DATA MINING – FINAL PROJECT

I. Title page

This is the project name is “bank”.

This project is done by student Lien Pham (ID: # 12251947).

II. Abstract

The project is called Portuguese banking project which consists of all data related to the characteristics of clients such as their ages, marital status, education, their current balances and others. The data also consists of the marketing campaign activities applied to a client, specifically the result of the previous campaign, number of contacts for that client in the campaign and other related information.

The purpose of this project is to explore the focus group of clients who are potential to deposit based on the above collected data.

There are four predictive models being used to analyze and predict the probability of the client who are potential to deposit, then a model that gives the best accuracy rate will be chosen. They are logistic model, tree classifier, bagging and random forest. Four models give the approximately similar accuracy rates, however the logistic model seems to be the best one with 87.9% on test data, 90.4% on validation data

III. Introduction

Data “bank” which contains data related with the direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls with the aim of assessing if the client would subscribe to a bank term deposit or not.

The question is to find out the focus group who are most likely to deposit based on their characteristics such as age, education, marital status, their credit loan, etc and based on the result of the marketing campaign, number of banker contacting the client in the campaign and others related.

IV. Methods

There are 4 methods being used in this project to find out the best model that has the highest accuracy of prediction. They are:

- Logistic model: is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables
- Tree classifier is a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

- Bagging: or Bootstrap Aggregation , is a simple and very powerful ensemble method. An ensemblemethod is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model
- Random Forest: is a term for an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random forests are collections of trees, all slightly different. It randomize the algorithm, not the training data

V. Results

1. Exploration bank data

Firstly, we will explore relation between all indepence variable to probability of client subscribed a term deposit.

Proportion of client subscribed a term deposit

```
> table(bank_data$y)/nrow(bank_data)

     no     yes
0.88476 0.11524
```

As the table above, percentage of clients who subscribed a term deposit is 11.5%.

Analysis age of clients who subscribed

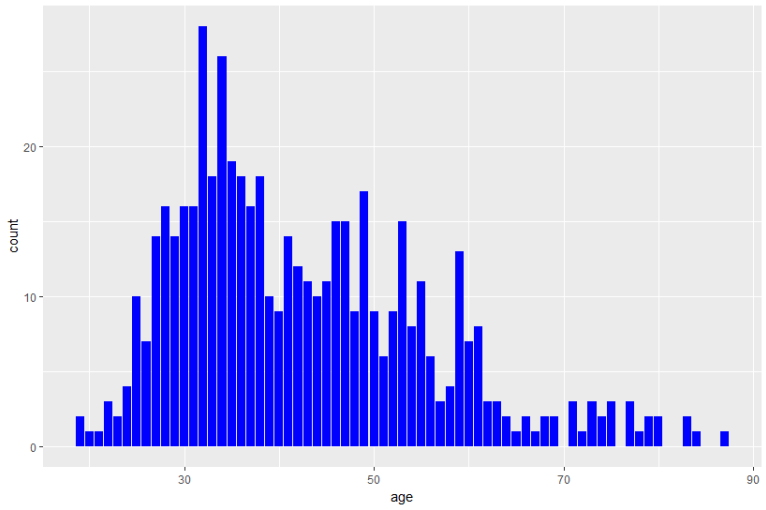

Figure 1 Age of clients subscribed term deposit

Comment: Majority of people who deposit are from 25 to 62 years old

Analysis clients who subscribed a term deposit by type of jobs

Comments: Although large number of clients who subscribed a bank term deposit work as ***manager****, probability subscription a bank termdeposit of clients who is* ***student or retired*** *are higher than others, more than 20%.*

Table 1 Job of clients subscribed term deposit

	no	yes
admin.	420	58
blue-collar	877	69
entrepreneur	153	15
housemaid	98	14
management	838	131
retired	176	54
self-employed	163	20
services	379	38
student	65	19
technician	685	83
unemployed	115	13
unknown	31	7

	no	yes
admin.	0.87866109	0.12133891
blue-collar	0.92706131	0.07293869
entrepreneur	0.91071429	0.08928571
housemaid	0.87500000	0.12500000
management	0.86480908	0.13519092
retired	0.76521739	0.23478261
self-employed	0.89071038	0.10928962
services	0.90887290	0.09112710
student	0.77380952	0.22619048
technician	0.89192708	0.10807292
unemployed	0.89843750	0.10156250
unknown	0.81578947	0.18421053

Analysis clients who subscribed a term deposit by marital status

Percentage of clients who subscribed a bank term deposit by marital status is given in figure below:

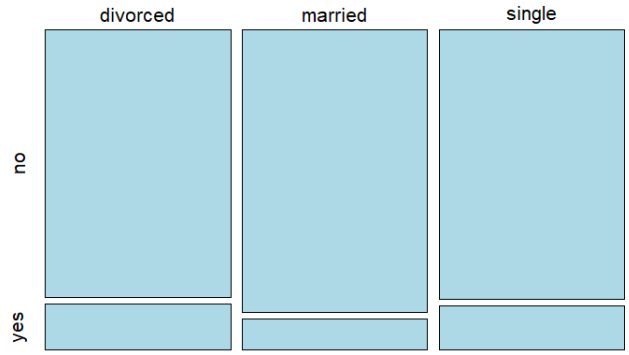

Figure 2 Marital status of clients who subscribed

Comment: Clients who is single or divorced have higher percentage of subscription bank term deposit than clients who married.

Analysis clients who subscribed a term deposit by education

Percentage of clients who subscribed a bank term deposit by education is given in figure below:

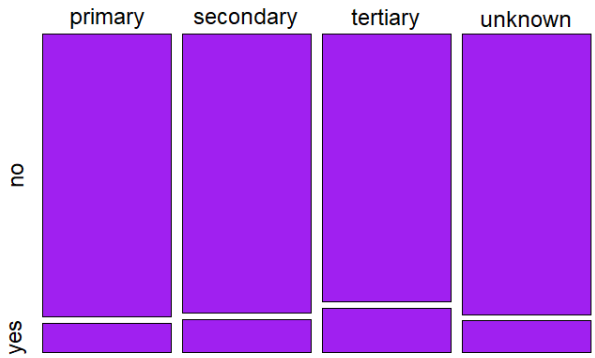

Figure 3 Education of clients who subscribed

Comment: People with tertiary education have the highest rate of deposit.

Analysis clients who subscribed a term deposit by credit default

Percentage of clients who subscribed a bank term deposit by credit default of clients is given in figure below:

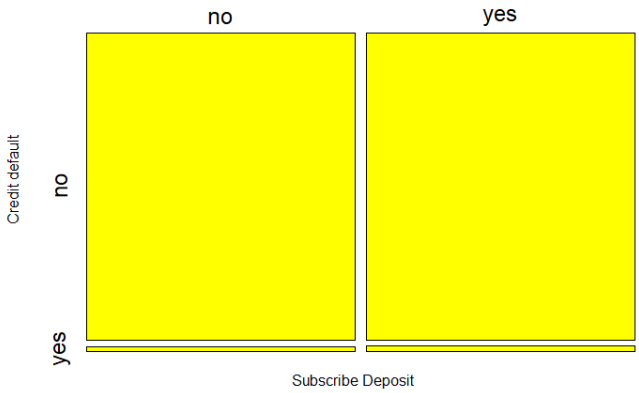

Figure 4 Credit default of clients who subscribed

Comment: 88.2% clients subscribed deposit is not default in credit.

Analysis clients who subscribed a term deposit by average yearly balance

Statistic about average of yearly balance of clients who subscription or not, as below:

```
> aggregate(balance ~ y, data = bank_data, mean) # yearly balance between 2 group
    y  balance
1  no 1403.212
2 yes 1571.956
```

Comment: clients who subscribed a term deposit have average yearly balance higher than others.

Analysis clients who subscribed a term deposit by housing loan

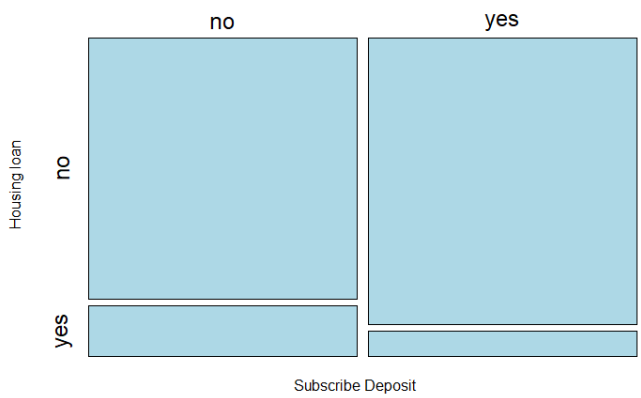

Figure 5 Housing loan of clients who subscribed

Comment: large proportion of client who subscribed termdeposit, don't have housing loan. it is likely that p eople who take housing loan they do not have money for deposit

Analysis clients who subscribed a term deposit by last contact day

Comment: Clients who are contacted in May, August, Jul have highest rate of deposit. And, clients who are contacted in day 5th, 12th, 18th and 30th have highest rate of deposit.

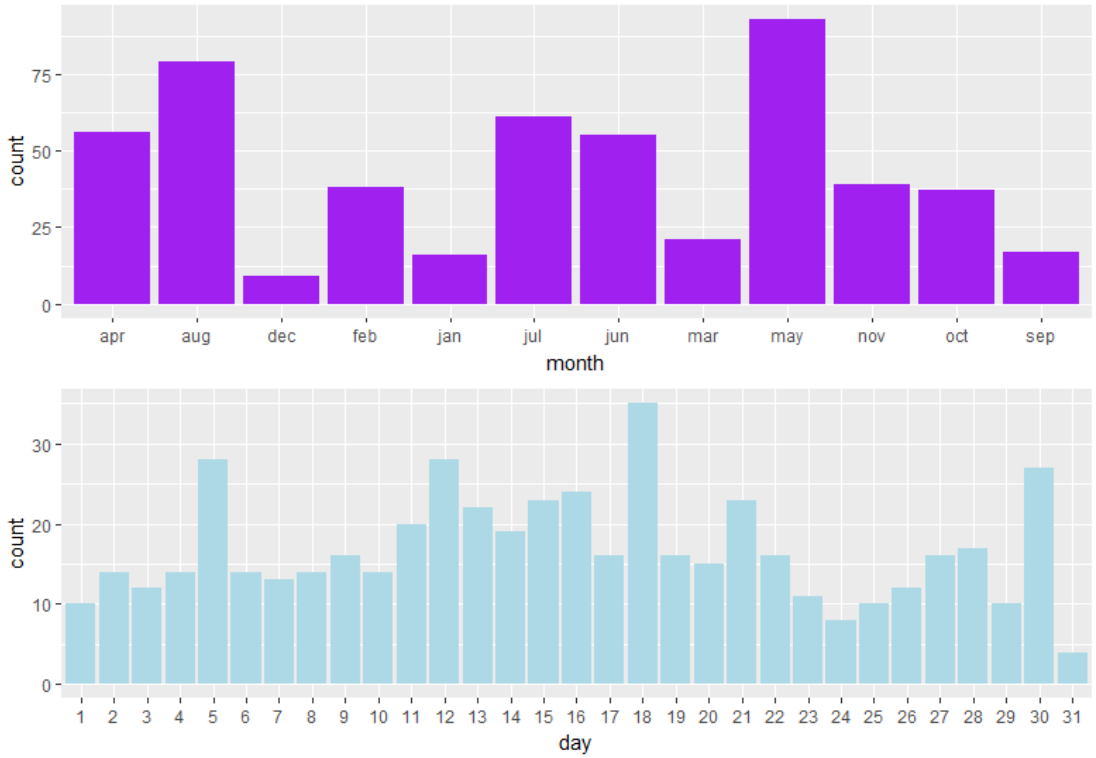

Figure 6 Last contact time of client subscribed

Analysis clients who subscribed a term deposit by number contact during this campaign

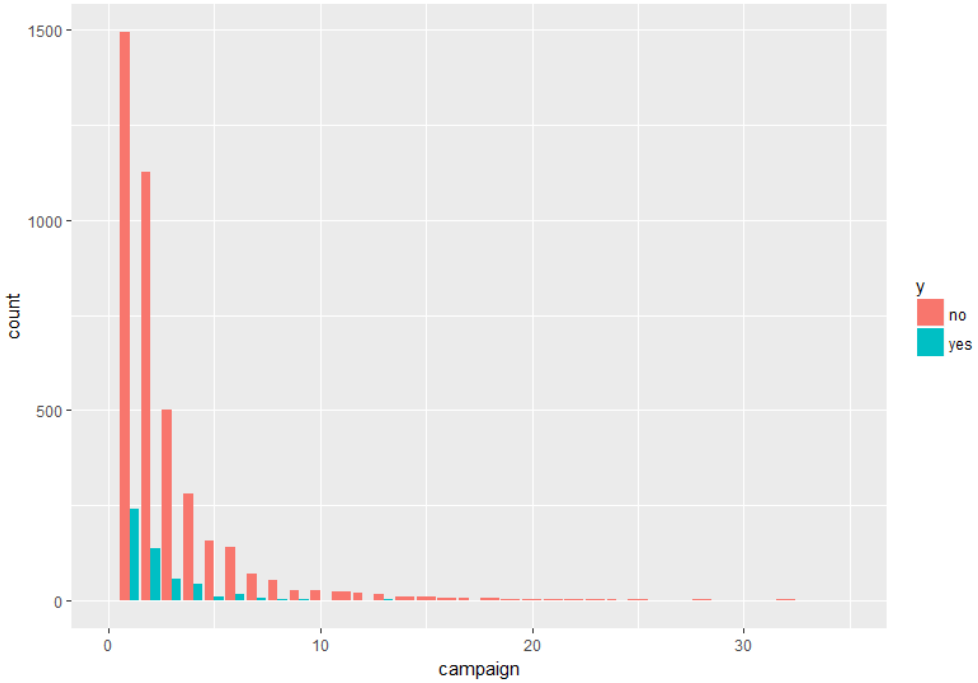

Figure 7 Number contact during this campaign

Comment: The more contacts perform in this campaign, the less people deposit

Analysis clients who subscribed a term deposit by impact of previous campaign

In bank data, “pdays” is number of days that passed by after the client was last contacted from a previous campaign. In this study, we will analysis impact of previous campaign to probability of client who subscribed a term deposit in stead of “pdays”. We use transformation from “pdays” to “pre_campaign”:

- If “pdays” = -1, pre_campaign = 0
- If “pdays” > 0, pre_campaign = 1

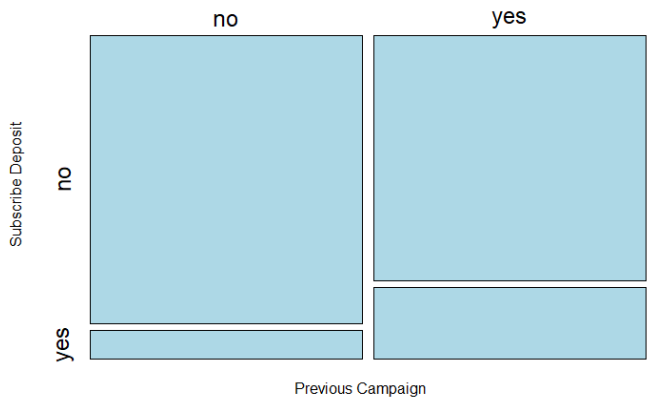

Figure 8 Impact of previous campaign

Comment: Previous campaign is not impact much to probability of client subscribed term deposit.

Analysis clients who subscribed a term deposit by number of contacts performed before this campaign

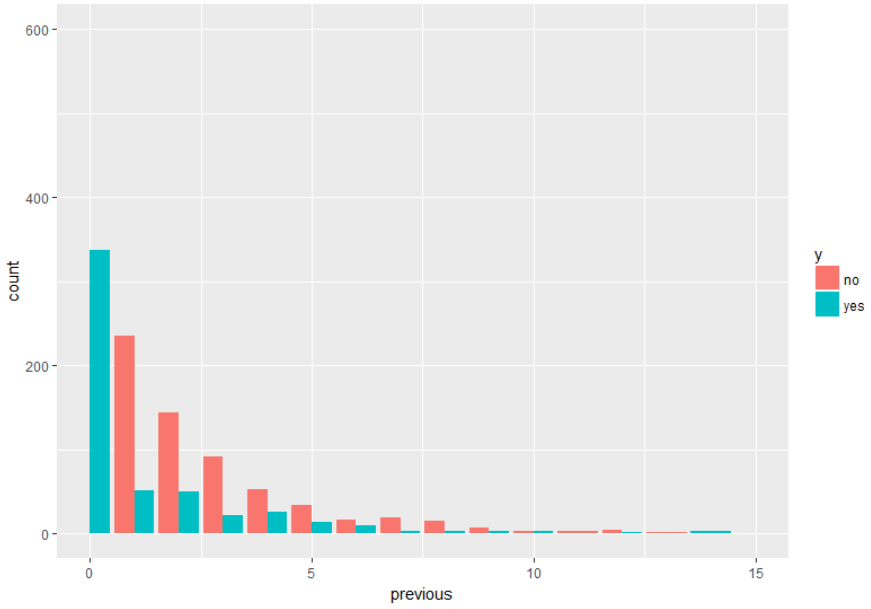

Figure 9 Number contact of previous campaign

Comment: The more contacts perform in previous campaign, the less people deposit

Analysis clients who subscribed a term deposit by outcome of the previous marketing campaign

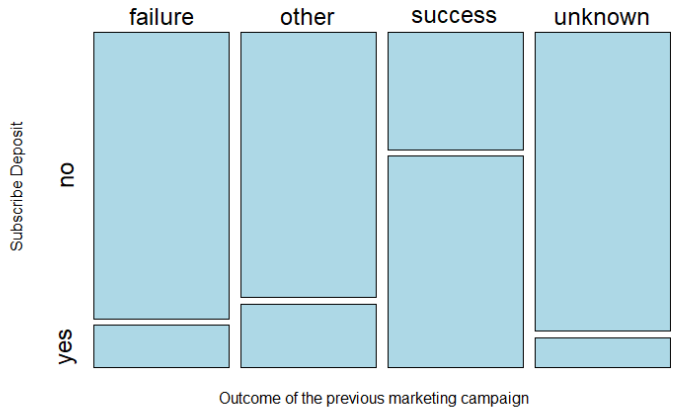

Figure 10 Impact of outcome of the previous marketing campaign

Comment: Group of clients who have outcome of the previous marketing campaign is "success", have bigger proportion of clients subscribed than others.

2. Fiting model

Secondly, we construct 4 predictive models to test the accuracy prediction of each model. We will pick the one that has the highest accuracy.

2.1 Logistic

Result of logistic regression as table below:

Table 2 Logistic result

	Coefficent	Standard error	P-value
(Intercept)	3.68E-01	3.39E-02	0.00E+00
age	-6.98E-04	3.58E-03	8.46E-01
jobblue-collar	-1.22E-01	3.83E-02	1.49E-03
jobentrepreneur	1.42E-01	4.01E-03	0.00E+00
jobhousemaid	-2.30E-01	1.64E-03	0.00E+00
jobmanagement	7.57E-02	4.23E-02	7.36E-02
jobretired	1.10E+00	2.99E-02	0.00E+00
jobself-employed	2.88E-01	5.00E-03	0.00E+00
jobservices	-1.81E-01	1.03E-02	0.00E+00
jobstudent	8.80E-01	8.33E-03	0.00E+00
jobtechnician	2.83E-02	7.82E-02	7.18E-01
jobunemployed	-2.97E-01	2.82E-03	0.00E+00
jobunknown	1.34E-01	7.34E-04	0.00E+00
maritalmarried	-5.83E-01	8.20E-02	1.17E-12
maritalsingle	-2.87E-01	6.60E-02	1.37E-05
educationsecondary	1.49E-01	6.50E-02	2.23E-02
educationtertiary	1.96E-01	6.07E-02	1.27E-03

	Coefficent	Standard error	P-value
educationunknown	-7.19E-01	4.50E-03	0.00E+00
defaultyes	-1.70E-01	1.59E-03	0.00E+00
balance	-2.20E-05	1.87E-05	2.39E-01
housingyes	-1.06E-01	7.38E-02	1.51E-01
loanyes	-4.10E-01	1.45E-02	0.00E+00
day2	-2.06E+00	1.37E-03	0.00E+00
day3	-2.28E+00	1.69E-03	0.00E+00
day4	-2.32E+00	3.43E-03	0.00E+00
day5	-1.52E+00	3.79E-03	0.00E+00
day6	-1.79E+00	3.22E-03	0.00E+00
day7	-2.00E+00	3.29E-03	0.00E+00
day8	-2.33E+00	1.64E-03	0.00E+00
day9	-2.11E+00	3.83E-03	0.00E+00
day10	-7.03E-01	4.23E-03	0.00E+00
day11	-1.89E+00	5.57E-03	0.00E+00
day12	-1.07E+00	4.52E-03	0.00E+00
day13	-1.31E+00	6.50E-03	0.00E+00
day14	-1.91E+00	2.63E-03	0.00E+00
day15	-1.38E+00	7.46E-03	0.00E+00
day16	-1.48E+00	4.97E-03	0.00E+00
day17	-2.34E+00	1.71E-03	0.00E+00
day18	-1.29E+00	4.97E-03	0.00E+00
day19	-2.15E+00	3.97E-03	0.00E+00
day20	-2.51E+00	2.08E-03	0.00E+00
day21	-1.47E+00	5.70E-03	0.00E+00
day22	-1.33E+00	3.64E-03	0.00E+00
day23	-1.68E+00	2.08E-03	0.00E+00
day24	-1.12E+00	1.24E-03	0.00E+00
day25	-1.36E+00	5.25E-03	0.00E+00
day26	-1.71E+00	2.63E-03	0.00E+00
day27	-1.72E+00	1.10E-03	0.00E+00
day28	-1.87E+00	1.21E-03	0.00E+00
day29	-2.29E+00	1.35E-03	0.00E+00
day30	-1.49E+00	4.59E-03	0.00E+00
day31	-2.33E+00	2.61E-03	0.00E+00
monthaug	-2.58E-01	7.01E-02	2.34E-04
monthdec	-1.23E-01	2.04E-03	0.00E+00
monthfeb	-2.09E-01	5.02E-03	0.00E+00
monthjan	-8.99E-01	3.00E-03	0.00E+00
monthjul	-4.71E-01	4.30E-02	0.00E+00
monthjun	-3.33E-01	1.81E-02	0.00E+00
monthmar	7.83E-01	2.12E-03	0.00E+00
monthmay	-9.61E-01	4.80E-02	0.00E+00
monthnov	-6.66E-01	4.24E-03	0.00E+00
monthoct	1.31E+00	3.83E-03	0.00E+00
monthsep	-1.45E-02	3.14E-03	4.07E-06

	Coefficent	Standard error	P-value
campaign	-4.53E-02	2.55E-02	7.59E-02
previous	-3.17E-02	3.97E-02	4.25E-01
poutcomeother	7.41E-01	1.76E-02	0.00E+00
poutcomesuccess	2.63E+00	2.02E-02	0.00E+00
poutcomeunknown	6.03E-02	8.14E-02	4.59E-01
pre_Campaignyes	3.08E-01	6.75E-02	5.10E-06

The “red sign” in p-value of “Logistic Result” show that all variable impact to subscription term deposit of clients with significant level 10%.

As the result, “age”, “housing loan”, “jobtechnician”, “average yearly balance”, “number of contacts performed before this campaign”, “poutcomeunknown” don’t impact to decision of deposit with significant level 10%.

Base on validation data, accuracy rate of Logistic method is **90.4%**

```
> acc.l
[1] 0.9041298
```

Perform of Logistic on test data is **87.9%**

```
> acc.l.t
[1] 0.8792342
```

2.2 Tree classification

Result of tree classification as figure below:

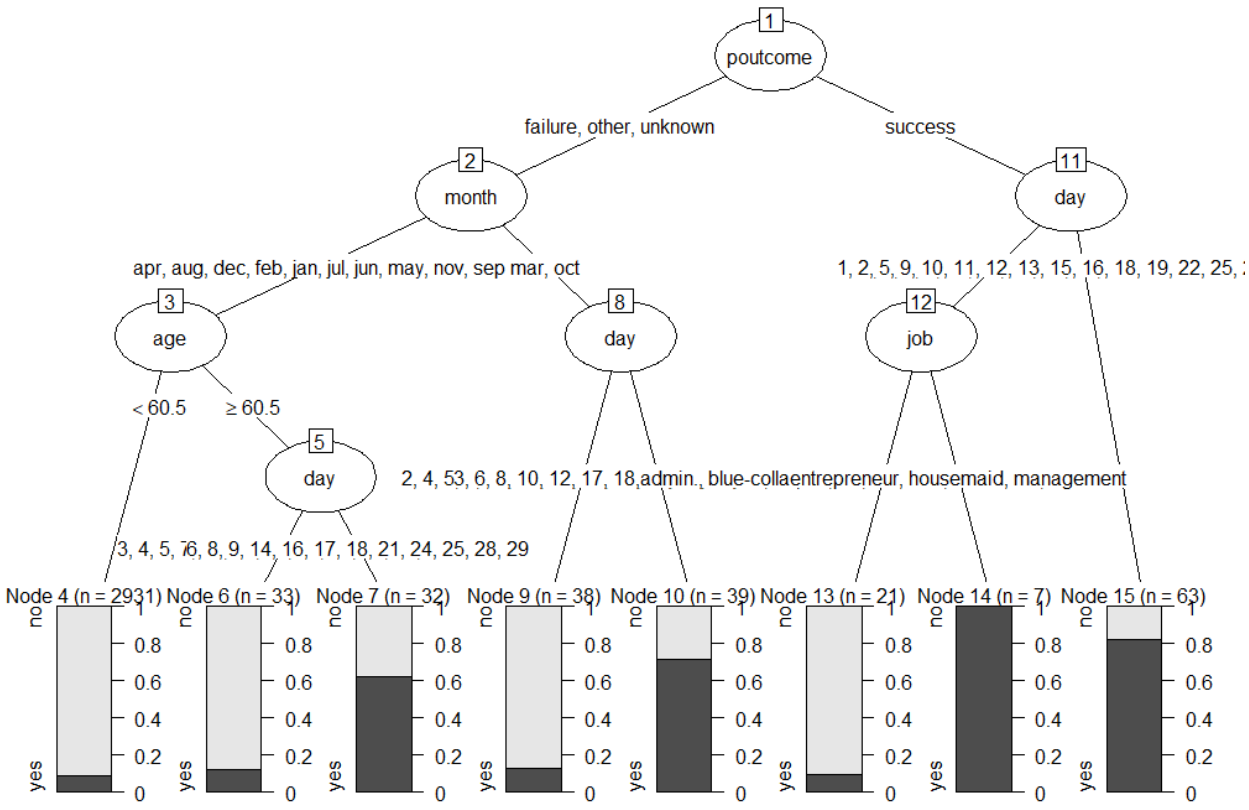

Figure 11 Tree classification

As tree classification result, variables that considered to be most important are: poutcome, month of year, day of month, age, job.

Base on validation data, accuracy rate of Tree classification method is **89.4%**

```
> acc.r
[1] 0.8938053
```

Perform of tree classification on test data is **88.4%**

```
> acc.r.t
[1] 0.8836524
```

2.3 Bagging

Base on validation data, accuracy rate of Bagging method is **89.7%**

```
> sum(diag(tab_bag))/sum(tab_bag)
[1] 0.8967552
```

Perform of Bagging on test data is **88.7%**

```
> sum(diag(tab_bagt))/sum(tab_bagt)
[1] 0.8865979
```

2.4 Random Forest

Base on validation data, accuracy rate of Random Forest model is **89.1%**

```
> sum(diag(tab_rf))/sum(tab_rf)
[1] 0.8908555
```

In Random forest model, 4 variables that considered to be most important are: **day**, **balance**, **month** and **age**.

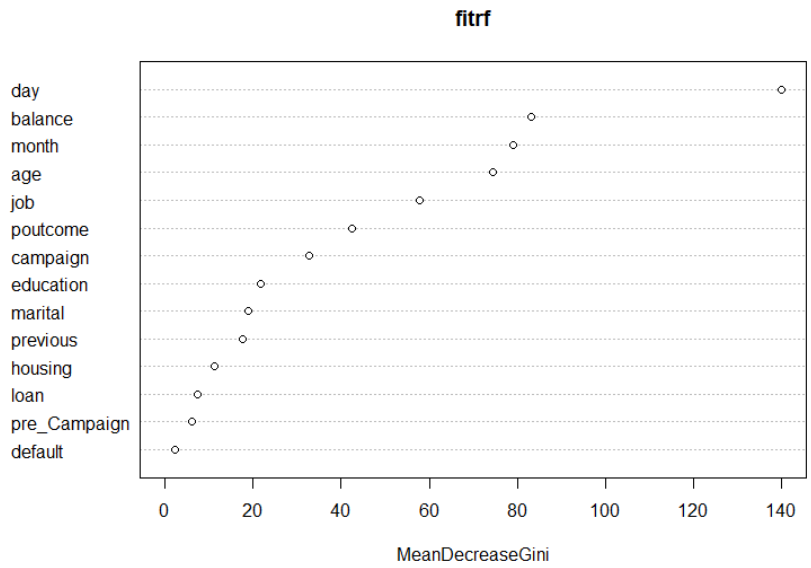

Figure 12 Important of each variable in random forest model

Perform of Random forest on test data is **88.7%**

```
> sum(diag(tab_rf_t))/sum(tab_rf_t)
[1] 0.8865979
```

VI. Discussion

Performing four forecast model, we get the results as below:

- Accuracy of logistic model: 87.9% on test data, 90.4% on validation data
- Accuracy of tree model: 88.4 % on test data, 89.4% on validation data
- Accuracy of bagging model: 88.7% on test data, 89.7% on validation data
- Accuracy of Randomforest: 88.7% on test data, 89.1% on validation data

We see that the accuracry rate of tree, bagging and random forest are approximately the same so the question what is the model should be chosen from?

According the losgistic model, there are the most important variables that have the strongest impact on the clients' decision to deposit:

- "Age", "housing loan", "jobtechnician", "average yearly balance", "number of contacts performed before this campaign",

While the Ramdom Forest model shows the most important variables:

- Day, Balance, Month, Age, Job

Performing basic statistic for Bank data, we find out the below statistic facts, these are groups that deposit the most

- Age: Clients from 25 to 62 years old
- Marital: divorce and single
- Job: student and retired
- Housing loan: without house loan
- Education: with tertiary education
- Month and day contacted: clients who are contacted in May, August, Jul, day 5th, 12th, 18th and 30th
- The more contacts perform in previous campaign, the less people deposit
- Outcome of previous campaign: Group of clients who have outcome of the previous marketing campaign is “success”, have bigger proportion of clients subscribed than others
- Previous campaign: does not impact much to probability of client subscribed term deposit

We see that logistic model gives the best predictive results and it shows the most important variables which are logical with the statistic exploration of bank data.

The logistic and random forest models BOTH show that age and balance are the most important variables

VII. Conclusion

Compare accuracy rate on validation data, Logistic model is the method that has higher accuracy rate on validation than other methods.

Perfom of Logistic method on test data

```
> tab.r.t

      no yes
  no  587  14
  yes  65  13
```

And **the accuracy rate is 87.9%.**

```
> acc.l.t
[1] 0.8792342
```

VIII. References

IX. Appendix

 # R code

```
# library in use
library(ggplot2)
library(data.table)
library(dplyr)
library(rpart)
library(partykit)
library(adabag)
library(randomForest)
library(nnet)
library(gridExtra)

# import data
bank_data <- read.csv("G:/Study/Data Analyst/Master of Data Analytic/Data Mining/Project/Submit Project/bank.csv", header = TRUE)

# structure data
str(bank_data)
View(bank_data)

# Transform data
bank_data$pre_Campaign <- ifelse(bank_data$pdays == -1,"no", "yes")
bank_data$pre_Campaign <- as.factor(bank_data$pre_Campaign)
str(bank_data)
bank_data <- bank_data[,-12]
bank_data$day <- as.factor(bank_data$day)

# Exploration data
table(bank_data$y)/nrow(bank_data) # Proportion of client subscribed is 11.5%
ggplot(bank_data[bank_data$y == 'yes',], aes(x=age)) + geom_bar(fill = "blue") ## explore age of clients who subscribed
ggplot(bank_data[bank_data$y == 'yes',], aes(x=job, fill = job)) + geom_bar() +theme(axis.text.x = element_blank()) ## explore job of client who subscribe
table(bank_data$job, bank_data$y)/rowSums(table(bank_data$job, bank_data$y))
table(bank_data$job, bank_data$y)
```

```
t <- table(bank_data$marital, bank_data$y)/rowSums(table(bank_data$marital, bank_data$y)) # explore marital status of clients who subscribed
plot(t, col = "lightblue", main = '')
t <- table(bank_data$education, bank_data$y)/rowSums(table(bank_data$education, bank_data$y)) # explore education of clients who subscribed
plot(t, col = "purple", main = '', cex = 1.5)
t <- table(bank_data$default, bank_data$y)/rowSums(table(bank_data$default, bank_data$y)) # explore default of clients who subscribed
plot(t, col = "yellow", main = '', cex = 1.5, xlab = "Credit default" , ylab = "Subscribe Deposit")
t2 <- table(bank_data$y, bank_data$default)/rowSums(table(bank_data$y, bank_data$default)) # explore default of clients who subscribed
plot(t2, col = "yellow", main = '', cex = 1.5, ylab = "Credit default" , xlab = "Subscribe Deposit")
aggregate(balance ~ y, data = bank_data, mean) # yearly balance between 2 group
t2 <- table(bank_data$y, bank_data$loan)/rowSums(table(bank_data$y, bank_data$loan)) # explore housing loan of clients who subscribed
plot(t2, col = "lightblue", main = '', cex = 1.5, ylab = "Housing loan" , xlab = "Subscribe Deposit")
p1 <- ggplot(data = bank_data[bank_data$y =="yes",], aes(x = month) ) + geom_bar(fill="purple") # explore month of year
p2 <- ggplot(data = bank_data[bank_data$y =="yes",], aes(x = day) ) + geom_bar(fill="lightblue") # explore day of month
grid.arrange(p1, p2, ncol=1)
ggplot(data = bank_data, aes(x = campaign, fill = y) ) + geom_bar(position='dodge') + xlim(0,35) # number of contact
t2 <- table(bank_data$y, bank_data$pre_Campaign)/rowSums(table(bank_data$y, bank_data$pre_Campaign)) # explore pre_Campaign of clients who subscribed
plot(t2, col = "lightblue", main = '', cex = 1.5, ylab = "Previous Campaign" , xlab = "Subscribe Deposit")
t3 <- table(bank_data$pre_Campaign, bank_data$y)/rowSums(table(bank_data$pre_Campaign, bank_data$y)) # explore pre_Campaign of clients who subscribed
plot(t3, col = "lightblue", main = '', cex = 1.5, xlab = "Previous Campaign" , ylab = "Subscribe Deposit")
ggplot(data = bank_data, aes(x = previous, fill = y) ) + geom_bar(position='dodge') + xlim(0,15) + ylim(0,600) # number of contact previous campaign
t3 <- table(bank_data$poutcome, bank_data$y)/rowSums(table(bank_data$poutcome, bank_data$y)) # explore poutcome of clients who subscribed
plot(t3, col = "lightblue", main = '', cex = 1.5, xlab = "Outcome of the previous marketing campaign" , ylab = "Subscribe Deposit")
# Modelling acuracy predict
# Comparing classifiers using training, validation and test
#Set seed of random number generator
set.seed(1000)
#Ensure that y is factor
is.factor(bank_data$y)

## Random Forest
```

```
#Split data into three sets
N<-nrow(bank_data)
trainind<-sort(sample(1:N,size=floor(N*0.70)))
nottestind<-setdiff(1:N,trainind)
validind<-sort(sample(nottestind,size=length(nottestind)/2))
testind<-sort(setdiff(nottestind,validind))


#Fit to test and asses performance on validation
fitrf<-randomForest(y~.,data=bank_data,subset=trainind)
pred<-predict(fitrf,type="class",newdata=bank_data)
tab_rf <-table(bank_data$y[validind],pred[validind])
tab_rf
sum(diag(tab_rf))
sum(diag(tab_rf))/sum(tab_rf)

#performance on test
tab_rf_t <-table(bank_data$y[testind],pred[testind])
tab_rf_t
sum(diag(tab_rf_t))
sum(diag(tab_rf_t))/sum(tab_rf_t)


varImpPlot(fitrf)


## Bagging

fitbag<-bagging(y~.,data=bank_data[trainind,])
pred<-predict(fitbag,type="class",newdata=bank_data[validind,])
tab_bag <-pred$confusion
tab_bag
sum(diag(tab_bag))
sum(diag(tab_bag))/sum(tab_bag)

#performance on test
pred<-predict(fitbag,type="class",newdata=bank_data[testind,])
tab_bagt <-pred$confusion
tab_bagt
sum(diag(tab_bagt))
sum(diag(tab_bagt))/sum(tab_bagt)
```

```
## Single classifier

#Fit a classifier to only the training data
fit.r <- rpart(bank_data$y~.,data=bank_data,subset=trainind)
plot(as.party(fit.r))

#Fit a logistic regression to the training data only too
fit.l <- multinom(bank_data$y~., data=bank_data,subset=trainind)
summary(fit.l)
z <- summary(fit.l)$coefficients/summary(fit.l)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
fit.l_res <- cbind(summary(fit.l)$coefficients, summary(fit.l)$standard.errors,p)
View(fit.l_res)
#Classify for ALL of the observations
pred.r <- predict(fit.r,type="class",newdata=bank_data)
pred.l <- predict(fit.l,type="class",newdata=bank_data)

#Look at table for the validation data only (rows=truth, cols=prediction)
tab.r <- table(bank_data$y[validind],pred.r[validind])
tab.r
tab.l <- table(bank_data$y[validind],pred.l[validind])
tab.l

#Work out the accuracy
acc.r <- sum(diag(tab.r))/sum(tab.r)
acc.l <- sum(diag(tab.l))/sum(tab.l)

acc.r
acc.l

#perform in test data
tab.r.t <- table(bank_data$y[testind],pred.r[testind])
tab.r.t
tab.l.t <- table(bank_data$y[testind],pred.l[testind])
tab.l.t
acc.r.t <- sum(diag(tab.r.t))/sum(tab.r.t)
acc.l.t <- sum(diag(tab.l.t))/sum(tab.l.t)

acc.r.t
acc.l.t
```