

**DATA MINING GROUP ASSIGNMENT**

# **PGP-BABI FEB18, GROUP-7**

**Team Members**

Vijay Velpula,

Nitesh Saurabh

Priyank Mehar

# **Problem Statement:**

**Build Neural Network and Decision tree model from the following Employee Attrition data:**

HR\_Employee\_Attrition\_Data.csv[View in a new window](https://classroom.greatlearning.in/courses/1421/files/89783/download?wrap=1)

HR\_Employee\_Attrition\_Metadata.xlsx[View in a new window](https://classroom.greatlearning.in/courses/1421/files/89782/download?wrap=1)

**Steps involved should be:**

**a) Data Import (Target variable is "Attrition" column)**  
**b) Split the data in Dev & Hold Out sample (70:30)**  
**c) Perform Exploratory Data Analysis(EDA)**  
**d) Identify columns which are of no use. drop those columns**  
**e) Write Hypothesis and validate the Hypothesis**  
**f) Build Neural Network Model (Development sample)**  
**g) Validate NN model on Hold Out. If need be improvise**  
**h) Build CART Model**  
**i) Validate CART Model**  
**j) Compare NN with CART**  
**k) Combine NN and CART into Ensemble Model**  
**l) Check whether Ensemble Model Performance outperforms the individual CART & NN model**

**Approach:**

In the given attrition dataset and the metadata target is to predict employee leaving the organization. In the dataset, field “Attrition” covers the same with values “Yes” and “No”.

So, the plan is to do the following:

* Create CART model and check the performance
* Create NN model and check the performance

1. **R-Code for Exploratory Data Analysis (EDA) & CART Model:**

# Importing HR\_Employee\_Attrition\_Data.csv data into R dataframe using read.csv function.

hrdata <- read.csv(file.choose())

View(hrdata)

## Exploratory Data Analysis (EDA)

# To check the variables, their data types and sample data

str(hrdata)

#To check the statistical summary of each variable

summary(hrdata)

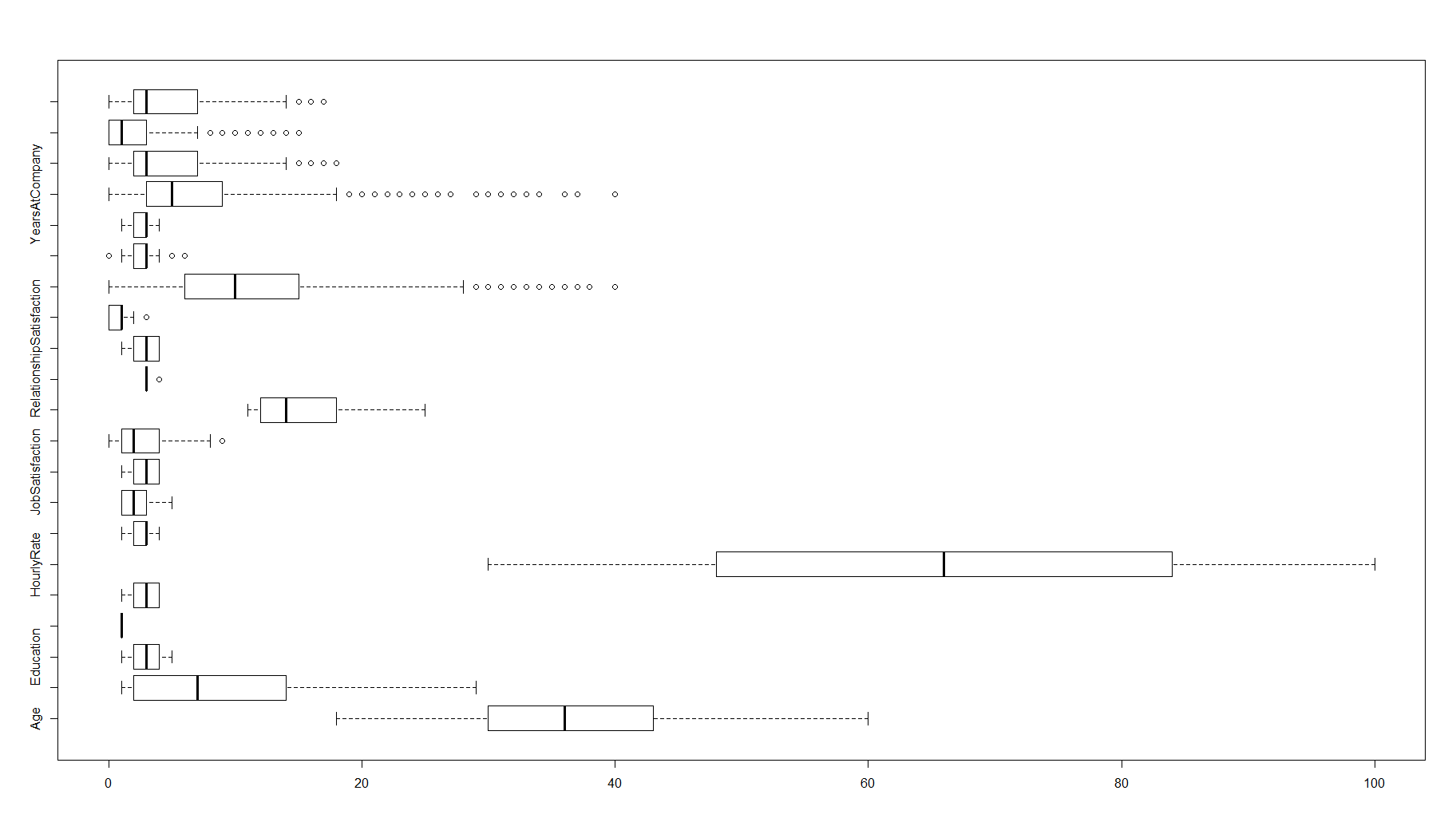
#To select all the numeric variables data

isnum <- unlist(lapply(hrdata, is.numeric))

nums <- hrdata[, nums]

#Exploratory Data Analysis-1

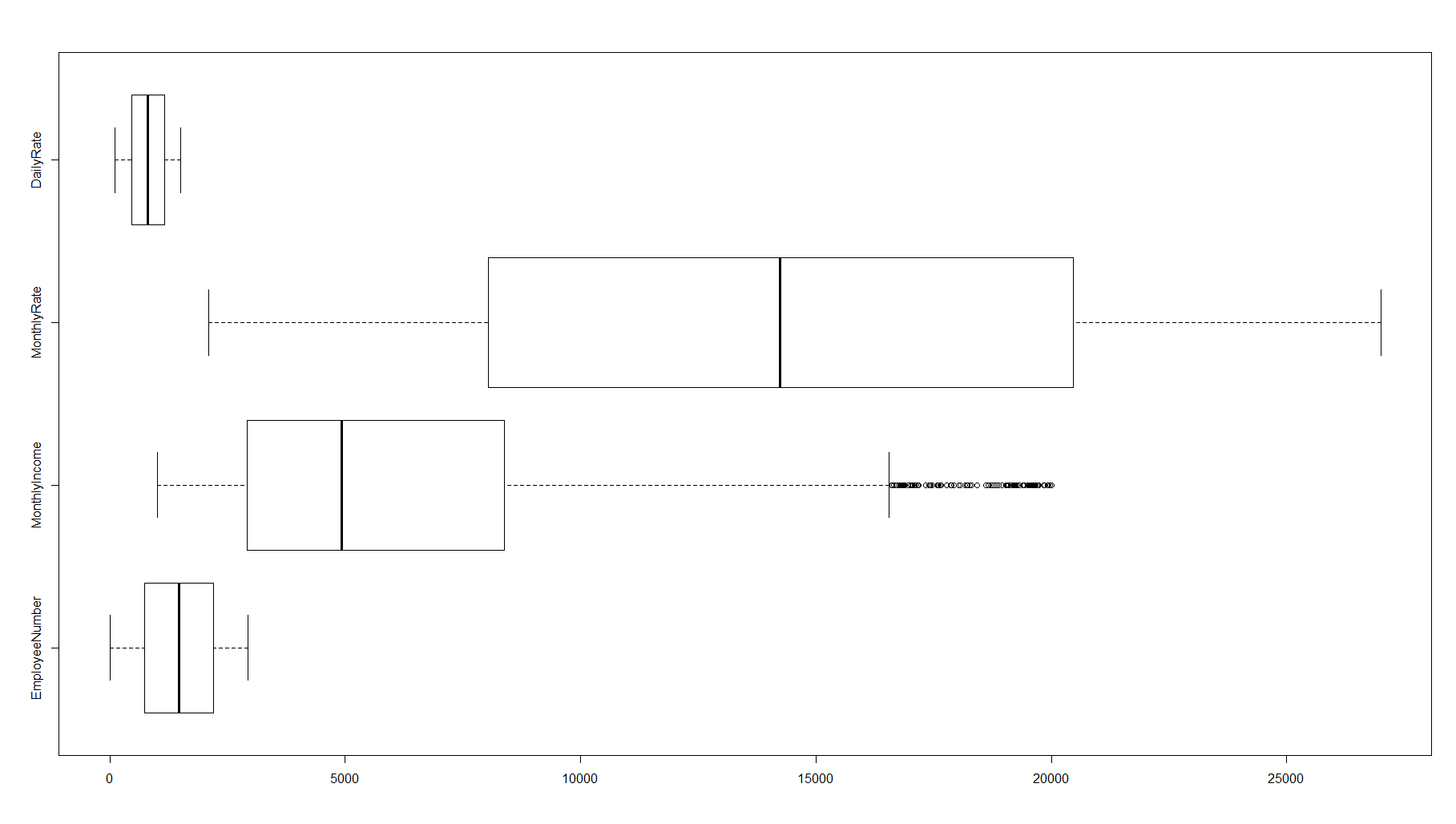
boxplot(nums[,c("Age","DistanceFromHome","Education","EmployeeCount","EnvironmentSatisfaction","HourlyRate","JobInvolvement","JobLevel","JobSatisfaction","NumCompaniesWorked","PercentSalaryHike","PerformanceRating","RelationshipSatisfaction","StockOptionLevel","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBalance","YearsAtCompany","YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager")], horizontal=TRUE)

****

**Fig1:** **Distribution of numeric data variables**

#Exploratory Data Analysis-2

boxplot(nums[,c("EmployeeNumber","MonthlyIncome","MonthlyRate","DailyRate")], horizontal=TRUE)

**Fig2:** **Distribution of numeric data variables (At different scale)**

#Checking for Null

sum(is.na(hrdata))

#Result: 0, Means there are No Null values in the data

#Checking for missing values

t(apply(is.na(hrdata), 2, sum))

sum(is.na(hrdata))

#Result: 0, Means there are No missing values in the data

#Checking for Duplicates

sum (is.na(duplicated(hrdata)))

#Result: 0, Means there are No Duplicates

#Checking the important factor contributing to attrition

prop.table(table(STG\_HR\_EADF$OverTime))

table(STG\_HR\_EADF$OverTime, STG\_HR\_EADF$Attrition)

#Result:

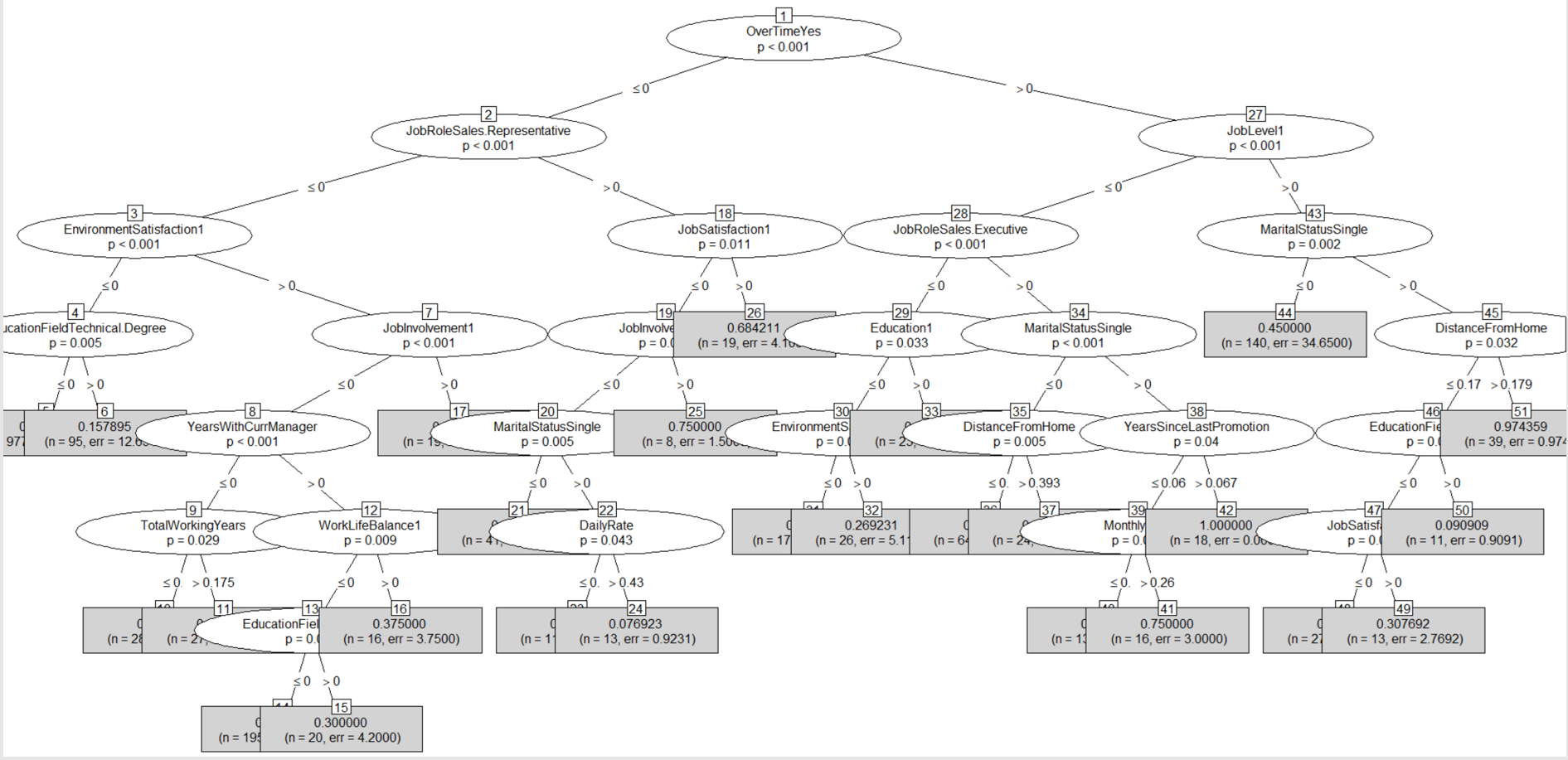
No Yes

0.7170068027 0.2829931973

No Yes

No 1888 220

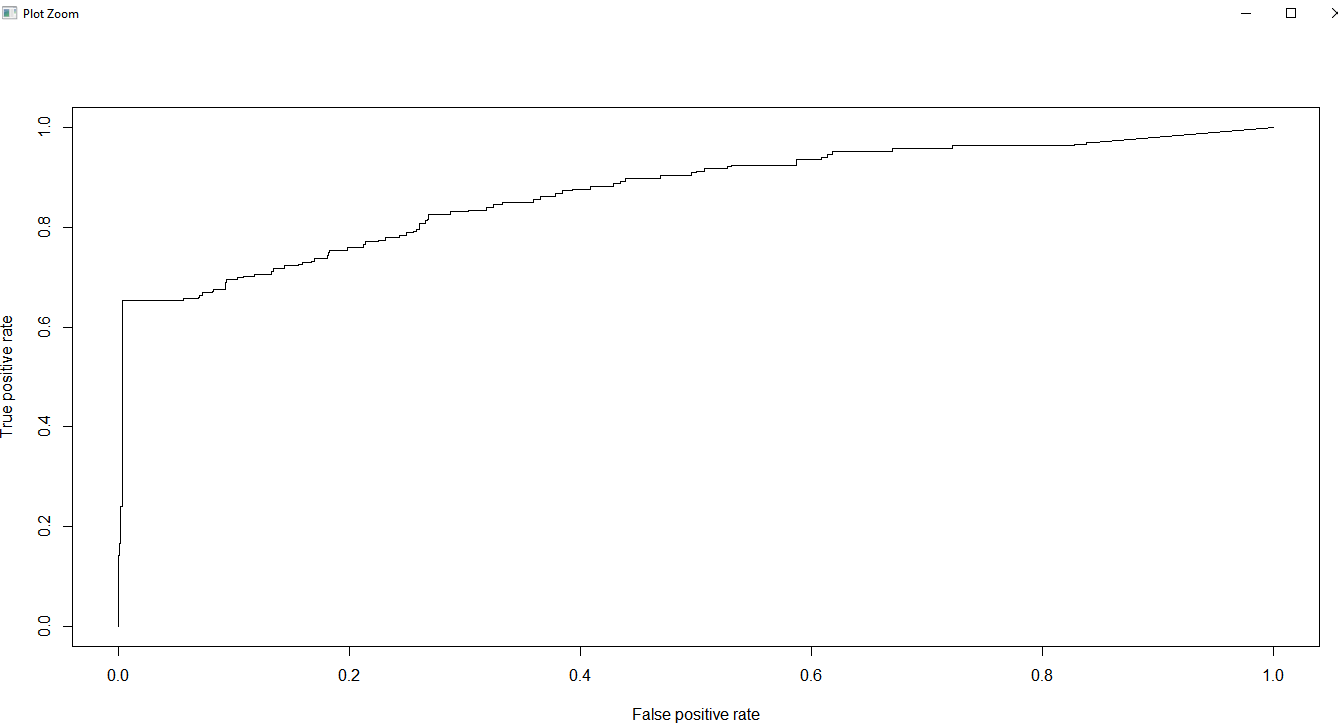
Yes 578 254



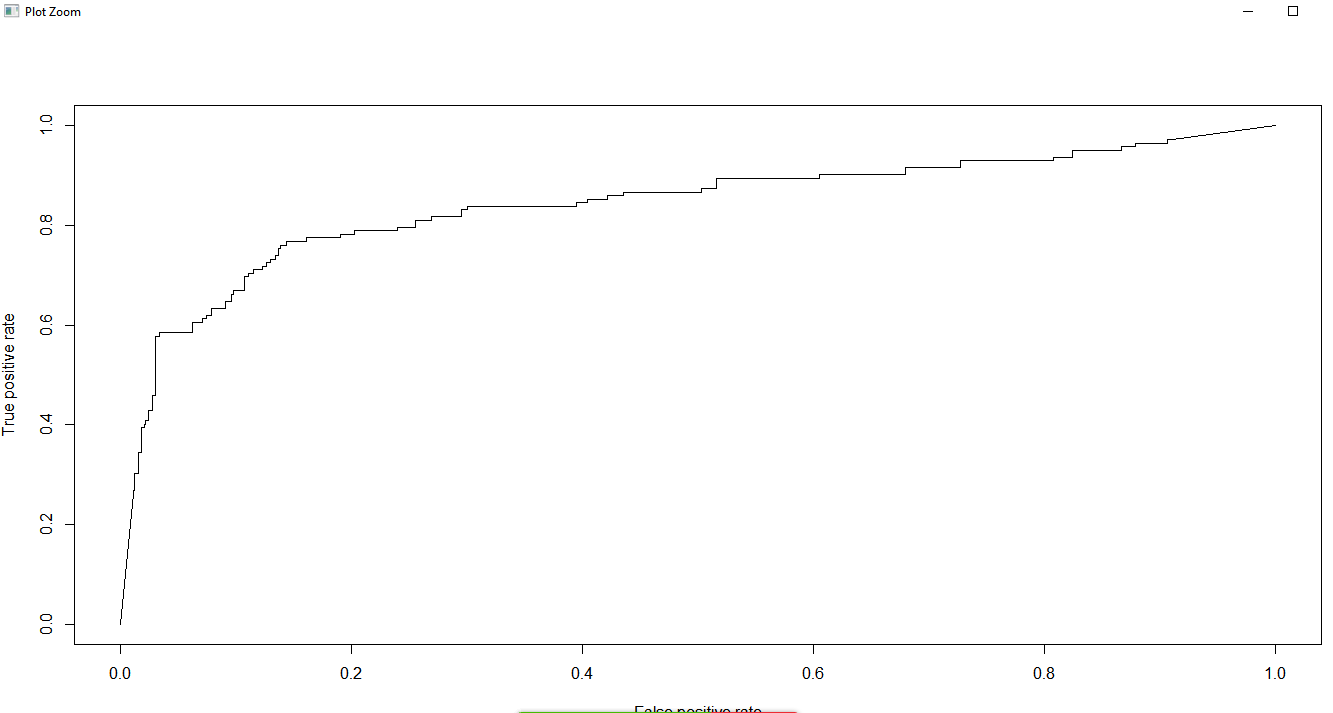
1. **NEURAL NETWORK**

****

ROC Output on Training Data



ROC Output for Test Data

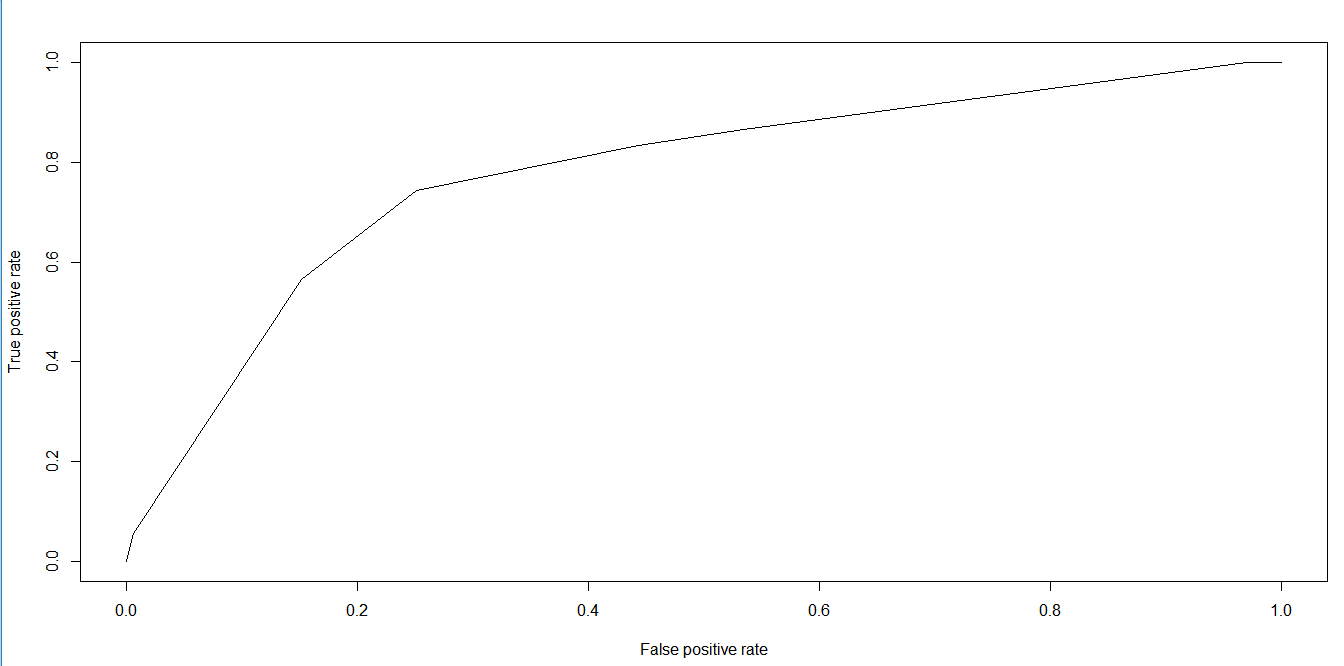


Comparing above ROC, it is clear, the model is predicting as expected.

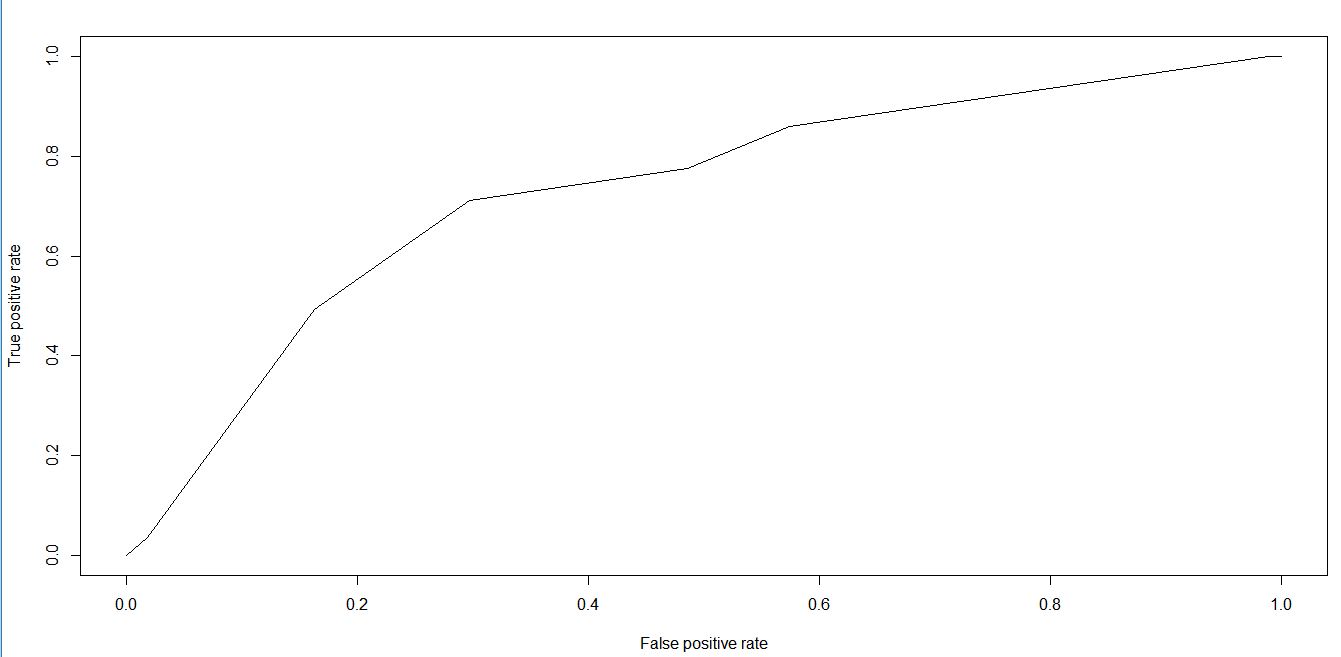
1. **CART**

****

ROC Output on Training Data



ROC Output for Test Data



Comparing above ROC, it is clear, the model is predicting as expected.

1. **Comparison of Neural Networks to CART model**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Neural Networks** | | | | **CART** | | | |
|  | AUC | KS | Gini | Error Rate | AUC | KS | Gini | Error Rate |
| **Train** | 0.871070028 | 0.650138 | 0.607604 | 0.059732 | 0.776236 | 0.492682 | 0.288176 | 0.339623 |
| **Test** | 0.841968024 | 0.624362 | 0.642722 | 0.1025 | 0.728807 | 0.414921 | 0.296357 | 0.41868 |

By looking at the above comparison, we conclude that Neural networks outperforms CART in predicting the model

Also we can draw some critical points:

1. Overtime is the singlemost important factor contributing towards attrition.
2. Young people with lesser years at the company contribute to the attrition.