

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID |  |
| Student Name | Suyog Sitaula |
| Module Name & CRN | Applied Machine Learning |
| Level | 5 |
| Assessment Name & Part No. | Part 2 |
| Project Title | Model Implementation for Fraud Detection |
| Data of Submission | 5/10/2024 |
| Course | Computing |
| Academic Year | 2024 |

Contents

[PART II 3](#_Toc166275772)

[Introduction 3](#_Toc166275773)

[Data splitting 3](#_Toc166275774)

[Modelling: 3](#_Toc166275775)

[Target Variable 3](#_Toc166275776)

[Train control 3](#_Toc166275777)

[1. Logistic Regression 3](#_Toc166275778)

[2. KNN 6](#_Toc166275779)

[3. Naïve Bayes 9](#_Toc166275780)

[4. Decision Tree 10](#_Toc166275781)

[5. Random Forest 13](#_Toc166275782)

[Model Comparison 14](#_Toc166275783)

[Model Interpretation 16](#_Toc166275784)

[Conclusion 16](#_Toc166275785)

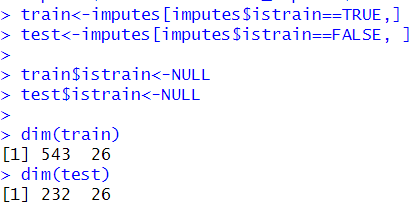
(NV5 Geospatial Solutions, n.d.) (Wein, n.d.) (Brownlee, 2023)

Model Implementation for Fraud Detection Introduction

The dataset given to us was made suitable to train and test a model by detecting the missing values, imputing the missing values with suitable methods, detecting outliers and many more. Outliers were detected in the combined dataset but were not removed as it contained large numbers of information within itself. The problem that occurred due to collinearity of variables were solved and the remedy of constant variable was applied during scaling. Currently the scaled dataset contains 775 obs and 27 variables. The dataset was made ready to be split and use it to train and test the models.

## Data splitting

We have split our combined dataset into two which are train and test dataset where we train our model using train dataset and test it using test dataset. The splitting process was easier as we have isTrain column which shows whether the row belongs to training or testing dataset using Boolean values.



# Modelling:

## Target Variable

Our training variable for our data set is Risk which is the 27th column of our previous dataframe.

## Cross Validation

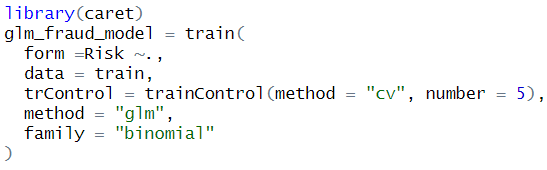
The sampling was not done in this dataset as there was not that many observations to work on and had pretty balanced data. K-Fold-Cross-Validation is used to check whether the result we got was by luck or was it the doing of the model. The K here refers to the number of groups that the sample that was given to use will be split into. Here we have selected the value of K as 5 so that we will get less error.

### Logistic Regression

Logistic regression is done in R with the help of Generalized Linear Method(glm). We can use it with the package named caret. The generalized linear model used here is binomial.

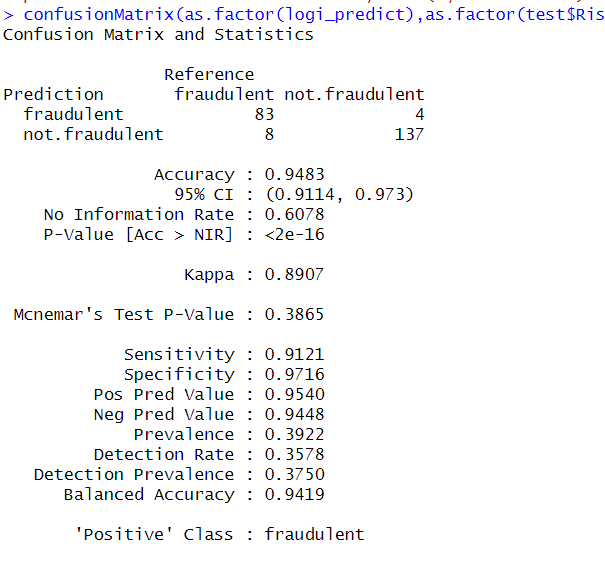
#### Model Training

The train dataframe is taken from where we will be taking the variable needed for the formula.

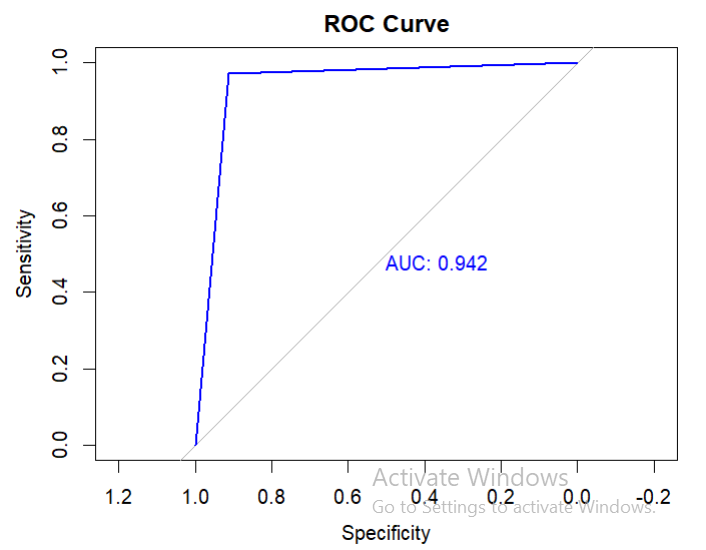


#### Model Evaluation

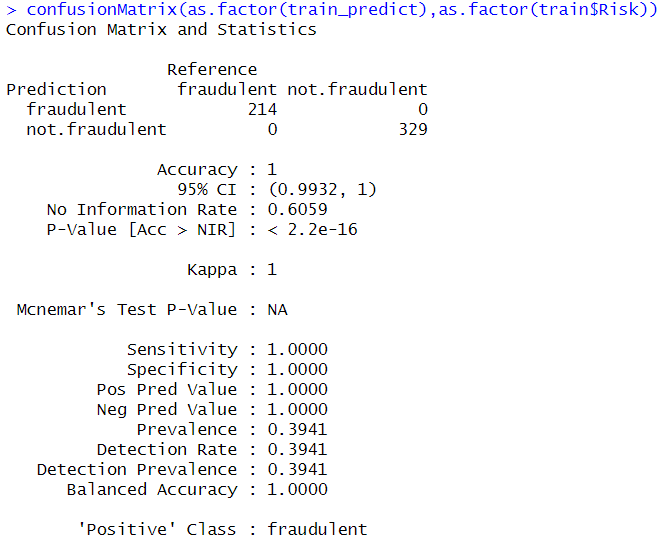
This is the confusion matrix for testing data. Here it shows the accuracy of the model which is 94% and kappa which shows the reliability and consistency of the model. Here kappa is 89% that means the classification made by model agrees 89% with the actual or true value. The sensitivity and specificity shown in below confusion matrix are 91% and 97% respectively which is pretty high.



The Receiver Operating Curve(ROC) here made shows the Area Under Curve(AUC) which is 0.942.

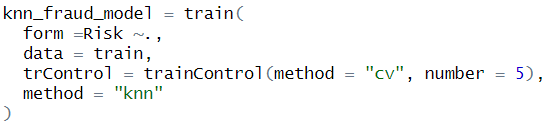


I have made the confusion matrix of Train data as well to see whether the model is overfitting, underfitting or a perfect model. The accuracy and kappa shown for it is 1 so both our training and testing dataset have high accuracy for the model. So, it can be considered as a slightly overfitting.



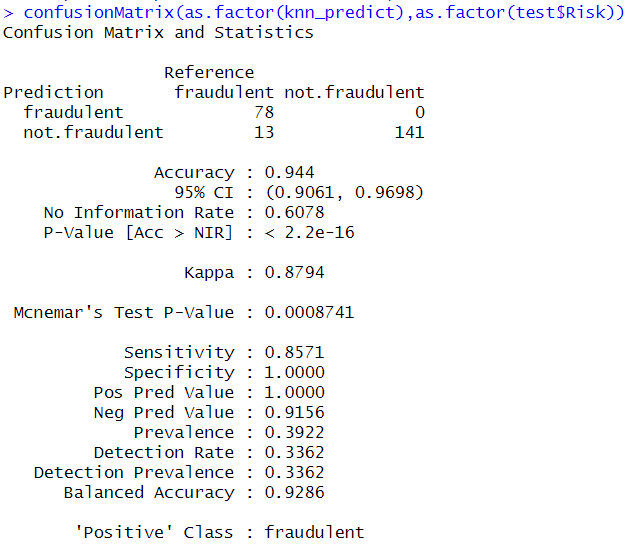
### KNN

#### Model Training

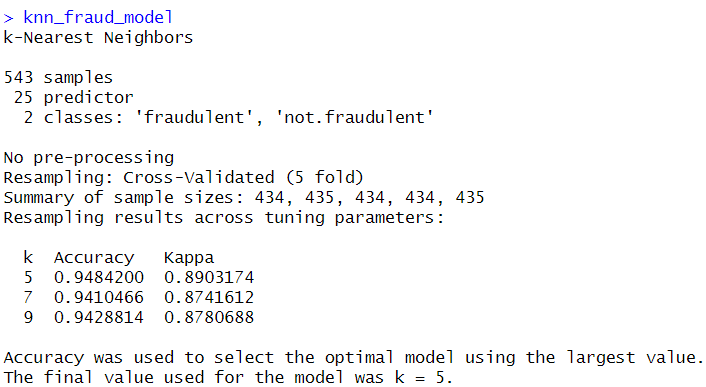
A KNN model was made using the knn method which can be found on caret package. 

#### Model Evaluation

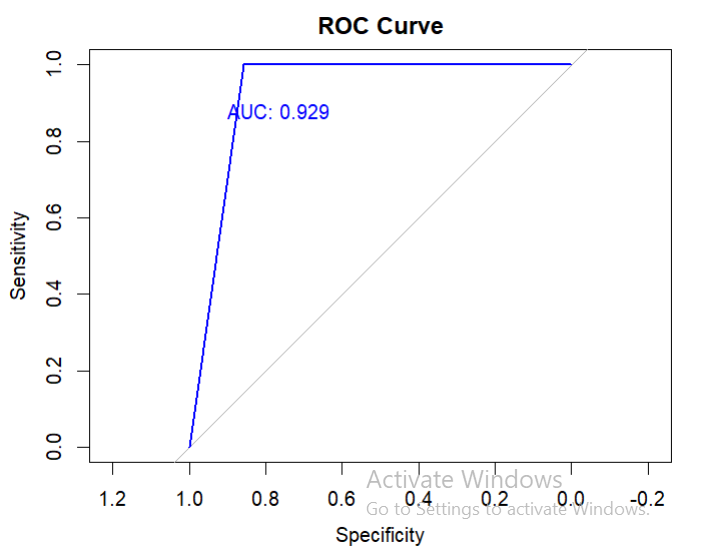
The confusion matrix for the KNN model is made where we can see the accuracy and kappa are 94% and 87% which are very good. We also don’t need to tune cause the model perform very well on both training and testing dataset. It is slightly overfitting compared to its train confusion matrix.



We can see the accuracy of the model with different k and the K with best accuracy and kappa was chosen.

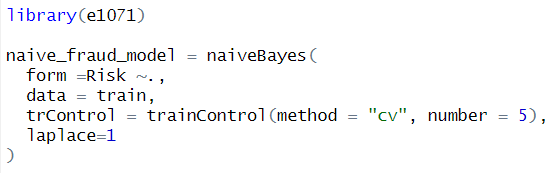


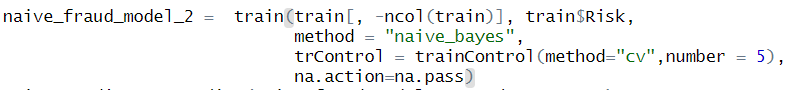
This is the ROC curve of the model where AUC is 0.929



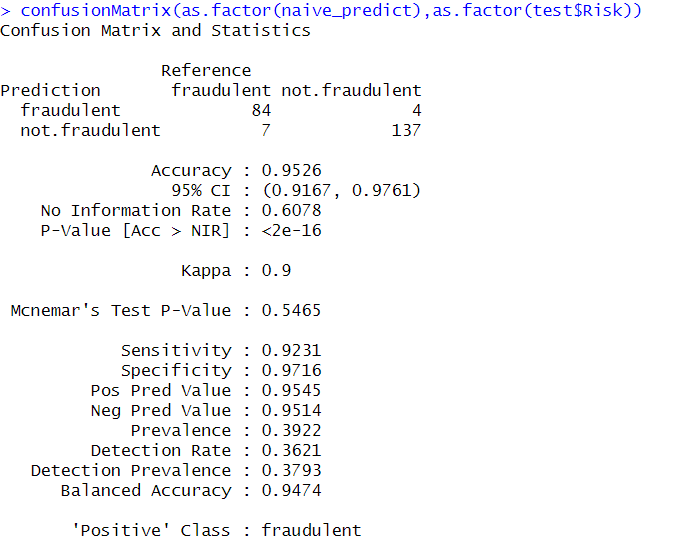
### Naïve Bayes

#### Model Training

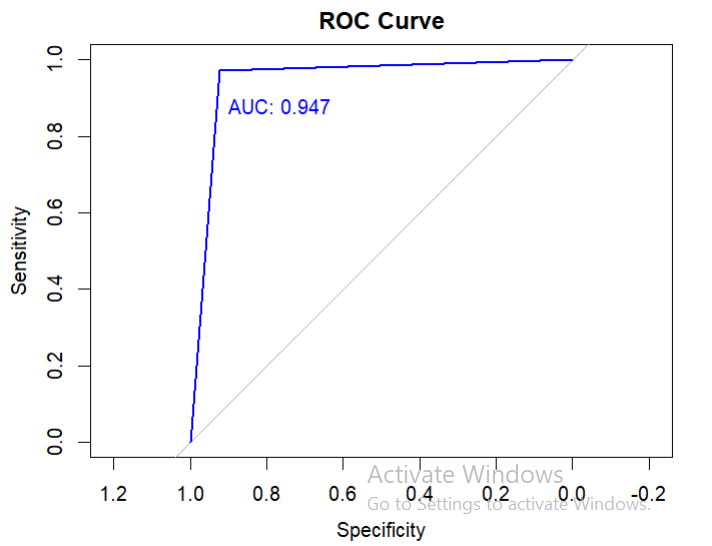
We have used naiveBayes from package e1071 to train model but also done it in caret 



#### Model Evaluation

The confusion matrix is made for test data set and when tried both naïve bayes model it did not show any difference in accuracy and kappa but we will be using the model made by using caret as it does not show the train control parameter. The accuracy of the model is 0.9526 and the kappa is 0.9 which can be considered as slightly overfitting. 

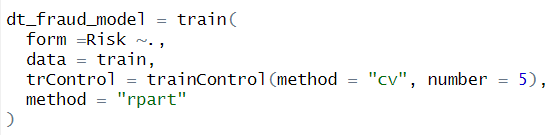
This is the ROC curve for the model where it has shown the AUC 0.947.



### Decision Tree

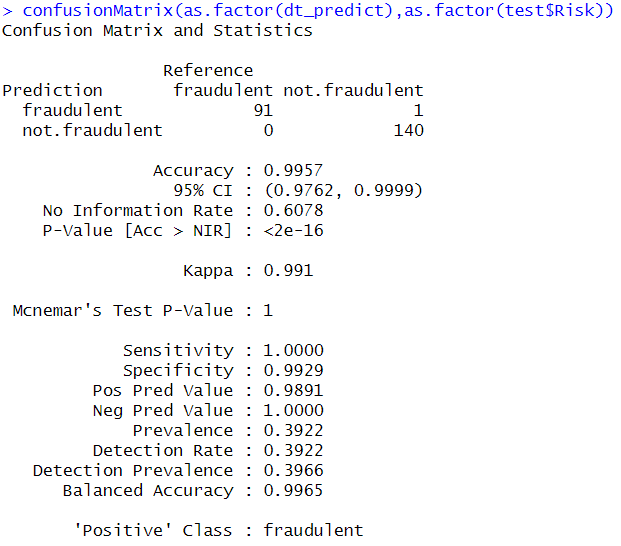
#### Model Training

This is the decision tree model where rpart is used as a method.

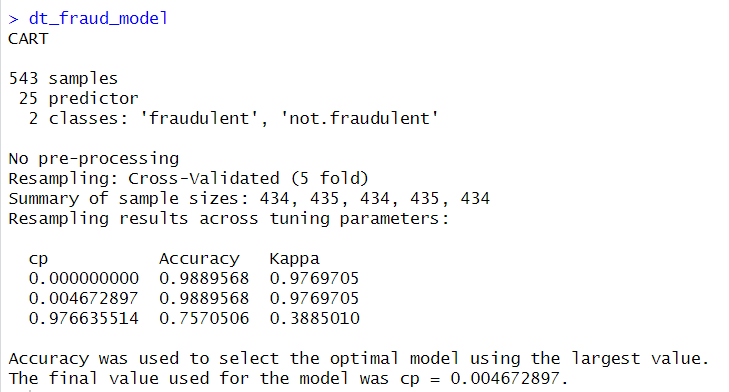


#### Model Evaluation

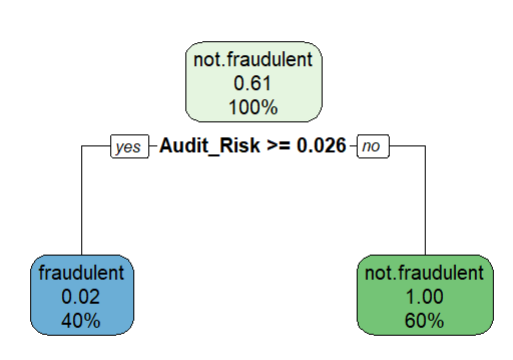
The confusion matrix below shows the accuracy and kappa of 99%. It shows that model was 99% sure to identify the actual value.



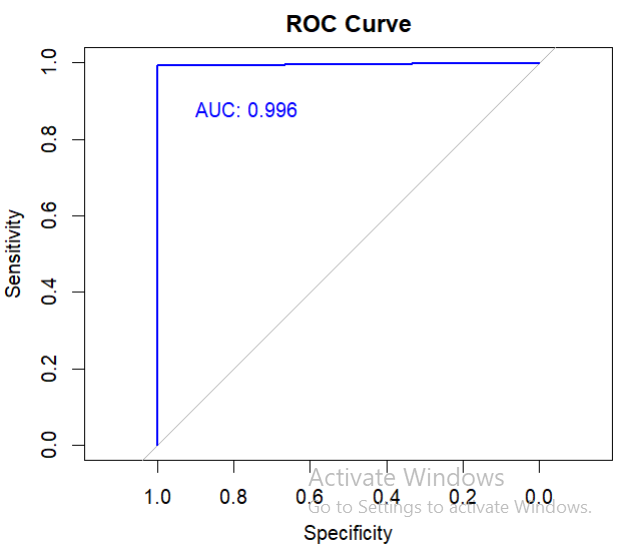
In bellow screenshot we can see three different complexity parameter and the one with best accuracy is chosen. In this case it was 0.004672897.



We also plot a decision tree using rpart.plot. The tree is very simple as it doesn’t have many other nodes to explore.

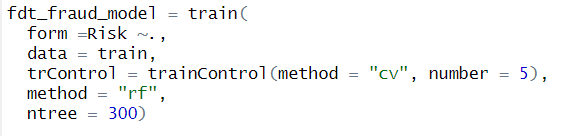


The ROC curve of the prediction was made and the AUC was 0.99 which shows us that it is a perfect model with very less error percentage.



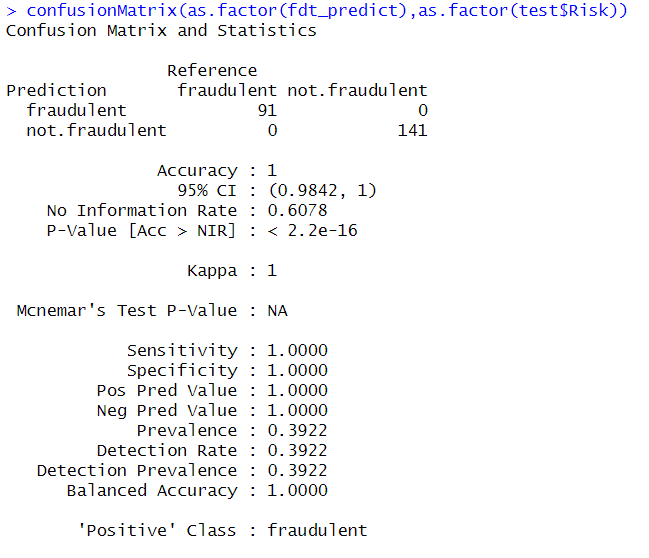
### Random Forest

#### Model Training

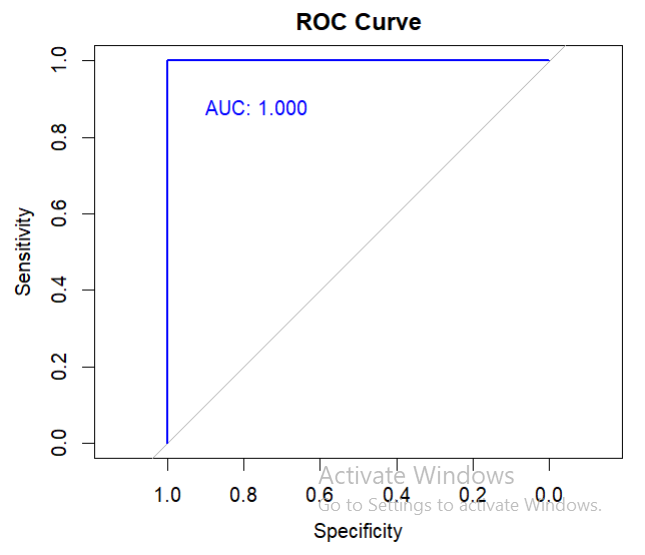
We have used rf method to train random forest model where the number of trees was set to 300. 

#### Model Evaluation

The confusion matrix for the Random Forest shows the accuracy and kappa of 1 which shows that the model is perfect.

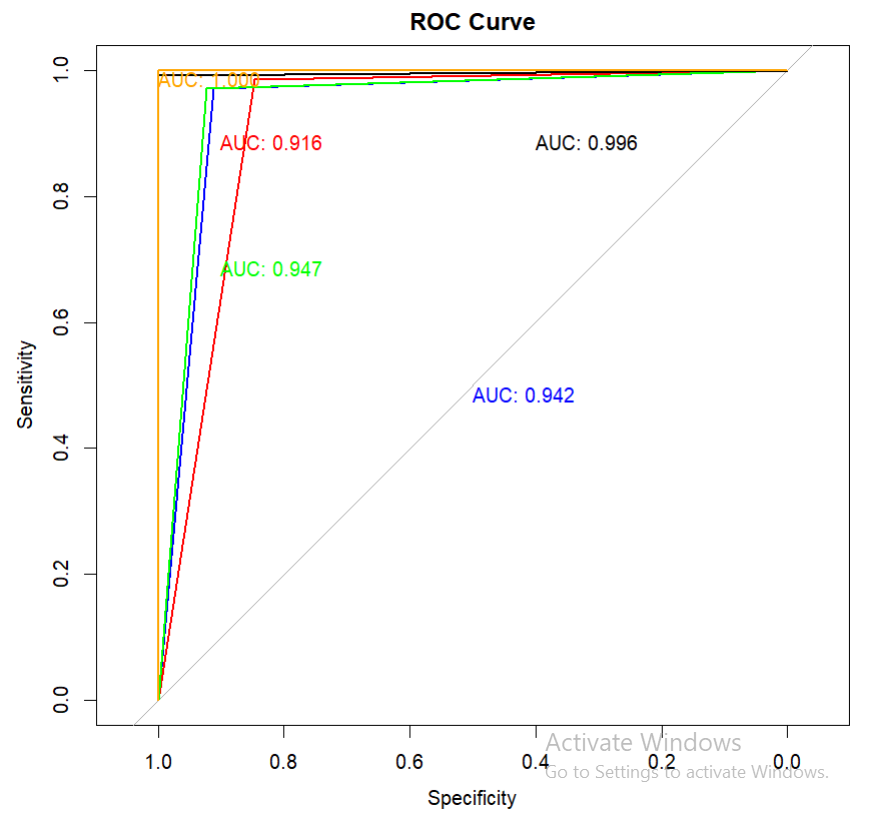


This is ROC curve for the model where it shows that the AUC is 1.



# Model Comparison

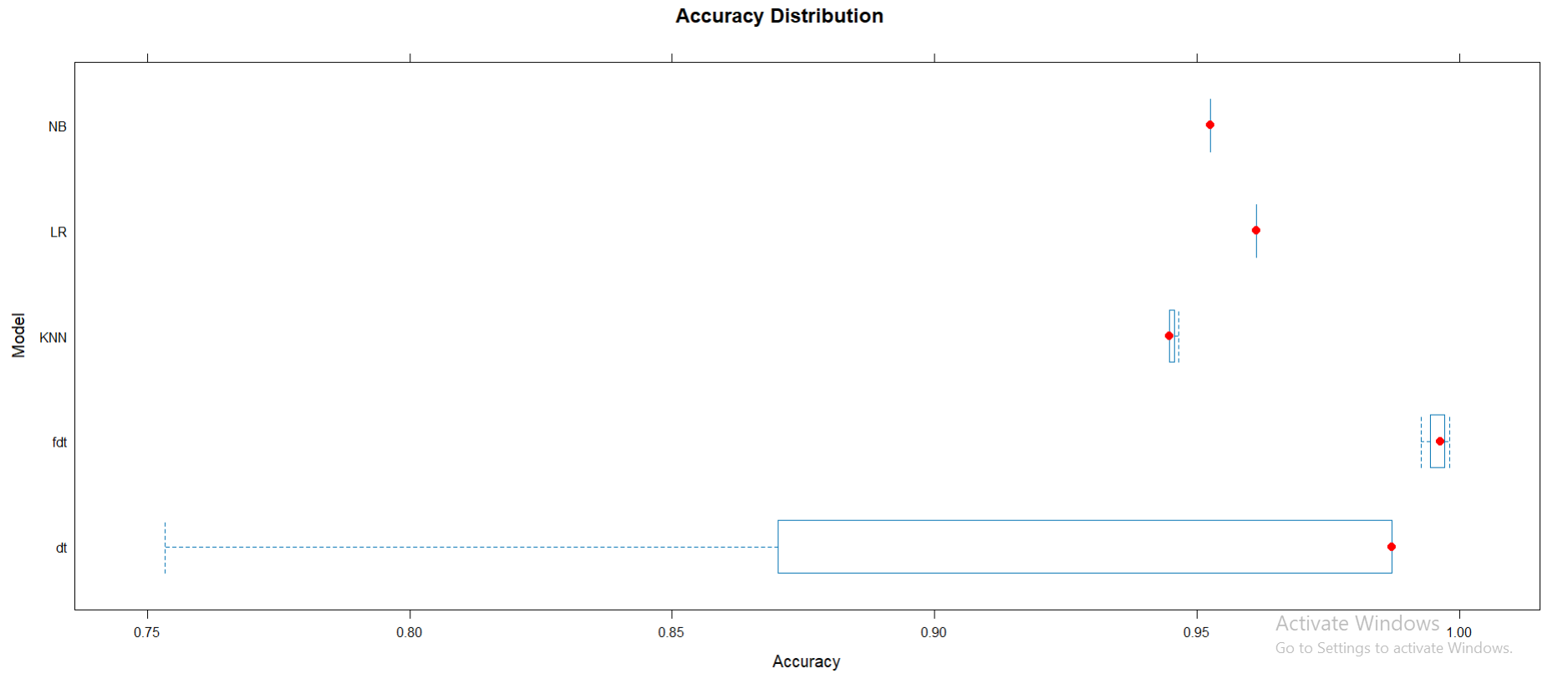
ROC of all model combined is shown below. The black one is for the decision tree model, blue one is for logistic regression model, the red one is for KNN model, the orange one is for random forest model and the green one is for naïve bayes.

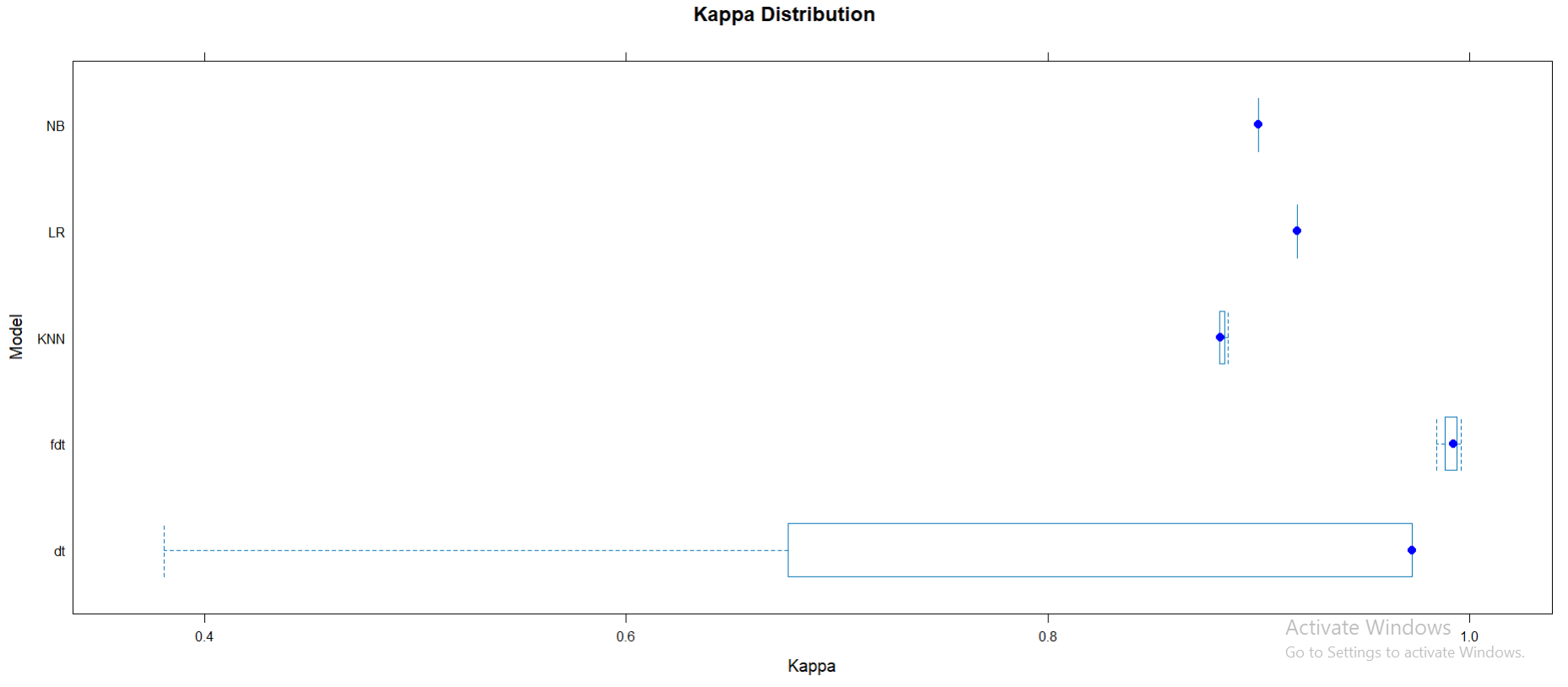


The below table shows the name of the model along with its accuracy, sensitivity, specificity, false positive, false negative, kappa and AUC. The trend is seen where the false negatives are seen more in numbers that false negative. This means that less people will be wrongly accused of fraud but very few will be able to get away with it looking at all model except rf .

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **FP** | **FN** | **Kappa** | **AUC** |
| LR | 0.9483 | 0.9121 | 0.9716 | 0.046 | 0.0552 | 0.8907 | 0.942 |
| KNN | 0.944 | 0.8571 | 1.0 | 0 | 0.0844 | 0.8794 | 0.929 |
| NB | 0.9526 | 0.9231 | 0.9716 | 0.0455 | 0.0486 | 0.9 | 0.947 |
| DT | 0.9957 | 1 | 0.9929 | 0.0109 | 0 | 0.991 | 0.996 |
| RF | 1 | 1 | 1 | 0 | 0 | 1 | 1 |

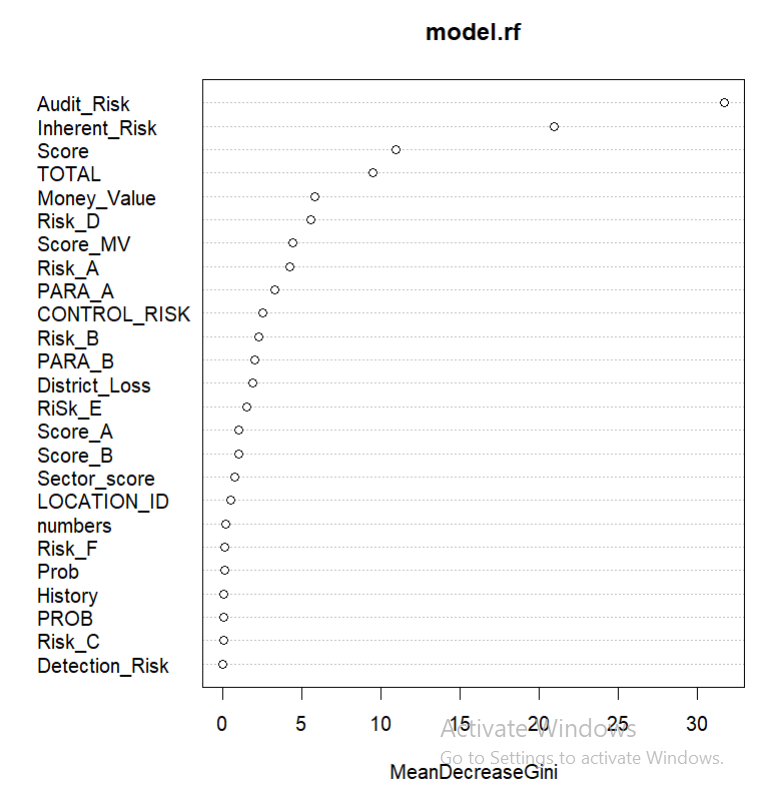
The box plot is made for accuracy and kappa to compare the model performance. It also helps to visualize the stability of these 5 model performances.





# Model Interpretation

According to the information we got from the above figure and table it shows that the Random Forest model is the best one to detect fraudulent companies as it has the highest accuracy and kappa among the other models which is 1. The below chart shows the mean decrease gini which indicates the importance of the predicted variables for classification. In this case the most important variable is Audit\_Risk. Decision Tree is also a viable and reliable model if a particular is concerned about resources. The random forest has used 300 number of trees so it might take a lot of resources to execute this model but decision tree have used 1/300th of its resources and gave the best accuracy compared to other model except random forest.



# 

# Conclusion

In conclusion the random forest model seems to be a better model compared to other 4 model we tried. Even if it is being compared to other algorithms, we discussed in literature review, it still is one of the most reliable models as its accuracy and kappa is 1. There is no room for error if this model is used in the dataset such as ours. The result showed good performance of random forest model on handling the dataset and giving the correct prediction of cases.

Despite Random Forest being the best one among the model decision tree model has also proven its worth by its remarkable performance considering it took 1/300th resources than Random Forest. It has the accuracy of 99.57% and kappa of 99%.

Through mean decrease gini we found out the variable that had the most influence on prediction. The variable was Audit\_Risk and it played a significant role on predictive accuracy. The other variables that played major role are Inherent\_Risk, Score, Total, etc. They all played a significant role on increasing the accuracy of the model.

# Bibliography

Brownlee, J., 2023. *A Gentle Introduction to k-fold Cross-Validation.* [Online]   
Available at: https://machinelearningmastery.com/k-fold-cross-validation/  
[Accessed 10 5 2024].

NV5 Geospatial Solutions, n.d. *Calculate Confusion Matrices.* [Online]   
Available at: https://www.nv5geospatialsoftware.com/docs/CalculatingConfusionMatrices.html#:~:text=The%20kappa%20coefficient%20measures%20the,of%200%20represents%20no%20agreement.  
[Accessed 10 5 2024].

Wein, T., n.d. *naiveBayes.* [Online]   
Available at: https://rdrr.io/rforge/e1071/man/naiveBayes.html  
[Accessed 10 5 2024].