A Comparative Study of Few Classifications Techniques

Dr. James Kurian

Associate Professor, Department of Statistics, Maharaja's College, Ernakulam, Kerala, India, PIN-682011.

DOI: https://doi.org/10.51244/IJRSI.2025.120700010

Received: 23 June 2025; Revised: 05 July 2025; Accepted: 09 July 2025; Published: 28 July 2025

ABSTRACT

A comparative study of the performance of three classifiers, Logistic regression, Discriminant analysis, Naïve Bayes' classifier was conducted using the 'Credit card defaulter' data. The relative comparison of the classifiers was done using measure of accuracy and precision obtained from the confusion matrix. Cross validation technique was used while constructing the confusion matrix. Study showed that Logistic regression provided better performance based on accuracy measure from the confusion matrix (77.88% accuracy) compared to the other two and the accuracy level of Bayes' classifier was the least (36.22%). The results of these study are limited to this particular data set and hence cannot be extended as a general result.

INTRODUCTION

A classification problem is the problem of assigning objects into two or more predefined groups based on the information of a number of variables related to it. In general, classification techniques are used to predict the membership category of individuals or data vectors, and also try to identify which characteristics of individuals or data vectors can efficiently predict their category membership. This means that the dependent variable is a categorical or nominal or non-metric variable and the independent variables are metric variables. Classification techniques found applications (Harris, R. J (2001)) in many fields including, Physics, Computer science, Life Science, Business Social media applications etc. There are many statistical techniques available for solving classification problems including classification trees, logistic regression, discriminant analysis, Naïve Bayes technique etc. But we have used only three, that is, Logistic regression, Discriminant analysis, Naïve Bayes' classifier, because these three are more statistical in nature.

Different classification methods

The discriminant function (see, Harris, R. J. (2001), Huberty, et. al. (1987), Johnson, N. and Wichern, D (2002)), is the linear combination of the two or more predictor variables that will discriminate objects into two or more in the groups. A linear discriminate function requires the Normality, Linearity and no-multicollinearity assumptions (Huberty, C. J. and Olejnik, S. (2006)). Proposed by Fisher (1936), it constructs a linear function of predictor variables which minimize the possibility of misclassification.

If S_i denote the sample variance-covariance matrix for population I, then the variance-covariance matrix Σ is estimated by the pooled variance-covariance matrix $S_p = \frac{\sum_{i=1}^{k}(n_i-1)S_i}{\sum_{i=1}^{k}(n_i-1)}$ and the Linear Score Function can be written as

$$\widehat{D}_i^L = -\frac{1}{2}\bar{x}_i' S_p^{-1}\bar{x}_i + \bar{x}_i' S_p^{-1} x + \log(p_i)$$

Thus, the linear score function $\widehat{D}_i^L$ is a function of the sample mean vectors, the pooled variance-covariance matrix, and prior probabilities for k different populations. The probabilities are computed. One limitation of linear discriminant function is that it can accommodate only quantitative variables.

Another popular statistical technique that can be used for discrimination is Logistic Regression model (Harrell, Frank E. (2001)). It has the advantage that, it can accommodate qualitative variable and does not require the

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

assumption of normality and linearity. There are situations in which the response variable in a regression problem takes only two possible values 0 and 1. Assume that the data $y_1, y_2, ..., y_n$ are independent with y_i is Binomial $B(n_i, \pi_i)$. Consider the general form $y=\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ or $y = x\beta$, where, $x = (1 \quad x_1 \quad x_2 \quad \dots \quad x_k)$, $\beta' = (\beta_0 \quad \beta_1 \quad .. \quad \beta_k)$. Since we assume that the response variable y_i is a Bernoulli random variable with probability distribution as $P(y = 1) = p$ and $P(y = 0) = 1 - p$ and hence $E(y) = x\beta'$ is just the probability that the response variable takes on the value 1. Therefore, such models are called linear probability models (Menard, Scott W. (2002)) and can be expressed in the form;

$$E(y) = p_i = \frac{\exp(x\beta')}{1+\exp(x\beta')}$$

Third popular technique used for classification is Bayes' classification or Naïve Bayes' (Webb. et.al. (2005)) technique. Assume that we have k populations and the i^{th} population is denoted as π_i and $p_i = P(\pi_i)$ is the probability that a randomly selected observation is in population. The idea behind this technique is, suppose we are interested to compute $P(\pi_i|x)$, the conditional probability that an observation came from population π_i given that the observed values of the vector of variables. Now classify an observation to the population for which the value of $P(\pi_i|\mathbf{x})$**,** is maximum. This is the most probable group, given the observed values of x. Let us assume $f(\mathbf{x}|\pi_i)$ as the conditional probability density function of the variable **x**. Then, using the Bayes' rule, the posterior probability of π_i is

$$P(\pi_i|\mathbf{x}) = \frac{p_i f(\mathbf{x}|\pi_i)}{\sum_{j=1}^{k} p_j f(\mathbf{x}|\pi_j)}$$

Then, the Bayes' classification or Naïve Bayes' (Hastie & Trevor (2001)) assigns, observation **x** to the population for which the posterior probability is the maximum.

Data

Data used for this study is the credit card defaulter's data provided by Yeh, I.C.,& Lien, C.H. (2009). This is a big data set consists of credit card payment and other details of 30000 users. Out of these 30000 samples 6636 were defaulters and 23364 were non defaulters. Hence, we can say that the data set was slightly imbalances as the number of non-defaulters outnumbered the number of defaulters. A bank is interested in knowing which customers are likely to default on loan payments. The bank is also interested in knowing what characteristics of customers may explain their loan payment behaviour. So it is very useful to categorize the clients as likely 'defaulters' and 'unlikely defaulters' based on their past data history. Therefore, a good statistical classification technique or discriminating technique is necessary to analyze this data. Therefore, in this study, I compare the relative performance of different discrimination techniques by analyzing the data. To comply with the assumptions of linear discriminate analysis, few categorical variables were eliminated from the original data set. The variables used are:

Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

BILL_AMT 1 -amount of bill statement in September (NT dollar)

BILL_AMT 2 -amount of bill statement in October (NT dollar)

BILL_AMT 3 -amount of bill statement in November (NT dollar)

BILL_AMT 4 -amount of bill statement in December (NT dollar)

BILL_AMT 5 -amount of bill statement in January (NT dollar)

BILL_AMT 6 -amount of bill statement in February (NT dollar)

PAY_AMT_1 -amount of previous payment paid in September (NTdollar)

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

PAY_AMT_2 -amount of previous payment paid in October (NTdollar)

PAY_AMT_3 -amount of previous payment paid in November (NTdollar)

PAY_AMT_4 -amount of previous payment paid in December (NTdollar)

PAY_AMT_5 -amount of previous payment paid in January (NTdollar)

PAY_AMT_6 -amount of previous payment paid in February (NTdollar)

Y -default payment next month (1 yes, 0 No)

AGE -Age (year)

In this study, I compare the confusion matrix of the three classification methods, that is Discriminate function, logistic regression and naïve Bayes' classification.

Data Analysis

The data summary provided by R output is shown below:

```
   LIMIT_BAL           AGE          BILL_AMT1         BILL_AMT2
 Min.   :  10000   Min.   :21.00   Min.   :-165580   Min.   :-69777
 1st Qu.:  50000   1st Qu.:28.00   1st Qu.:   3559   1st Qu.:  2985
 Median : 140000   Median :34.00   Median :  22382   Median : 21200
 Mean   : 167484   Mean   :35.49   Mean   :  51223   Mean   : 49179
 3rd Qu.: 240000   3rd Qu.:41.00   3rd Qu.:  67091   3rd Qu.: 64006
 Max.   :1000000   Max.   :79.00   Max.   : 964511   Max.   :983931
   BILL_AMT3         BILL_AMT4         BILL_AMT5         BILL_AMT6
 Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603
 1st Qu.:   2666   1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256
 Median :  20089   Median :  19052   Median : 18105   Median :  17071
 Mean   :  47013   Mean   :  43263   Mean   : 40311   Mean   :  38872
 3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.: 50191   3rd Qu.:  49198
 Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664
    PAY_AMT1          PAY_AMT2          PAY_AMT3          PAY_AMT4
 Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0
 1st Qu.:  1000   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296
 Median :  2100   Median :   2009   Median :  1800   Median :  1500
 Mean   :  5664   Mean   :   5921   Mean   :  5226   Mean   :  4826
 3rd Qu.:  5006   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013
 Max.   :873552   Max.   :1684259   Max.   :896040   Max.   :621000
    PAY_AMT5           PAY_AMT6           default
 Min.   :     0.0   Min.   :     0.0   Min.   :0.0000
 1st Qu.:   252.5   1st Qu.:   117.8   1st Qu.:0.0000
 Median :  1500.0   Median :  1500.0   Median :0.0000
 Mean   :  4799.4   Mean   :  5215.5   Mean   :0.2212
 3rd Qu.:  4031.5   3rd Qu.:  4000.0   3rd Qu.:0.0000
 Max.   :426529.0   Max.   :528666.0   Max.   :1.0000
```

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

The data was analyzed using the three classification methods logistic regression, discriminant analysis, Naïve Bayes techniques and the confusion matrices were computed. A confusion matrix is a table that is used to describe the performance of a classification model on a data set for which the true values are known. The results are provided below:

The confusion matrix provided by Bayes' classification:

Table-1: Confusion Matrix and Statistics

```
          Reference

Prediction     0     1

         0  4911   681

         1 18453  5955

Accuracy : 0.3622

95% CI : (0.3568, 0.3677)
```

(ii) The confusion matrix provided by linear discriminating function:

Table-2: Confusion Matrix and Statistics

```
          Reference

Prediction     0     1

         0 23360  6636

         1     4     0

Accuracy : 0.7787

95% CI : (0.7739, 0.7834)
```

(iii) The confusion matrix provided by logistic regression function:

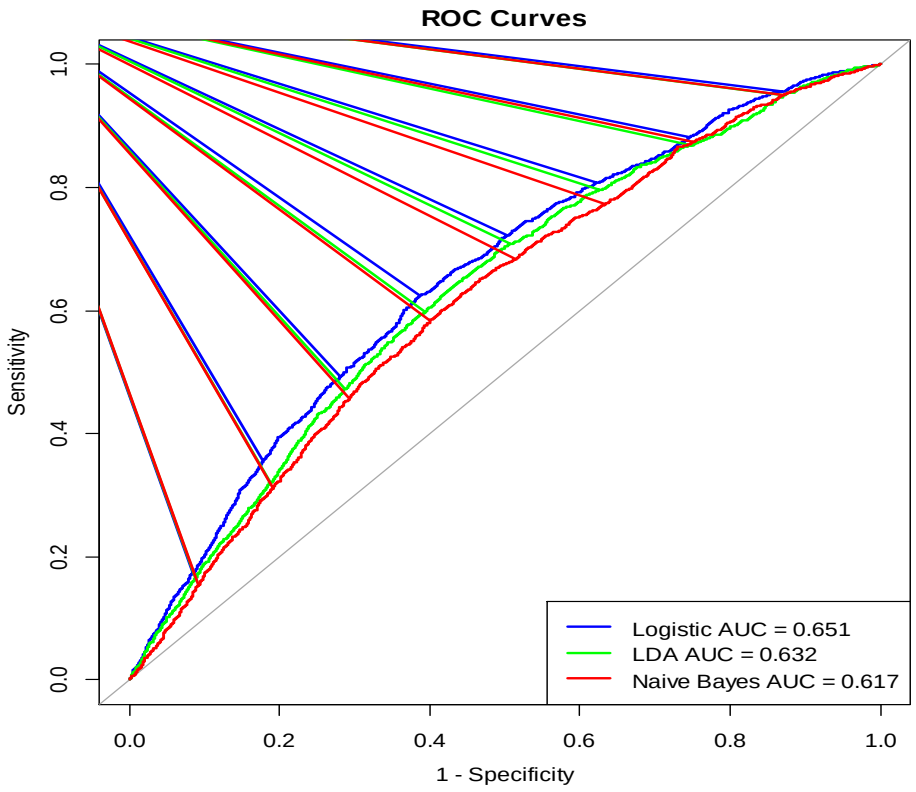
Table-3: Confusion Matrix and Statistics

```
          Reference

Prediction     0     1

         0 23363  6636

         1     1     0

Accuracy : 0.7788

95% CI : (0.774, 0.7835)
```

Results in the above three tables' shows that the best accuracy for the predicted values is provided by logistic regression. Logistic regression model had an accuracy level 77.88%, the second highest level of accuracy is for linear discriminant function with accuracy level 77.87%. Surprisingly, Bayes' classification method performed (36.22%) poorly for this data.

ROC curves for the above analysis is shown below.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025



Assumptions check

Logistic Regression

One of the important assumptions of the logistic regression is that, there is no multicollinearity in the data. Let me check this assumption through the VIF values.

BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6
32.638	51.477	36.345	33.012	35.975	20.986

The VIF values of the above three variables are extremely high which might be due to the multicollinearity

Linear Discriminant Analysis (LDA)

We assume the multivariate normality for the predictor variables and can be checked through Shapiro test for normality. The computed value of the test statistic and the p value are reported below

data: x, sample size 1000, dimension 14, replicates 100

E-statistic = 279.45, p-value < 2.2e-16

A test for multivariate normality was rejected by a sample data from this data set. Samples data was used because of the very large size of the data. This means that, the assumptions of LDA also might be violated for this data set

Naïve Bayes

For Naïve Bayes classification, we assume the conditional independence of predictors. This was tested through the correlations among the predictors. Because of the large size of the matrix, entire results are not reproduced here. But the correlation matrix shows that the BILL_AMT1 is highly correlated with BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5 and BILL_AMT6. Hence, there is a serve assumption violation in the case of Naïve Bayes classification

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

CONCLUSION

While logistic regression model had the highest accuracy level based on confusion matrices, Discriminant function has the second highest accuracy. The better performance of logistic regression model was expected because of the weak set of assumptions required. Even though the assumption of no multicollinearity is not fully satisfied, logistic regression is more robust to assumption violations. Bayes' classification performance was poor compared to the other two methods. The reason might be the huge asymmetry in the number of observations in the two categories. Another reason for the poor performance of the Bayes' classification method was that, it assumes independence among predictors, but not satisfied for such a financial data set. The reason for the poor performance of the linear discriminant analysis is that, it assumes multivariate normality and equal class covariances of the data, but unfortunately for this data set, these two assumptions are not well suited. Since this study is based on a particular data set, the study is a limited one, and a general conclusion cannot be arrived.

REFERENCE

1. Asparoukhov, O. K., Krzanowski, W. J. (2001). A comparison of discriminant procedures for binary variables. Comput. Stat. Data Anal. 38, 139–160.
2. Harris, R. J. (2001). A Primer of Multivariate Statistics, 3rd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
3. Harrell, Frank E. (2001). Regression Modeling Strategies (2nd ed.). Springer-Verlag.
4. Hastie, Trevor. (2001). The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations.
5. Tibshirani, Robert., Friedman, J. H. (Jerome H.). New York: Springer.
6. Huberty, C. J., Wisenbaker J. W., and J. C. Smith (1987). Assessing Predictive Accuracy in Discriminant Analysis. Multivariate Behavioral Research 22.
7. Huberty, C. J. and Olejnik, S. (2006). Applied MANOVA and Discriminant Analysis, Second Edition. Hoboken, New Jersey: John Wiley and Sons, Inc.
8. Johnson, N., and D. Wichern (2002). Applied Multivariate Statistical Analysis, 5th ed. Upper Saddle River, NJ: Prentice Hall.
9. Menard, Scott W. (2002). Applied Logistic Regression (2nd ed.). SAGE
10. Webb, G. I.; Boughton, J.; Wang, Z. (2005). "Not So Naive Bayes: Aggregating One-Dependence Estimators". Machine Learning. 58
11. Yeh, I.C.,& Lien, C.H. (2009). The comparison of data mining techniques for the predictive accuracy of the probability of default of credit card clients. Expert systems with Applications 36(2) 2473-2480.