**Project Report**

Contents

[Customer Churn Classification 2](#_Toc109134258)

[1-Scatter plot matrix and heatmap 2](#_Toc109134259)

[2-Ensure data is in the correct format for downstream processes 3](#_Toc109134260)

[3-Split the dataset into 80 training/20 test set and fit Decision tree 4](#_Toc109134261)

[4-Try different ways to improve the decision tree algorithm 5](#_Toc109134262)

[5-Classify the data using the XGBoost model 7](#_Toc109134263)

[6-Train a deep neural network using Keras with 3 dense layers 8](#_Toc109134264)

[7-Compare the performance using (precision, recall, accuracy, F-measure) 9](#_Toc109134265)

[8-ROC graph to compare the performance of the DT, XGboost & DNN techniques 12](#_Toc109134266)

# Customer Churn Classification

## 1-Scatter plot matrix and heatmap

Graphical user interface, application

Description automatically generated

Chart, box and whisker chart

Description automatically generated

## 2-Ensure data is in the correct format for downstream processes

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

## 3-Split the dataset into 80 training/20 test set and fit Decision tree

Graphical user interface, text, application, email

Description automatically generated

A picture containing diagram

Description automatically generated

As shown in the figure We can deduce that our model chooses the most important 6 features and decide to make his decisions on them.

Cases where the model decide to say YES are:

1-ContractMonth.toMonth=no and internetServiceFiber.optic=1 and tenure <6 and techSupportNo =0

2- ContractMonth.toMonth=no and internetServiceFiber.optic=0 and tenure >=16

3- ContractMonth.toMonth=no and internetServiceFiber.optic=0 and tenure <16 and PaymentMethodElectronic.check =0 and totalCharge >=3084

Cases where the model decide to say NO are:

1- ContractMonth.toMonth=no and internetServiceFiber.optic=0 and tenure <16 and PaymentMethodElectronic.check =0 and totalCharges <3084

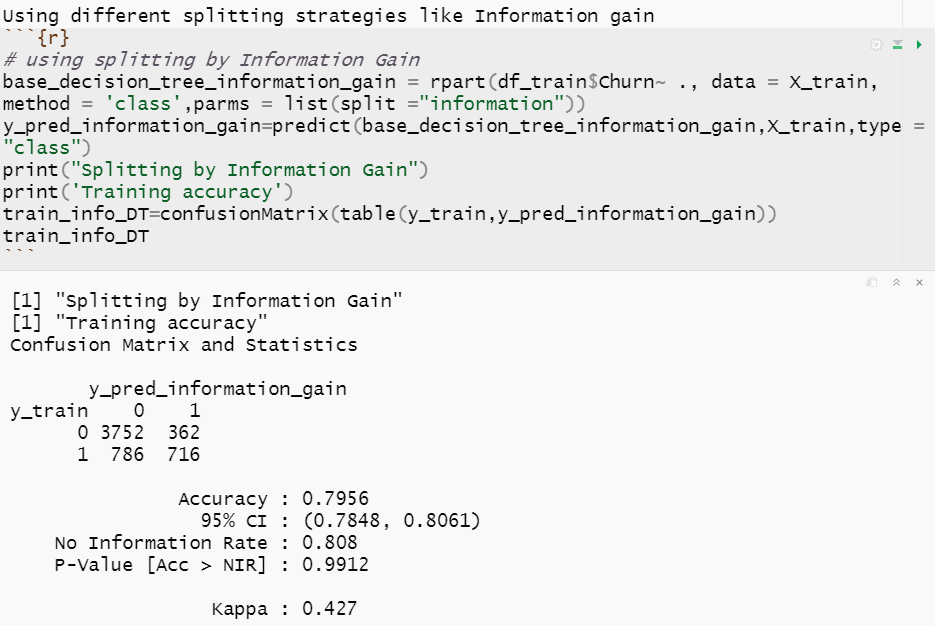
2- ContractMonth.toMonth=no and internetServiceFiber.optic=0 and tenure <16 and PaymentMethodElectronic.check =1

3- ContractMonth.toMonth=no and internetServiceFiber.optic=1 and tenure >=6 and techSupportNo =1

4- ContractMonth.toMonth=no and internetServiceFiber.optic=1 and tenure <6

5- ContractMonth.toMonth=yes

## 4-Try different ways to improve the decision tree algorithm



Text

Description automatically generated

We can deduce that the information gain accuracy is almost the same as Geni index

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

preprunning increased the accuracy of training data, but the accuracy of testing decreased which means overfitting

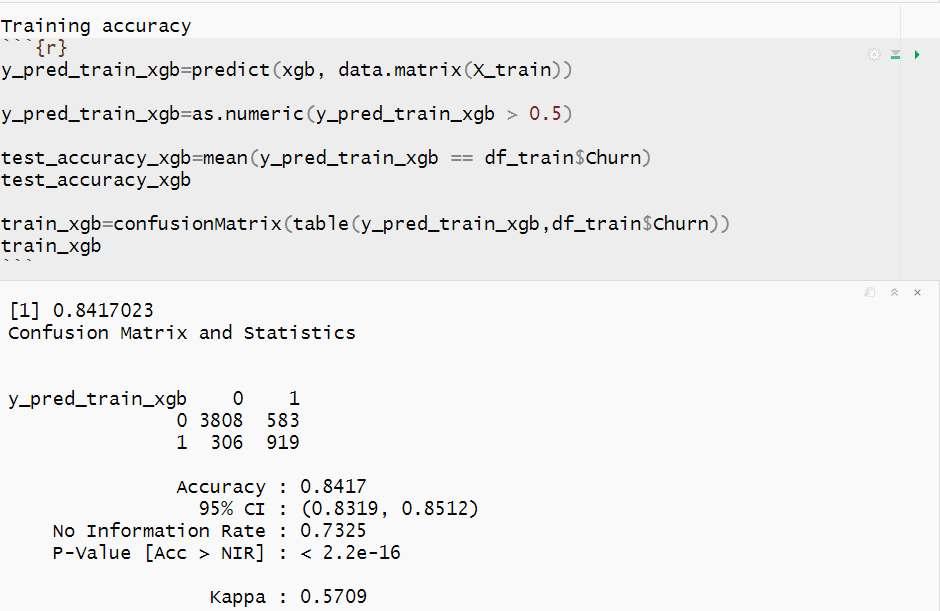
Graphical user interface, text, application

Description automatically generated

## 5-Classify the data using the XGBoost model

Graphical user interface, application

Description automatically generated

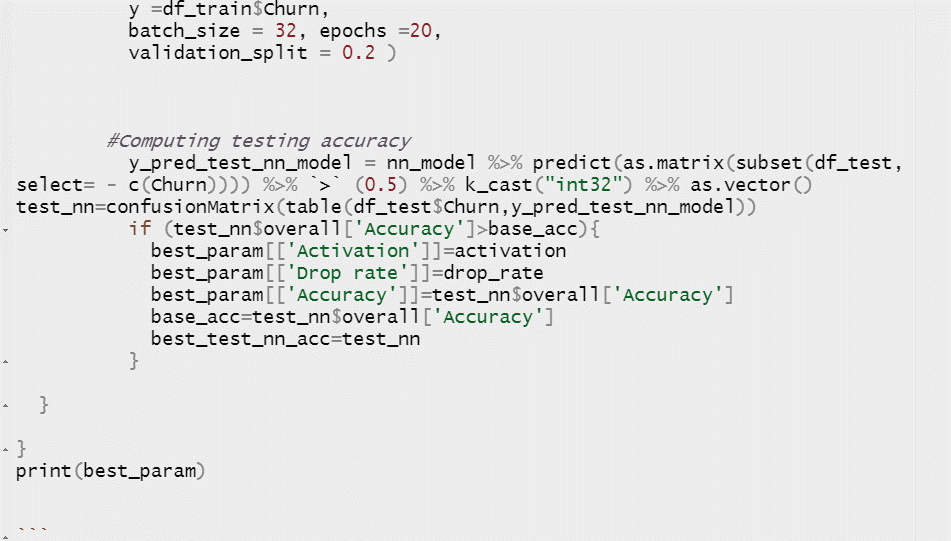


Text

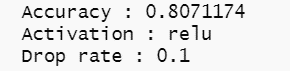
Description automatically generated

We can notice that there is **no overfitting**

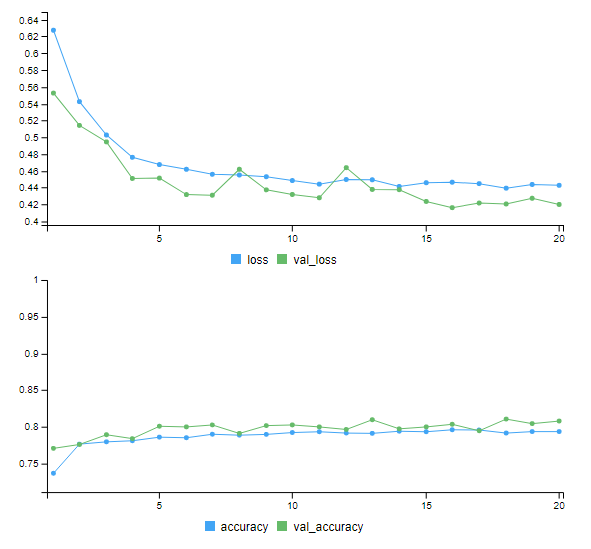
## 6-Train a deep neural network using Keras with 3 dense layers



From all these experiments I got the best parameters



This is the visualization of the experiment with the best parameter (relu ,0.2)



## 7-Compare the performance using (precision, recall, accuracy, F-measure)

To loop over all models, I stored them in a list as following

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

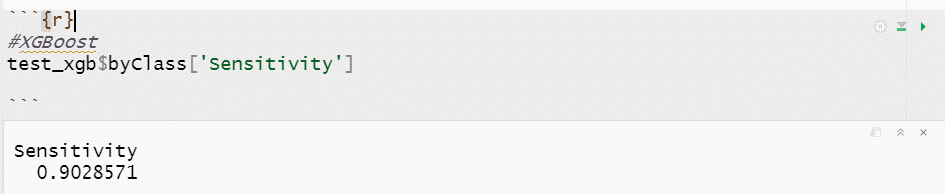
Description automatically generated



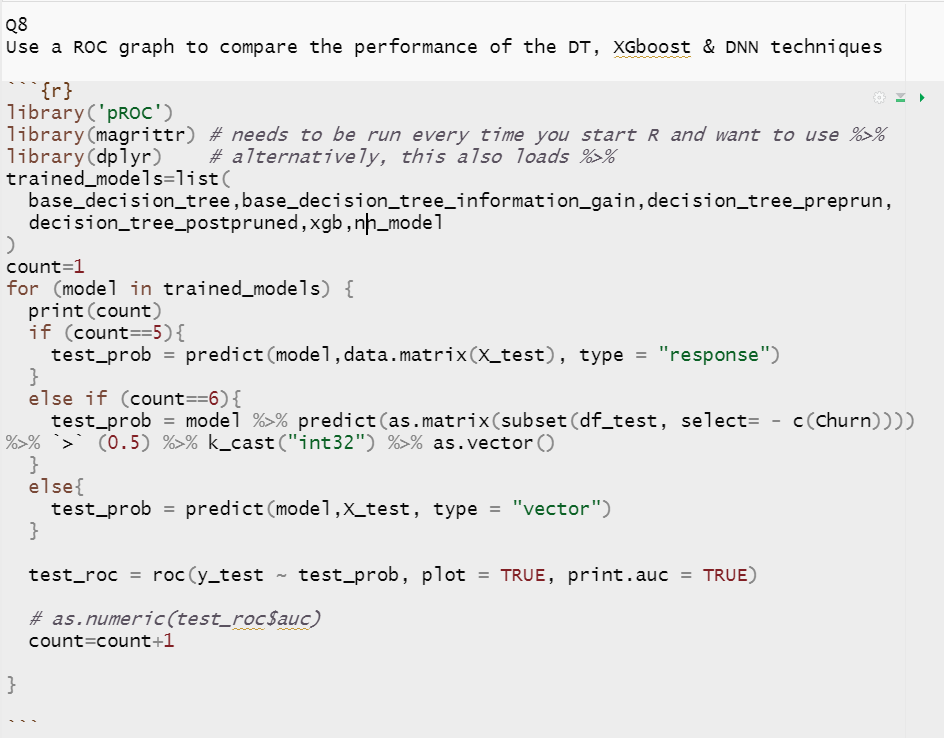
Choosing the Best model

According to this business case:

The highest recall model is the best because it cares about the false negative cases and this is the most important because predicting that the customers who will leave will stay are critical and harmful, but predicting that the customers who will stay they will leave are not harmful, so the best model will be **XGBoost**

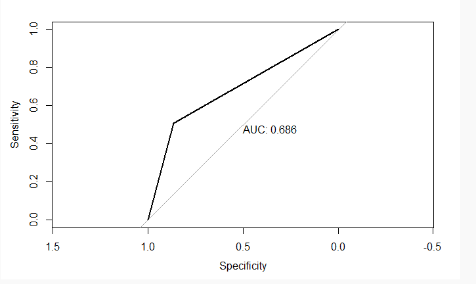
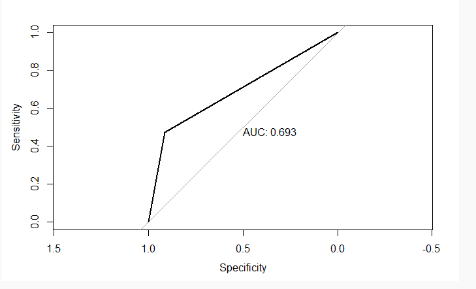


## 8-ROC graph to compare the performance of the DT, XGboost & DNN techniques

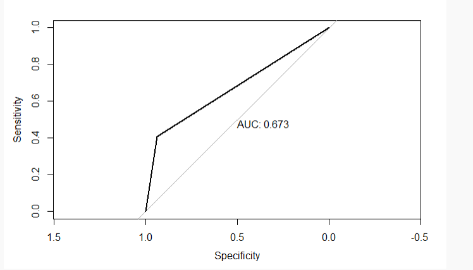
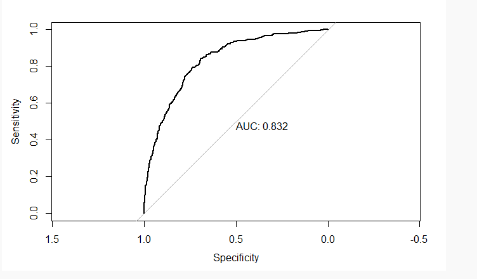
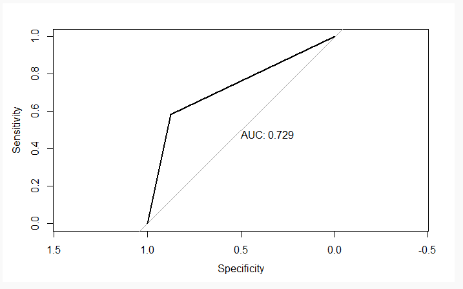


Base DT Base DT with info Pre-pruning DT

A picture containing diagram

Description automatically generated

Post-pruning DT XGBoost DNN model



**The best ROC accuracy is XGBoost with AUC =0.832**