IDS 575 Project

Objectives

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y). We have to know first what are the factors that affect the decision of the customers. Then, we can use those variables to classify if which customers would subscribe the deposits and which customers would not. The steps of finding the best model to predict whether the client will subscribe a term deposit or not as follows:

Dataset Information

Bank client data:

1. age
2. job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. marital : marital status (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','unkn
5. default: has credit in default? (categorical: 'no','yes','unknown')
6. housing: has housing loan? (categorical: 'no','yes','unknown')
7. loan: has personal loan? (categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

8. contact: contact communication type (categorical: 'cellular','telephone')
9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed.

Other attributes:

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3 month rate - daily indicator (numeric)

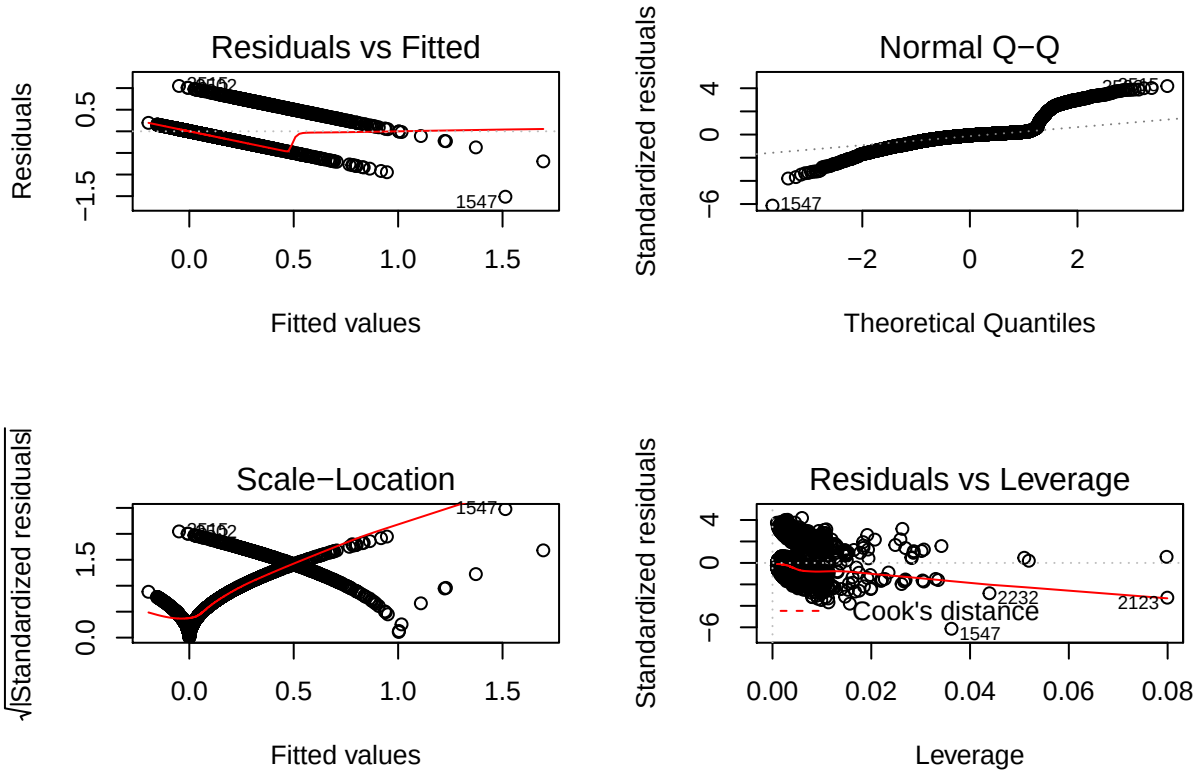
20. nr.employed: number of employees - quarterly indicator (numeric) ###Output variable (desired target):
21. y - has the client subscribed a term deposit? (binary: 'yes','no')

Linear regression

I use linear regression to find out which parameters have direct relationpship with the variable y which is a term deposit.

- Forward stepwise selection is used to get the best variables affecting variable y the most.
- Here are the statiscally signicant variables:

duration, nr.employe,pdays, month, cons.conf.idx, poutcome, previous, contact,ons.price.idx, education, emp.var.rate



- According to the graphs, we can see that linear regression may not be appropriate to use for classifying variable y.

Logistic regression

Since variable y is a nominal variable (0 and 1), it is better to use logistic regression to make a classification. First of all, we have divided data into two sets, training and test data. Also, we have to find which variables are statistically significant to variable y. We use the same method as mentioned.

- Then, the variables that have importance the most are duration, nr.employed, pdays, month, cons.conf.idx, contact, poutcome,campaign.

```
##
## bank.glm.pred    0    1
##             0 2563  180
```

```
##             1   58  132
```

- Here are the training confusion matrix. We still can see that the model still cannot predict y equal to 1. The specificity of this model is only 0.456 percent. This is still too low since the objective is to classify "y" correctly.
- Since the number of people who subscribe for a bank term is much lower than the number of people who do not subscribe. Therefore, I duplicate the data of y equal 1 and repeat it for 6 times in order to balance the two classed. Consequently, I construct a model by still using logistic model
- After duplicating the number of rows, the variables which are important to variable y are still the same as mentioned above.

```
##
## data.glm.pred    0    1
##             0 2236  316
##             1  339 1905
```

- From the above confusion matrix. we can see that we have got a better model after adding more rows having varible y. I used cross validation to find the more precise error rate of this model, which is 0.0983669 The over all accuracy is 86.34 percent, considered very high and the specificity is now 0.8577

```
##
## data.glm.test.pred   0   1
##                  0 860 729
##                  1 233 207
```

- The confusion matrix above shows that the specificity of the test data is 0.2233 which is relatively low. This model is overfitting.

Decision Tree

Decision tree is used to predict qualitative response. Rpart package is used for constructing the model. First of all, I use only cp equal 0 so that the tree is fully grown and I find the minimum error and select cp later.

```
##          Actual
## Predicted    0    1
##         0 2233  133
##         1  342 2088
```

- After the tree is prunned, cp is 0.0022512. We can see that decision tree results has much better specificity rate which is 0.9356

```
##          Actual
## Predicted   0   1
##         0 922  63
##         1 171 873
```

- The specificity for the test data is 0.9357 which is similar to the number of training data. Therefore, there is no overfitting issue for this model.

Bagging method

- Bagging can improve predictions for many regression methods. I used 500 number of trees for growing the model and use the number of columns which is 20 for paramenter "mtry"

```
## Warning in randomForest.default(m, y, ...): invalid mtry: reset to within
## valid range
```

```
##           Actual
## Predicted    0    1
##         0 2575    0
##         1    0 2221
```

- From the confusion matrix above, we can see that the model can predict y equal to 1 as well as 0 correctly. Therefore, the specificity rate is 1

```
##           Actual
## Predicted    0    1
##         0 1003    0
##         1   90  936
```

- Trying the test data with the bagging model. It can predict variable y equal to 1 perfectly but it has missclassified y equal to 0 for 92 out of 1081. However, the accuracy rate of this model is 95.467.

Random Forests

Random forests provide an improvement over bagged trees by way of a random small tweak that decorrelates the trees. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose m equivalent $\sqrt{p}$. Therefore, for this data set, m is around 4.5825757

```
##           Actual
## Predicted    0    1
##         0 2574    0
##         1    1 2221
```

- We can see that random forest could give 100 percent of accuracy, recall, as well as specificity rates.

```
##           Actual
## Predicted    0    1
##         0 1001    0
##         1   92  936
```

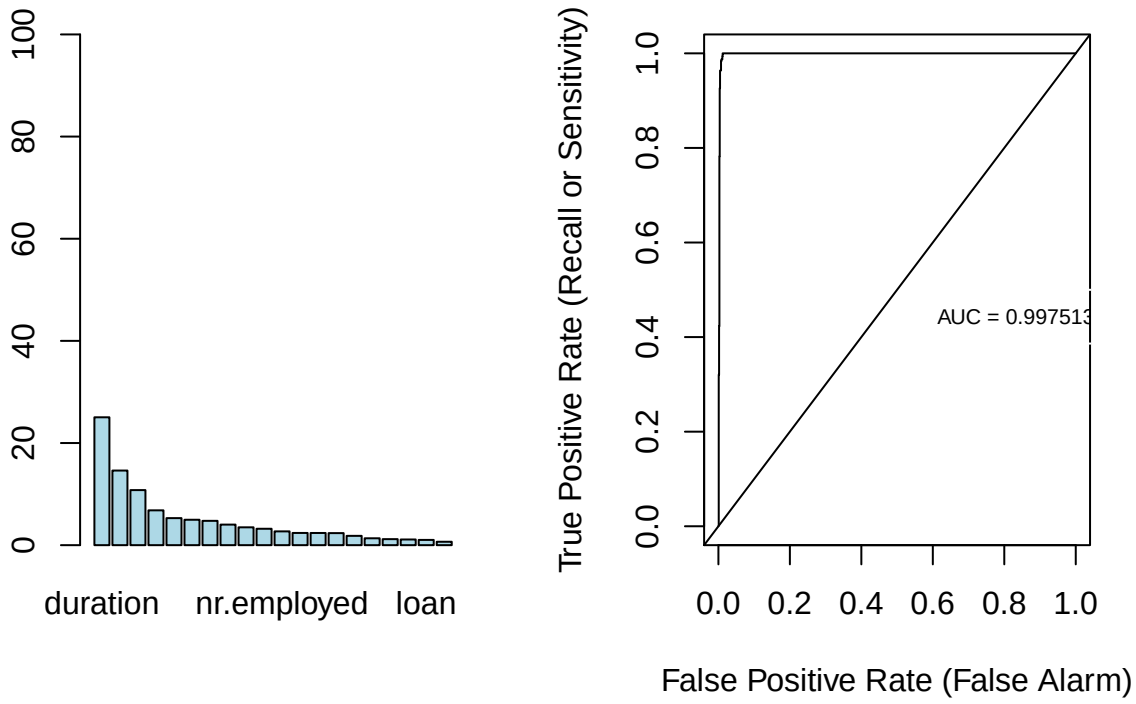
- However, the model with test data results in less percentage of recall rate but it gives the a better overall result than the bagging model.

Boosting

The term 'Boosting' refers to a family of algorithms which converts weak learner to strong learners.It works on similar method as discussed above. It fits a sequence of weak learners on different weighted training data. It starts by predicting original data set and gives equal weight to each observation. If prediction is incorrect using the first learner, then it gives higher weight to observation which have been predicted incorrectly.

```
##                Observed Class
## Predicted Class    0    1
##               0 2575    0
##               1    0 2221
```

- Adaboost or boosting can make a classification of variable y almost perfectly



- We can see that which varaibles are the most important, which affects the decision to subscribe the program or the y variable.
- The second graph is AUC graph whic the x axis represents False Postive Rate and they axis is True Positive Rate or Recall rate. AUC is approximatlely 0.99. This shows that this model could possibly classify the variable y perfectly

SVM

A support vector machine is a classification method. Its principle is fitting a boundary to a region of points which are all alike (that is, belong to one class). Once a boundary is fitted (on the training sample), for any new points (test sample) that need to be classified, we must simply check whether they lie inside the boundary or not.

- Advantage The advantage of SVM is that once a boundary is established, most of the training data is redundant. All it needs is a core set of points which can help identify and set the boundary. These data points are called support vectors because they "support" the boundary.

```
##
##        0    1
##   0 2575    0
##   1    0 2221
```

- We can see that support vector machine for the training data could classify the the data into two classes completely.
- Then, we will try this model with the test data.

```
##
## prd    0    1
##   0 1091    0
##   1    2  936
```

- The test data is almost perfectly classified. It could classify y equal to 1 perfectly.

Hypothesis

- H1: Loan has no direct effect on the customer's decision of a subscription on a term of deposit.

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## loan           1    0.7  0.7141   2.873 0.0901 .
## Residuals   6823 1696.0  0.2486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since the p value is 0.246, we can accept the null hypothesis.
- H2: Education has no direct relationship with a subscription on a term of deposit.

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## education      1   19.8  19.848   80.76 <2e-16 ***
## Residuals   6823 1676.8   0.246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- P-value is 2e-16; therefore, we fail to reject the null hypothesis.

Summary

- We can see that the variable y does not have a linear relationship with other variables. Therefore, the linear regresion and linear discriminant analysis cannot classify the variable y perfectly.
- Although logistic model could provide a better result; however, the false alam rate is still not good enough for the bank to identify which customer would subscribe for the bank program. Since logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression, the model cannot perform well.
- The best model should use for this dataset is bagging, random forests,boosting or svm. This is because both model gave almost the perfect classification. The error rate in total is so minimal and the specificity rate is as high as 100 percent. This is the reason why these models are the most suitable for this project.