Data Mining on Bank Marketing for Term Deposits

STAT 515 (Final Project)

VAMSHIDAR REDDY CHAULAPALLY
ANIRUDH MYAKALA
ARUN REDDY BOLLAM

1 TABLE OF CONTENTS

2 INTRODUCTION

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no').
Objective of the project is to build a classification model which can correctly classify the customers who are most likely to say yes to term deposits so that those customers can be targeted and to classify customers who are likely to say no so that they can be ignored. We will be building set of models and applying them to data and take the best most which has less misclassification error or with high sensitivity.

3 DATA

Data set contains 21 variables. The target variable is last column(Y) which indicates whether a term deposit has been made or not. The remaining variables are the factors which influence the most on likeliness of a term deposit. This is how dataset looks like.

age	job	marital	education	default	housing	loan	contact	month	day_of_w	duration	campaign	pdays	previous	poutcome	emp.var.r	cons.price	cons.conf.	euribor3m	nr.employ	y
56	housemaic	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
57	services	married	high.schoc	unknown	no	no	telephone	may	mon	149	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
37	services	married	high.schoc	no	yes	no	telephone	may	mon	226	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
56	services	married	high.schoc	no	no	yes	telephone	may	mon	307	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
59	admin.	married	profession	no	no	no	telephone	may	mon	139	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
24	technician	single	profession	no	yes	no	telephone	may	mon	380	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
25	services	single	high.schoc	no	yes	no	telephone	may	mon	50	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
25	services	single	high.schoc	no	yes	no	telephone	may	mon	222	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
29	blue-collar	single	high.schoc	no	no	yes	telephone	may	mon	137	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
57	housemaic	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
54	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
46	blue-collar	married	basic.6y	unknown	yes	yes	telephone	may	mon	440	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
50	blue-collar	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
39	manageme	single	basic.9y	unknown	no	no	telephone	may	mon	195	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
30	unemploye	married	high.schoc	no	no	no	telephone	may	mon	38	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
55	blue-collar	married	basic.4y	unknown	yes	no	telephone	may	mon	262	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
55	retired	single	high.schoc	no	yes	no	telephone	may	mon	342	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
41	technician	single	high.schoc	no	yes	no	telephone	may	mon	181	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
37	admin.	married	high.schoc	no	yes	no	telephone	may	mon	172	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
35	technician	married	university.	no	no	yes	telephone	may	mon	99	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
59	technician	married	unknown	no	yes	no	telephone	may	mon	93	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no
39	self-emplo	married	basic.9y	unknown	no	no	telephone	may	mon	233	1	999	0	nonexister	1.1	93.994	-36.4	4.857	5191	no

3.1 Attribute Information

Total of 21 attributes (20 Predictor variables + 1 Target variable)

S. No	Variables	Type
1	age	Numeric
2	job	Categorical:'admin.','bluecollar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown
3	marital	Categorical:'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed
4	education	Categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course', 'university.degree','unknown'
5	default	Categorical: 'no','yes','unknown' has credit in default?
6	housing	Categorical: 'no','yes','unknown has housing loan?
7	loan	Categorical: 'no','yes','unknown' has personal loan?
8	contact	Categorical: 'cellular','telephone' contact communication type
9	month	Categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec last contact month of year
10	day_of_week	Categorical: 'mon','tue','wed','thu','fri' last contact day of the week
11	duration	Numeric last contact duration, in seconds
12	campaign	Numeric, includes last contact
13	pdays	Numeric; 999 means client was not previously contacted number of days that passed by after the client was last contacted from a previously
14	previous	Numeric: number of contacts performed before this campaign and for this client
15	poutcome	Categorical: 'failure','nonexistent','success' outcome of the previous marketing campaign
16	emp.var.rate	Numeric: quarterly indicator
17	con.price.idx	Numeric: monthly indicator
18	cons.conf.idx	Numeric: monthly indicator
19	euribor3m	Numeric: daily indicator euribor 3 month rate
20	nr.employed	Numeric: quarterly indicator number of employees
21	**y**	**Binary: 'yes','no' ----------has the client subscribed a term deposit?**

3.2 Data Sources

The dataset has been collected from UCI machine Learning Repository.

- http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

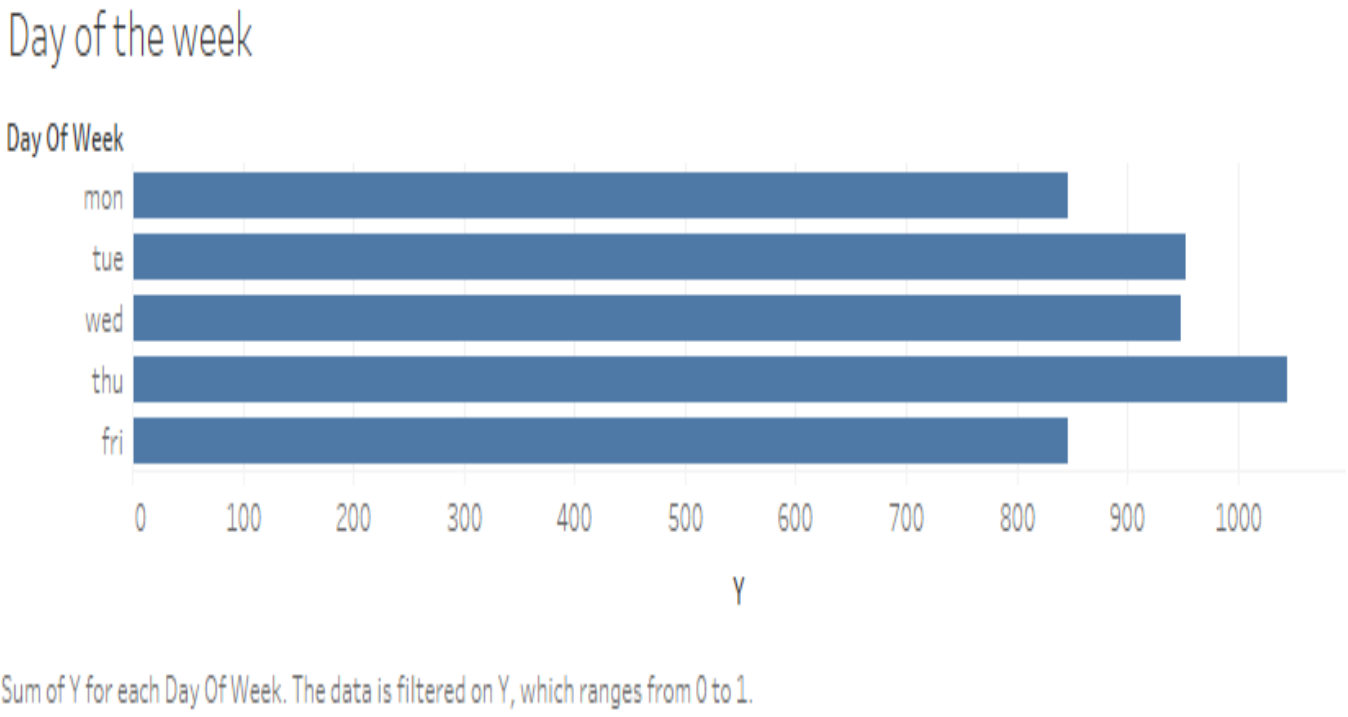
4 Data Cleansing

- The first data preprocessing included formatting the data using excel and making it into a readable csv format for loading in R.
- The data has lot of NA's and missing values
- Excel has been used for most of the cleaning.
- Apart from it the target variable was converted into numeric from categorical for the visualization purposes
- The data had lot of no's (80%) than yes's (20%) i.e. data is not normalized
- To normalized the data appropriate changes were done.
- Some of the unimportant predictors were removed.

5 Visualizations

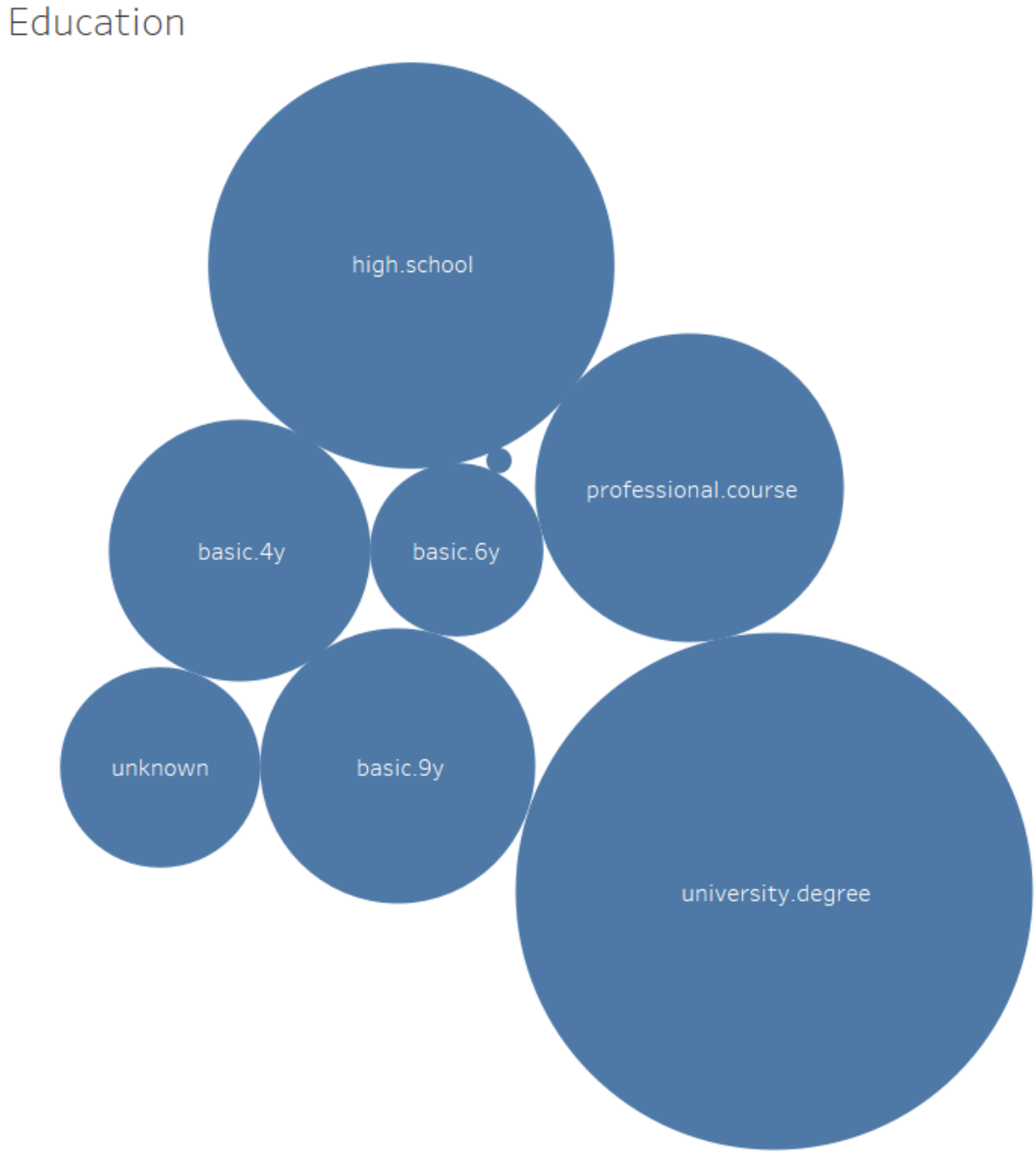
V1: Number of people who said yes for term deposits according to day of the week.

As seen in below graph, there are highest number of people who said yes when called them on Thursday.



V2: Categorizing all people in dataset per their education

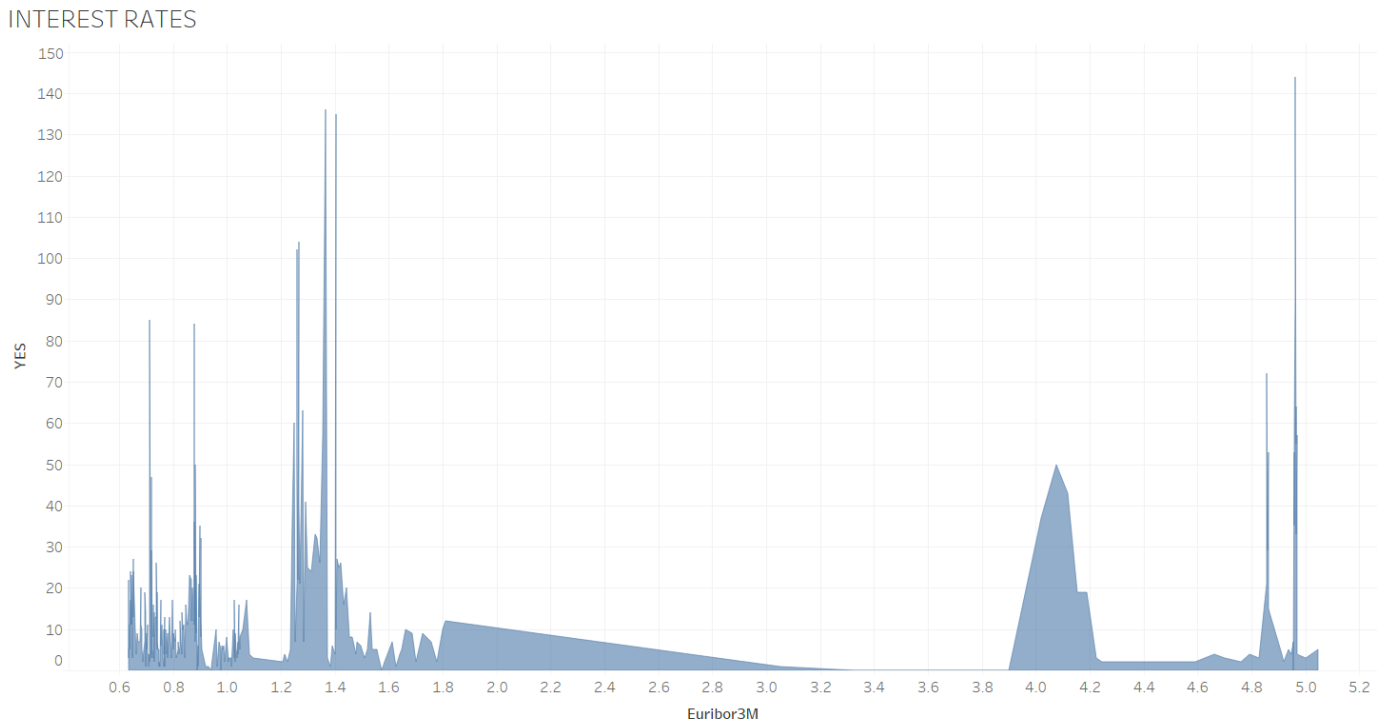
As seen from below graph, most of the people bank called was in university degree, high school and professional course.



V3: Number of people who said yes vs Europe interbank offer rate (Euribor3M)

We can see that there is no relation between 3-month interest rate and the number of people accepting for term deposits. But there are lot of people who said yes for term deposits with in a range of 0.6 – 1.8%.

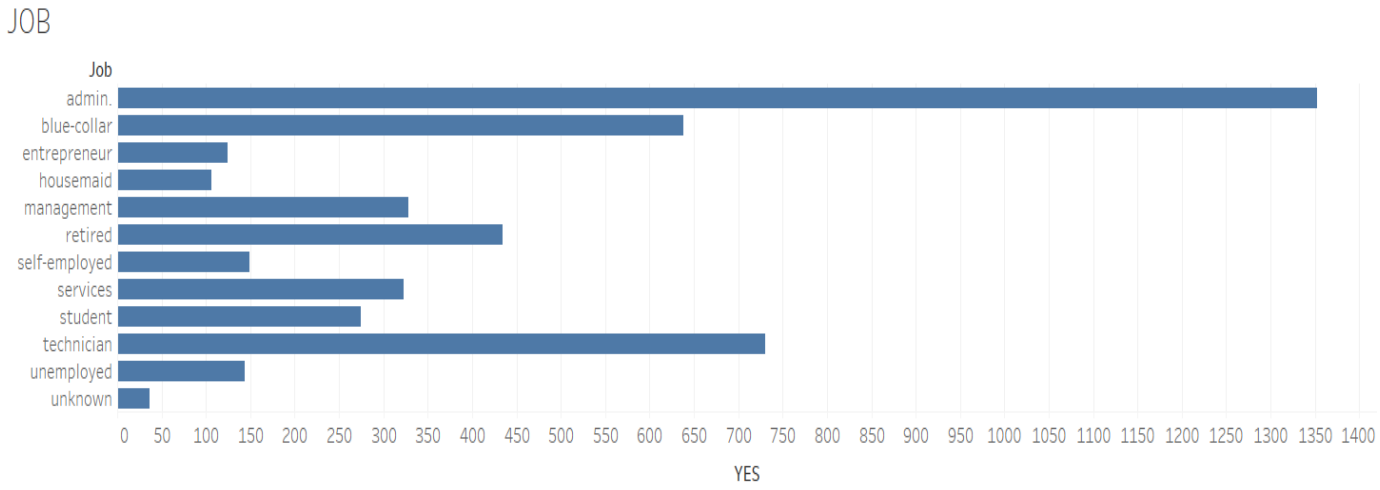
-



The plot of sum of Y for Euribor3M.

V4: Number of people accepting depending on their job profile

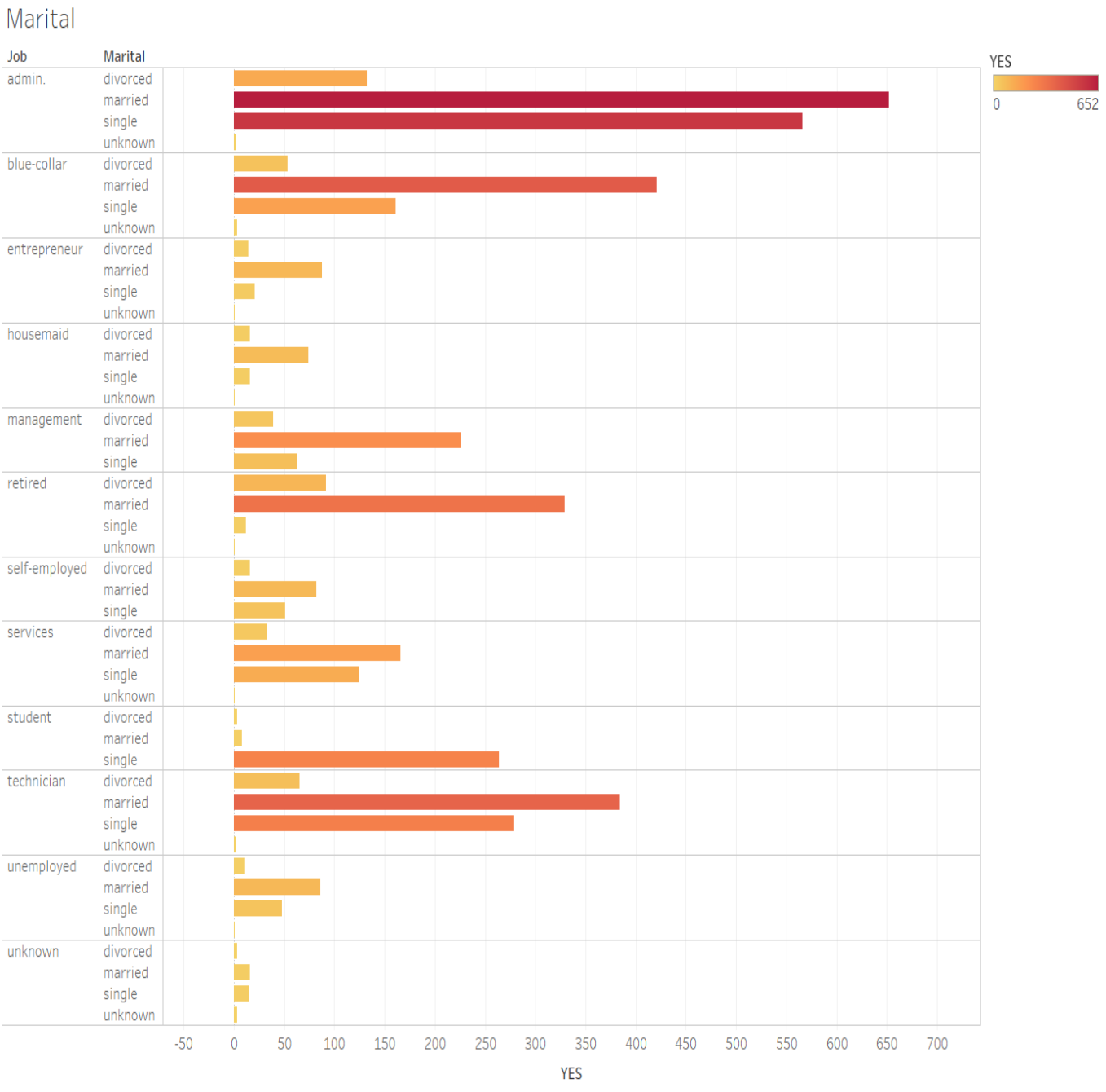
People with jobs like admin, blue-collar and technician have contributed more for banks term deposits.



Sum of Y for each Job. The data is filtered on Action (Education), which keeps 8 members.

V5: Number of people (clients) saying yes depending on their marital status

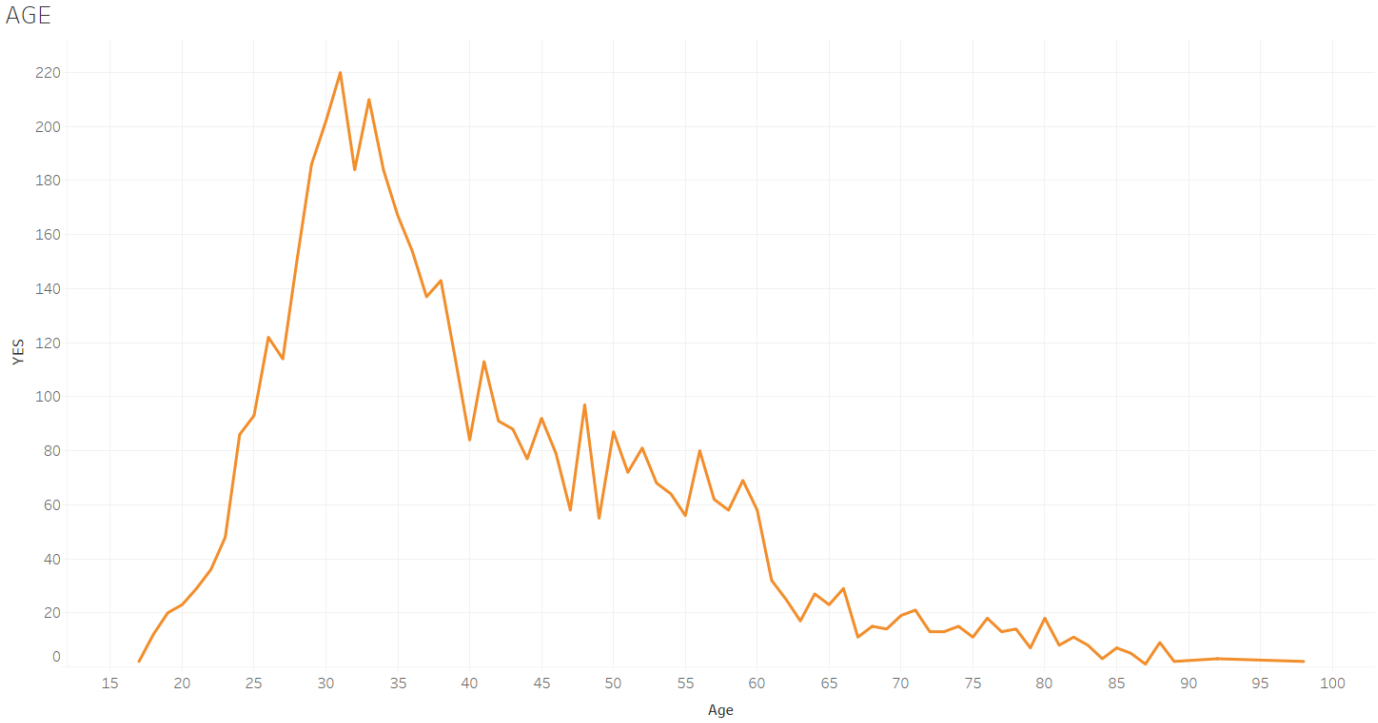
In almost every job, people who got married has highest number of acceptance for term deposits compared to divorced and single.



Sum of Y for each Marital broken down by Job. Color shows sum of Y.

V6: Number of people accepting for term deposits vs age of people

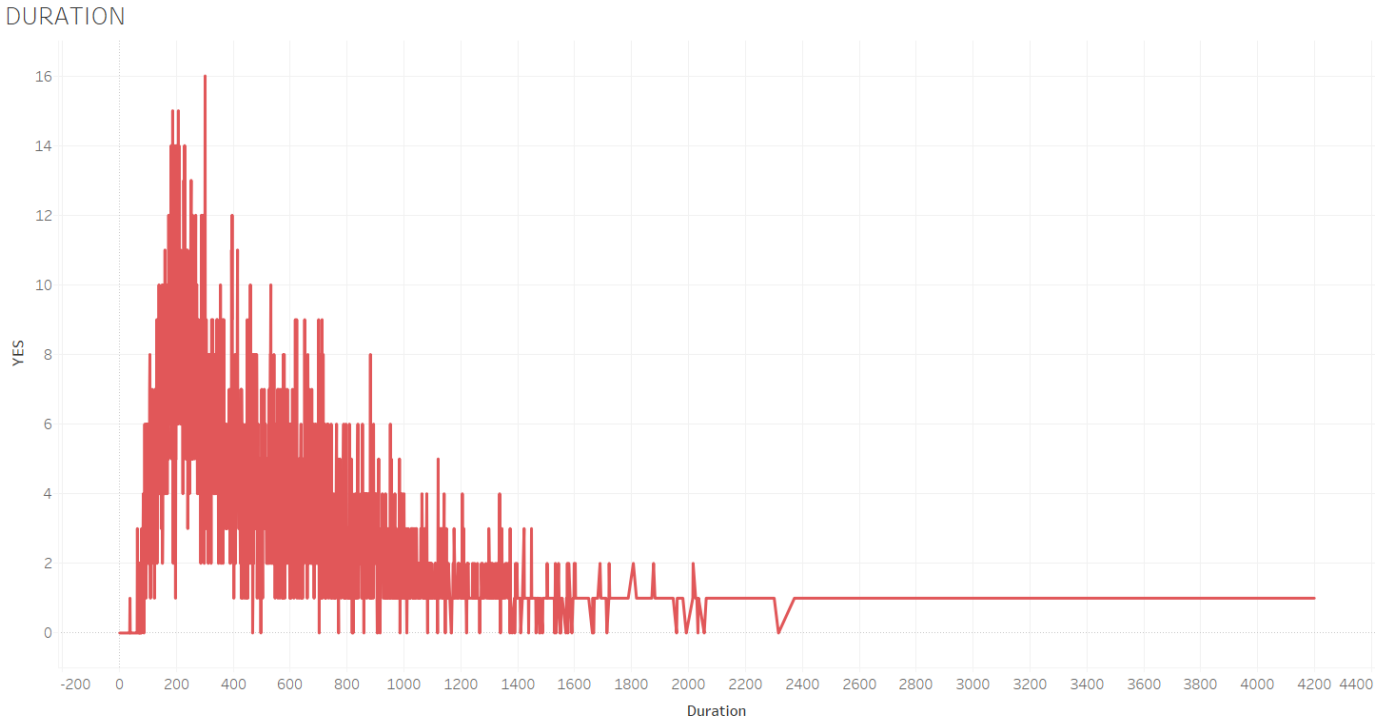
Highest number of people accepting term depositing are between the age of 25 – 40.



The trend of sum of Y for Age.

V7: Duration of phone call vs number of people accepting term deposits

If a phone call duration is between 200 – 400 seconds, there is highest chance of saying yes.



The trend of sum of Y for Duration.

6 ANALYSIS

Cleary our response variable is a binary type. So, we cannot go for linear regression. The best option is to go for classification. So here we tried two models

- ✓ Logistic Regression
- ✓ Decision Trees

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead.

Decision tree is a common and popular data mining technique that uses the tree hierarchy for data classification and rule inductions. The internal nodes of the tree represent the attributes' tests, the branches hold the resulted tests' values, and the leaf nodes represent the class labels for the decision attributes. For new object classification, a path for the attribute values of that object are examined according to the decision tree nodes and branches, starting from the root node till reach to the leaf node that holds the class label, such class label is considered as the class prediction for the new object

6.1 Logistic Regression

Logistic Regression is applied on train data using backward selection.

BEST PREDICTORS

- ✓ Euribor3m
- ✓ Contacttelephone
- ✓ Defaultyes
- ✓ jobstudent
- ✓ campaign
- ✓ emp.var.rate

CONFUSUSION MATRIX

	obs	
pred	0	1
0	1970	531
1	251	833

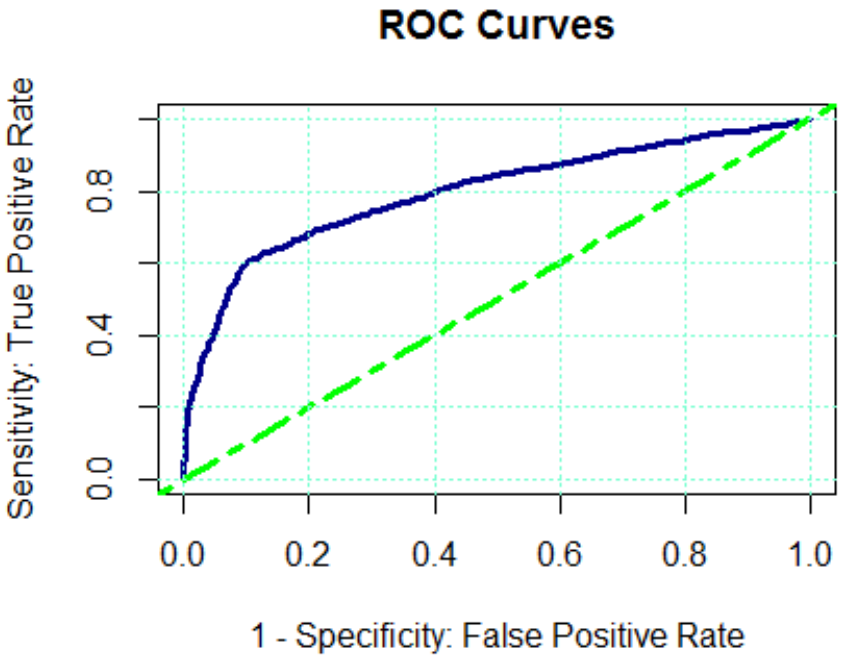
From the above confusion matrix, we can say that it has predicted 1970 times correctly that a customer will not make a deposit and 833 times correctly that a customer is going to make a deposit. So, this model is more helpful in knowing which customers won't make a deposit rather than who are going to make a deposit.

AccuMeasures

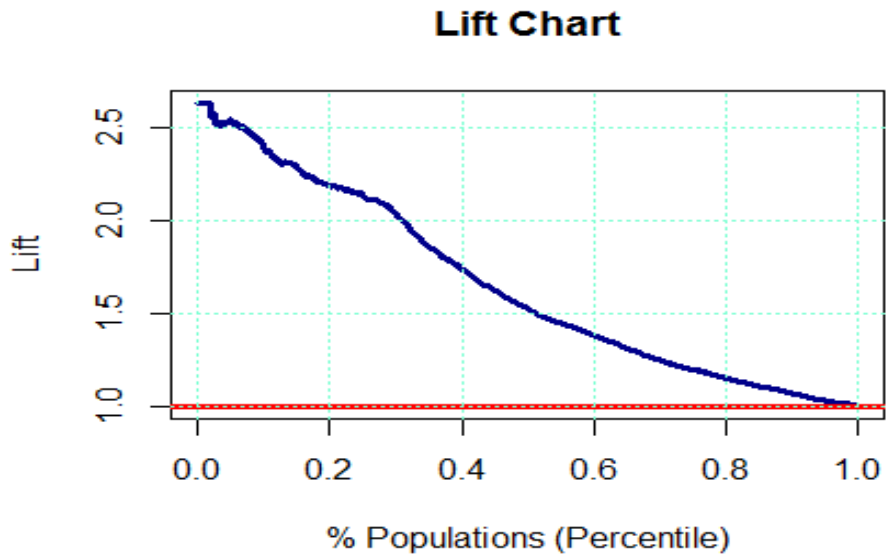
threshold	0.5
AUC	0.7248936
sensitivity	0.5227273
specificity	0.9270599
prop.correct	0.7732218

The high specificity value tells that the model is excellent in predicting true negative i.e. correct predictions of who are not going to deposit. This was an expected result as we already concluded that from the confusion matrix. which was already concluded from the confusion matrix. The sensitivity value (reasonably good) shows the model is good at even predicting the True Positive value i.e. customers who are likely to make a deposit.

ROC Curve:



Lift Chart:



Gains Table:

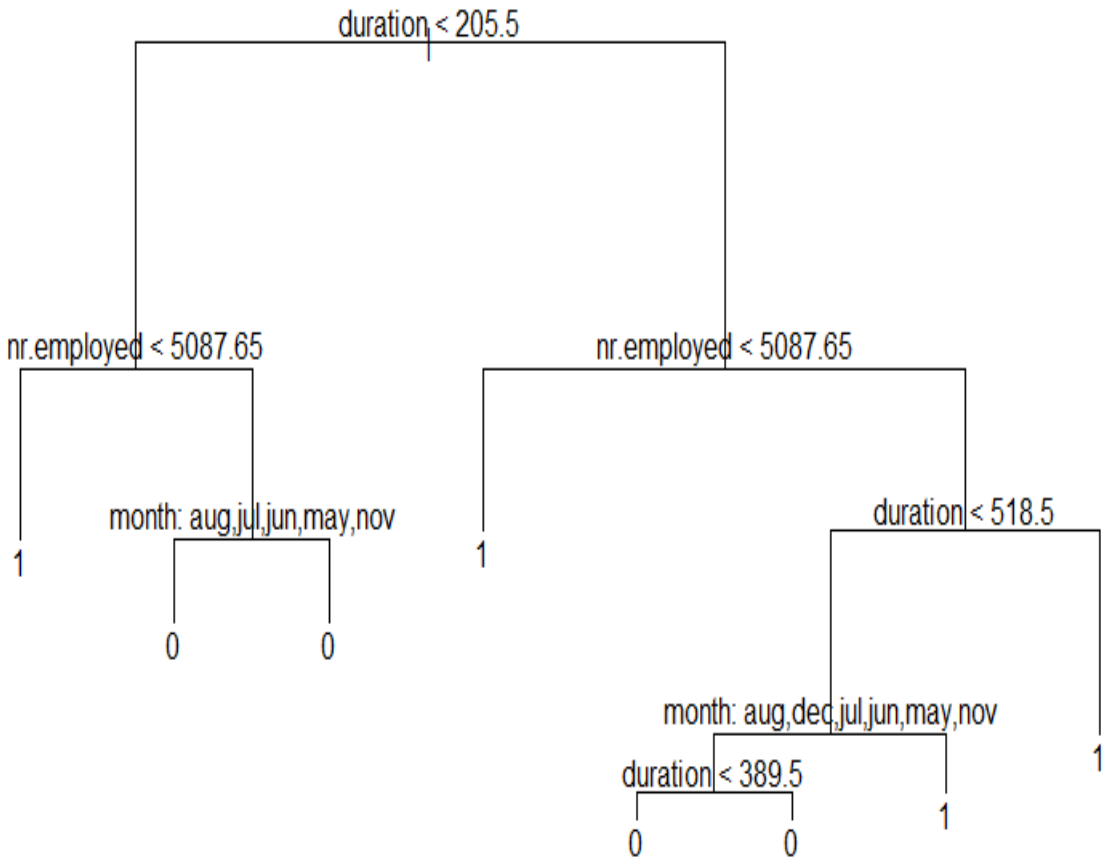
Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	358	358	0.91	0.91	23.9%	239	239	2.61
20	359	717	0.75	0.83	43.7%	198	218	1.25
30	360	1077	0.66	0.77	61.0%	172	203	0.47
40	357	1434	0.32	0.66	69.4%	85	174	-0.48
50	358	1792	0.26	0.58	76.2%	68	153	-0.91
60	359	2151	0.26	0.53	83.0%	67	138	-1.12
70	358	2509	0.16	0.47	87.2%	43	125	-1.28
80	359	2868	0.18	0.44	91.9%	47	115	-1.45
90	359	3227	0.17	0.41	96.3%	44	107	-1.65
100	358	3585	0.14	0.38	100.0%	37	100	-1.91

From the gains table we can say that with 50% of data only 76.2% good predictions can be made.

6.2 Decision Tree

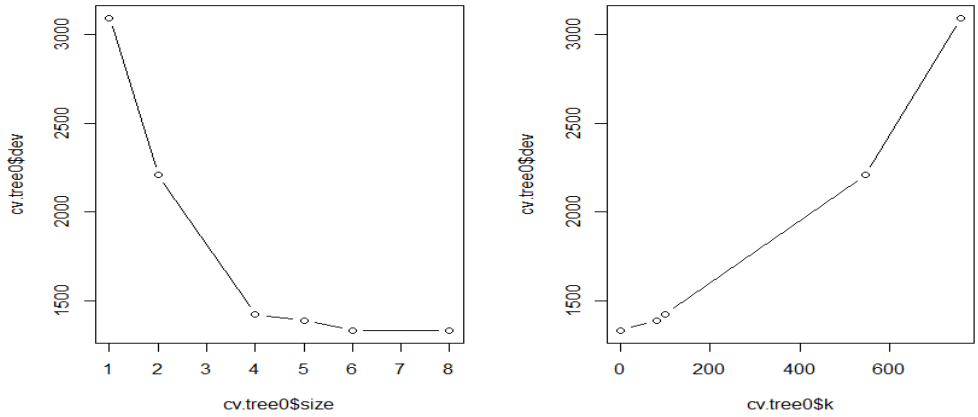
The decision tree shown below is final classification tree after pruning. Accuracy of original tree has not improved but there has been significant reduction in overfitting of data to the model after pruning.

Before Pruning:



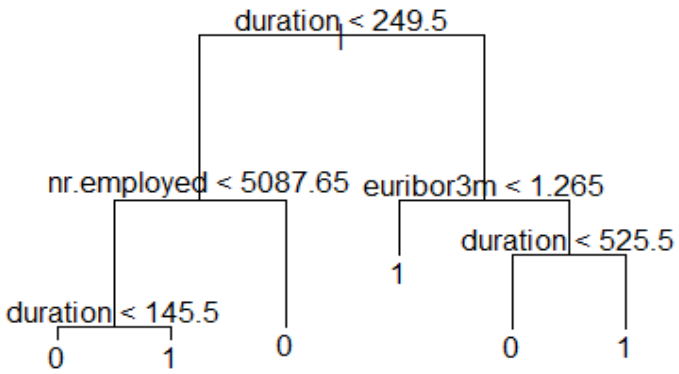
From the above tree, we can say that duration is the best variable. Second important variable is no.of people employed.

Before pruning tree, we have to decide the exact number of splits in a tree in order to achieve least error. Therefore cv.tree() function is used to calculate errors for different splits and plot using plot() function.



According to above graphs, least error is achieved when tree is pruned at n=6.

After Pruning:



CONFUSION MATRIX

tree0.pred	No	Yes
0	1994	309
1	227	1055

ACCURACY : 0.850

From the above confusion matrix, we can say that it has predicted 1994 times correctly that a customer will not make a deposit and 1055 times correctly that a customer is going to make a deposit. So, this model is good at predicting both the customers who won't make a deposit and also who are going to make a deposit. Now this is obvious much better than logistic model.

7 CONCLUSION

- Reasoning about banking and marketing knowledge base is one of the most challenging issues due to the vast growth of data records and transactions.
- Accordingly, various data mining approaches are there to make better decision making. In this project, we focused on decision tree and logistic regression as their abilities for classification of new objects.
- Real marketing banking data, which has been obtained from Portuguese marketing campaign related to bank deposit subscription, is used in both techniques for extracting significant decision rules, and discovering the significant features that discriminate between objects.
- And finally, we can say that classification model has outperformed logistic regression model.
- Also, both models could predict better on the customers who are not likely to make a deposit than the customers who are most likely to make a deposit.

8 REFERENCES

- http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
- S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
- S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]

9 APPENDIX

```
install.packages("caret")
install.packages("SDMTools")
install.packages("boot")
install.packages("ROCR")
install.packages("gains")
install.packages("tree")
library(SDMTools)
library(caret)
library(boot)
library(ROCR)
library(gains)
library(tree)

#reading the data
bank <- read.csv("bank-additional-full.csv")

#dividing the data into training and testing data
set.seed(9)
trainingindex <- sample(nrow(bank), 0.7*nrow(bank))
bank1 <- bank[trainingindex,]
bank2 <- bank[-trainingindex,]
nrow(bank1)
nrow(bank2)


#############################################################################
#################################
#################  LOGISTIC REGRESSION  ########################

logit <- glm(data = bank1, y ~.-duration,family = "binomial")

#backward selection loigistic regression model
logit1 <- step(logit, direction = "backward")
summary(logit1)

#testing the logit1 model on the test data
bankp <- predict(logit1,bank2)

#building confusion matrix
matrix1 <- confusion.matrix(bank2$y, bankp,threshold = 0.1)
matrix1

#Accuracy Measures
AccuMeasures1 <- accuracy(bank2$y, bankp)
```

AccuMeasures1

```
# creating a range of values to test for accuracy
thresh=seq(0,1,by=0.05)

# Initializing a 1*20 matrix of zeros to save values of accuracy
acc = matrix(0,1,20)

# computing accuracy for different threshold values from 0 to 1 step by 0.05
for (i in 1:21){
  matrix = confusion.matrix(bank2$y,bankp,threshold=thresh[i])
  acc[i]=(matrix[1,1]+matrix[2,2])/nrow(bank2)
}

# print and plot the accuracy vs cutoff threshold values

print(c(accuracy= acc, cutoff = thresh))
plot(thresh,acc,type="l",xlab="Threshold",ylab="Accuracy", main="Validation Accuracy for
Different Threshold Values")



# create the ROC, LIFT and GAIN Curves

logit_scores <- prediction(predictions=bankp, labels=bank2$y)

#PLOT ROC CURVE
logit_perf <- performance(logit_scores, "tpr", "fpr")

plot(logit_perf,
     main="ROC Curves",
     xlab="1 - Specificity: False Positive Rate",
     ylab="Sensitivity: True Positive Rate",
     col="darkblue",  lwd = 3)
abline(0,1, lty = 300, col = "green",  lwd = 3)
grid(col="aquamarine")


# AREA UNDER THE CURVE
logit_auc <- performance(logit_scores, "auc")
as.numeric(logit_auc@y.values)  ##AUC Value


# Getting Lift Charts in R
# For getting Lift Chart in R, use measure="lift", x.measure="rpp" in the performance function.
# Get data for ROC curve and create lift chart values
logit_lift <- performance(logit_scores, measure="lift", x.measure="rpp")
```

```
plot(logit_lift,
    main="Lift Chart",
    xlab="% Populations (Percentile)",
    ylab="Lift",
    col="darkblue", lwd = 3)
abline(1,0,col="red",  lwd = 3)
grid(col="aquamarine")

## GAINS TABLE

# gains table
# Three most important parameters in this functions are
# (1) actual vector with observed target values,
# (2) predicted vector with predicted probabilities and
# (3) groups to mention number of groups to be created.
gains.cross <- gains(actual=bank2$y , predicted=bankp, groups=10)
print(gains.cross)

##########################################################################
#####################################
##################        DECISION TREE        ##########################

bank$y = ifelse(bank$y==1,"Yes","No")

# bank1 <- bank[trainingindex,]
# bank2 <- bank[-trainingindex,]

par(mfrow=c(1,1))
tree0 = tree(as.factor(y)~.,bank1)
summary(tree0)
plot(tree0)
text(tree0,pretty=0)   # pretty=0 includes the category names for any qualitative predictors
tree0
bank2$y = ifelse(bank2$y==1,"Yes","No")
tree.pred = predict(tree0 ,bank2,type="class")
table(tree.pred,bank2$y)



set.seed(3)
cv.tree0 = cv.tree(tree0,FUN=prune.misclass)
names(cv.tree0)
cv.tree0
par(mfrow=c(1,2))
plot(cv.tree0$size,cv.tree0$dev,type="b")
plot(cv.tree0$k,cv.tree0$dev,type="b")
```

```
par(mfrow=c(1,1))
prune.tree0 = prune.misclass(tree0,best=6)
plot(prune.tree0)
text(prune.tree0,pretty=0)

tree0.pred = predict(prune.tree0,bank2,type="class")
table(tree0.pred, bank2$y)
```