# **Part A: Classification:**

Customer churn rate is an important performance metric in the Telecoms industry due to the highly competitive markets. The churn rate enables companies to understand why their customers are leaving. Churn dataset were used that contains randomly collected data from a telecom company’s database. ML models were developed to help the retention team predict high risk churn customers before they leave.

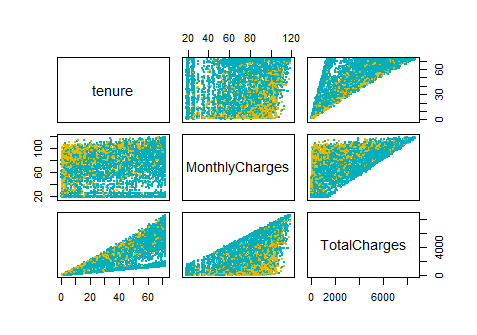
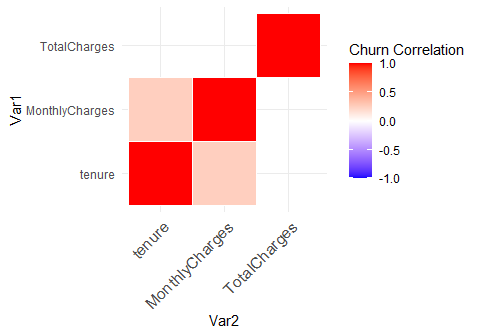
1. 1 Scatter plot matrix was shown between variables.

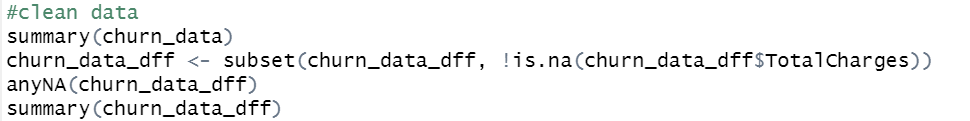
Figure 1: Scatter plot between numerical variables

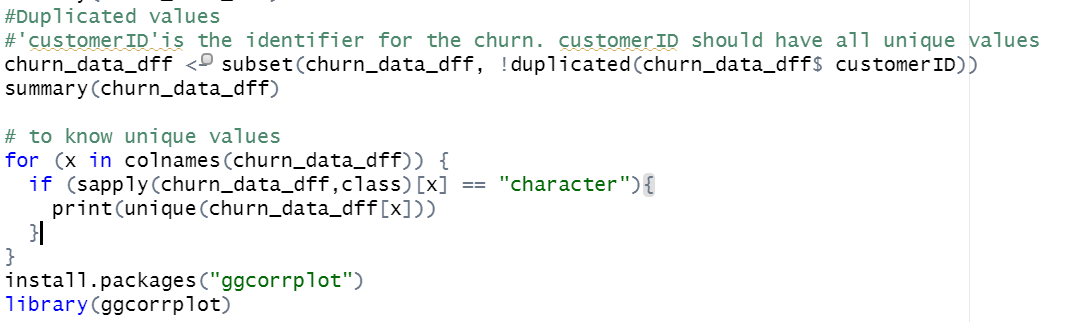
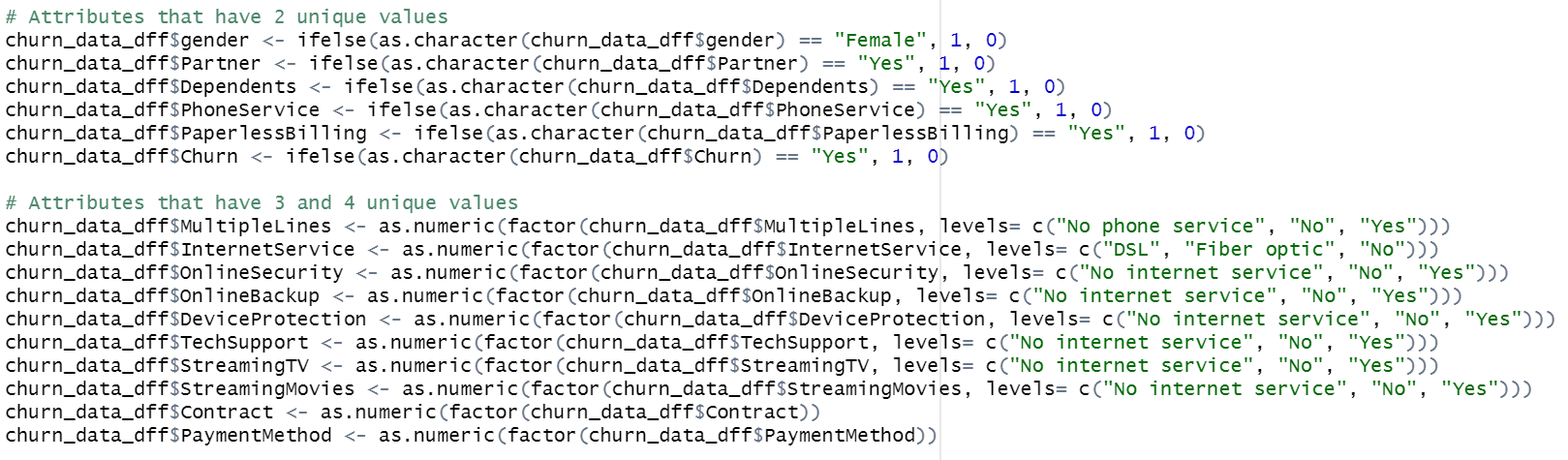
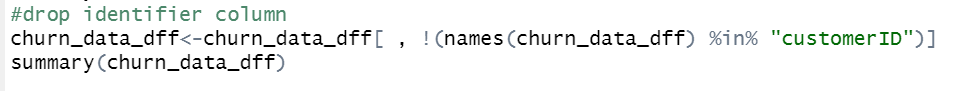
1. 2 heatmap to determine correlated attributes

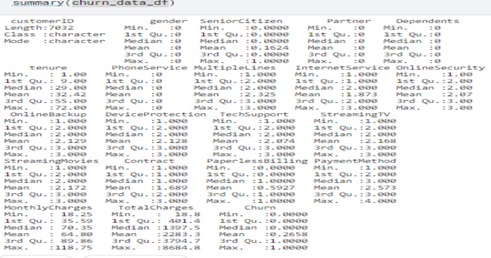


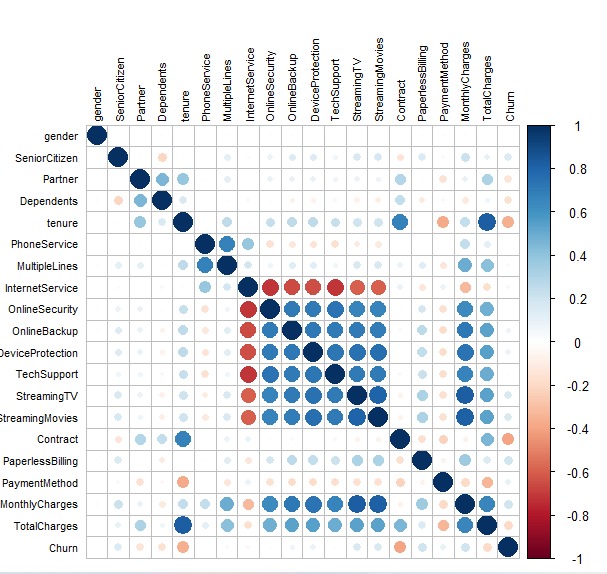
## 2 **Data Preprocessing**

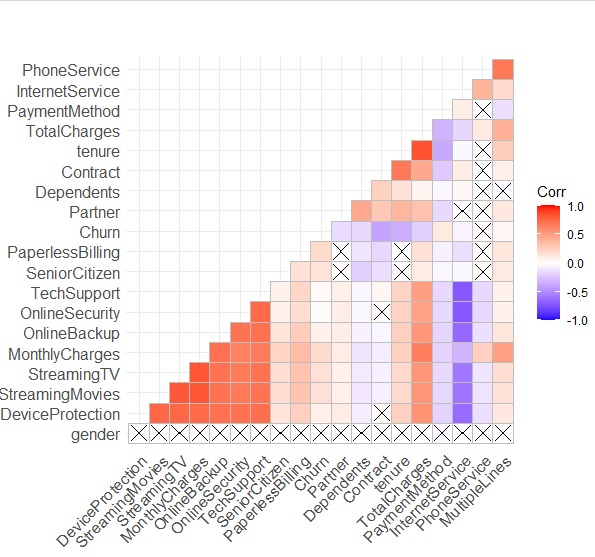
1. 1 **Remove null values**



1. 2 **Removing redundant information**
   1. **Convert categorical data into numerical labels**
   2. **Drop out identifier column**
   3. **Data frame summary after preprocessing**

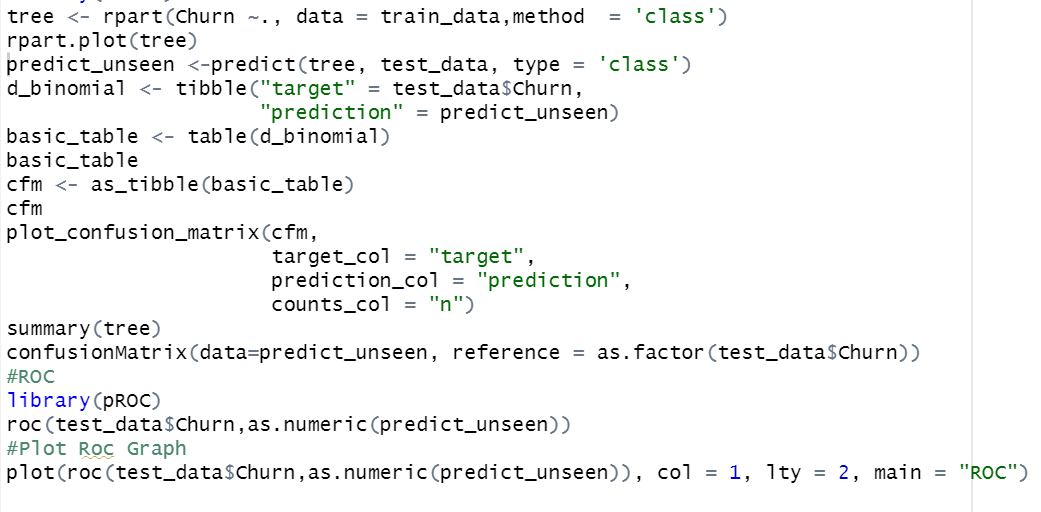


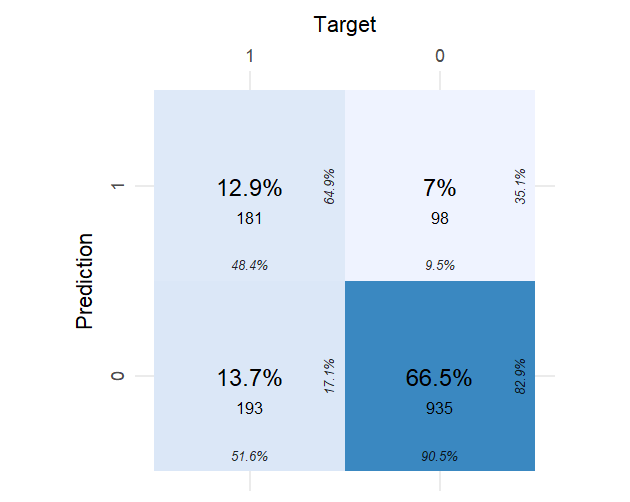
* 1. **Correlated features after data cleaning**

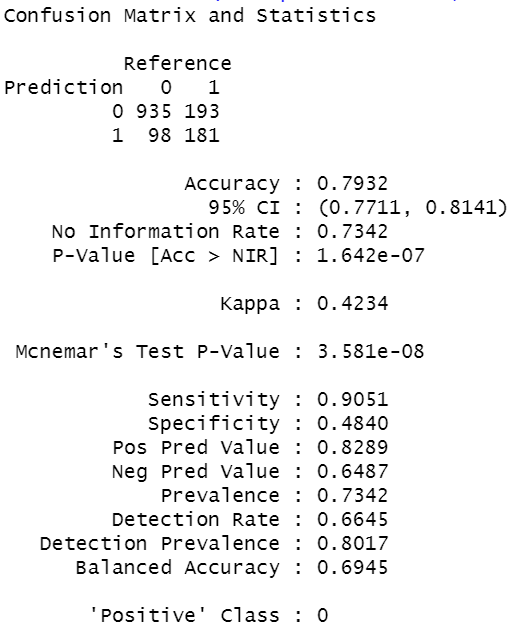


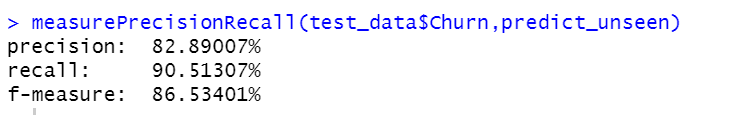
## **1. Split the dataset into 80 training/20 test set and fit a decision tree to the training data**

## **2 Plot the Decision tree**









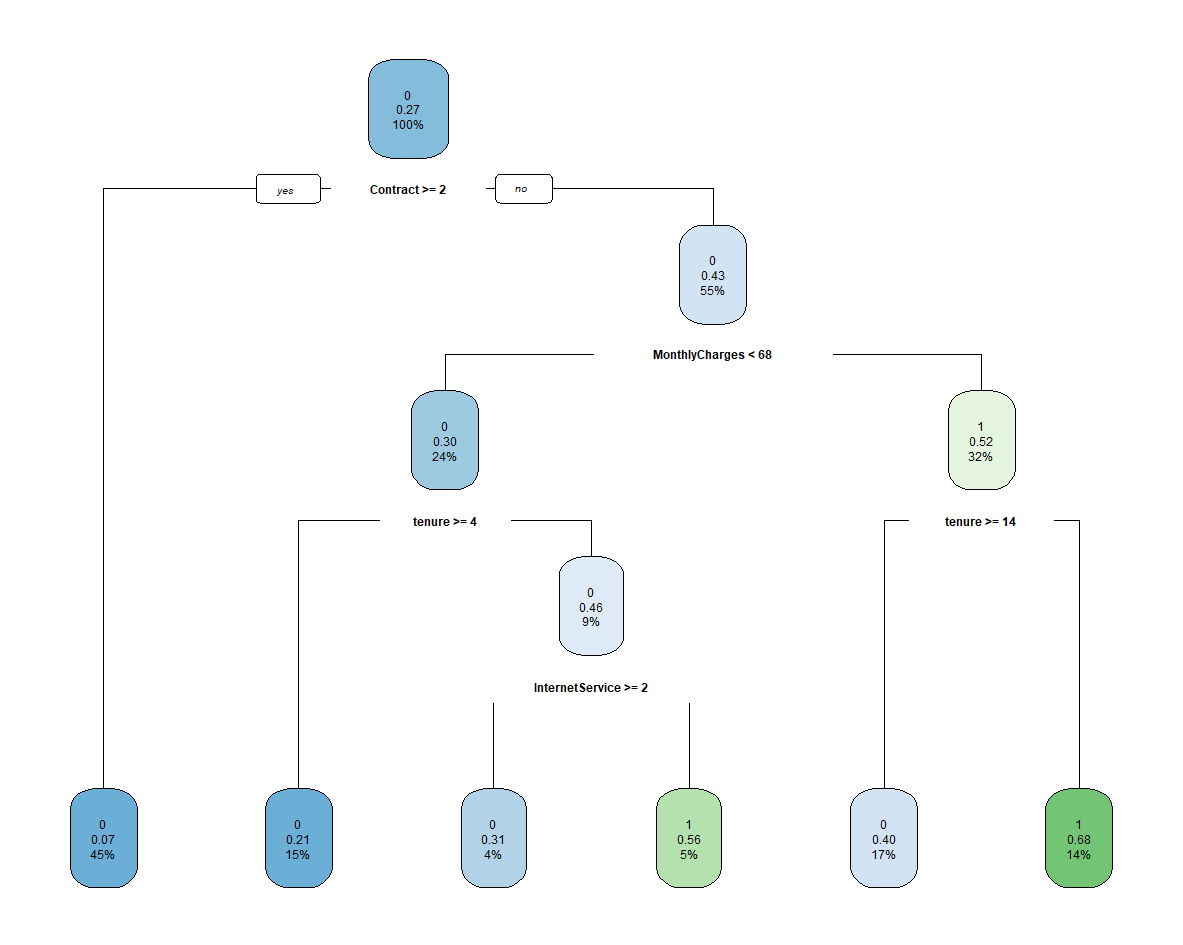
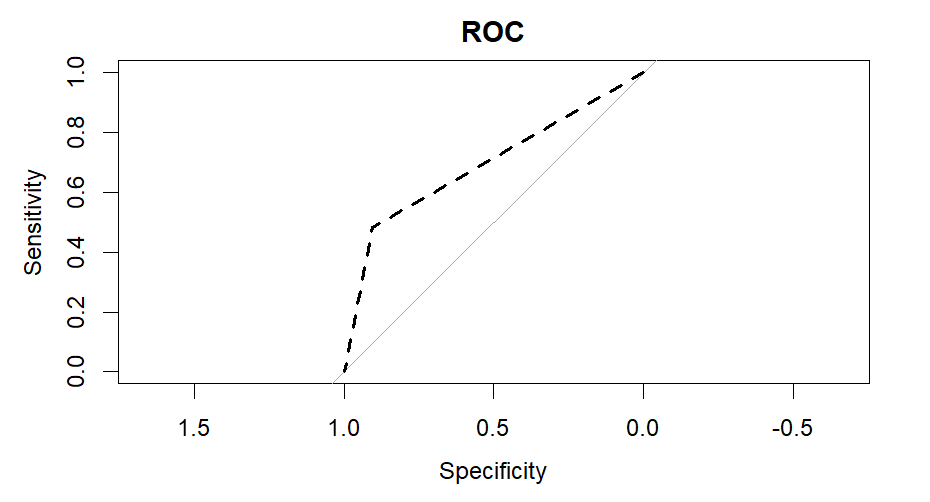
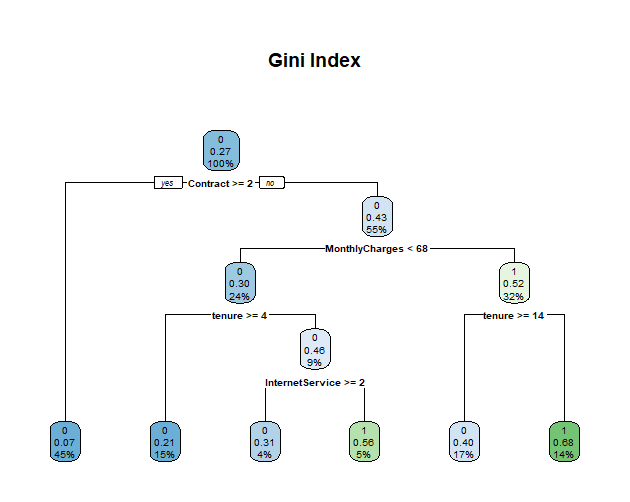
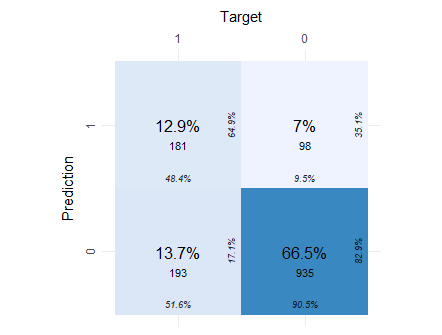


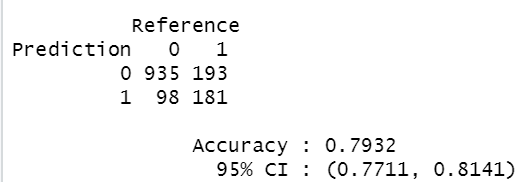
Figure 2 Decision Tree



## **Try different ways to improve the decision tree algorithm**

**4.1 Gini splitting strategy**





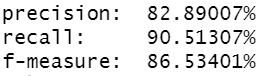
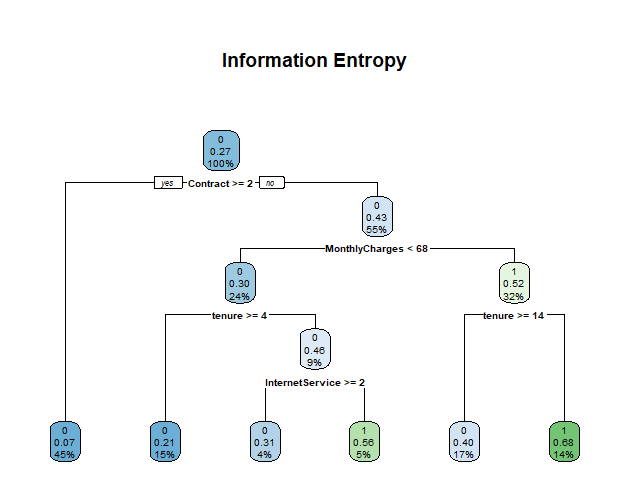
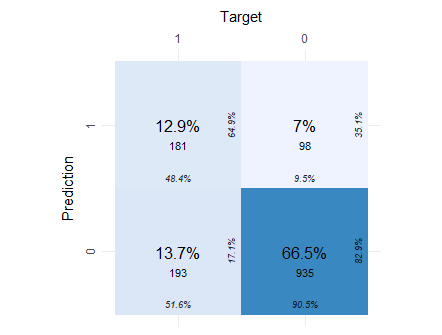


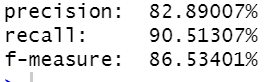
Figure 3 Accuracy, precision Recall

Figure 4 Gini index confusion matrix

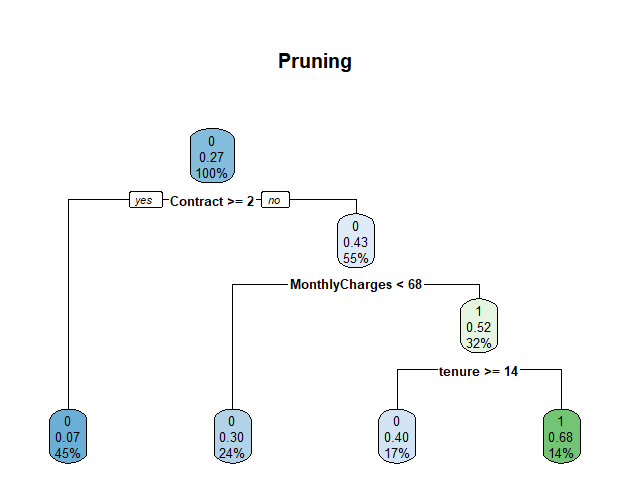
* 1. **Information splitting strategy**

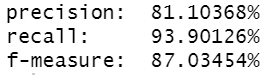
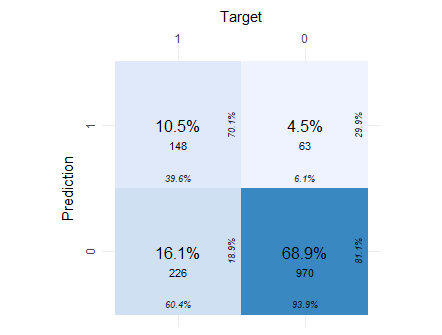






**Gini index and information gain splitting strategies is used but it shows no improvement in the accuracy**

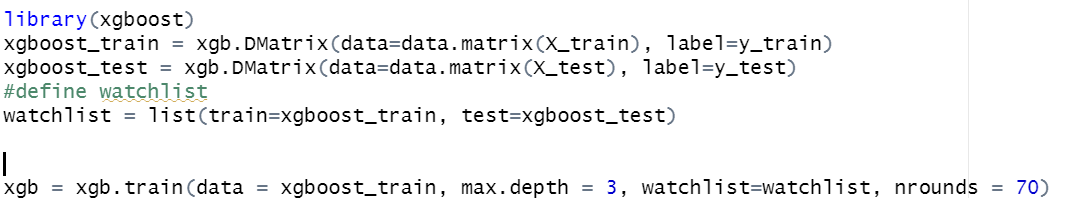
* 1. **Pruning**

****

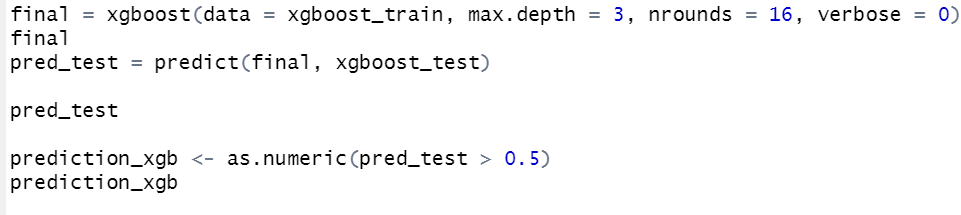
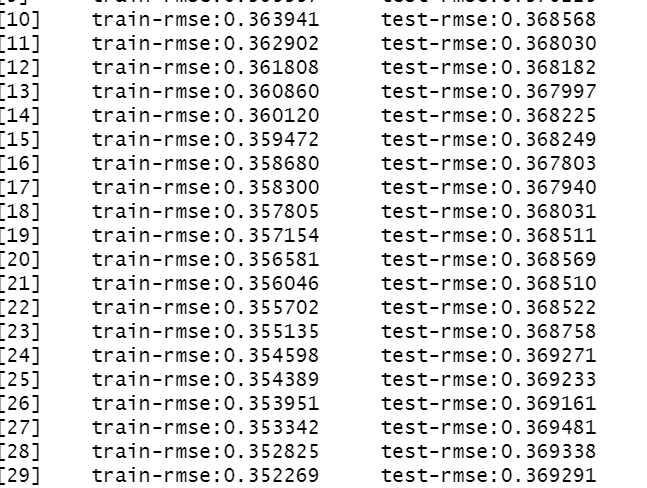
**After pruning the tree, the accuracy slightly improved with accuracy 79.46%**

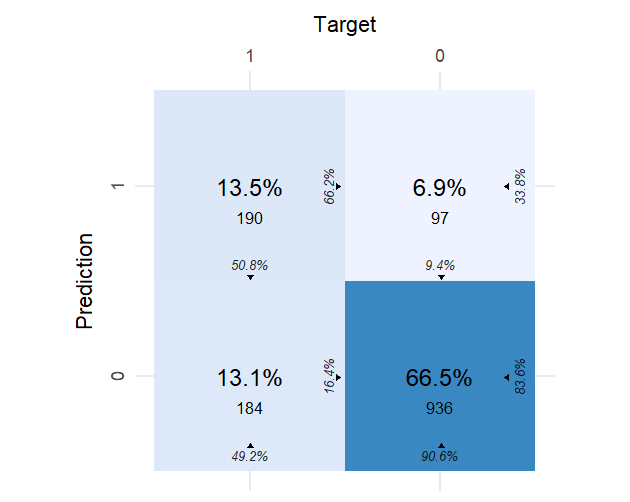
## 5 **Classify the data using the XGBoost model**

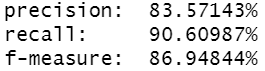
Fit XGBoost model and display training and testing data at each round

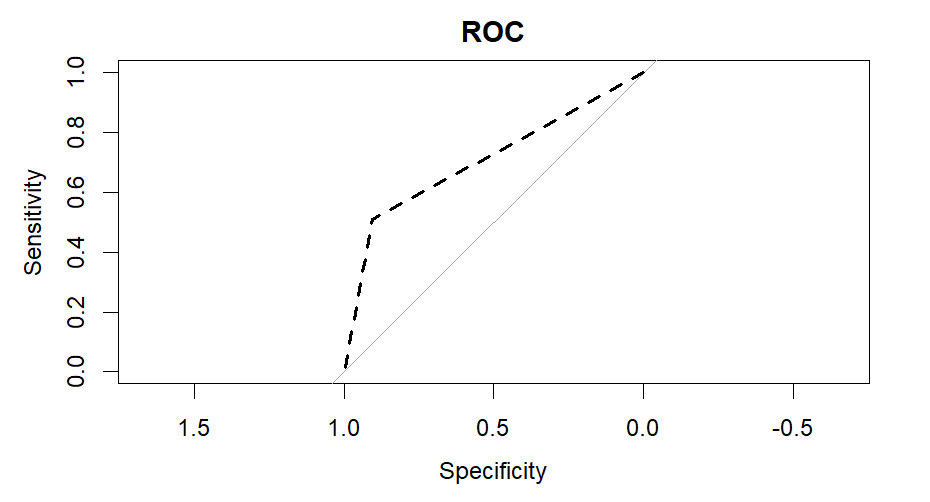


From the output we can see that the minimum testing RMSE is achieved at 16 rounds. Beyond this point, the test RMSE actually begins to increase, which is a sign that we’re overfitting the training data. define final model



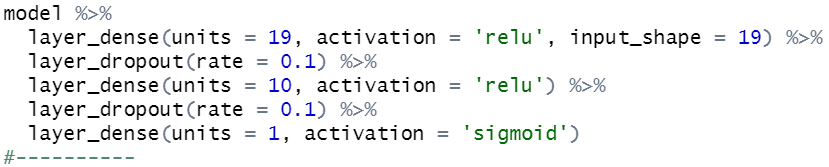


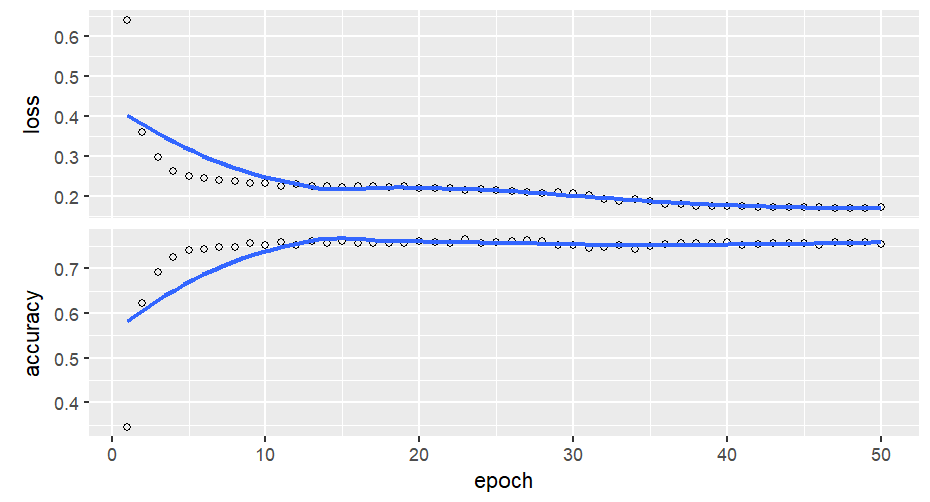
 

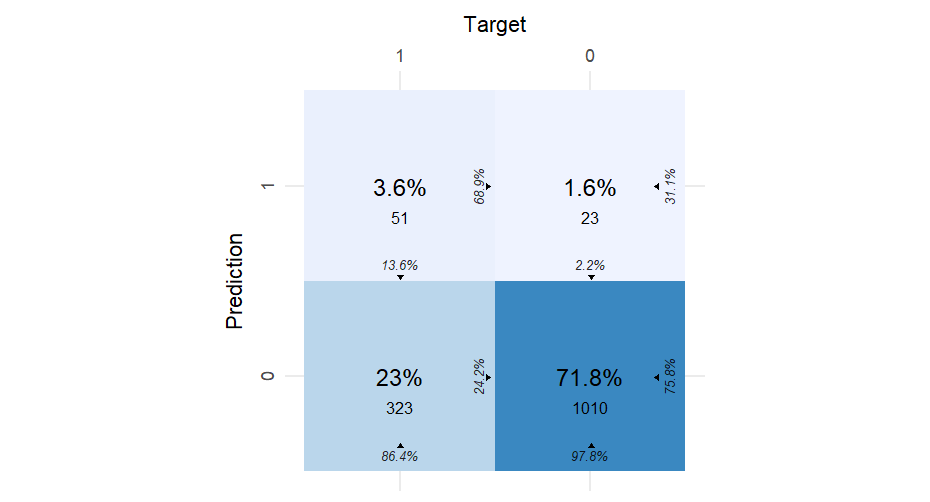


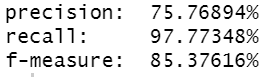
## 5 **Train a deep neural network using Keras with 3 dense layers.**

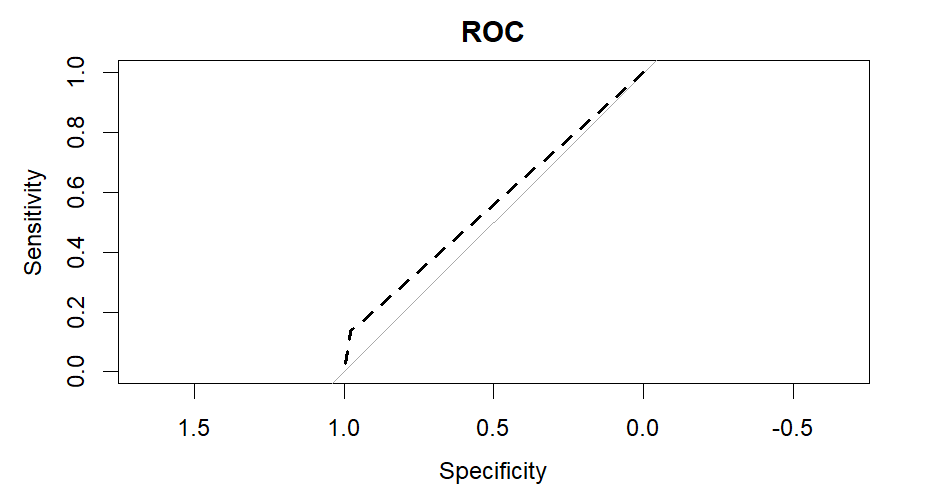
**DNN Model 1**

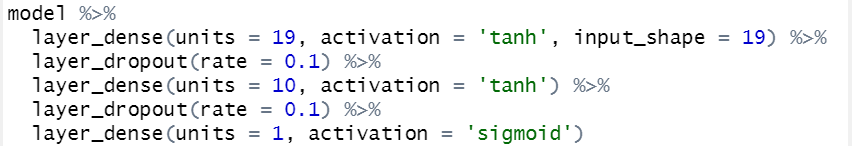




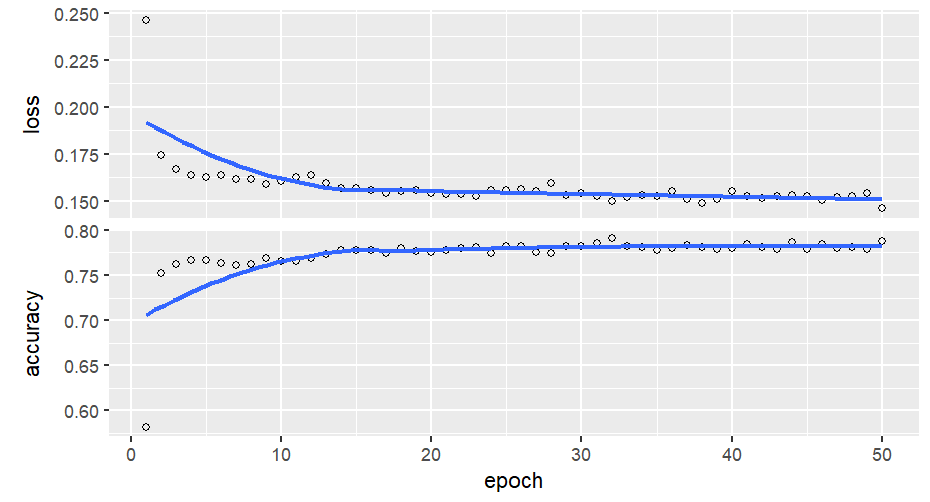


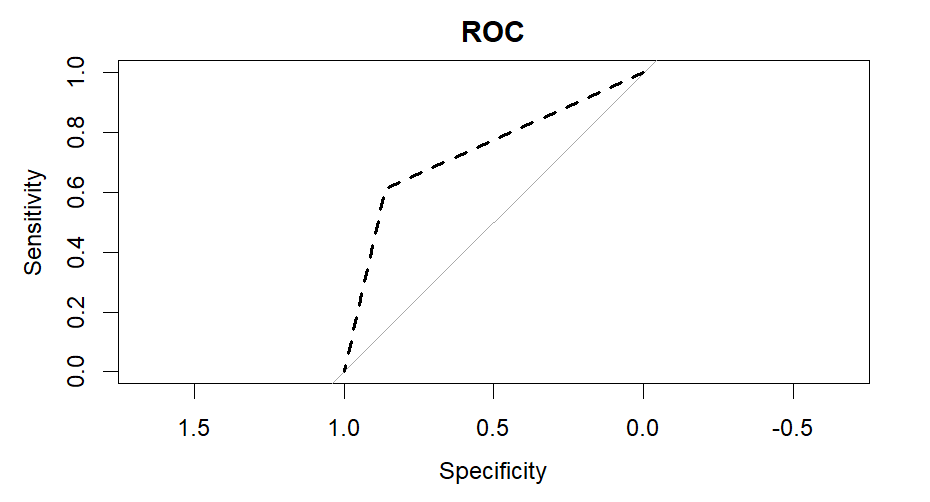
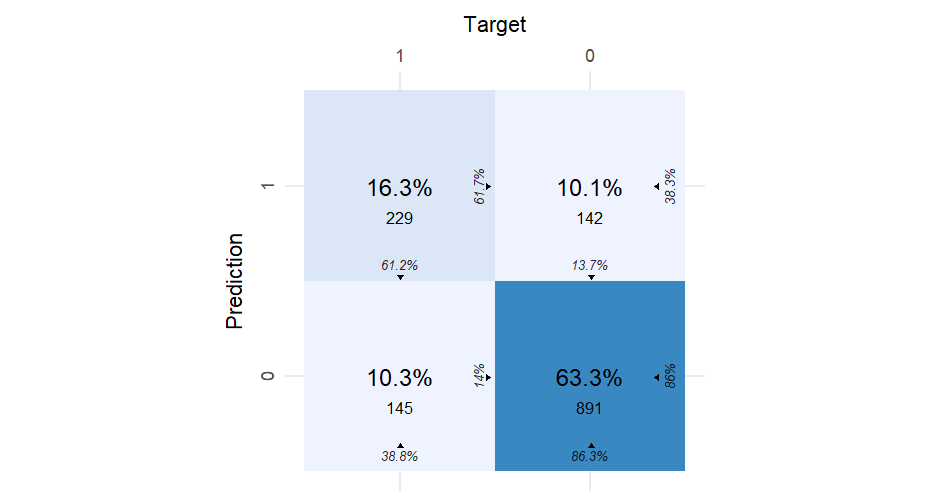
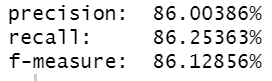


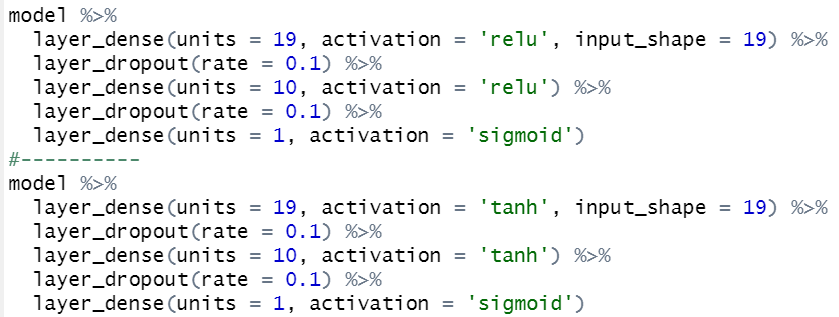


5.2 **Try changing the activation function or dropout rate.**

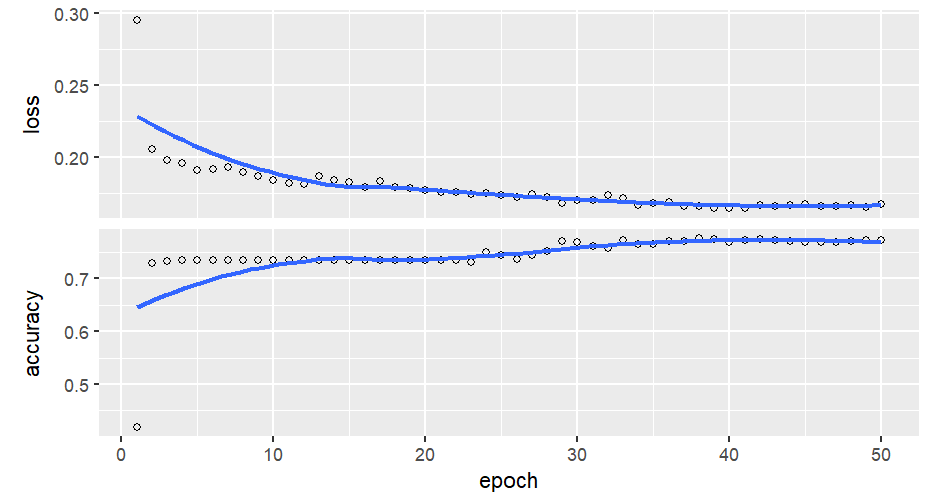
**DNN Model 2**

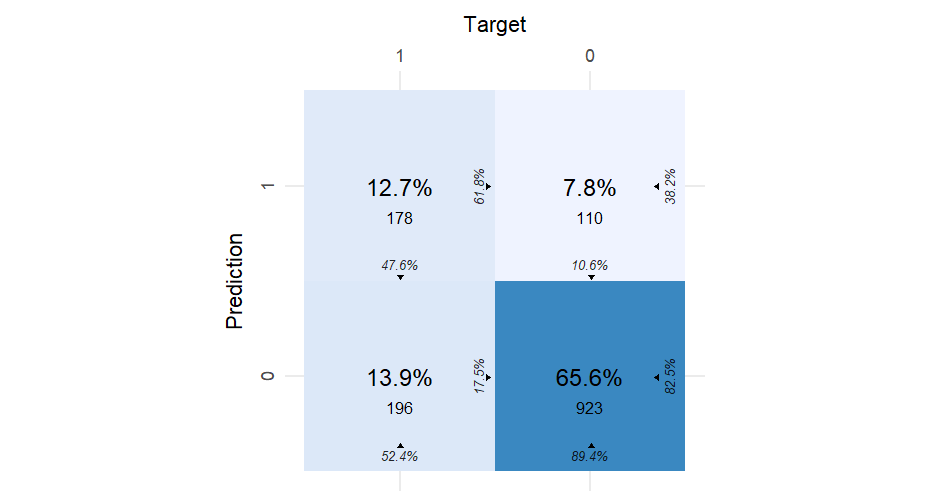
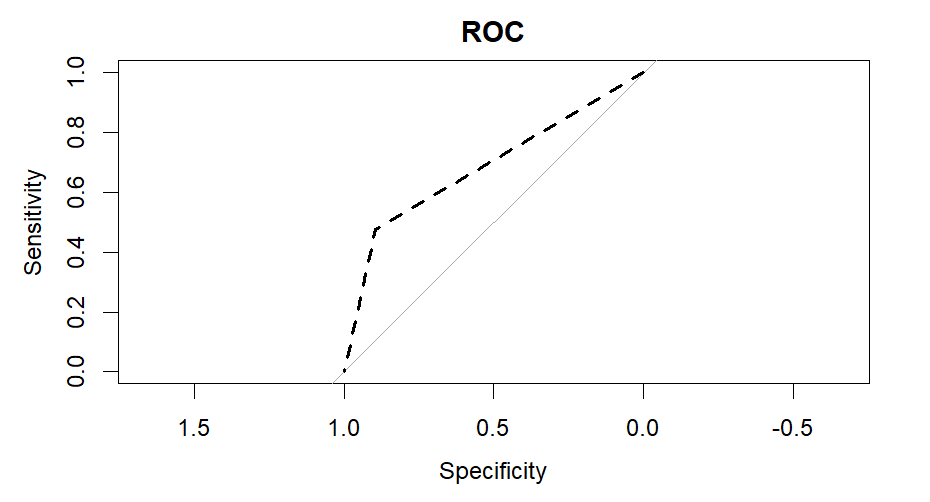


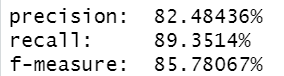




**DNN Model 3**



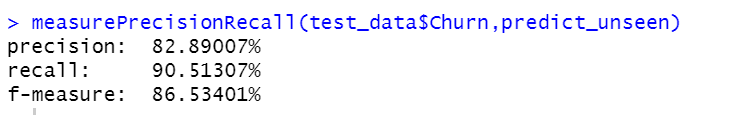
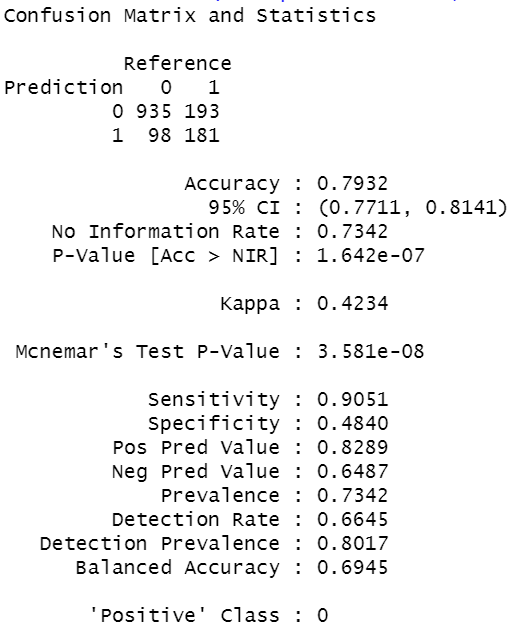


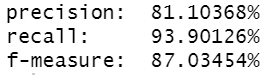


v

## 7 **Compare the performance**

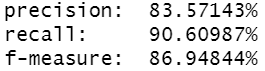
**Decision tree model**



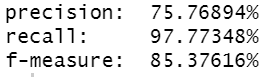
**Pruning DT**

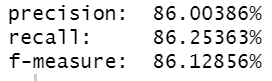


**XGboost**



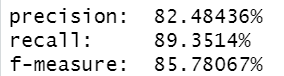
**DNN model1**



**DNN model2**



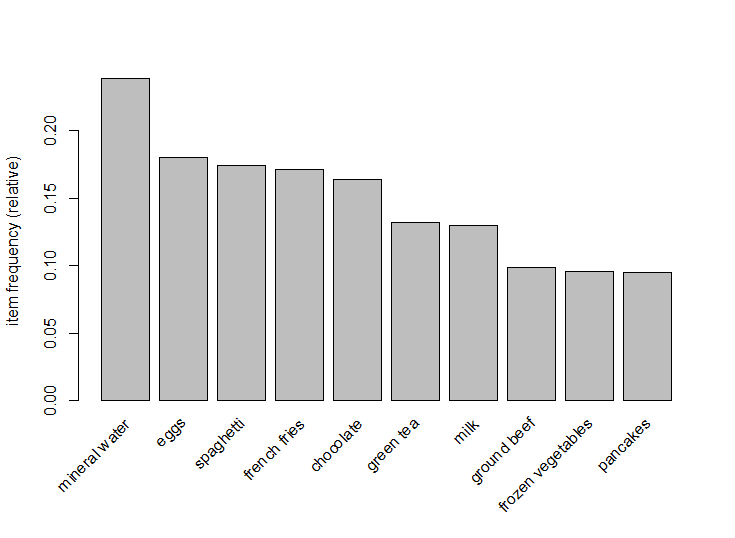
**DNN model3**

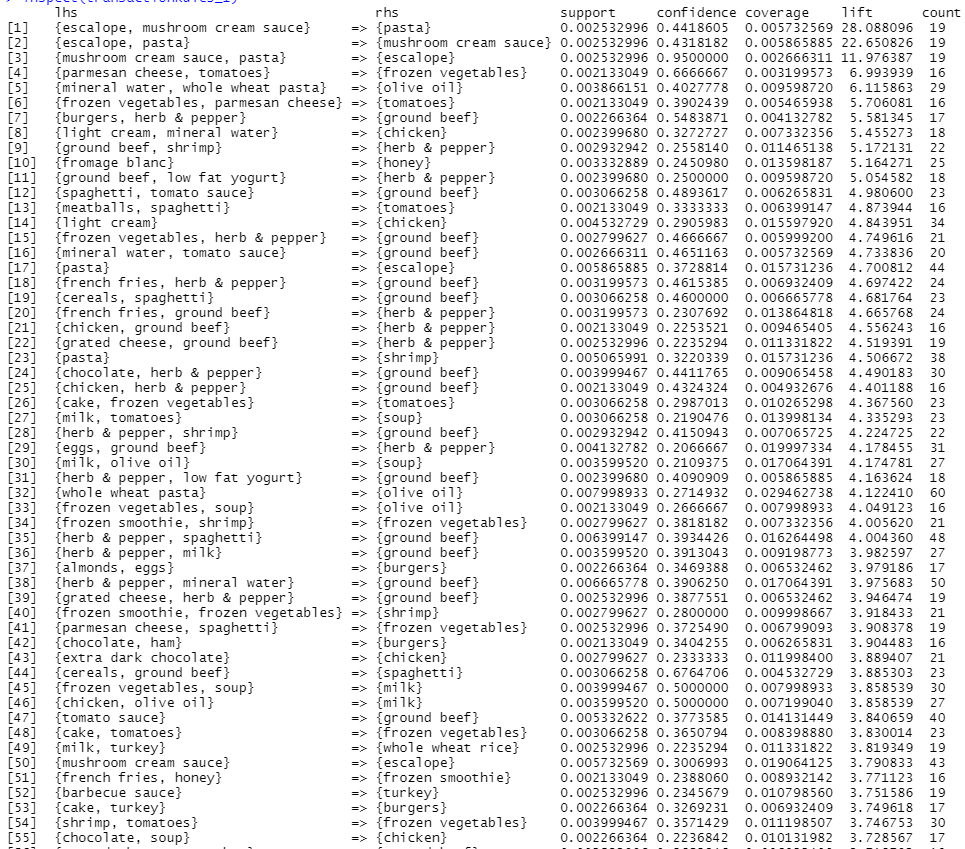


Best Model **: XGBoost**

Worst Model: **DNN model1**

**Part B)**

a)

b)

c)

1- Rule\_number\_1 when maximum length =3:



2- Rule\_number\_2 when maximum length =2:



1. Rule\_number\_1 has higher lift value = 28.0881, but Rule\_number\_2 has a higher support = 0.00333.
2. I would select the Rule\_number\_1, because the lift is much higher than Rule\_number\_2 meaning that this rule is more important. Also the confidence for Rule\_number\_1 is higher.