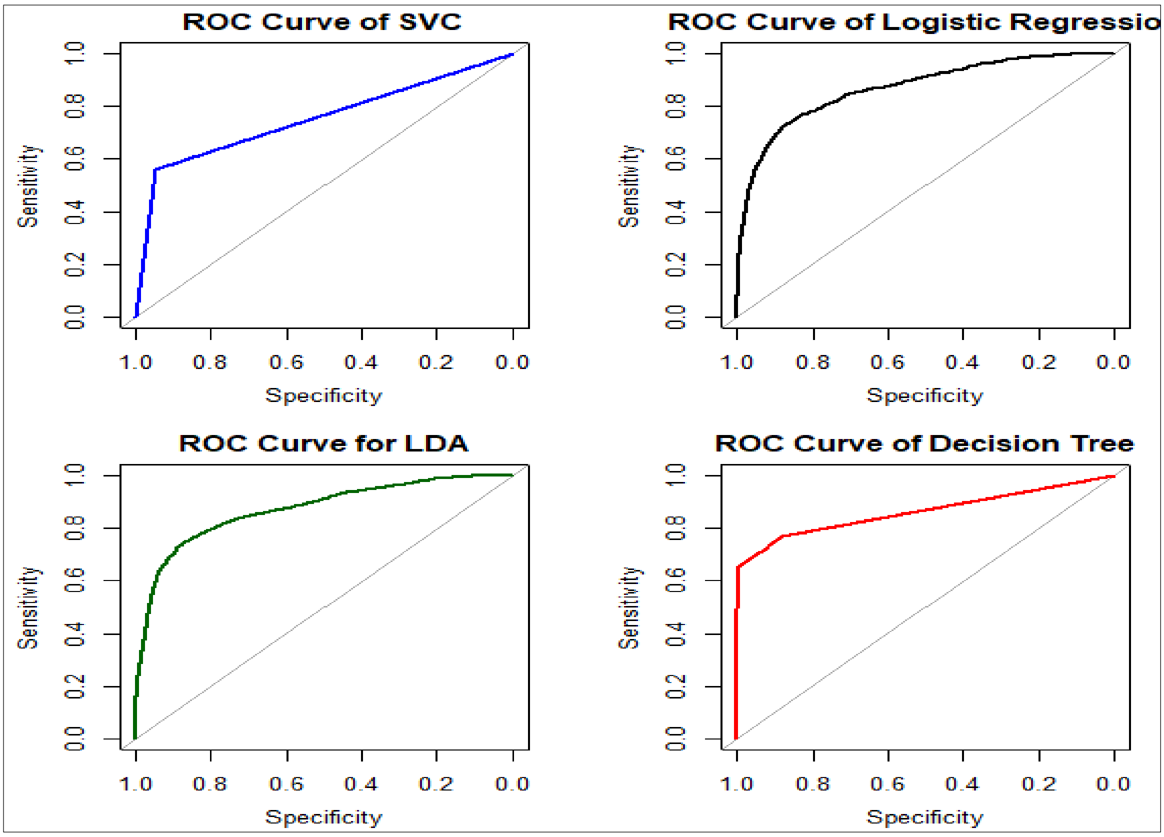

**Fig 3:** ROC Curves of different Models.

**Checking Account Balance:** The odds of creditworthiness increase with higher checking account balances. Compared to those with a zero balance, the odds are higher by factors of $e^{\beta_{0-1}}$= 1.662, $e^{\beta_1}$= 3.015, and $e^{\beta_0}$= 6.25 for various account balances. Notably, those with no checking account have odds $e^{\beta_0+\beta_{0-1}}$= 3.76 times higher than the 0-1 balance group, which suggests an unexpected trend where no checking account is associated with higher creditworthiness.

**Personnel Variable:** Consumers whose personnel category is "purchase of a personnel" have odds of creditworthiness $e^{\beta(personnel)-\beta(education)}$= 5.497 times higher compared to those whose personnel category is "education."

**Loan Duration:** Each additional month of loan duration reduces the odds of creditworthiness by a factor of $e^{\beta(education)}$= 0.958 reflecting the increased default risk associated with longer loans.

**Credit History**
Consumers with a "critical" credit history show significant increases in creditworthiness odds.

**Table 7:** Variable importance based on SVC, LR and LDA.

| Variables | Variables Importance | | |
|---|---|---|---|
| | SVC | LR | LDA |
| Person_home_ownership_rent | 0.3056 | 0.8330 | 0.0432 |
| Loan_percent_income | 0.1240 | 1.2949 | 0.2164 |
| Person_home_ownership_own | 0.0742 | 1.3715 | 0.0200 |
| Loan_intent_medical | 0.0626 | 0.2383 | 0.0170 |
| Loan_intent_home_improvement | 0.0609 | 0.4963 | 0.0142 |
| Loan_ntent_debit_consolidated | 0.0568 | 0.4247 | 0.0176 |
| Loan_interest_rate | 0.0558 | 0.9209 | 0.1812 |
| Loan_intent_venture | 0.0435 | 0.6156 | 0.0074 |
| Person_home_ownership_other | 0.0429 | 0.4566 | 0.0007 |
| Person_income | 0.0349 | 0.0264 | 0.1667 |
| Person_home_ownership_mortagage | 0.0269 | 0.0831 | 0.0288 |
| Cb_person_default_on_N | 0.0218 | 0.0746 | 0.0092 |
| Person_employment_length | 0.0200 | 0.0456 | 0.0608 |
| Loan_intent_education | 0.0188 | 0.4064 | 0.0098 |
| Person_age | 0.0158 | 0.0098 | 0.0544 |
| Loan_intent_personal | 0.0143 | 0.1359 | 0.0084 |
| Loan_amount | 0.0114 | 0.5650 | 0.0915 |
| Cb_person_credit_history_length | 0.0099 | 0.0021 | 0.0408 |
| Cb_person_default_on_Y | 0.0000 | 0.0758 | 0.0118 |

From table 7, Variable selected based on given variable importance following table shows min and max values of importance with respect to each model.

**Table 8:** Maximum and Minimum Selected Variables.

| Variable Importance | SVC | LR | LDA |
|---|---|---|---|
| Min Value | 0 | 0.0007 | 0.0021 |
| Max Value | 0.3056 | 0.2164 | 1.3715 |

We created six different set of variables using different thresholds.

1. **SVC_TOP_S1:** Variables having SVC importance > 0.05
2. **LR_TOP_S1:** Variables having LR importance > 0.5
3. **RF_TOP_S1:** Variables having RF importance > 0.05

So, we refer Table.6 variable importance bold values indicate higher variables importance than respective thresholds.

**Table 9:** Model Evaluation based on SVC_TOP_S1.

| Classifier | Train Accuracy | Test Accuracy | Train F-1 Accuracy | Test F-1 Accuracy | Train ROC-AUC | Test ROC-AUC | AUC DROP |
|---|---|---|---|---|---|---|---|
| GNB | 0.8336 | 0.8425 | 0.6145 | 0.6291 | 0.7515 | 0.7615 | -0.0100 |
| LR | 0.8397 | 0.8458 | 0.5471 | 0.5574 | 0.6966 | 0.7023 | -0.0057 |
| LDA | 0.8400 | 0.8440 | 0.5508 | 0.5553 | 0.6989 | 0.7017 | -0.0028 |
| XGB | 0.9075 | 0.8937 | 0.7463 | 0.7051 | 0.8045 | 0.7832 | 0.0213 |
| KNN | 0.9081 | 0.8646 | 0.7661 | 0.6547 | 0.8285 | 0.7668 | 0.0617 |
| RF | 0.9586 | 0.8655 | 0.8986 | 0.6580 | 0.9150 | 0.7691 | 0.1459 |
| DT | 0.9586 | 0.8516 | 0.8986 | 0.6373 | 0.9094 | 0.7620 | 0.1474 |

**Table 10:** Model Evaluation Based on LR_TOP_S1.

| Classifier | Train Accuracy | Test Accuracy | Train F-1 Accuracy | Test F-1 Accuracy | Train ROC-AUC | Test ROC-AUC | AUC DROP |
|---|---|---|---|---|---|---|---|
| GNB | 0.8186 | 0.8243 | 0.5838 | 0.5970 | 0.7330 | 0.7441 | -0.0111 |
| LDA | 0.8402 | 0.8474 | 0.5539 | 0.5720 | 0.7008 | 0.7115 | -0.0107 |
| SVC | 0.8758 | 0.8793 | 0.6455 | 0.6525 | 0.7463 | 0.7509 | -0.0046 |
| XGB | 0.9171 | 0.8923 | 0.7838 | 0.7138 | 0.8338 | 0.7944 | 0.0394 |
| KNN | 0.9180 | 0.8616 | 0.7963 | 0.6501 | 0.8510 | 0.7653 | 0.0857 |
| RF | 0.9897 | 0.8858 | 0.9761 | 0.7066 | 0.9799 | 0.7959 | 0.1841 |
| DT | 0.9897 | 0.8530 | 0.9759 | 0.6621 | 0.9776 | 0.7857 | 0.1919 |

**Table 11:** Model Evaluation Based on LDA_TOP_S1.

| Classifier | Train Accuracy | Test Accuracy | Train F-1 Accuracy | Test F-1 Accuracy | Train ROC-AUC | Test ROC-AUC | AUC DROP |
|---|---|---|---|---|---|---|---|
| GNB | 0.8245 | 0.8322 | 0.5900 | 0.6065 | 0.7352 | 0.7477 | -0.0125 |
| LR | 0.8372 | 0.8451 | 0.5410 | 0.5599 | 0.6935 | 0.7043 | -0.0107 |
| SVC | 0.8400 | 0.8461 | 0.5428 | 0.5571 | 0.6938 | 0.7020 | -0.0082 |
| XGB | 0.8768 | 0.8790 | 0.6381 | 0.6416 | 0.7397 | 0.7423 | -0.0027 |
| KNN | 0.9393 | 0.9090 | 0.8430 | 0.7548 | 0.8689 | 0.8149 | -0.0540 |
| RF | 0.9124 | 0.8565 | 0.7787 | 0.6301 | 0.8372 | 0.7513 | -0.0858 |
| DT | 0.9997 | 0.9031 | 0.9994 | 0.7379 | 0.9995 | 0.8048 | -0.1947 |

Among statistical models (LDA, Logistic Regression and SVC), SVC has higher ROC-AUC based on LR_TOP_S1 variable set. Thus, SVC is best classifier with consistent evaluation metrics across train and test split.

Drop-out-loss has been calculated for support vector classification (SVC) and logistic regression for top S1 variables. Loan percent income, person home ownership rent, loan interest rate and loan amount these are the most significant variables.

**Table 12:** From base line top four nearest values are selected our criteria.

| Variables | SVC Drop-out-loss | SVC Rank | LR Drop-out-loss | LR Rank |
|---|---|---|---|---|
| Baser line | 0.4974 | -- | 0.5117 | -- |
| Loan percent income | 0.4097 | 1 | 0.3419 | 1 |
| Person home ownership rent | 0.3189 | 2 | 0.1589 | 4 |
| Loan interest rate | 0.3136 | 3 | 0.2619 | 2 |
| Loan amount | 0.2636 | 4 | 0.1787 | 3 |
| Loan intent home improvement | 0.2478 | 5 | 0.1496 | 8 |
| Person home ownership own | 0.2443 | 6 | 0.1574 | 5 |
| Loan intent consolidation | 0.2438 | 7 | 0.1493 | 9 |
| Loan intent education | 0.2413 | 8 | 0.1502 | 7 |
| Loan intent venture | 0.2398 | 9 | 0.1548 | 6 |
| Person home ownership other | 0.2363 | 10 | 0.1469 | 10 |
| Full model | 0.2356 | -- | 0.1470 | -- |

The dalex package is used to calculate drop-out loss. Based on drop-out loss, the best variables are selected.

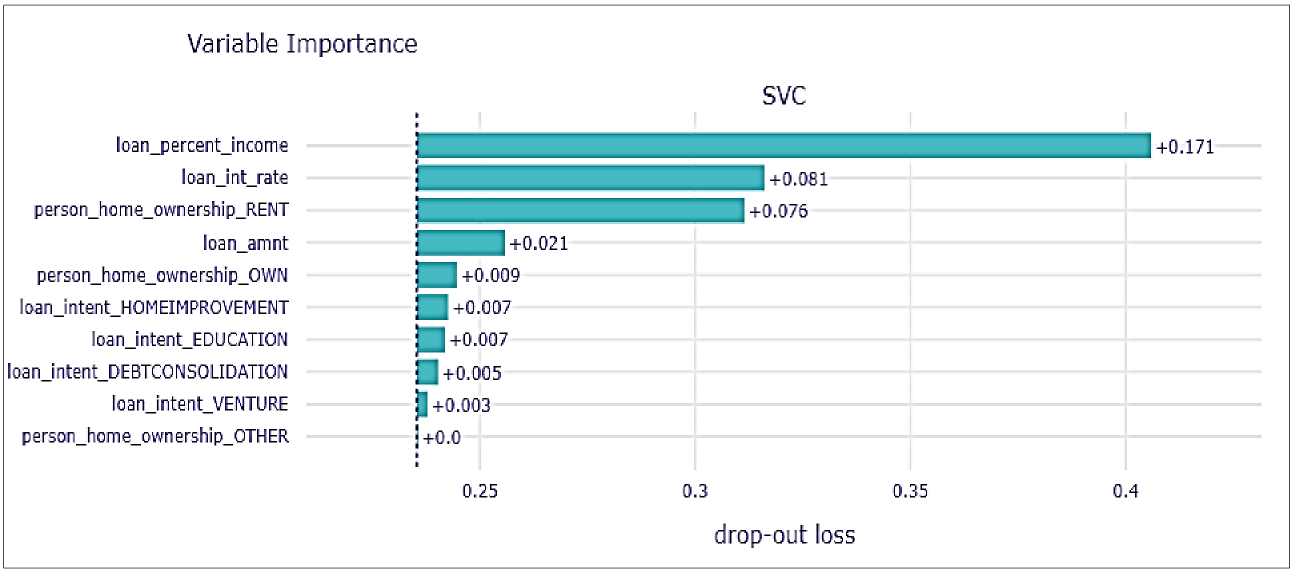

**Fig 4:** Variable importance using SVC.

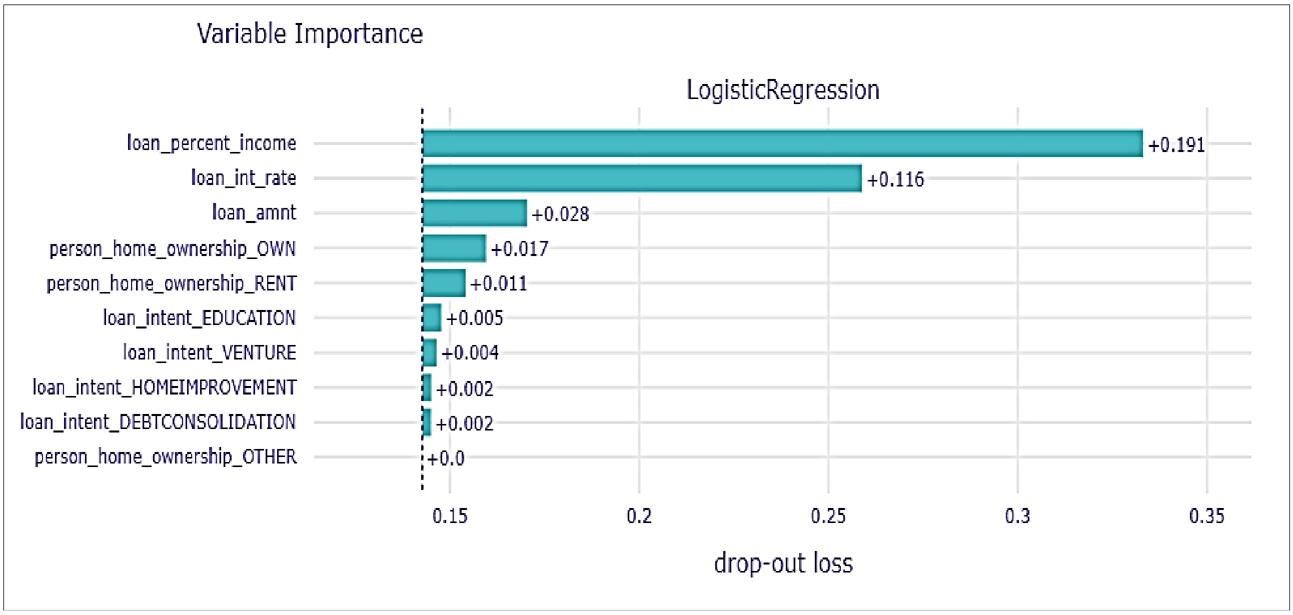

**Fig 5:** Variable importance using Logistic Regression.

Table 12 and figure 4 and 5 together, gives that drop-out loss against the variables of importance. We see that in both figures; the top four variables of importance are the same. It indicates loan percent income, loan interest rate, loan amount and person home ownership own are the most significant and contributing variables towards sanction of the loan.

## Conclusions

The credit scoring enables lenders to create a scorecard where each characteristic is assigned a weight, and the aggregate score determines an individual's creditworthiness. The decision to approve or reject an applicant is made by setting a cut-off level (threshold) corresponding to a specific value of the estimated probability of default (PD). Applicants with a PD above this threshold are not granted credit. Loan percent income, person home ownership RENT, loan int rate, loan amount these are the key characteristics of a customer to support credit decision. Based on the analysis, we conclude that, for SVC, LR, and RF, the bold values indicate higher variables of importance. We also conclude that, the loan percent income is most significant variable and home ownership, the other is the least significant. Among statistical models (LDA, LR, GNB), Gaussian Naïve Bayes has higher ROC AUC based on LR_TOP_S1 variable set. The SVC classifier is considered as the best classifier with consistent evaluation metrics across train and test split.

Drop out loss has been calculated for support vector classification (SVC) and logistic regression for LR_TOP_S1 variable. The four variables such as loan_percent_income, person_home_ownership_RENT, loan_int_rate, and loan_amnt are the most significant variables.

**Conflict of interest**: The authors declare that they have no conflict of interest.

## References

1. Kumar K. Banking profitability and asset management. International Journal of Banking and Finance. 1998;22(2):89-102.
2. Bhattacharya P. The impact of rigorous scrutiny in credit card issuance. Journal of Financial Services Research. 2021;50(1):123-139.
3. Thomas LC, Edelman DB, Crook JN. Credit scoring and its applications. Philadelphia: SIAM; 2002.

4. Abdou H, Pointon J. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management. 2011;18(2-3):59-88.
5. Halima H, Humira H. Decision-making in consumer credit: Techniques and applications. Journal of Consumer Credit. n.d.;27(4):212-225.
6. Altman EI, Saunders A. Credit risk measurement: Developments over the last 20 years. Journal of Banking and Finance. 1998;21(11-12):1721-1742.
7. Dastile X, Celik T, Potsane M. Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing. 2020;91:106263.
8. Hand DJ, Henley WE. Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society. Series A, Statistics in Society. 1997;160(3):523-541.
9. Tsai CF, Wu JW. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications. 2008;34(4):2639-2649. doi:10.1016/j.eswa.2007.05.019
10. Saunders A, Allen L. Credit risk management in and out of the financial crisis: New approaches to value at risk and other paradigms. Hoboken, NJ: Wiley; 2010.