Comparison of different Classification Models on Given Dataset

(Group No. 18)
Vinay kumar Sharma(191171)
Rajat Agarwal (191104)

November 2020

1 Acknowledgement

We would like here to express our heartfelt gratitude to Dr. Shankar Prawesh for helping us in every difficulty which we face during this project. It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course : IME 692, Advanced Statistical Methods for Business Analytics.

We also take this opportunity to thank the authors and publishers of the various books ,journals and blogs we have consulted. Without those this task would not have been completed.

We would also like to thank our parents for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time-period.

Contents

2 Introduction

Here we are comparing the performance various classification model on a particular dataset with 200 observation in the training dataset and 1000 observation in testing dataset . Each observation has 20 independent variable and a binary response or dependent variable . The performance measure of each model is measured in terms of mis-classification error on test dataset .Here we consider the model to best having misclassification error less than 0.2.

3 Objective

Among various classification model we have to choose best model . We have decided to fit following model on data based on the data specification(we have all numerically feature in data set and binary response variable) :-

Logistic Regression model

LDA (Linear Discriminant Analysis)

QDA (Quadratic Discriminant Analysis)

SVM(Support Vector Machine)

KNN Classifier (K-nearest neighbour)

4 Data Description

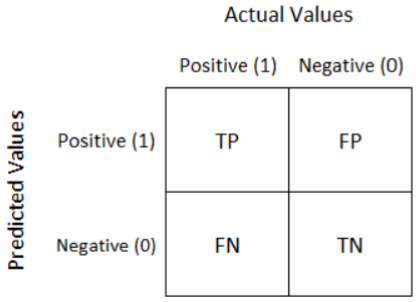
train datasets has 200 observation and each observation is comprises of 20 predictor variable and a binary response variable

test datasets has 1000 observation and each observation is comprises of 20 predictor variable and a binary response variable

5 Performance Metrics :-

5.1 Confusion Matrix :-

A confusion matrix is a matrix that is used to describe the performance (by different performance metrices) of a classification model or a classifier on a given set of data for which the true values are known. Confusion matrix for binary classifier is depicted below :-



This are a lot of rates that are often computed from a confusion matrix for a binary classifier. but we will focus on two main rate here to measure model perfromance :
1. Accuracy = (TP+TN)/total ; total = TP + TN + FP + FN
Overall, how often is the classifier correct!
2. Misclassification Error Rate = (FP+FN)/total = 1 - Accuracy
Overall, how often is it wrong!

5.2 ROC Curve :-

ROC curve is commonly used graph that summarizes the performance of a classifier over all the possible thresholds by using area under it(called AUC). It is generally plotted with TPR(True Positive Rate) on Y-axis against the FPR(False positive rate) on X-axis, as you vary the threshold for assigning observations to a given class.

False Positive Rate(FPR)= FP/(FP + TN)

True Negative Rate(TNR) = TN/(FP + TN)

6 Model Description

6.1 Logistic regression

Logistic regression models the probablity that response(y) belong to a particular category i.e. 0 or 1 given features variable for binary classification and similarily for multi-class classification we have multiple category . The basic assumptions of logistic regression is that the classes are linearly separable . In logistic regression model , we use the logistic function as :-

$$p(X) = \frac{e^{beta^T * X}}{(1 + e^{beta^T * X})}$$

p(X) denotes the probablity that response belongs to a particular category given feature variable
X denote the matrix of features variable
beta denotes the unknown coefficient to be estimated while training the model
we can also write the logistic regression model as:-

$$log(\frac{p(X)}{1-p(X)}) = beta^T * X$$

the left hand side is abbreviated as log-odds or logit thus we can say that logit are linear in X

Fitting of Logistic regression model

In our train dataset, we fitted Logistic regression model
Below down is our confusion matrix on **test data-set.**

$$\text{M}= \begin{pmatrix} 418 & 73 \\ 95 & 414 \end{pmatrix}$$

clearly here TP=418 ,TN=414, FN=95 and FP=73. Therefore **accuracy** = 0.832 and **missclassification error** = 0.168

6.2 LDA

LDA stands for linear discriminant analysis . Basically , LDA is typically example of bayes theorem classifier as here we model the distribution of predictor (X) separately in each of the response class, and then use Bayes theorem tricks to flip these around into estimates for Pr(Y = k|X = x) . Basic assumption of LDA is that feature vector (X) are multivariate gaussian(normal) with class specific mean vector and a common variance-covariance matrix .

Here, the word linear in classifier's name stems from the fact that discriminant functions are linear function's of X . discriminant function is written as :-

$$\delta_k(x) = x^T \sum^{-1} \mu_k - \mu_k^T \sum^{-1} \mu_k + log(\pi_k)$$

$\sum$ denote the common covariance matrix and μ_k denote the class specific mean
π_k denote the probablity that response belong to the k^{th} class

Fitting of Linear Discriminant Analysis model

In our train dataset, we fitted LDA model
Below down is our confusion matrix on **test data-set.**

$$\text{M}= \begin{pmatrix} 419 & 72 \\ 86 & 423 \end{pmatrix}$$

clearly here TP=419 ,TN=423, FN=86 and FP=72. Therefore **accuracy** = 0.842 and **missclassification error** = 0.158

6.3 QDA

QDA stands for Quadratic discriminant analysis .QDA is quite similar to LDA just a liitle relaxation in assumption of covariance matrix is here that in QDA feature vecctor have class specific covariance matrix . here the dicrminant function are quadratic function of X and is written as :-

$$\delta_k(x) = \frac{-1}{2} * (x - \mu_k)^T \sum_k^{-1} (x - \mu_k) + log(\pi_k)$$

$\sum_k$ denote the class specific covariance matrix and μ_k denote the class specific mean
π_k denote the probality that response belong to the k^{th} class

Fitting of Quadratic Discriminant Analysis model
In our train dataset, we fitted QDA model
Below down is our confusion matrix on **test data-set.**

M= $\begin{pmatrix} 359 & 132 \\ 126 & 383 \end{pmatrix}$

clearly here TP=359 ,TN=383, FN=126 and FP=132. Therefore **accuracy** = 0.742 and **missclassification error** = 0.258

6.4 SVM

(we use the reference of shuzhanfan: model-understanding-mathematics-behind-support-vector-machines and link is given in bibliography),

Support vector machines (SVM) are powerful and flexible supervised machine learning algorithm which can used for both classification and regression problems. But generally, they are used in classification problems.

Working of SVM
SVM model is basically a representation of different classes in a hyper plane in multidimensional space. The hyper plane will be generated in an iterative manner by SVM so that the error can be minimized. The Aim of SVM is to divide the data points into classes to find a maximum marginal hyper plane.

The followings are some concepts in SVM
Support Vectors: Data points that are closest to the hyper plane is called support vectors.
Hyper plane : It is a decision plane or space which is divided between a set of objects having different classes.
Margin: It is defined as the gap between two lines on the nearest data points of different classes and can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as good margin and small margin is considered as bad margin so our task is to maximise the margin

The main task of SVM is to divide the data points into classes to find a maximum marginal hyper plane and it can be done in the following two steps
(i) SVM will generate hyper planes iteratively that separates the classes in best way.
(ii) It will choose the hyperplane that separates the classes correctly.

SVM Kernels SVM algorithm are implemented with kernel that transforms an input data space into the required forms. SVM uses a algorthmic technique called the kernel trick, In which kernel undertake a low dimensional input space and transforms it into a higher dimensional space or In simple words, kernel converts non-separable points into separable points by adding more dimensions to it. That makes SVM more powerful, flexible and accurate. The following are some types of kernels that we used in our data.

Linear Kernel It is used as a dot product between any two observations. The formula of linear kernel :

$$K(x_i, x_j) = x_i \cdot x_j$$

In practice, linear kernel works well for text classification problems.

Polynomial Kernel It is more generalized form of the linear kernel and distinguish curved or nonlinear input space. Following below is the formula for polynomial kernel

$$K(x_i, x_j) = (x_i \cdot x_j + c)^d$$

This kernel contains two parameters: a constant 'c' and a degree 'd'. A value with d=1 is just the linear kernel and generally its represent the degree of polynomial.

Radial Basis Function (rbf) Kernel RBF kernel, mostly used in SVM classification, maps input space in indefinite dimensional space. Following below formula explains it mathematically

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$$

The rbf (Radial Basis Function) kernel is also called the Gaussian(normal) kernel. It will results in a more complex decision boundary. The rbf kernel contains a parameter γ. A small value of γ will make the model behave as like a linear SVM. A large value of γ will make the model heavily impacted by the given support vectors.

In practice, it is recommended to try rbf kernel first because it normally performs well.

Fitting of SVM in our data set

SVM with linear kernel

In our dataset, we fit SVM with linear kernel, Below down is our confusion matrix on **test data-set.**

$$\text{M}= \begin{pmatrix} 426 & 65 \\ 103 & 406 \end{pmatrix}$$

clearly here TP=426 ,TN=406, FN=103 and FP=65. Therefore **accuracy** = 0.832 and **missclassification error** = 0.168

SVM with polynomial kernel

In our dataset, we fit SVM with polynomial kernel, Below down is our confusion matrix on **test data-set.**

$$\text{M}= \begin{pmatrix} 370 & 121 \\ 119 & 390 \end{pmatrix}$$

clearly here TP=370 ,TN=390, FN=119 and FP=121. Therefore **accuracy** = 0.76 and **missclassification error** = 0.24

SVM with rbf kernel

In our dataset, we fit SVM with polynomial kernel, Below down is our confusion matrix on **test data-set.**

$$\text{M}= \begin{pmatrix} 409 & 82 \\ 101 & 408 \end{pmatrix}$$

clearly here TP=409 ,TN=408, FN=101 and FP=82. Therefore **accuracy** = 0.817 and **missclassification error** = 0.183

6.5 KNN

K-Nearest Neighbour(K-NN) is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new data points and available data points, and put the new case into the category that is most similar to the categories which are available. It is commonly used for Regression as well as for Classification problems but mostly it is used for the Classification problems.

Working Algorithm of K-NN

Step-1: Select number K , which represent the number of neighbors.(here k is selecting using hit and trail to achieve maximum accuracy)
Step-2: Calculating the Euclidean distance of K number of neighbors.
Step-3: Take K nearest neighbors as per calculated Euclidean distance.
Step-4: Among all these k neighbors, count the number of the points in each category.
Step-5: Assign the new points(data) to that category for which the number of the neighbor is maximum.
Step-6: Now our model is ready for predictions.

How to select the parameter K ?

Actually there is no particular well defined way is available to select K. So we try some values(generally hit and trail) to find the best out of them also we can use loop to check for K which gives minimum missclassification error.

Advantages of KNN Algorithm

(i) K-NN is a non-parametric algorithm, that means it does not make any assumption on underlying data.
(ii) It is more effective if the training data is large.
(iii) It is simple to implement.

Disadvantages of KNN Algorithm:

(i) Always needs to determine the value of K which may be complex some time.
(ii) The computation cost is high because of calculating the distance between the data points for all the training samples.

In our dataset, we fit K-NN model with 30 neighbors (K=30) by using some hit and trail and observe that at this value of K our model gives some satisfactory accuracy.
Below down is our confusion matrix on **test data-set.**

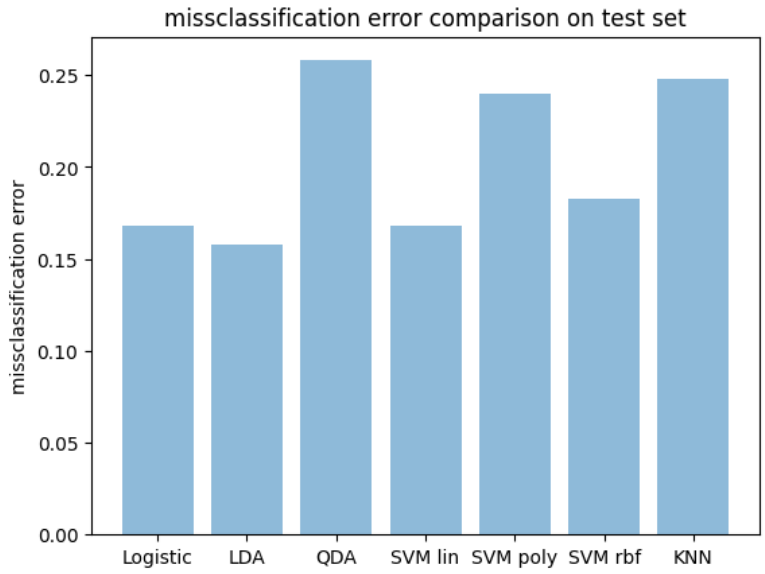
$$\text{M}= \begin{pmatrix} 369 & 122 \\ 126 & 383 \end{pmatrix}$$

clearly here TP=369 ,TN=383, FN=126 and FP=122. Therefore **accuracy** = 0.752 and **mis-classification error** = 0.248

7 Comparision of different models

The test error or mis-classification error on test data using various model is depicted below :-

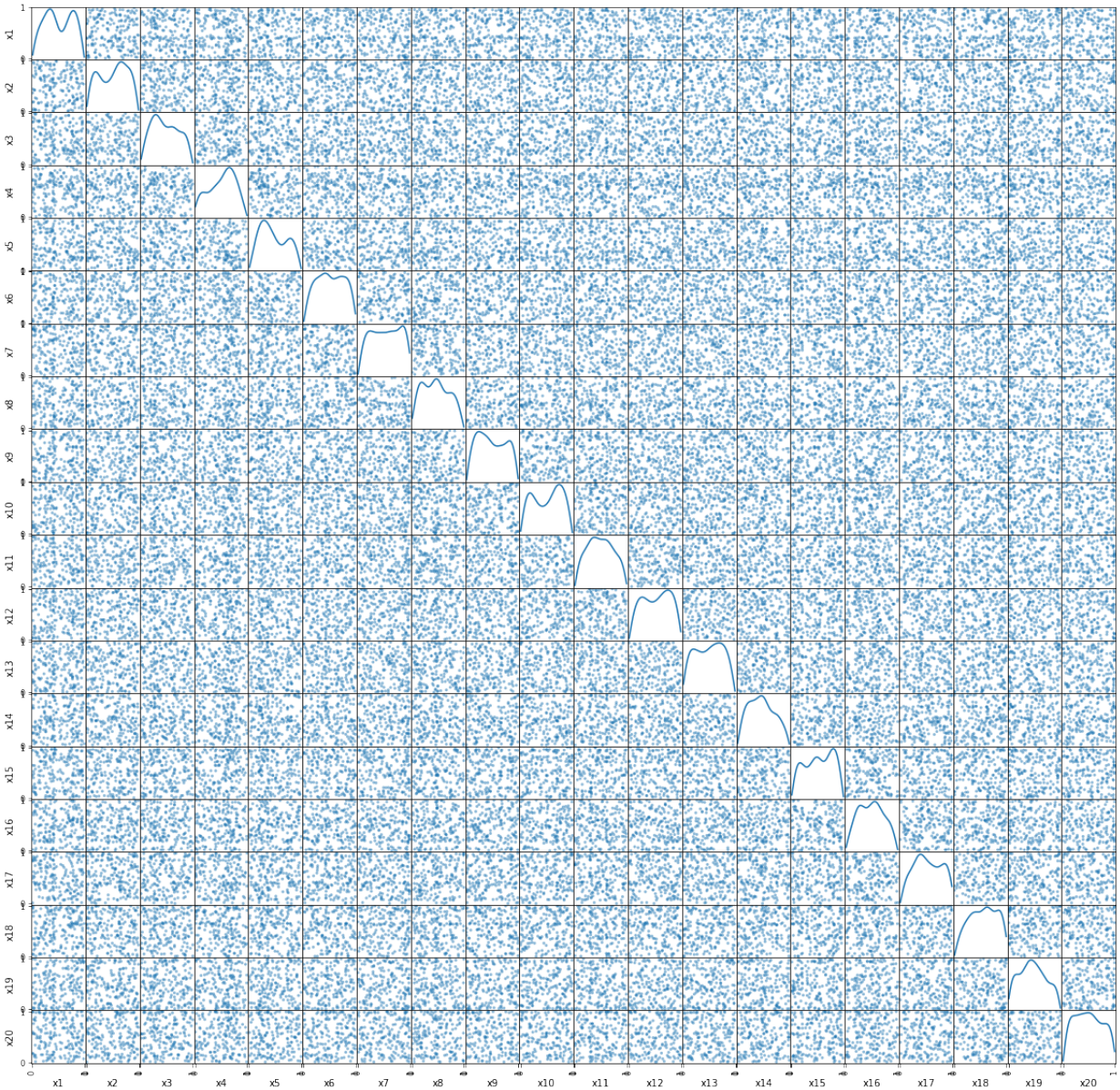
models	Accuracy(test set)	misclassification error (test set)
Logistic	0.832	0.16800000000000004
LDA	0.842	0.15800000000000003
QDA	0.742	0.258
SVM lin	0.832	0.16800000000000004
SVM poly	0.76	0.24
SVM rbf	0.817	0.18300000000000005
KNN	0.752	0.248



It is clearly visible from the above bar plot that the LDA classifier has lowest misclassification error rate on test dataset .

8 Explanation for the best model and ROC curve

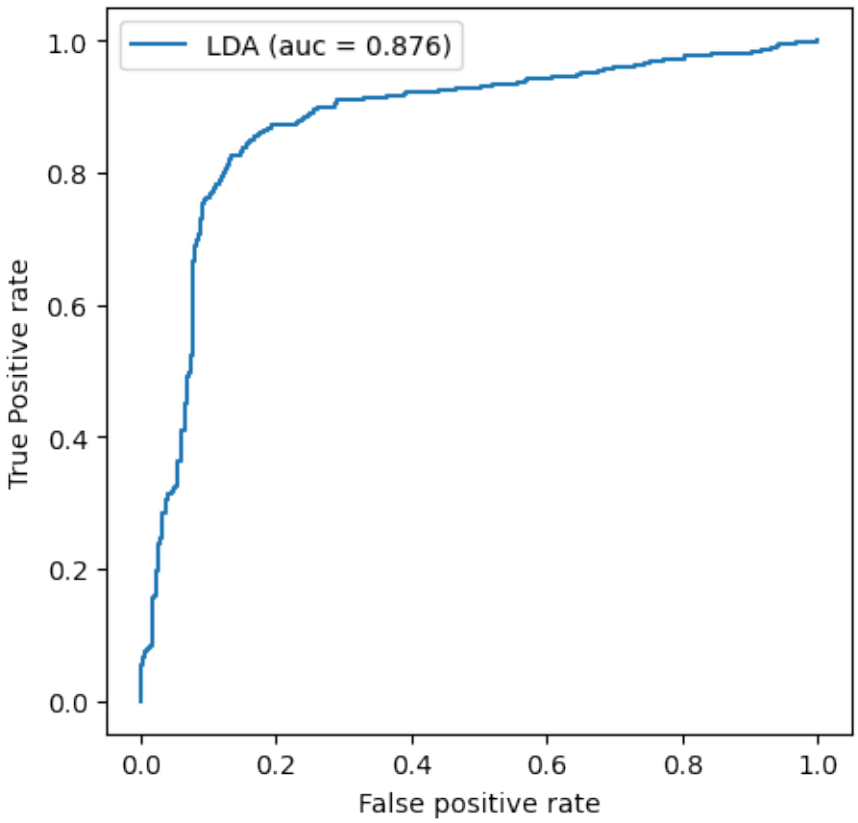
We found after comparing the misclassification error on test data that LDA model has lowest misclassification error on test data comapred to all other fitted model . Thus , LDA is best model for the classification problem. The reason for LDA model to perform best among all the model we have fitted so far is due to fact that the training data satisfies the basic assumption LDA model i.e. the assumption of gaussian feature and common covariance matrix . the below scatter plot helps us to visualize that asssumptions of LDA holds true for the given dataset.


x1
x2
x3
x4
x5
x6
x7
x8
x9
x10
x11
x12
x13
x14
x15
x16
x17
x18
x19
x20

The graph shows that density plot(diagonal) of all the 20 feature is showing approximately normal behaviour and all feature are independent from each other is clear from scatter plot (off diagonal) .

ROC Curve for LDA model

AUC(Area under curve) captures the extent to which the curve is up in the Northwest corner. An higher AUC is good. we have got ROC curve with area under cuve 0.876 indicating that the model is good enough for predicting our response with good accuracy. ROC curve for LDA model is plotted below :-



9 Conclusion

(*) Among the all classification models we fit in our data, it is clearly seen that the misclassification error on test data is lowest in case of LDA (linear discriminant analysis) that is 0.1580 (below 0.20 that means a good model).
(**) Here LDA is best among all the given models because our training data satisfy the basic assumption of LDA that is data has gaussian feature and common covariance.
(***) In LDA the value of AUC (area under ROC) is 0.876, that is close to 1 means the model better in predicting 0s as 0s and 1s as 1s.

Taking all the results we conclude that for our data-set, Linear discriminant analysis (LDA) classification model is best among all classification models we perform.

10 Bibliography

1. An Introduction to Statistical Learning,with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
2. The Elements of Statistical Learning:Data Mining, Inference, and Prediction by Trevor Hastie,Robert Tibshirani,Jerome Friedman
3. www.analyticsvidhya.com
4. https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/
5. https://shuzhanfan.github.io/2018/02/model-evaluation-metrics/

11 Contrbutions

contribution of group members:-
(Rajat agarwal) :-
a.) Written python code for Logistic model , LDA model , QDA model ,checking missing observation in data ,
b.) written model description of logistic model , LDA model , QDA model
(Vinay Kumar sharma):-
a.) Written python code for SVM model with kernel's as linear , poly and radial and KNN model , plotted histogram for misclassification error rate .
b.) Writen model description of SVM (kernel :- linear , poly , radial) and KNN model .
Combined work :- Written combinedly the framework of project report and Identified the reason for LDA to be best performing model . Written conclusion of our analysis.