A Comparative Study of Classification Algorithms on Multivariate Data

(Project of MTH514A)

Submitted by,

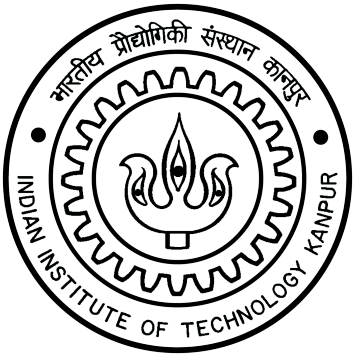
Rajdeep Saha (201380)
Sagnik Dey (201397)
Saumyadip Bhowmick (201408)
Shuvam Gupta (201421)
Soumik Karmakar (201428)

Supervised by,

Dr. Minerva Mukhopadhyay

Submitted on,

21^{st} April, 2022



Contents

1 Introduction **3**

2 Different Classification Algorithms **3**
2.1 Logistic Regression . . . 3
2.2 Linear Discriminant Analysis (LDA) . . . 3
2.3 Quadratic Discriminant Analysis (QDA) . . . 4
2.4 Decision Tree . . . 5
2.5 Random Forest . . . 5

3 Application **7**
3.1 PIMA Indian Diabetes Dataset . . . 7
3.1.1 Logistic Regression . . . 8
3.1.2 Decision Tree . . . 9
3.1.3 Random Forest . . . 9
3.1.4 Evaluating Performances of the Models . . . 10
3.1.5 Final Performance . . . 11
3.2 Caravan Insurance Dataset . . . 13
3.2.1 Logistic Regression . . . 19
3.2.2 Decision Tree . . . 20
3.2.3 Random Forest . . . 20
3.2.4 Evaluating Performances of the Models . . . 21
3.2.5 Final Performance . . . 22
3.3 Banknote Authentication Dataset . . . 24
3.3.1 Logistic Regression . . . 25
3.3.2 Decision Tree . . . 26
3.3.3 Random Forest . . . 26
3.3.4 Evaluating Performances of the Models . . . 27
3.3.5 Final Performance . . . 28
3.4 Weekly Dataset . . . 30
3.4.1 LDA . . . 31
3.4.2 QDA . . . 32
3.4.3 Logistic Regression . . . 32
3.4.4 Decision Tree . . . 33
3.4.5 Random Forest . . . 33
3.4.6 Evaluating Performances of the Models . . . 34
3.4.7 Final Performance . . . 37

4 Conclusion **38**

5 Appendix **38**
5.1 R Codes . . . 38
5.2 Synthetic Minority Oversampling Technique (SMOTE) . . . 38
5.3 F-Score . . . 39
5.4 ROC-AUC Curve . . . 39

6 References **39**

1 Introduction

In real life scenarios, the Classification problem arises more frequently than the basic regression one.Suppose we have to classify if patient is diabetic or not , a patient will or will not respond to the drug , a tumour indicates breast cancer or not based on the dataset.Now, there are several classfiers available for our use but not all of them will provide good fit and thus accurate prediction to the datasets. In our project, we are going to study the accuracy of several classifiers like Logistic Regression ,Linear Discriminant Analysis,Quadratic Discriminant Analysis,Decision , and Random Forest on 4 separate datasets i.e. PIMA Indian Diabetes Dataset, Caravan Insurance Dataset, Banknote Authentication Dataset, Weekly Dataset .Then we will provide a comparative study based on our results for all 4 classifiers on the 4 datasets.

2 Different Classification Algorithms

2.1 Logistic Regression

We here use Logistic Regression for binary classification.Here I use the sigmoid function to predict the class for a data point. For a model with k parameters and if I have to classify datapoints into two classes classes viz 0 and 1 . We denote p = P(Y=1|X=$\mathbf{x}$) where Y is predictor variable ,We can write

$$p = \frac{1}{1 + e^{-\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}$$

Now, Y is predicted to belong to class 1 if p $\geq$ 0.5 or predicted to belong to class 0 otherwise.

2.2 Linear Discriminant Analysis (LDA)

We now the concept of the LDA classifier in the case of multiple predictors. Here, we will assume that $X = (X_1, X_2, ..., X_p)$ is drawn from a multivariate Gaussian distribution, with a specific multivariate mean vector and a common covariance matrix.The multivariate Gaussian density is defined as

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))}$$

Also, it has no hyperparameters which require tuning. LDA can be derived from simple probabilistic models which model the class conditional distribution of the data $P(X \mid y = k)$ for each class k. Predictions can then be obtained by using Bayes' theorem, for each training sample $x \in R^d$

$$
\begin{aligned}
P(y = k \mid x) &= \frac{P(x \mid y = k)P(y = k)}{P(x)} \\
&= \frac{P(x \mid y = k)P(y = k)}{\sum_l P(x \mid y = l)P(y = l)} \\
&= \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_l f_l(x)}
\end{aligned}
$$

In the case of$p > 1$ predictors, the LDA classifier assumes that theobservations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \Sigma)$, whereμ_k is a class-specific mean vector, and Σ is a covariance matrix that is common to all K classes. Plugging the density function for the kth class, $f_k(X = x)$, into Baye's Classifier and performing a little bit of algebra reveals that the Bayes classifier assigns an observation X = x to the class for which

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

is the largest.

2.3 Quadratic Discriminant Analysis (QDA)

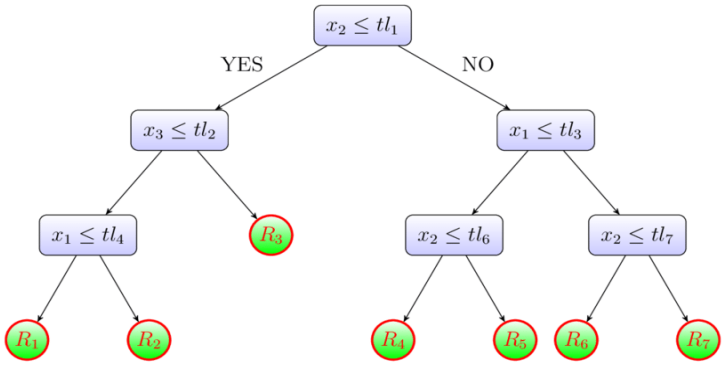
As we have seen, LDA has the assumption that the observations within each class are being drawn from a multivariate Normal distribution with a class specific mean vector and a covariance matrix that is common to all the classes. Quadratic discriminant analysis (QDA) provides an alternative approach. QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' rule in order to perform prediction.. However, unlike LDA, QDA assumes that each class has its own covariance matrix.That is, it assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the kth class. Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

$$
\begin{aligned}
\delta_k &= -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k \\
&= -\frac{1}{2} x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)
\end{aligned}
$$

is largest.The QDA classifier involves plugging estimates for $\Sigma_k, \mu_k, and \pi_k$ into (the above equation, then assigning an observation $X = x$ to the class for which this quantity is largest.

2.4 Decision Tree

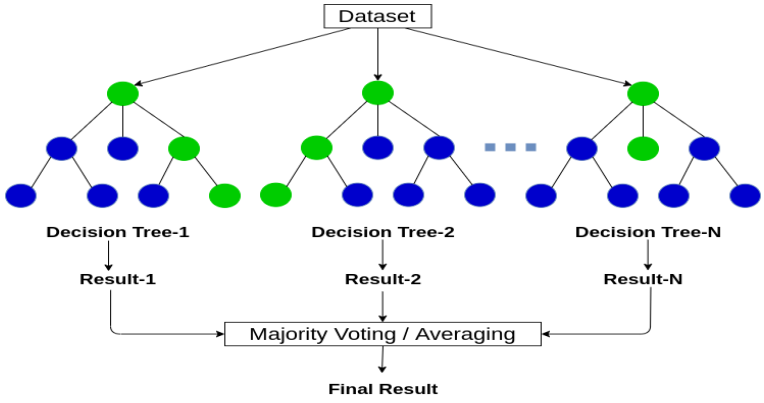
A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.



It is like an example of a multistage decision process.Rather than using the complete set of features jointly to make a output decision, different subsets of features are used at different levels of the tree.

2.5 Random Forest

A random forest algorithm consists of many decision trees. A random forest is a machine learning technique that's used to solve regression and classification problems.It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.In other words, in building a random forest, at each split in the tree,the algorithm is not even allowed to consider a majority of the available predictors.



Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees.

Decision trees suffer from high variance problem. Toget rid of this high variance problem, we use random forest. In this method, we use the concept of **"Averaging independent observations reduce variance"**. In random forest, we take a bootstrapped sample of size N from the training data and grow a decision tree with these N samples. The only difference is that, in this case, in each split, we take a random sample of total number of predictors, preferable sample size is the square root of the number of predictors. We repeat this process, say, M number of times.

For classification setting, we note down the class predicted by each of the 'M' trees and for final prediction, we take the majority of these 'M' outputs.

The advantage of Random Forest lies in the fact that it **decorrelates** the trees. Suppose there is a strong predictor in our dataset. Then there will be a tendency to use this predictor in each top split. So the trees may look similar. In Random Forest, we are using a random sample of predictors and in some cases, the most important variables may even not be considered. We can think this as decorrelating the trees, thereby making predictions less variable.

3 Application

3.1 PIMA Indian Diabetes Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- Pregnancies: Number of times pregnant.
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Here our goal is to predict whether or not a patient has diabetes. First we have checked for **missing values** in the dataset since missing values can cause problems in modeling purpose. In this dataset, there are no missing values. So, we proceed to check whether our dataset is **imbalanced** or not, since imbalanced dataset may often results in a erroneous model. We have seen in our dataset 65% observations have no diabetes and 35% observations have diabetes. So, we have implemented **Synthetic Minority Oversampling Technique(SMOTE)** to make our dataset balanced.

After all the preprocessing with our data, we have split our dataset into 3 parts, namely, **train set, cross validation set, test set**. We have trained logistic regression, decision tree, random forest using train set and checked whether our dataset validates the assumption of lda and qda.

After training the models with our training set, we have applied them on cross validation set and to evaluate the performances of the model, we have looked for **F-Score** and **AUC** of **ROC** curve. We have chosen that model to be the best for which we have obtained the highest F-Score and the highest AUC.

After obtaining the best model, we have applied it on a completely new test dataset and checked for how it performs on test set and for evaluating the performance of the chosen model, we have taken the same evaluation metric F-Score and AUC.

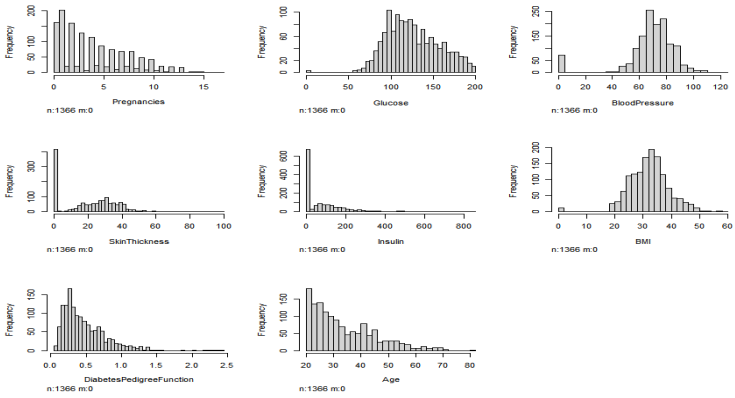


Figure 1: Histogram of Predictors

It is evident from the above figure that the assumption of LDA and QDA has been violated and for this reason, we cannot apply Discriminant Analysis method for this dataset. So, we proceed for applying all other methods.

3.1.1 Logistic Regression

After training logistic regression using training dataset, we have applied it on the cross validation dataset and obtained the following confusion matrix,

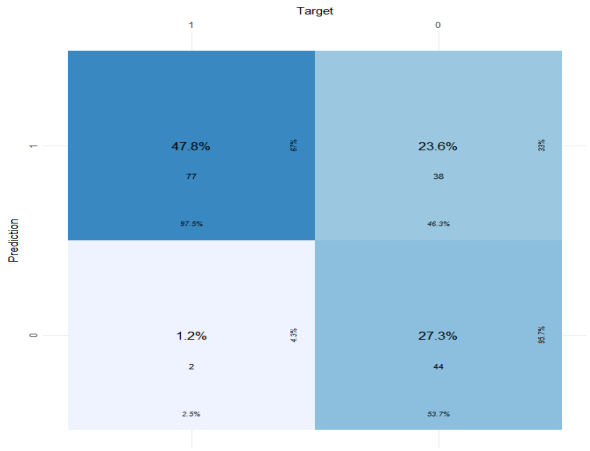


Figure 2: Confusion Matrix of Logistic Regression

3.1.2 Decision Tree

After training decision tree model using training set, we have pruned the tree and obtained the **minimum misclassification error rate** for the tree with 8 terminal nodes. we have applied it on the cross validation dataset and obtained the following confusion matrix,

Figure 3: Confusion Matrix of Decision Tree

3.1.3 Random Forest

We have 8 predictors in our dataset and to train random forest model, we have chosen $\sqrt{8} \approx 3$ predictors in each split. After training Random Forest Model using training set, we have applied it on cross validation set and obtained the following confusion matrix,

Figure 4: Confusion Matrix of Random Forest

3.1.4 Evaluating Performances of the Models

To evaluate and compare the performances of our models, we have used the evaluation metric **F-Score** and **AUC of the ROC Curve**. The AUC curves are as follows,

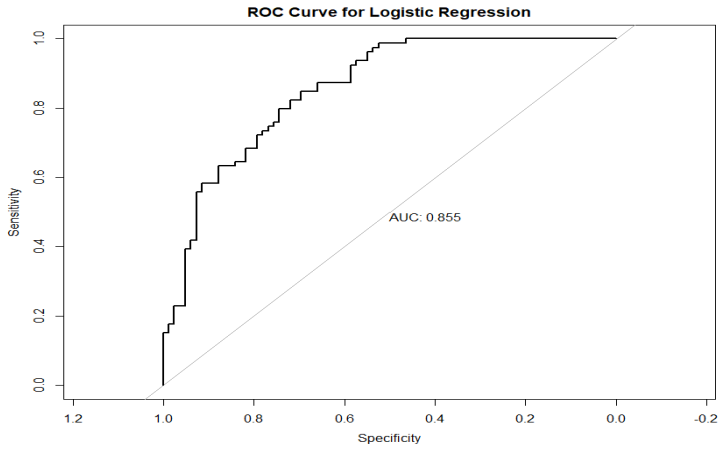


Figure 5: ROC Curve for Logistic Regression

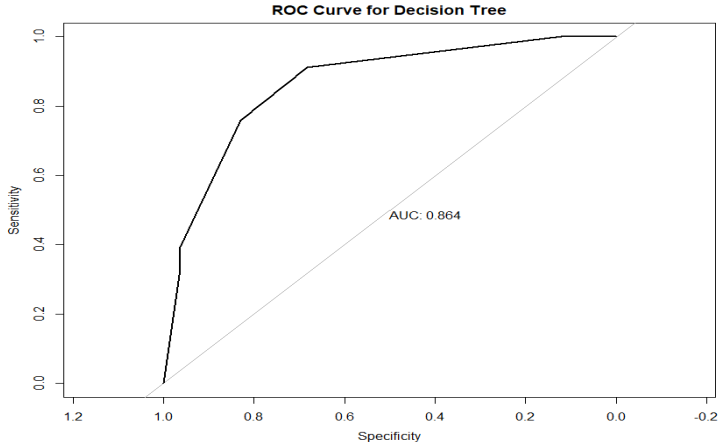


Figure 6: ROC Curve for Decision Tree

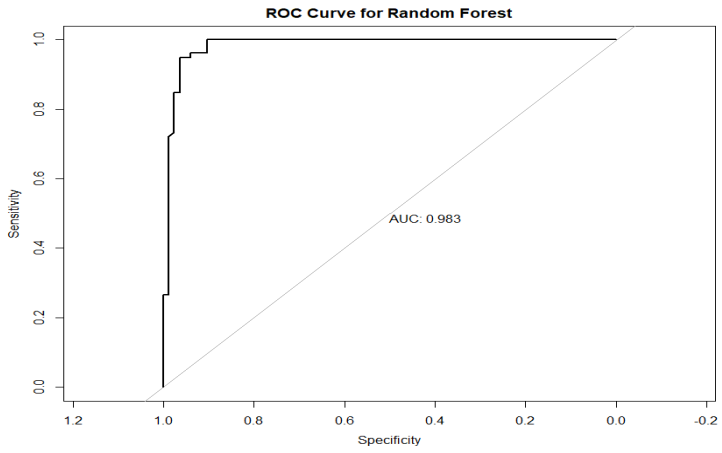


Figure 7: ROC Curve for Random Forest

From the graphs, it is evident that **Random Forest outperforms all the other models in that case**. After that, we have also compared F-Score and AUC to choose the final model.

	F-Score	AUC
Logistic Regression	0.79	0.86
Decison Tree	0.78	0.86
Random Forest	0.94	0.98

Table 1: F-Score and AUC for Various Models

From the above table, also, we can see that **Random Forest has the highest F-Score and AUC**. So, we have chosen random forest model to predict if a patient has diabetes or not.

3.1.5 Final Performance

After having chosen Random Forest to be the best model, we have applied it on test set and obtained the confusion matrix to visualize the correct classification. After that, to assess the performance of the model on this completely new dataset, we have obtained F-Score and AUC respectively.

Figure 8: Confusion Matrix of Final Model

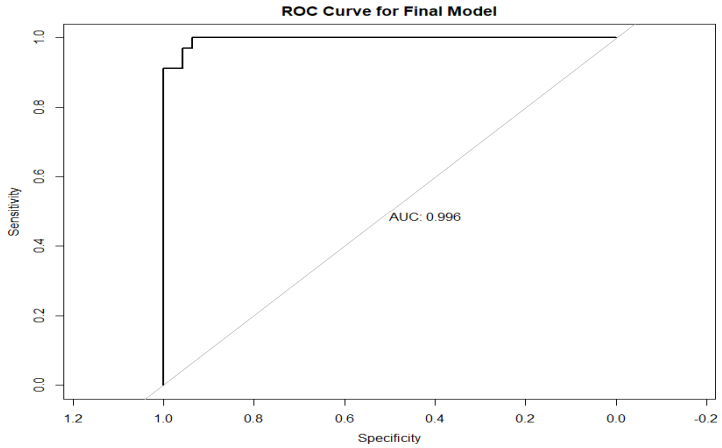


Figure 9: ROC Curve for Final Model

For the final model, we have obtained **F-Score of 0.96** and **AUC of 0.99**. So, we can conclude **it is satisfactory to use random forest model to predict whether a patient has diabetes or not.**

3.2 Caravan Insurance Dataset

This dataset is owned and supplied by the Dutch datamining company Sentient Machine Research, and is based on real world business data. This data set includes 85 predictors that measure demographic characteristics for 5,822 individuals. The response variable is Purchase, which indicates whether or not a given individual purchases a caravan insurance policy. In this data set, only 6 % of people purchased caravan insurance. The variables are as follows,-

- MOSTYPE: Customer Subtype
- MAANTHUI: Number of houses 1 - 10
- MGEMOMV: Avg size household 1 - 6
- MGEMLEEF: Avg age; see L1
- MOSHOOFD: Customer main type; see L2
- MGODRK: Roman catholic
- MGODPR: Protestant ...
- MGODOV: Other religion
- MGODGE: No religion
- MRELGE: Married
- MRELSA: Living together
- MRELOV: Other relation
- MFALLEEN: Singles
- MFGEKIND: Household without children
- MFWEKIND: Household with children
- MOPLHOOG: High level education
- MOPLMIDD: Medium level education
- MOPLLAAG: Lower level education
- MBERHOOG: High status
- MBERZELF: Entrepreneur
- MBERBOER: Farmer
- MBERMIDD: Middle management
- MBERARBG: Skilled labourers

- MBERARBO: Unskilled labourers
- MSKA: Social class A
- MSKB1: Social class B1
- MSKB2: Social class B2
- MSKC: Social class C
- MSKD: Social class D
- MHHUUR: Rented house
- MHKOOP: Home owners
- MAUT1: 1 car
- MAUT2: 2 cars
- MAUT0: No car
- MZFONDS: National Health Service
- MZPART: Private health insurance
- MINKM30: Income <30.000
- MINK3045: Income 30-45.000
- MINK4575: Income 45-75.000
- MINK7512: Income 75-122.000
- MINK123M: Income >123.000
- MINKGEM: Average income
- MKOOPKLA: Purchasing power class
- PWAPART: Contribution private third party insurance
- PWABEDR: Contribution third party insurance (firms) ...
- PWALAND: Contribution third party insurane (agriculture)
- PPERSAUT: Contribution car policies
- PBESAUT: Contribution delivery van policies
- PMOTSCO: Contribution motorcycle/scooter policies
- PVRAAUT: Contribution lorry policies
- PAANHANG: Contribution trailer policies

- PTRACTOR: Contribution tractor policies
- PWERKT: Contribution agricultural machines policies
- PBROM: Contribution moped policies
- PLEVEN: Contribution life insurances
- PPERSONG: Contribution private accident insurance policies
- PGEZONG: Contribution family accidents insurance policies
- PWAOREG: Contribution disability insurance policies
- PBRAND: Contribution fire policies
- PZEILPL: Contribution surfboard policies
- PPLEZIER: Contribution boat policies
- PFIETS: Contribution bicycle policies
- PINBOED: Contribution property insurance policies
- PBYSTAND: Contribution social security insurance policies
- AWAPART: Number of private third party insurance 1 - 12
- AWABEDR: Number of third party insurance (firms) ...
- AWALAND: Number of third party insurance (agriculture)
- APERSAUT: Number of car policies
- ABESAUT: Number of delivery van policies
- AMOTSCO: Number of motorcycle/scooter policies
- AVRAAUT: Number of lorry policies
- AAANHANG: Number of trailer policies
- ATRACTOR: Number of tractor policies
- AWERKT: Number of agricultural machines policies
- ABROM: Number of moped policies
- ALEVEN: Number of life insurances
- APERSONG: Number of private accident insurance policies
- AGEZONG: Number of family accidents insurance policies
- AWAOREG: Number of disability insurance policies

- ABRAND: Number of fire policies
- AZEILPL: Number of surfboard policies
- APLEZIER: Number of boat policies
- AFIETS: Number of bicycle policies
- AINBOED: Number of property insurance policies
- ABYSTAND: Number of social security insurance policies
- CARAVAN: Number of mobile home policies 0 - 1

Here our goal is to predict whether or not a given individual purchases Caravan Insurance Policy. First we have checked for **missing values** in the dataset since missing values can cause problems in modeling purpose. In this dataset, there are no missing values. So, we proceed to check whether our dataset is **imbalanced** or not. Since imbalanced dataset may often results in a erroneous model. Since our dataset is imbalanced, we have implemented **Synthetic Minority Oversampling Technique(SMOTE)** to make our dataset balanced.

After all the preprocessing with our data, we have split our dataset into 3 parts, namely, **train set, cross validation set, test set**. We have trained logistic regression, decision tree, random forest using train set and checked whether our dataset validates the assumption of lda and qda.

After training the models with our training set, we have applied them on cross validation set and to evaluate the performances of the model, we have looked for **F-Score** and **AUC** of **ROC** curve. We have chosen that model to be the best for which we have obtained the highest F-Score and the highest AUC.

After obtaining the best model, we have applied it on a completely new test dataset and checked for how it performs on test set and for evaluating the performance of the chosen model, we have taken the same evaluation metric F-Score and AUC.

To see, whether we can apply Discriminant Analysis method to this dataset, we have plotted histogram of the predictor variables. In each of the plots, we have taken 17 variables and after obtaining the plots, we have checked whether the assumption of normality holds or not.

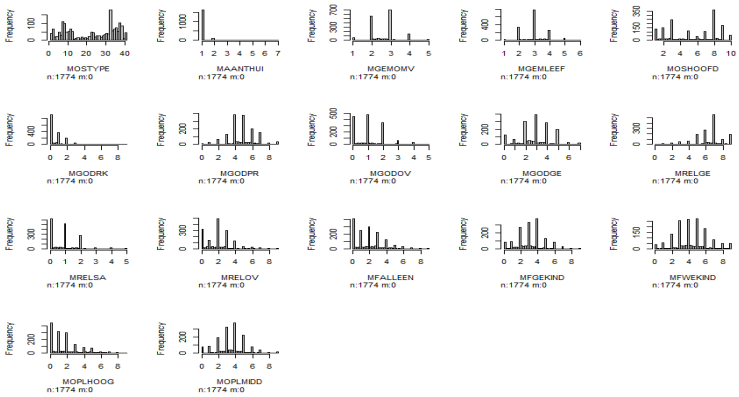


Figure 10: Histogram of Predictors

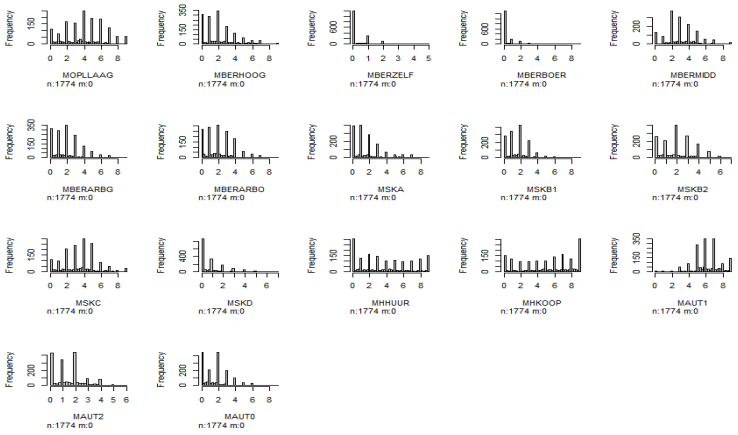


Figure 11: Histogram of Predictors

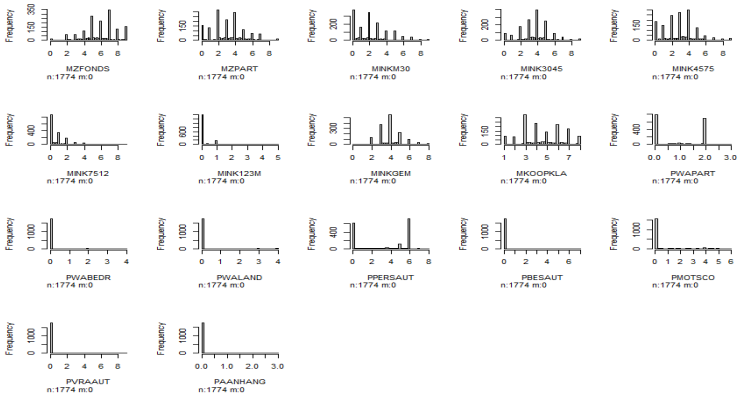


Figure 12: Histogram of Predictors

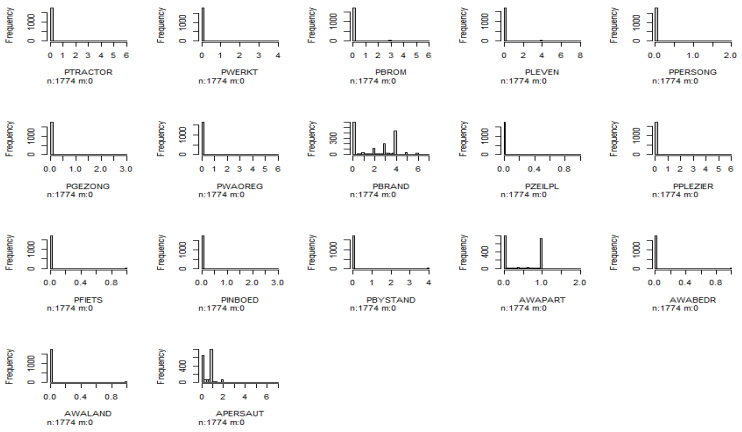


Figure 13: Histogram of Predictors

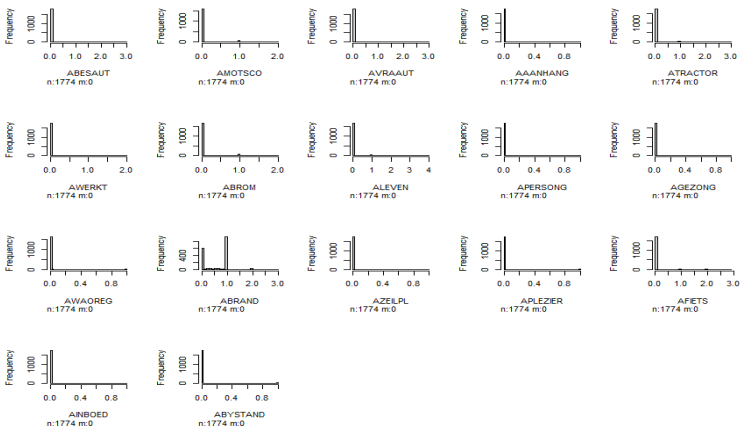


Figure 14: Histogram of Predictors

It is evident from the above figure that the assumption of LDA and QDA has been violated and for this reason, we cannot apply Discriminant Analysis method for this dataset. So, we proceed for applying all other methods.

3.2.1 Logistic Regression

After training logistic regression using training dataset, we have applied it on the cross validation dataset and obtained the following confusion matrix,

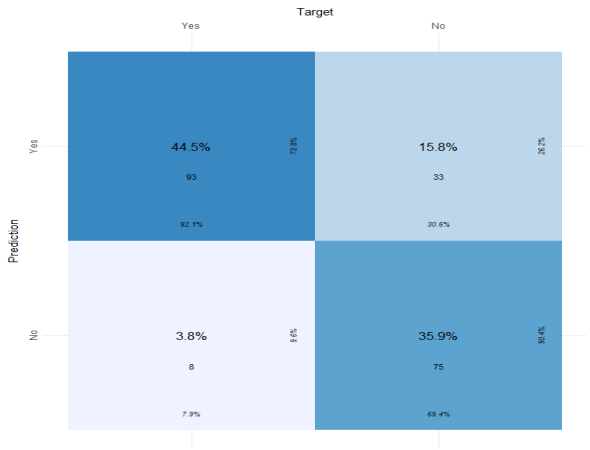


Figure 15: Confusion Matrix of Logistic Regression

3.2.2 Decision Tree

After training decision tree model using training set, we have pruned the tree and obtained the **minimum misclassification error rate** for the tree with 11 terminal nodes. we have applied it on the cross validation dataset and obtained the following confusion matrix,

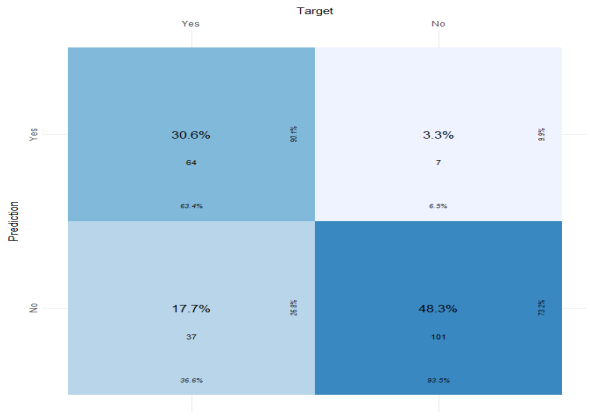


Figure 16: Confusion Matrix of Decision Tree

3.2.3 Random Forest

We have 8 predictors in our dataset and to train random forest model, we have chosen $\sqrt{85} \approx 9$ predictors in each split. After training Random Forest Model using training set, we have applied it on cross validation set and obtained the following confusion matrix,

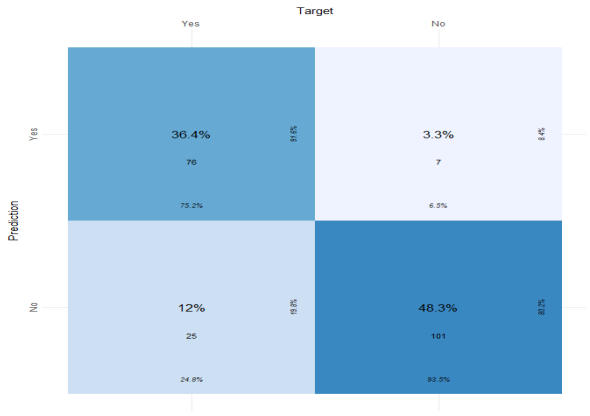


Figure 17: Confusion Matrix of Random Forest

3.2.4 Evaluating Performances of the Models

To evaluate and compare the performances of our models, we have used the evaluation metric **F-Score** and **AUC of the ROC Curve**. The AUC curves are as follows,

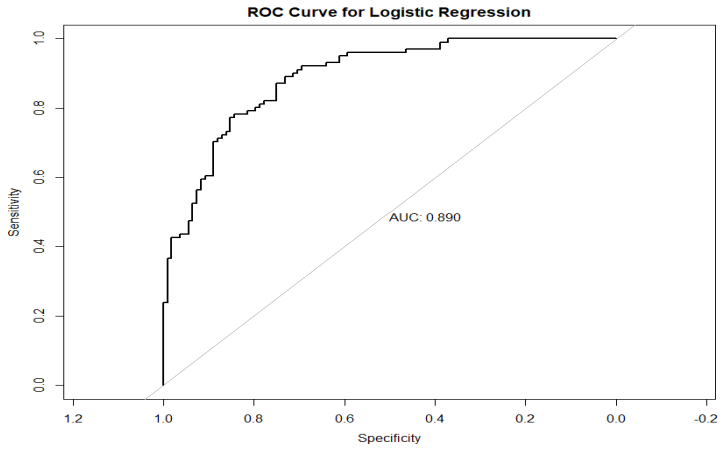


Figure 18: ROC Curve for Logistic Regression

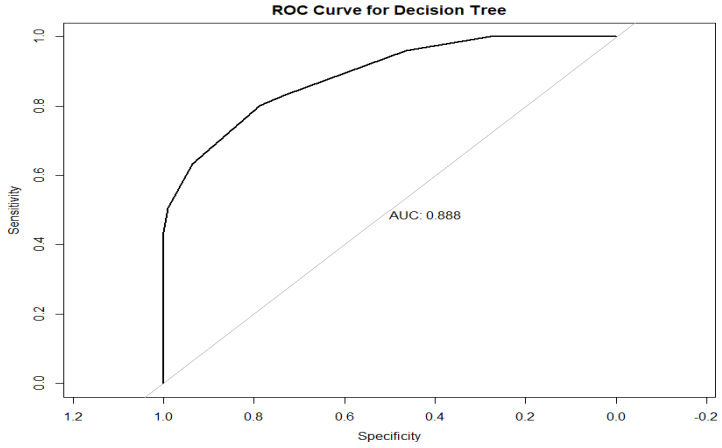


Figure 19: ROC Curve for Decision Tree

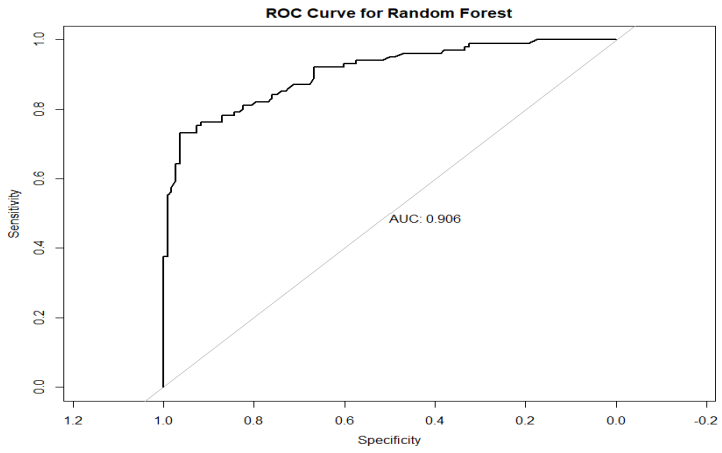


Figure 20: ROC Curve for Random Forest

From the graphs, it is evident that **Random Forest outperforms all the other models in that case**. After that, we have also compared F-Score and AUC to choose the final model.

	F-Score	AUC
Logistic Regression	0.81	0.89
Decison Tree	0.74	0.88
Random Forest	0.81	0.91

Table 2: F-Score and AUC for Various Models

From the above table, also, we can see that **Random Forest has the highest F-Score and AUC**. So, we have chosen random forest model to predict whether a given individual purchases a caravan insurance policy or not.

3.2.5 Final Performance

After having chosen Random Forest to be the best model, we have applied it on test set and obtained the confusion matrix to visualize the correct classification. After that, to assess the performance of the model on this completely new dataset, we have obtained F-Score and AUC respectively.

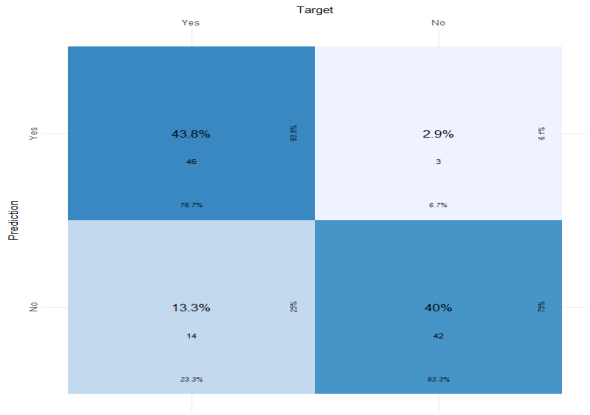


Figure 21: Confusion Matrix of Final Model

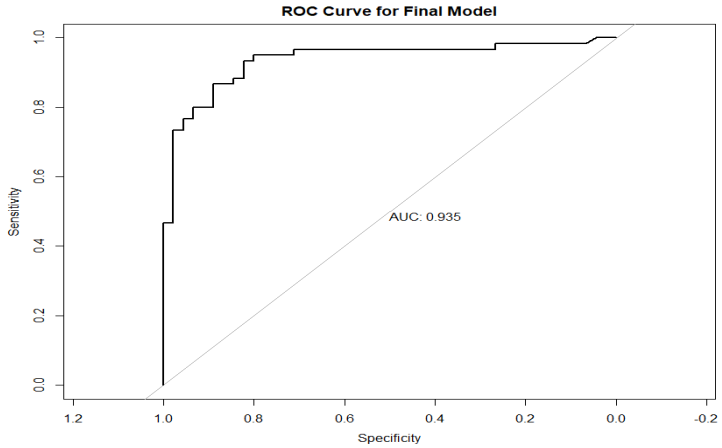


Figure 22: ROC Curve for Final Model

For the final model, we have obtained **F-Score of 0.84** and **AUC of 0.94**. So, we can conclude **it is satisfactory to use random forest model to predict whether a given individual purchases a caravan insurance policy or not.**

3.3 Banknote Authentication Dataset

In this dataset, data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400× 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images. The variables are as follows,-

- variance: variance of Wavelet Transformed image.
- skewness: skewness of Wavelet Transformed image.
- curtosis: kurtosis of Wavelet Transformed image.
- entropy: entropy of image.
- class: 0 represents real and 1 represents fake banknote.

Here our goal is to predict whether a banknote is fake or real. First we have checked for **missing values** in the dataset since missing values can cause problems in modeling purpose. In this dataset, there are no missing values. So, we proceed to check whether our dataset is **imbalanced** or not. In our dataset, we have 56 % real banknote and 44 % fake banknote. So, in this case, our dataset is balanced. After all the preprocessing with our data, we have split our dataset into 3 parts, namely, **train set, cross validation set, test set**. We have trained logistic regression, decision tree, random forest using train set and checked whether our dataset validates the assumption of lda and qda.

After training the models with our training set, we have applied them on cross validation set and to evaluate the performances of the model, we have looked for **F-Score** and **AUC** of **ROC** curve. We have chosen that model to be the best for which we have obtained the highest F-Score and the highest AUC.

After obtaining the best model, we have applied it on a completely new test dataset and checked for how it performs on test set and for evaluating the performance of the chosen model, we have taken the same evaluation metric F-Score and AUC.

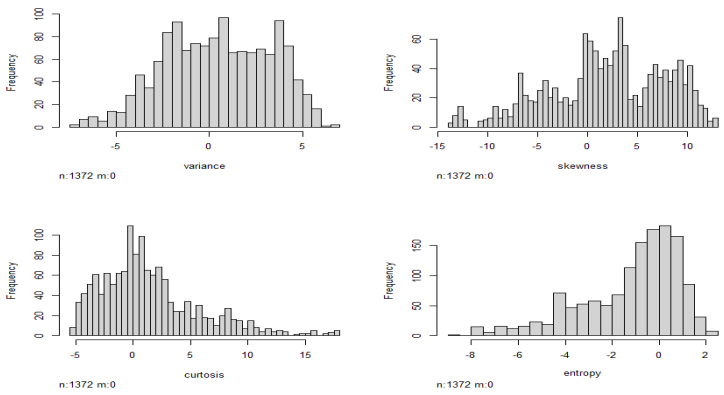


Figure 23: Histogram of Predictors

It is evident from the above figure that the assumption of LDA and QDA has been violated and for this reason, we cannot apply Discriminant Analysis method for this dataset. So, we proceed for applying all other methods.

3.3.1 Logistic Regression

After training logistic regression using training dataset, we have applied it on the cross validation dataset and obtained the following confusion matrix,

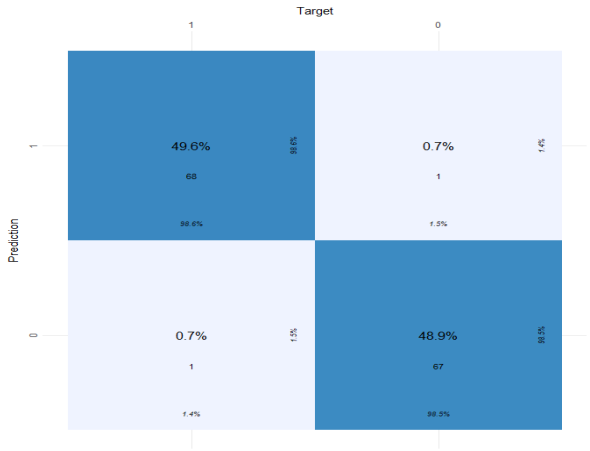


Figure 24: Confusion Matrix of Logistic Regression

3.3.2 Decision Tree

After training decision tree model using training set, we have pruned the tree and obtained the **minimum misclassification error rate** for the tree with 11 terminal nodes. we have applied it on the cross validation dataset and obtained the following confusion matrix,

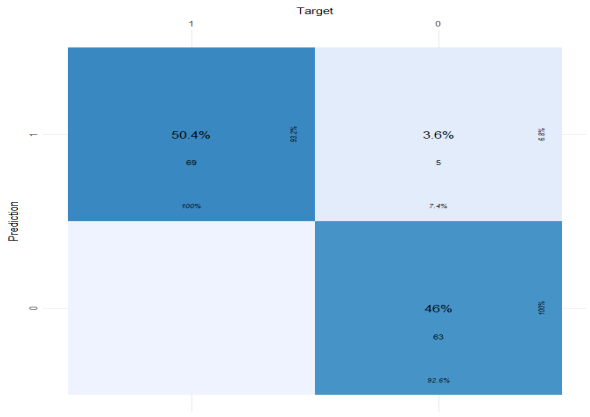


Figure 25: Confusion Matrix of Decision Tree

3.3.3 Random Forest

We have 8 predictors in our dataset and to train random forest model, we have chosen $\sqrt{4} = 2$ predictors in each split. After training Random Forest Model using training set, we have applied it on cross validation set and obtained the following confusion matrix,

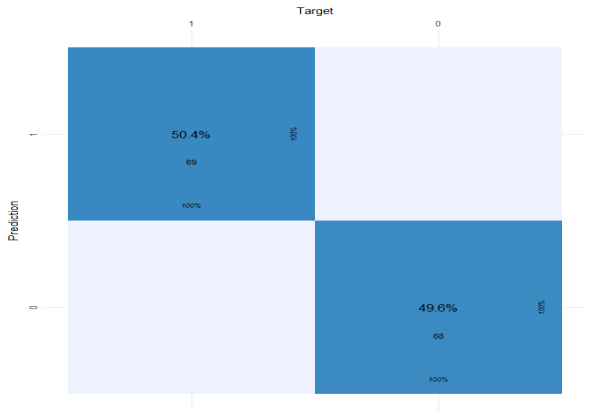


Figure 26: Confusion Matrix of Random Forest

3.3.4 Evaluating Performances of the Models

To evaluate and compare the performances of our models, we have used the evaluation metric **F-Score** and **AUC of the ROC Curve**. The AUC curves are as follows,

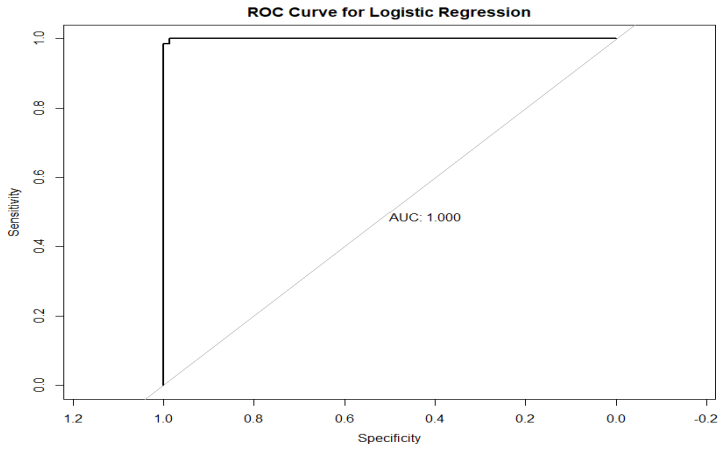


Figure 27: ROC Curve for Logistic Regression

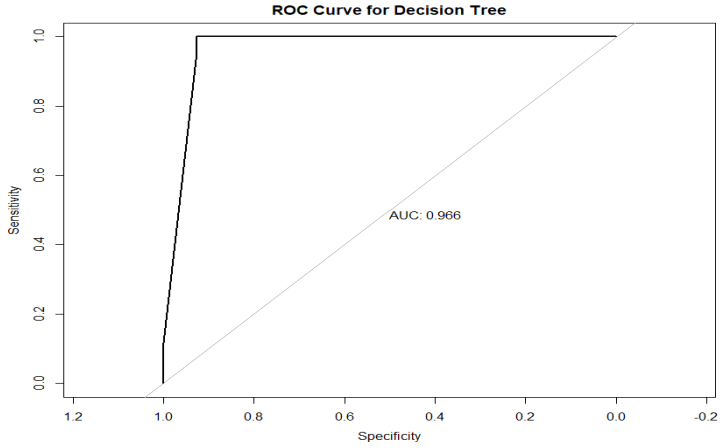


Figure 28: ROC Curve for Decision Tree

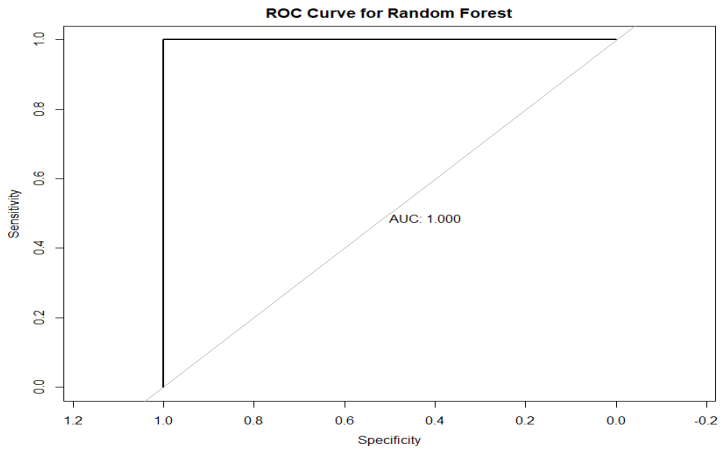


Figure 29: ROC Curve for Random Forest

From the graphs, it is evident that **Random Forest outperforms all the other models in that case**. After that, we have also compared F-Score and AUC to choose the final model.

	F-Score	AUC
Logistic Regression	0.9855	0.99979
Decison Tree	0.9650	0.96590
Random Forest	1.0000	1.00000

Table 3: F-Score and AUC for Various Models

From the above table, also, we can see that **Random Forest has the highest F-Score and AUC**. So, we have chosen random forest model to predict whether a given banknote is fake or real.

3.3.5 Final Performance

After having chosen Random Forest to be the best model, we have applied it on test set and obtained the confusion matrix to visualize the correct classification. After that, to assess the performance of the model on this completely new dataset, we have obtained F-Score and AUC respectively.

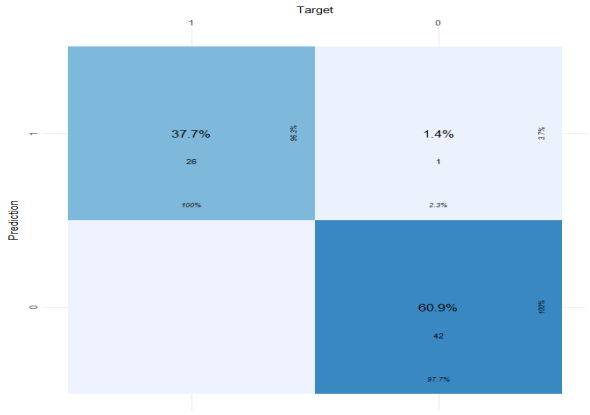


Figure 30: Confusion Matrix of Final Model

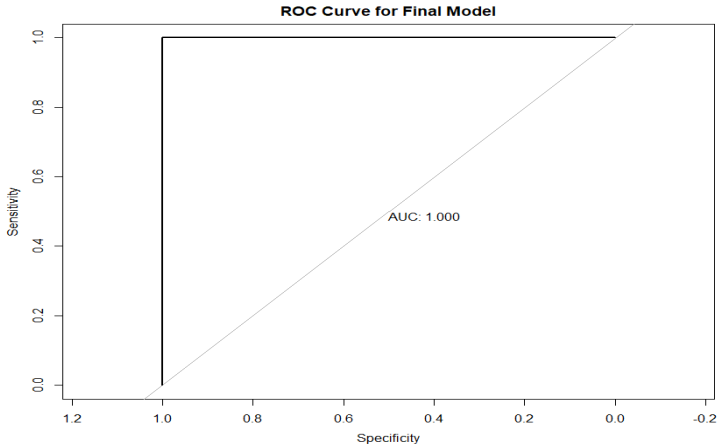


Figure 31: ROC Curve for Final Model

For the final model, we have obtained **F-Score of 0.98** and **AUC of 1**. So, we can conclude **it is satisfactory to use random forest model to predict whether a given banknote is fake or real.**

3.4 Weekly Dataset

In this dataset, weekly percentage returns for the S and P 500 stock index between 1990 and 2010 are given. The variables are as follows,

- Year: the year that the observation was recorded.
- Lag1: percentage return for previous week.
- Lag2: percentage return for two weeks previous.
- Lag3: percentage return for three weeks previous.
- Lag4: percentage return for four weeks previous.
- Lag5: percentage return for five weeks previous.
- Volume: Volume of shares traded (average number of daily shares traded in billions).
- Today: Percentage return for this week.
- Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week.

Here our goal is to predict whether a market will have a positive or negative return on a given week. First we have checked for **missing values** in the dataset since missing values can cause problems in modeling purpose. In this dataset, there are no missing values. So, we proceed to check whether our dataset is **imbalanced** or not. In our dataset, we have 56 % real banknote and 44 % fake banknote. So, in this case, our dataset is balanced. After all the preprocessing with our data, we have split our dataset into 3 parts, namely, **train set, cross validation set, test set**. We have trained logistic regression, decision tree, random forest using train set and checked whether our dataset validates the assumption of lda and qda.

After training the models with our training set, we have applied them on cross validation set and to evaluate the performances of the model, we have looked for **F-Score** and **AUC** of **ROC** curve. We have chosen that model to be the best for which we have obtained the highest F-Score and the highest AUC.

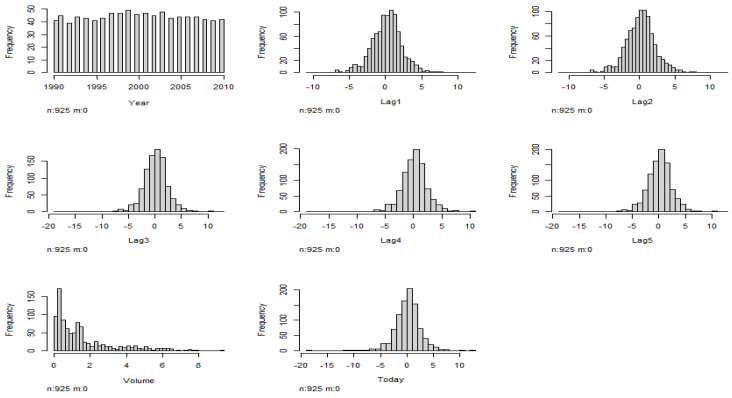


Figure 32: Histogram of Predictors

We will take Lag1, Lag2, Lag3, Lag4, Lag5, Today for applying LDA and QDA, since the histograms of these predictors seems symmetric, so we asume normality in these cases.

3.4.1 LDA

After training LDA using training dataset, we have applied it on the cross validation dataset and obtained the following confusion matrix,

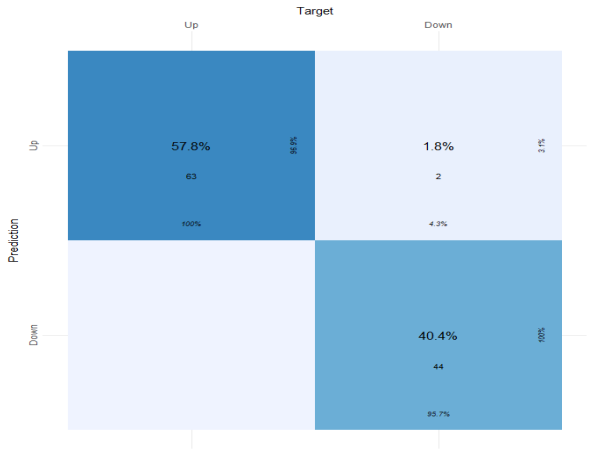


Figure 33: Confusion Matrix of LDA

3.4.2 QDA

After training QDA using training dataset, we have applied it on the cross validation dataset and obtained the following confusion matrix,

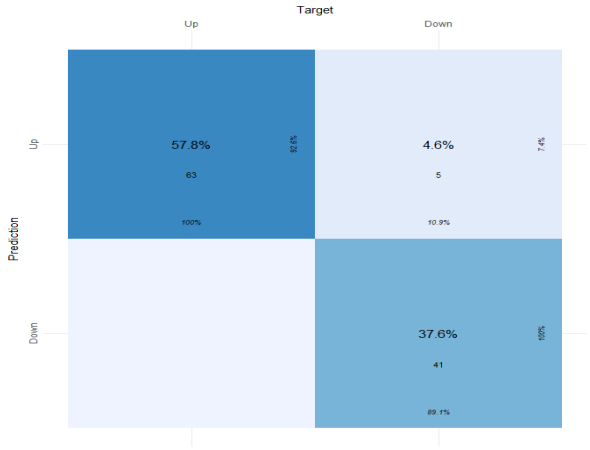


Figure 34: Confusion Matrix of QDA

3.4.3 Logistic Regression

After training logistic regression using training dataset, we have applied it on the cross validation dataset and obtained the following confusion matrix,

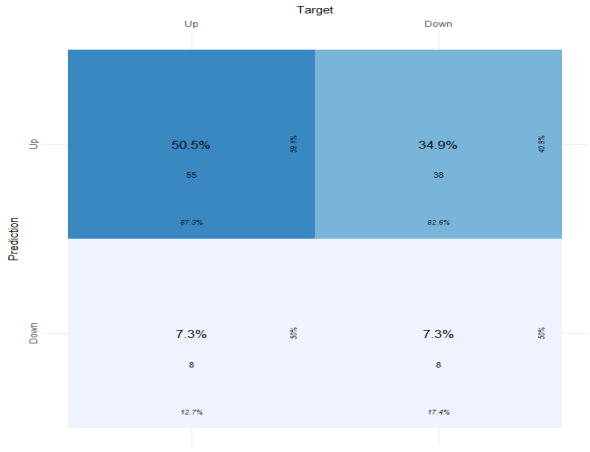


Figure 35: Confusion Matrix of Logistic Regression

3.4.4 Decision Tree

After training decision tree model using training set, we have pruned the tree and obtained the **minimum misclassification error rate** for the tree with 2 terminal nodes. we have applied it on the cross validation dataset and obtained the following confusion matrix,

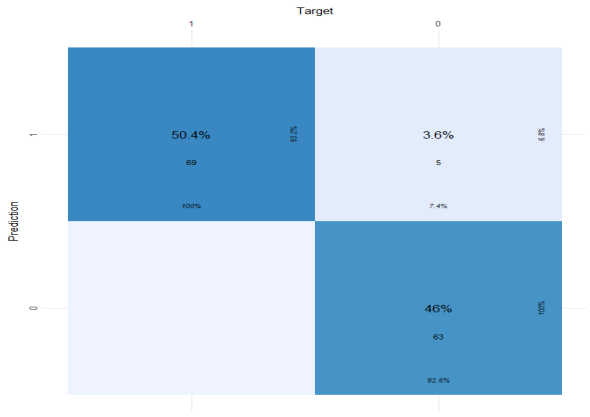


Figure 36: Confusion Matrix of Decision Tree

3.4.5 Random Forest

We have 7 predictors in our dataset and to train random forest model, we have chosen $\sqrt{7} \approx 2$ predictors in each split. After training Random Forest Model using training set, we have applied it on cross validation set and obtained the following confusion matrix,

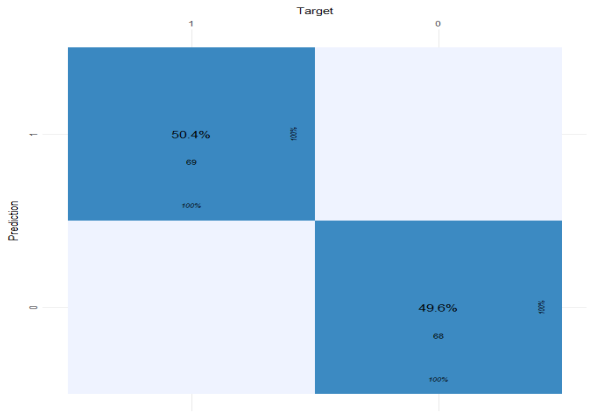


Figure 37: Confusion Matrix of Random Forest

3.4.6 Evaluating Performances of the Models

To evaluate and compare the performances of our models, we have used the evaluation metric **F-Score** and **AUC of the ROC Curve**. The AUC curves are as follows,

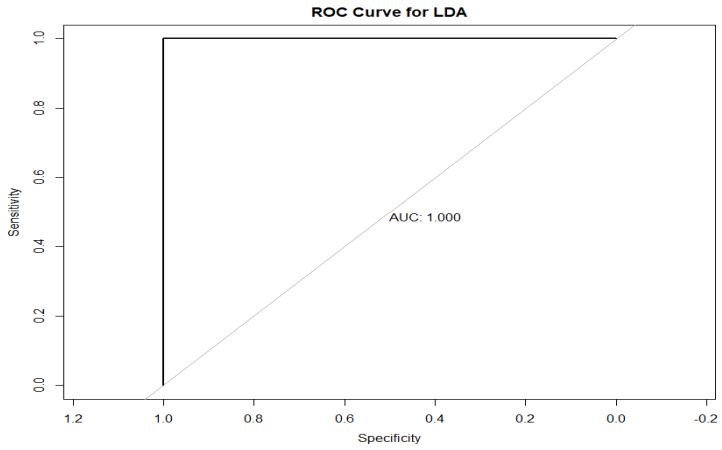


Figure 38: ROC Curve for LDA

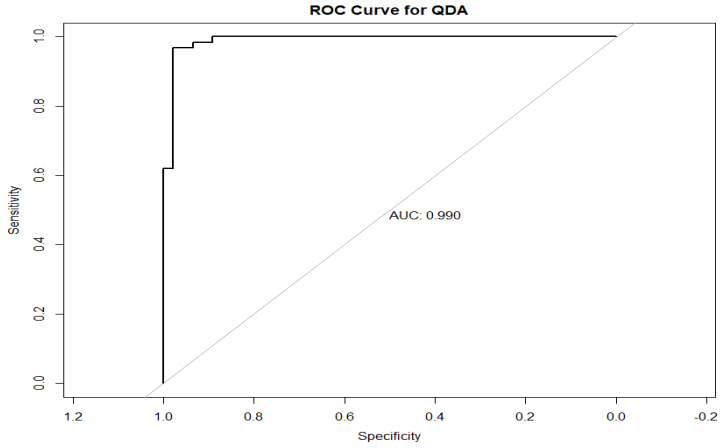


Figure 39: ROC Curve for QDA

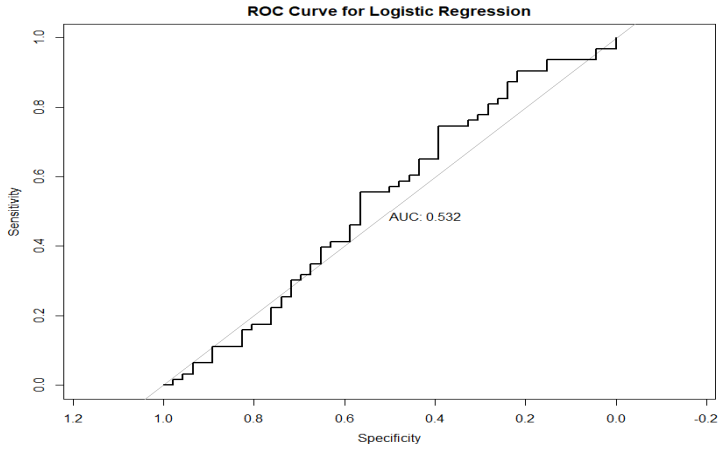


Figure 40: ROC Curve for Logistic Regression

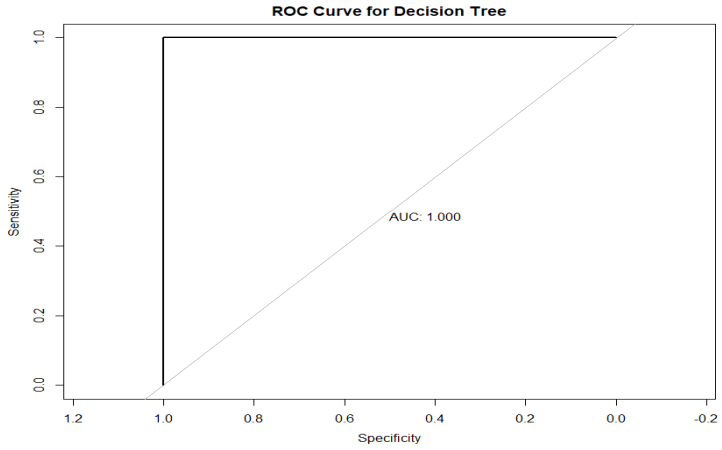


Figure 41: ROC Curve for Decision Tree

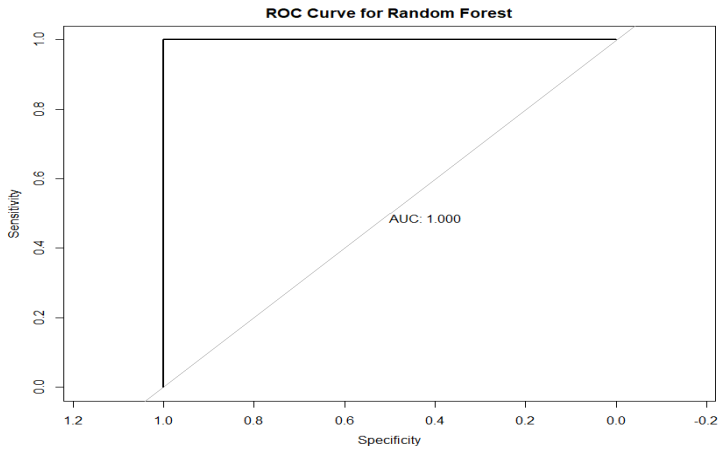


Figure 42: ROC Curve for Random Forest

From the graphs, it is evident that **Random Forest outperforms all the other models in that case**. After that, we have also compared F-Score and AUC to choose the final model.

	F-Score	AUC
LDA	0.98	1.00
QDA	0.96	0.99
Logistic Regression	0.71	0.53
Decison Tree	1.00	1.00
Random Forest	1.00	1.00

Table 4: F-Score and AUC for Various Models

From the above table, we can choose either Decision Tree or Random Forest either. But, decision trees are sometimes prone to overfitting. So, we choose random forest to predict whether a market will have a positive or negative return on a given week.

3.4.7 Final Performance

After having chosen Random Forest to be the best model, we have applied it on test set and obtained the confusion matrix to visualize the correct classification. After that, to assess the performance of the model on this completely new dataset, we have obtained F-Score and AUC respectively.

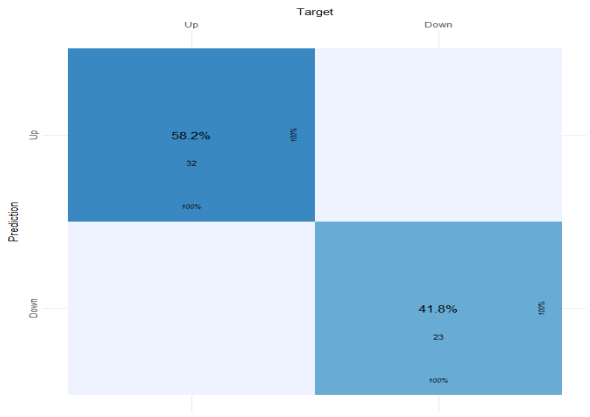


Figure 43: Confusion Matrix of Final Model

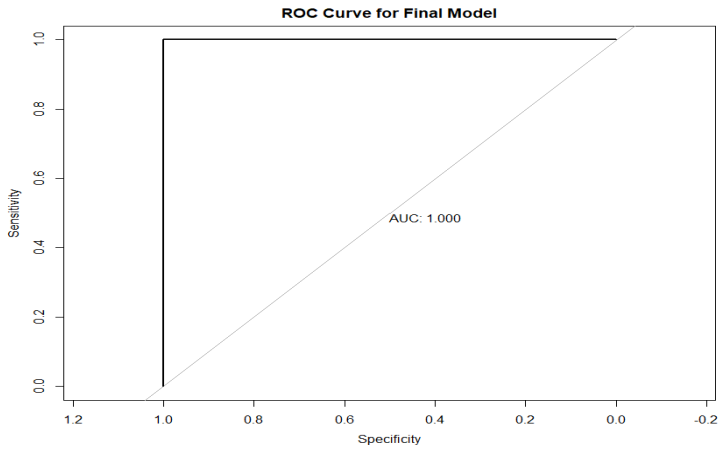


Figure 44: ROC Curve for Final Model

For the final model, we have obtained **F-Score of 1** and **AUC of 1**. So, we can conclude **it is satisfactory to use random forest model to predict whether a market will have a positive or negative return on a given week.**

4 Conclusion

The main findings of the project are as follows,

1. PIMA Indian Diabetes Dataset: For this dataset, Random Forest outperforms all the other models. F-Score and AUC for this model is close to 1 and the model is fitting the test dataset also very well. So, we can be assured that it is not overfitting.

2. Caravan Insurance Dataset: For this dataset, F-Score for Random Forest and Logistic Regression are coming out to be same, but, AUC for Random Forest is slightly better than that of Logistic Regression and for that reason we have chosen Random Forest to be the best model for this dataset.

3. Banknote Authentication Dataset: For this dataset, Random Forest outperforms all the other models. F-Score and AUC for this model is 1 and the model is fitting the test dataset also very well. So, we can be assured that it is not overfitting.

4. Weekly Dataset: For this dataset, both the Decision Tree and Random Forest models have F-Score and AUC of 1. But, since, random forest assures to be free from overfitting by decorrelating the trees, we have chosen this model to be the final one.

5 Appendix

5.1 R Codes

The R Codes for respective datasets can be found here.

5.2 Synthetic Minority Oversampling Technique (SMOTE)

A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary.

One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

An improvement on duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space. This procedure can be used to create as many synthetic examples for the minority class as are required.

5.3 F-Score

To define F-Score, we have to define **Precision** and **Recall** first.

$$\text{Precision (P)} = \frac{True Positive}{Total Number of Predicted Positive}$$
$$\text{Recall (R)} = \frac{True Positive}{Total Number of Actual Positive}$$

Then, F-Score is given by,-

$$\text{F-Score} = \frac{2PR}{P+R}$$

The higher the F-Score of a model, the greater the accuracy is.

5.4 ROC-AUC Curve

It is a curve of **Sensitivity vs Specificity**, where,

$$\text{Specificity} = \frac{True Negative}{Total Number of Actual Negative}$$
$$\text{Sensitivity} = \frac{True Positive}{Total Number of Actual Positive}$$

The area under this curve is known as AUC (Area Under the Curve). The higher the AUC of an ROC curve, the better the model is.

6 References

1. An Introduction to Statistical Learning with Applications in R Book- Gareth M. James, Trevor Hastie, Daniela Witten, and Robert Tibshirani.
2. https://towardsdatascience.com/
3. https://www.analyticsvidhya.com/
4. https://stackoverflow.com/