**Title: Speed dating experiment**

**Group Number: 130**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| First Name | Last Name | Online Students? (Y or N) | Monday or Tuesday | Shared with ITMD 525? (Y or N) |
| Vigneshwaarar | CR | N | Tuesday | N |
| Sridhar | Srinivasan | N | Tuesday | N |
| Vinod | Selvam | N | Tuesday | N |

Note, if you are an online student or remote student from India, you belong to Monday class

Table of Contents

[**1.** **Introduction and Motivations** 2](#_Toc481533059)

[**2.** **Data Description** 2](#_Toc481533060)

[**3.** **Research Problems and Solutions** 3](#_Toc481533061)

[**4.** **Model Learning** 4](#_Toc481533062)

[4.1. Data Processing 4](#_Toc481533063)

[4.2. Data Analytics Tasks and Processes 4](#_Toc481533064)

[**5.** **Evaluations and Results** 17](#_Toc481533065)

[5.1. Evaluation Methods 17](#_Toc481533066)

[5.2. Results and Findings 20](#_Toc481533067)

[**6.** **Conclusions and Future Work** 22](#_Toc481533068)

[6.1. Conclusions 22](#_Toc481533069)

[6.2. Limitations 22](#_Toc481533070)

[6.3. Potential Improvements or Future Work 22](#_Toc481533071)

# **1. Introduction and Motivations**

* Banking domains initiates various kinds of marketing campaigns to increase their client base. However, in recent times, its effect on the public has reduced. This has lead marketing managers of various banks to invest on directed campaigns by targeting specific customer group based on BI and statistical techniques.
* The motivation behind the project is to find the model that can explain the best possible ways to increase the client base.

# **2. Data Description**

Data was gathered from participants in experimental speed dating events from 2002-2004. During the events, the attendees would have a four minute "first date" with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. The dataset also includes questionnaire data gathered from participants at different points in the process. These fields include: demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information.

* Source of the dataset: [https://www.kaggle.com/annavictoria/speed-dating-experiment](https://www.kaggle.com/annavictoria/speed-dating-experiment%20)
* The data is composed of 8,378 observations and 195 variables.

Features:

iid: unique subject number, group (wave id gender)

id: subject number within wave

gender: Female=0, Male=1

idg: subject number within gender, group (id gender)

condtn: 1=limited choice, 2=extensive choice

wave#: 1 to 21 based on date, preference scale, variation, no of males and females.

round: number of people that met in wave

position: station number where met partner

order: the number of date that night when met partner

partner: partner’s id number the night of event

pid: partner’s iid number

match 1=yes, 0=no

int\_corr: correlation between participant’s and partner’s ratings of interests in

samerace: participant and the partner were the same race. 1= yes, 0=no

age\_o: age of partner

race\_o: race of partner

pf\_o\_att: partner’s stated preference at Time 1 (attr1\_1) for all six axis attributes

dec\_o: decision of partner the night of event

attr\_o: rating by partner the night of the event, for all six axis attributes

Survey filled by participant:

Age: Age of the participant

field: field of study

mn\_sat: Median SAT score for the undergraduate institution where attended.

imprace: scale (1-10) that a person participant date of the same racial/ethinic background?

imprelig: scale (1-10) that a person participant date be of the same religious background?

Zipcode: Place I grew up

income: Median household income

goal: primary goal in finding a partner

date: how frequently participant date

goout: how often participant go out.

career: intended career

career\_c: career coded.

scale (1-10): interests in terms of sports, hiking, gaming. etc

exphappy: scale(1-10)

what participant look for in the opposite sex?

Waves 6-9: Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 1-5 and 10-21: 100 points among the following attributes

Attr1\_1: Attractive

Sinc1\_1: Sincere

Int1\_1: Intelligent

Fun1\_1: Fun

Amb1\_1: Ambitious

Shar1\_1: Shared Interests

Opposite sex looking for a date?

Waves 6-9: Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 1-5 and 10-21: 100 points among the following attributes

attr2\_1: Attractive

sinc2\_1: Sincere

int2\_1: Intelligent

fun2\_1: Fun

amb2\_1: Ambitious

shar2\_1: Shared Interests

How does participant think that the participant measure up?

Participant’s attributes, on a scale of 1-10 (be honest!):

Attr3\_1: Attractive

Sinc3\_1: Sincere

Int3\_1: Intelligent

Fun3\_1: Fun

Amb3\_1: Ambitious

Shar3\_1: Shared Interests

what participant think MOST of participant fellow men/women look for in the opposite sex.

Attr4\_1: Attractive

Sinc4\_1: Sincere

Int4\_1: Intelligent

Fun4\_1: Fun

Amb4\_1: Ambitious

Shar4\_1: Shared Interests

And finally, how do you think others perceive you?

Attr5\_1: Attractive

Sinc5\_1: Sincere

Int5\_1: Intelligent

Fun5\_1: Fun

Amb5\_1: Ambitious

Shar5\_1: Shared Interests

# **3. Research Problems and Solutions**

* Determine the best model to predict the decision of date based on the ratings of the attributes from the given dataset.
* Along with the logistic regression techniques, KNN model, Naïve Bayes model, Decision Tree, Random Forest model techniques have been applied to determine various models for better efficiency.
* Determine the impact of AGE, RACE and INTERESTS in getting a date
* Determine the difference in men and women regarding overestimating and underestimating rates of own attributes.

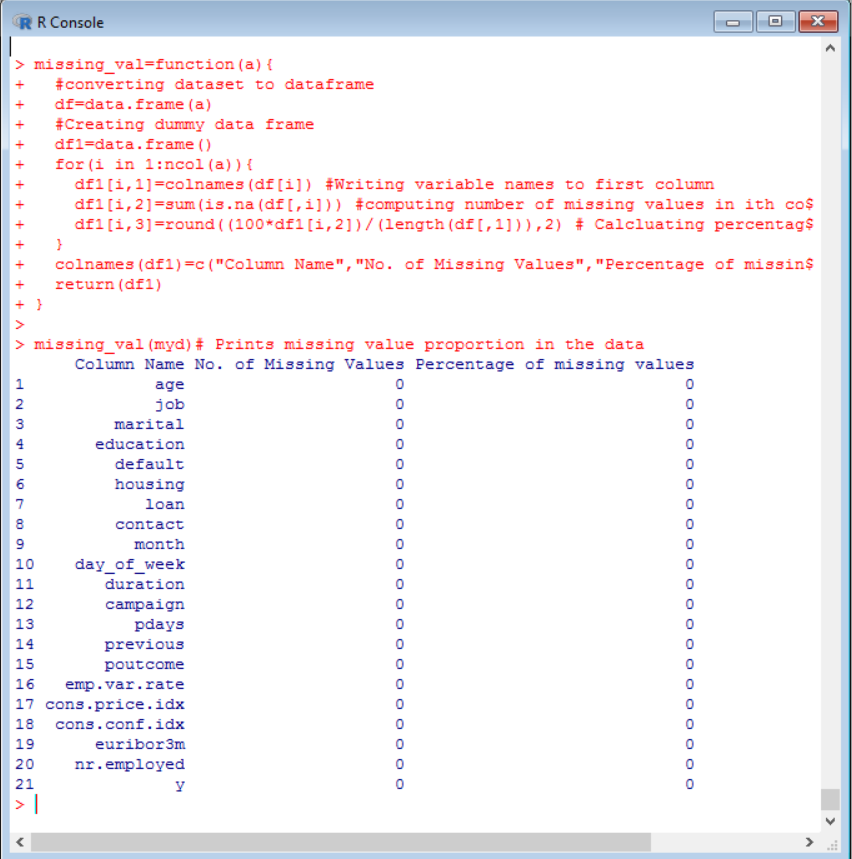
# **4. Model Learning**

## 4.1. Data Processing

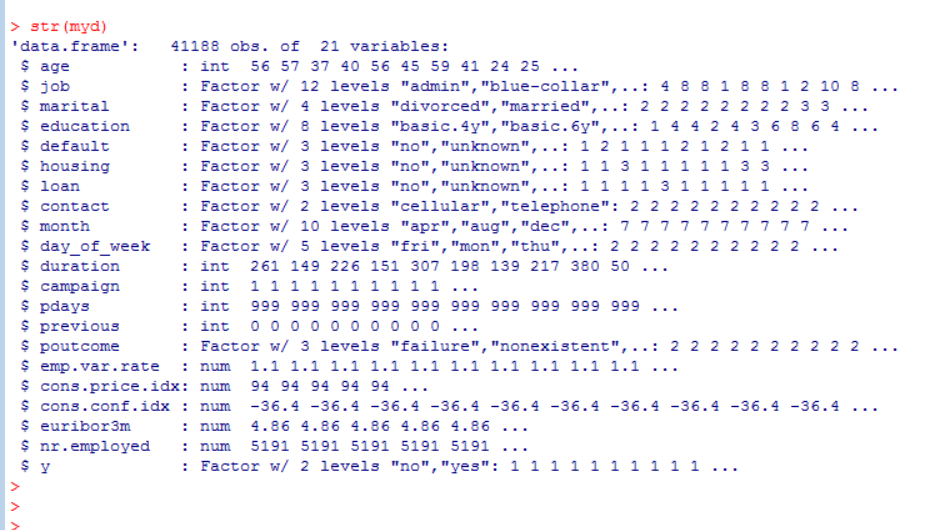
* The following data processing is done to the given dataset.
  + Data cleaning – Searching and removing unwanted characters like “.” “,” “-“ etc. from the obtained CSV file of data.
  + Missing values – Function to detect the missing values in each column and displaying the number of missing values in each row and the percentage of missing values in each parameter.

## 4.2. Data Analytics Tasks and Processes

* Missing value function and its result/output for the given dataset: We can observe that there are no missing values in the dataset.

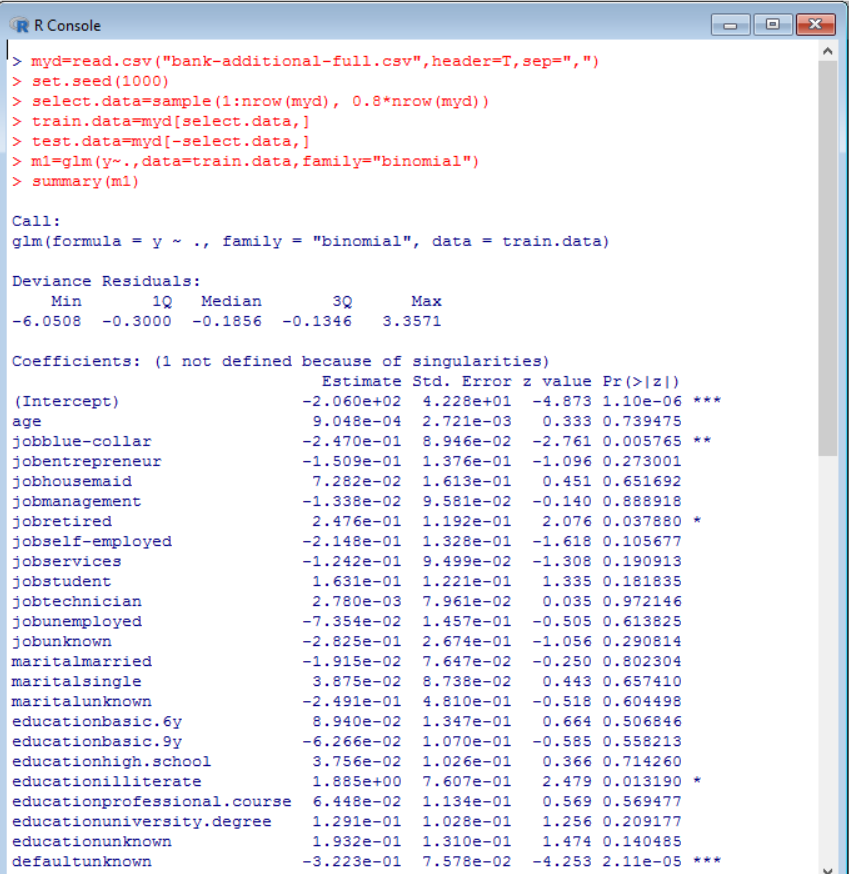


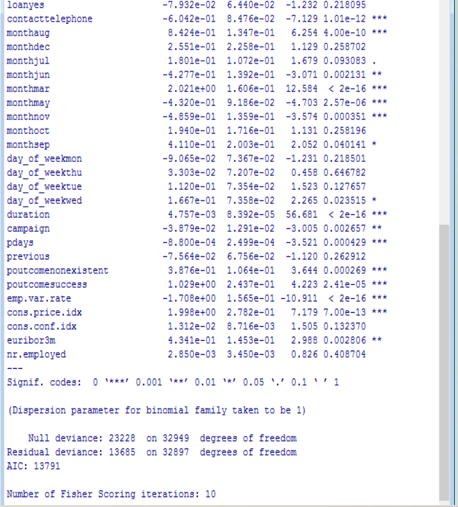
* We have described the structure of the dataset in the below snapshot:



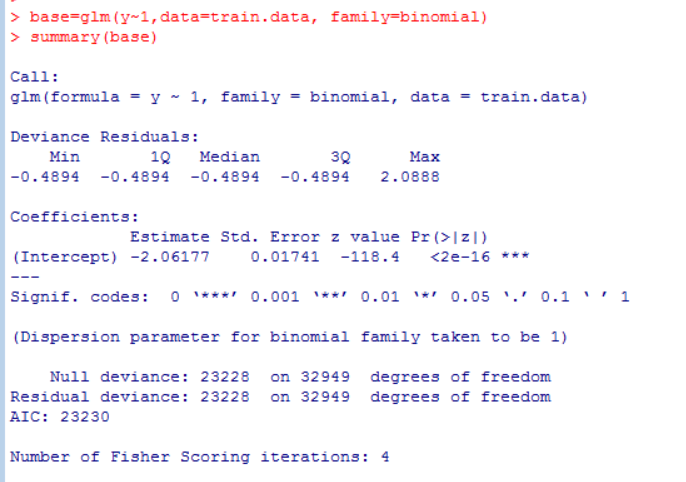
**LOGISTIC REGRESSION**

* + Loading the data from the CSV file and building logistic regression model as shown below. Here we split the data into 80% Training and 20% Testing. We build our models on the training data and use test dataset for predictions and evaluations.

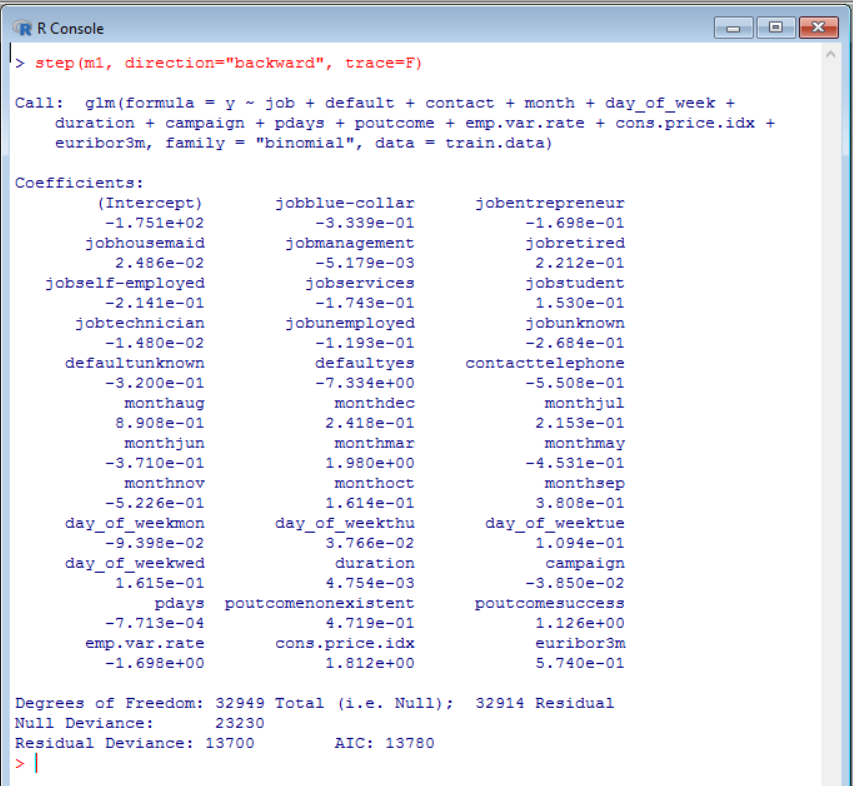




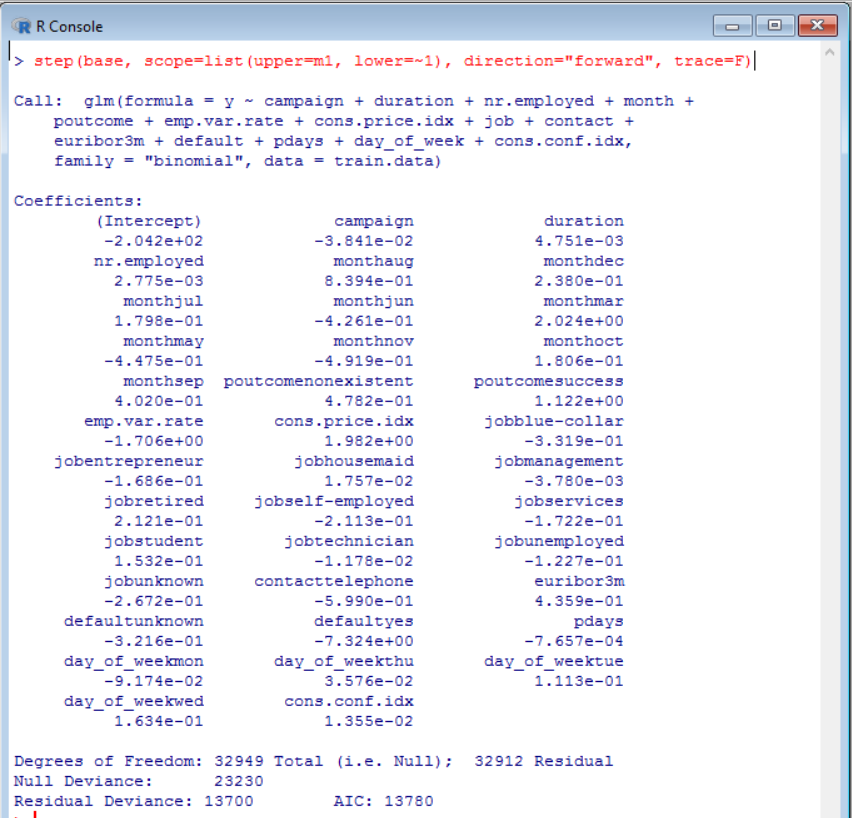
* + Model selection: Creating base model with a no variable (Null model). This base model is further used in step function to build models.



* + Model selection: Backward Elimination is used in the below snapshot. We can observe that there are 12 significant variables with AIC = 13780



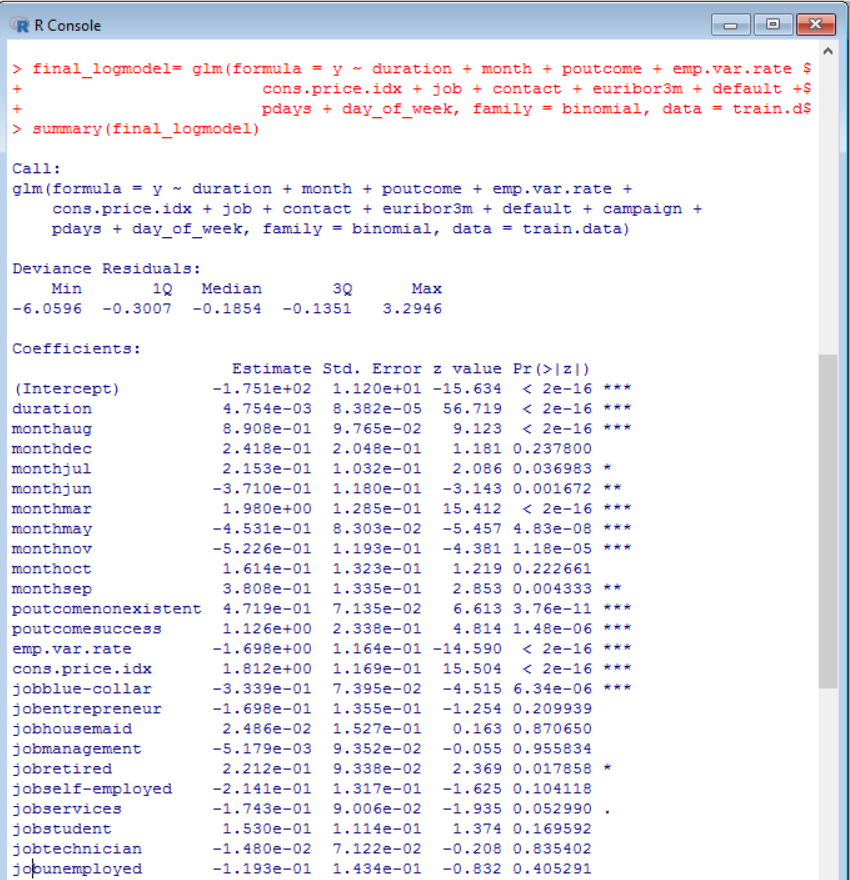
* + Model Selection: Forward Selection is used in the below snapshot. We can observe that there are 14 significant variables with AIC = 13780

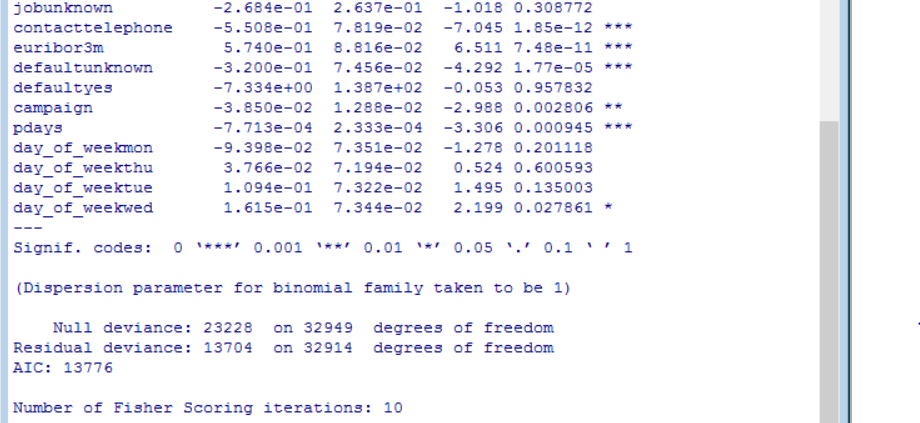


* + Model Selection: Both function is used in the below snapshot. We can observe that there are 12 significant variables with AIC = 13780

# 

* Next, we build the best model based on the backward elimination model. We obtained0 12 significant variables with an AIC of 13780.
  + Fitting the best model as final\_logmodel

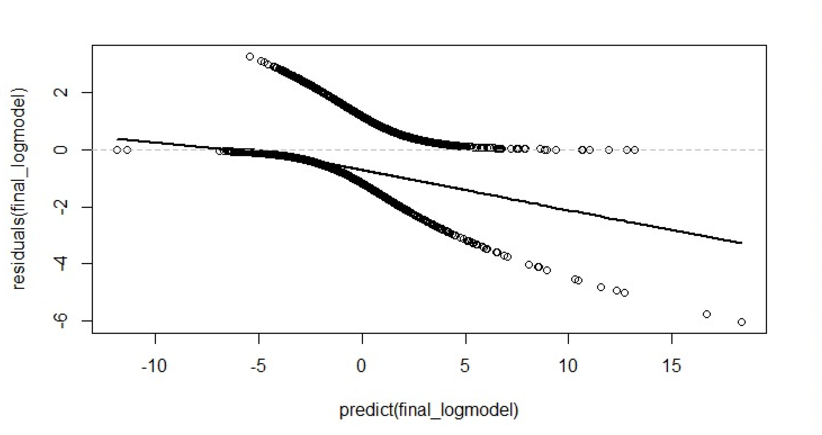




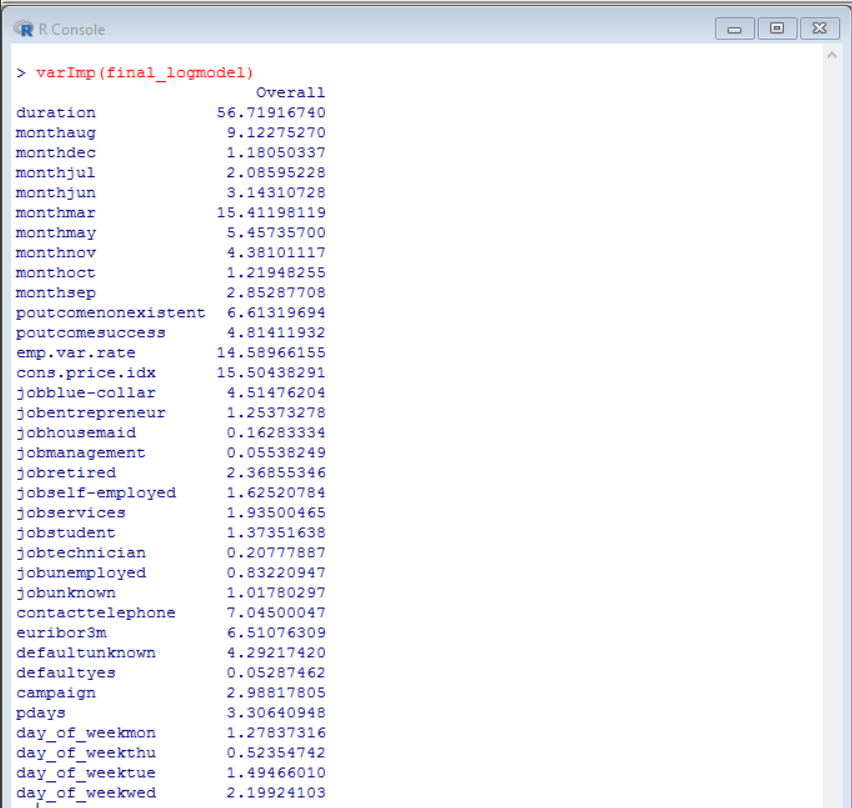
* + Fitted Logistic regression model equation:

log(odds) = -175.1 - 0.0385(campaign) + 0.0047(duration) + 0.8908(monthaug) + 0.2418(monthdec) + 0.2153(monthjul) - 0.371(monthjun) + 1.98(monthmar)- 0.4531(monthmay) - 0.5226(monthnov) + 0.1614(monthoct) + 0.3808(monthsep)+ 0.4719(poutcomenonexistent) + 1.126(poutcomesuccess ) - 1.698(emp.var.rate ) + 1.812(cons.price.idx) - 0.3339(jobblue-collar ) - 0.1698(jobentrepreneur )+ 0.0248(jobhousemaid) - 0.0051(jobmanagement) + 0.2212(jobretired) - 0.2141(jobself-employed)- 0.1743(jobservices) + 0.153(jobstudent) - 0.0148(jobtechnician) - 0.1193(jobunemployed) - 0.2684(jobunkown) - 0.5508(contacttelephone) + 0.574(euribor3m) - 0.32(defaultunknown)- 7.334(defaultyes) - 0.0007(pdays) - 0.0939(day\_of\_weekmon) + 0.03766(day\_of\_weekthu)+ 0.1094(day\_of\_weektue) + 0.1615(day\_of\_weekwed) + e

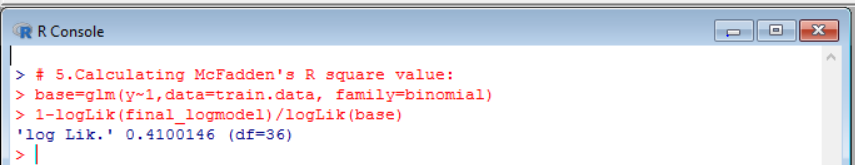
* + Residual Analysis of the obtained Logistic Regression Model: Here we predict the probability for a variable taking values 0 or 1 (Yes or No). If the value is 0, then we always predict more, and residuals must be negative. If the true value is 1, then we underestimate, and the residuals must be positive.
  + The plain line is somewhat assumed to be straight and depicts the local regression.



* + Finding variable importance of the obtained logistic regression model: Using the varImp function, we obtain the variable importance values for each significant variable as follows:

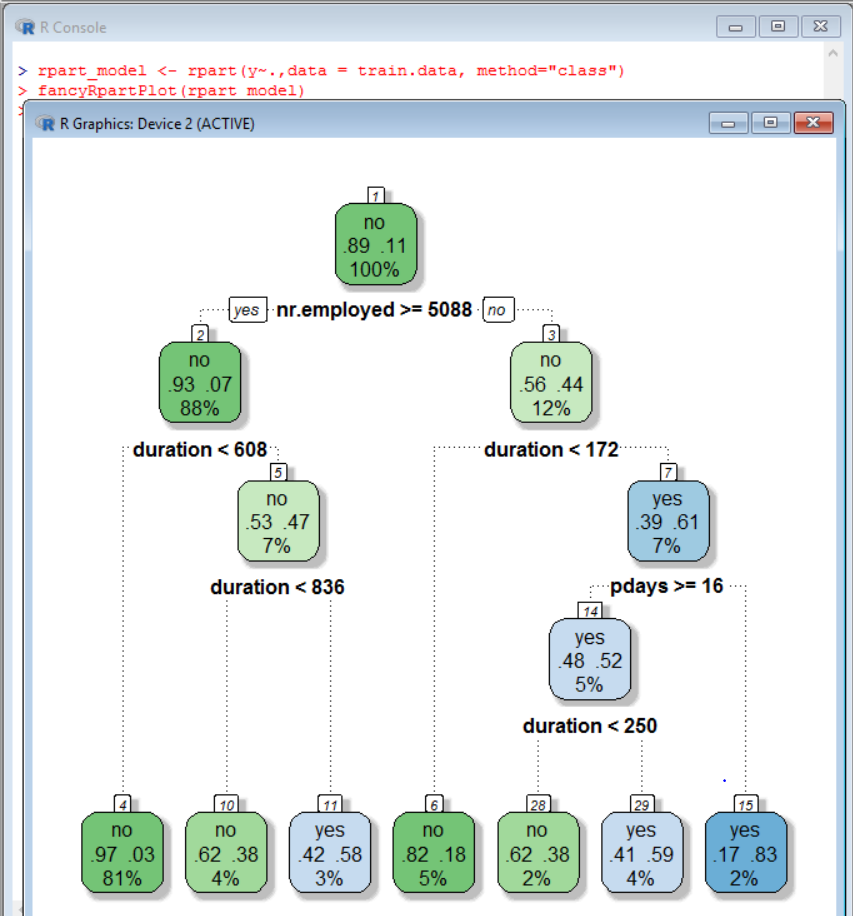


* + McFadden’s R-Square value: We can observe that the R square value is 41.0% for the logistic regression model.



**DECISION TREE**

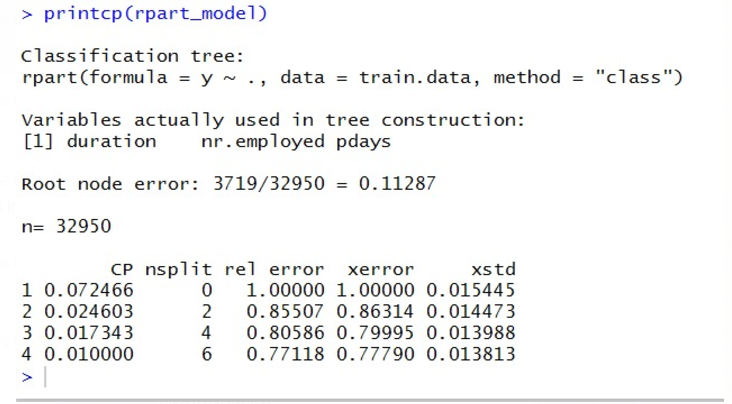
* A tree like structured algorithm in which the internal node represents test on an attribute, each branch represents the outcome of the test and each leaf node represents class label which is the decision taken after computing all attributes.
* Here we build catogorical variable decision tree as the target variable is catogorical(Yes or No). In R we use rpart function to build decision tree which uses Gini Index algorithm to split the nodes.



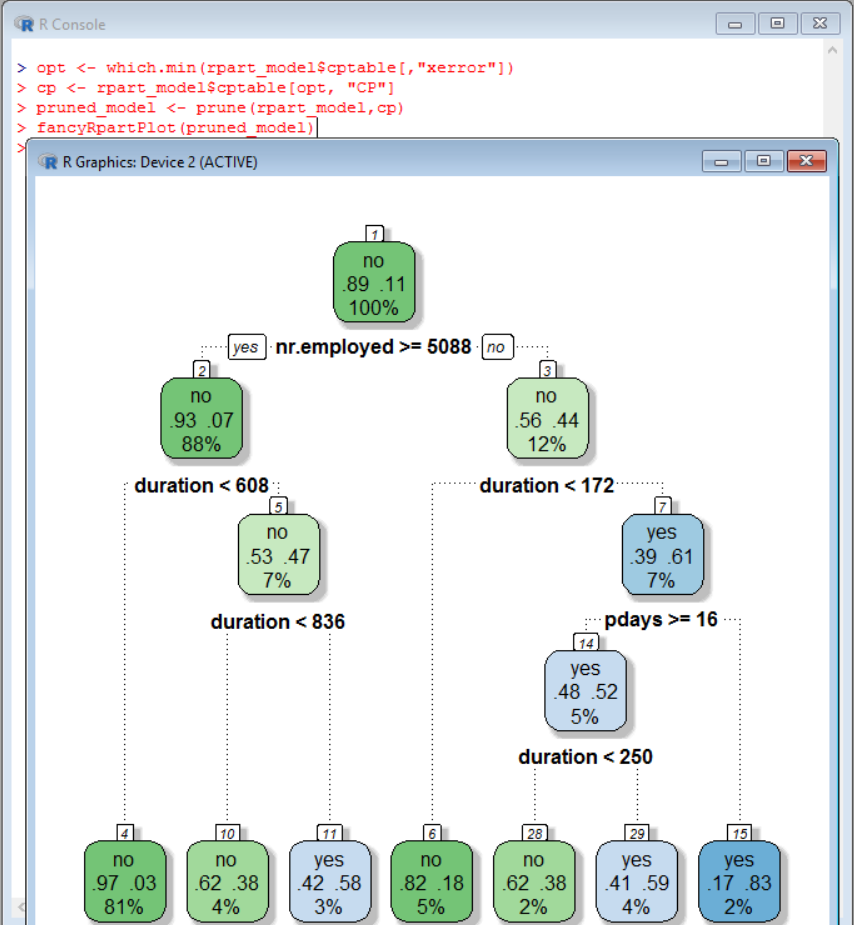
* In our model, which includes 20 variables, the root node considered by the rpart function is nr.employed, the branch nodes are: duration, pdays and the leaf nodes being the y variable (Yes or No).
* We can predict the possible outcome of a new customer contacted by knowing the root, branch and leaf attribute and tracing the above tree in top down approach.
* For example: Consider a customer who is contacted for 200 seconds, 20 days after the previous campaign and when the number of employees employed where 5000. From the above decision tree, we can predict that this customer is most likely to decline the subscription.

Decision Tree Pruning:

* In a decision tree, sometimes it generates unwanted and meaningless rules as it grows deeper. This is called overfitting problem.
* Pruning selects a tree size that minimizes the cross validated error, the **xerror** in cp table. The cp value against the lowest xerror is passed to rpart function to build the pruned tree.

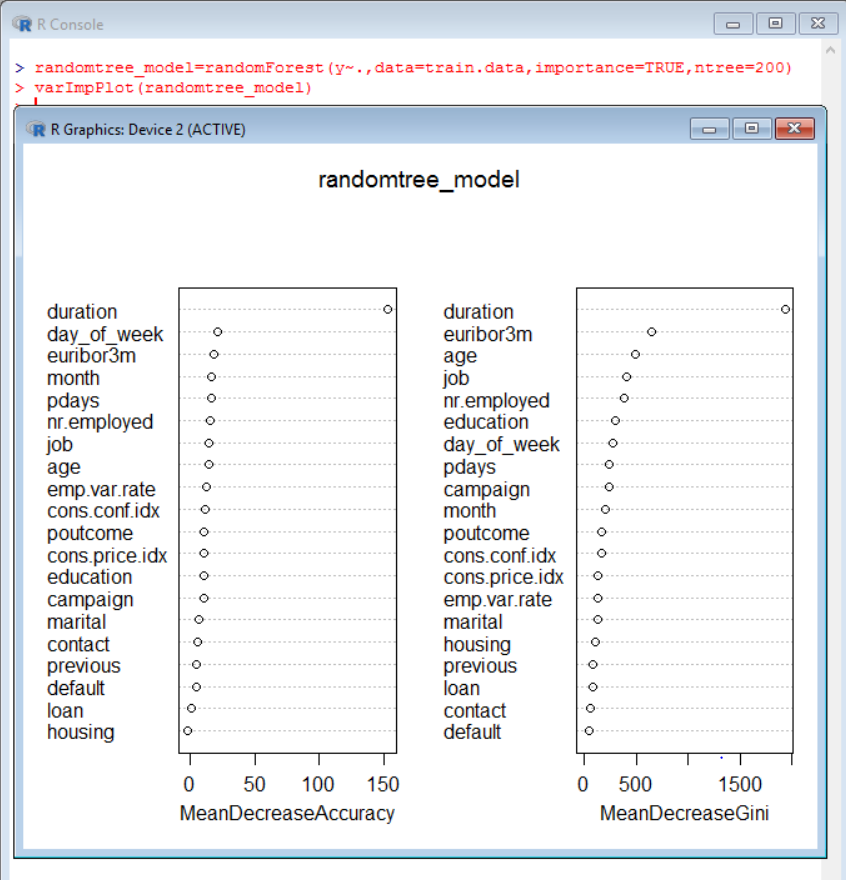


* As seen above, the cp value against the lowest xerror (0.77790) is 0.01
* The rpart function by default chooses 0.01 to build the decision tree. Hence, our pruned decision tree is same as unpruned as show below:



**RANDOM FOREST**

* In random forest, we grow multiple trees to classify a new object based on attributes. Each tree gives a classification with multiple trees for that class. The forest chooses the classification having the most votes.
* When the dataset is large, the number of trees to be chosen must be small. Here we chose to grow 200 trees (ntree=200).
* Random forest can handle large dataset with higher dimensionality. It can handle 1000s of input variables and identify the most significant variables. Therefore, it is considered as one of the dimensionality reduction method.
* Owing to the above advantages, random forest algorithm suits the best for our dataset as it large and has 20 independent variables.



* As shown above, the model outputs the significant variables in order of their importance which is a very handy feature.

# **5. Evaluations and Results**

## 5.1. Evaluation Methods

**Confusion Matrix:** It is a table used to describe the performance of classification model on a set of test data for which true values are known.

True positives (TP): These are cases in which we predicted yes (subscribed)

True negatives (TN): We predicted no, and they did not subscribe.

False positives (FP): We predicted yes, but they don't subscribe.

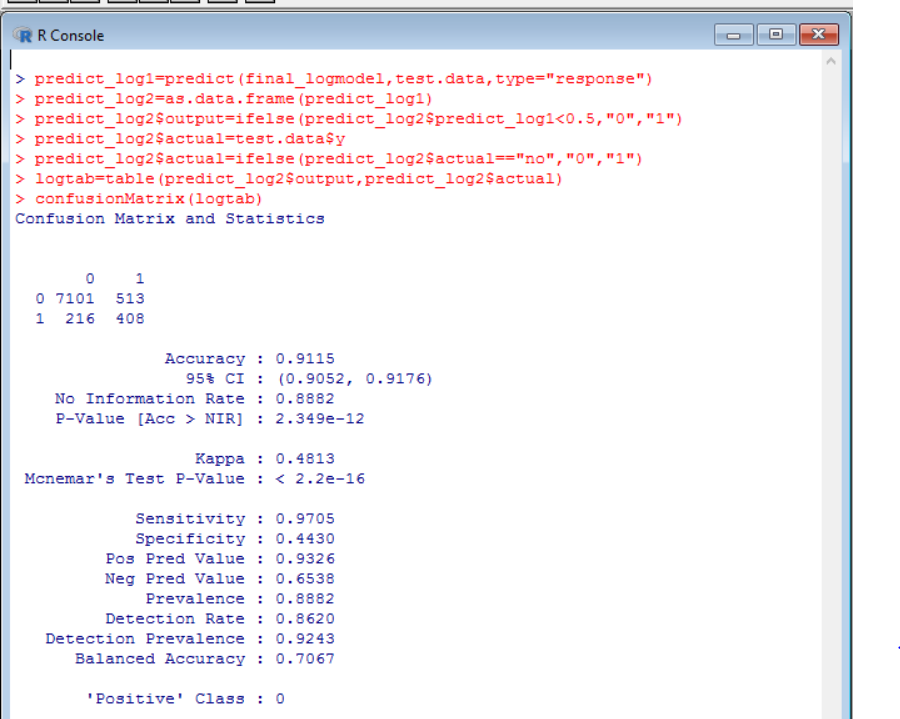
(Also, known as a "Type I error.")

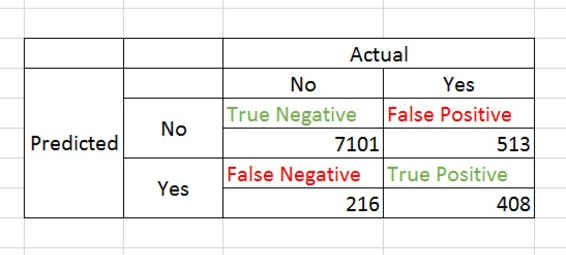
False negatives (FN): We predicted no, but they do subscribe.

(Also, known as a "Type II error.")

**Prediction and Confusion Matrix – Logistic Regression Model**

* The decision outcome of logistic regression model when tested on real world dataset is as shown below:

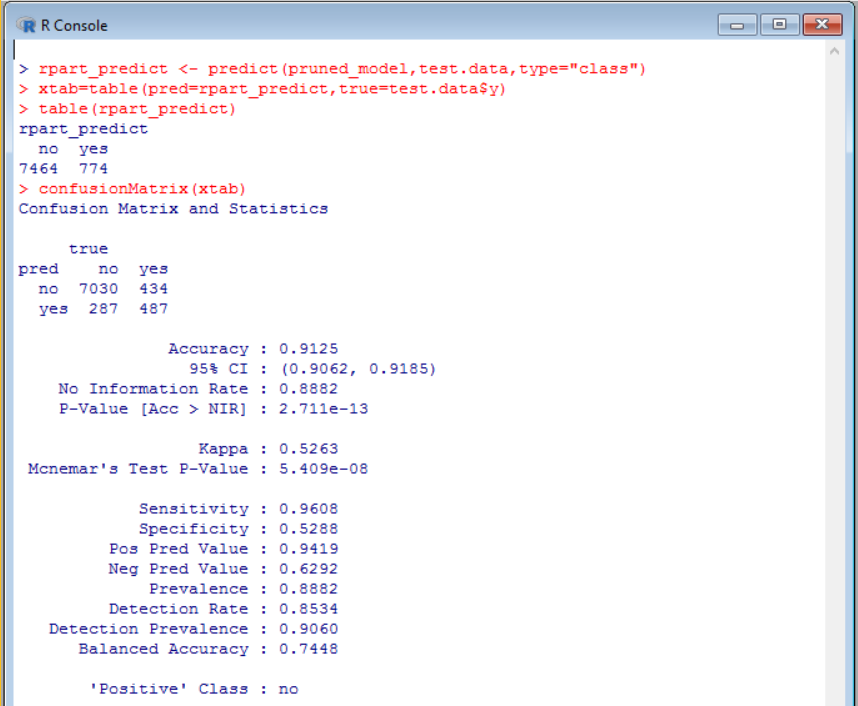


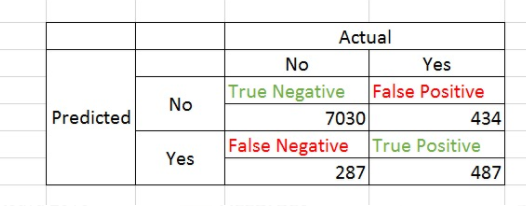


* The classifier made a total of 8238 predictions. Out of those, the classifier predicted Yes for 624(FN + TP) and No for 7614(TN + FP). In reality, 921(FP + TP) subscribed for the term deposit and 7317(TN + FN) did not subscribe.

**Prediction and Confusion Matrix – Decision Tree Model**

* The decision outcome of decision tree model when tested on real world dataset is as shown below:

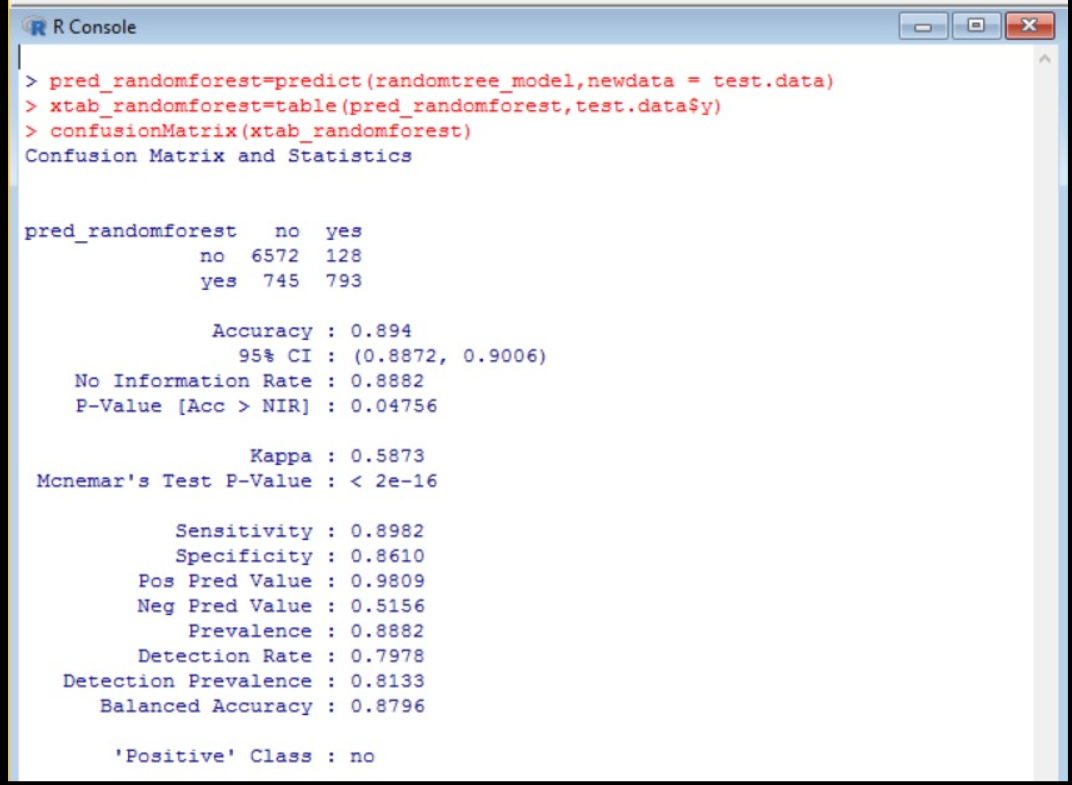


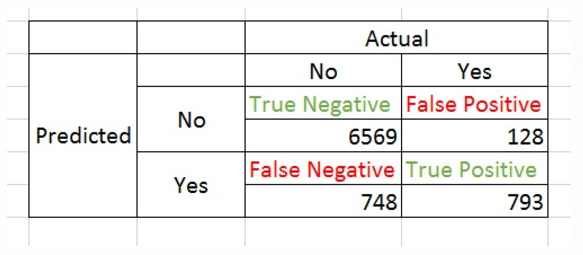


* The classifier made a total of 8238 predictions. Out of those, the classifier predicted Yes for 774(FN + TP) and No for 7464(TN + FP). In reality, 921(FP + TP) subscribed for the term deposit and 7317(TN + FN) did not subscribe.

**Prediction and Confusion Matrix – Random Forest**

* The decision outcome of random forest model when tested on real world dataset is as shown below:

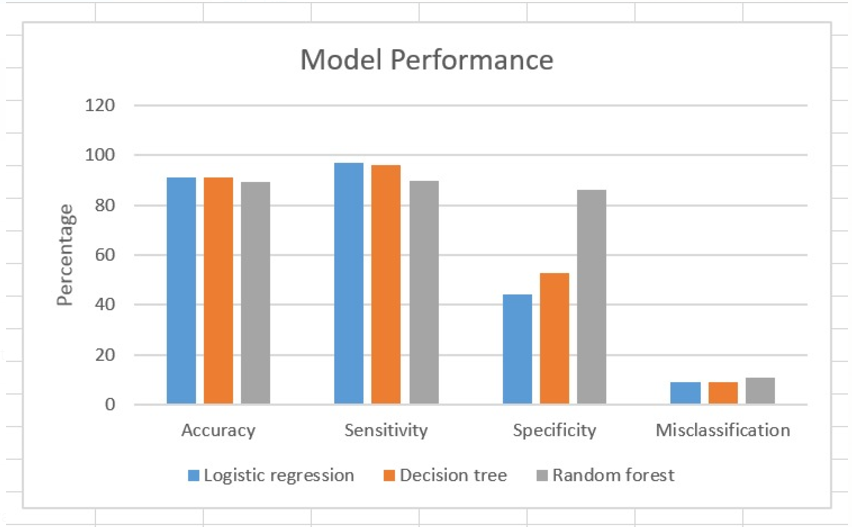




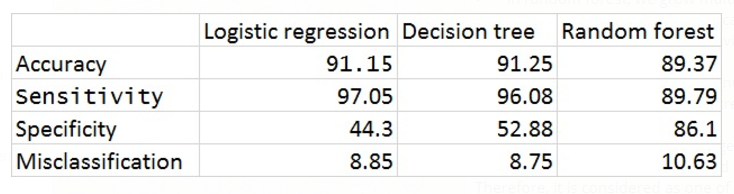
* The classifier made a total of 8238 predictions. Out of those, the classifier predicted Yes for 1541(FN + TP) and No for 6697(TN + FP). In reality, 921(FP + TP) subscribed for the term deposit and 7317(TN + FN) did not subscribe.

## 5.2. Results and Findings

* Model performance graph for all three models:

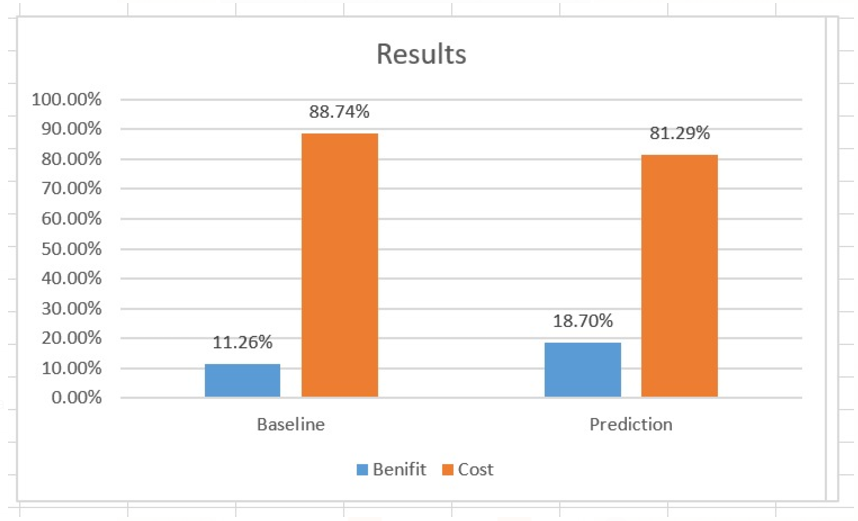


* Table depicting the final outcomes for each model in terms of Accuracy, Sensitivity, Specificity and Misclassification:



* In terms of Accuracy, Logistic regression and Decision tree model are slightly higher compared to Random forest.
* Random forest model also provides better accuracy with relatively low misclassification rate and higher sensitivity, specificity rates. Therefore, we choose Random forest model for further predictions.

**Findings:**



* Baseline Model: Benefit of calling 4640 customers who would subscribe to term deposit minus Cost of calling 36548 customers who would not subscribe.
* Prediction Model: Benefit of calling 793 customers who would subscribe to term deposit minus Cost of calling 128 customers who would not subscribe.
* The cost of calling the customers who would not subscribe decreases by 7.45% when we use the random forest prediction model.
* The marketing management could consider this factor before calling customers for subscriptions.
* Our model can increase the subscription rate by finding the potential customers from 11.26% to 18.70%.

# **6. Conclusions and Future Work**

## 6.1. Conclusions

* To evaluate the above models, we make use of confusion matrix and conclude that random forest model performs the best when compared to others.
* The cost of calling the customers who would not subscribe decreases by 7.45% when we use the random forest prediction model. Our model can increase the subscription rate by finding the potential customers from 11.26% to 18.70%.

## 6.2. Limitations

* Accuracy for Random forest model can be enhanced by using other parameters like mtry while building the model.

## 6.3. Potential Improvements or Future Work

* We can use machine learning techniques such as neural networks and SVM.