IMARTICUS LEARNING

Project Report on European Bank

BY

NIKHIL AIVALLI

UNDER THE GUIDANCE OF

MRS.THARANGINI VIJAYKUMAR

ASSISTANT VICE PRESIDENT

CONTENTS

1. ABSTRACT

"FINAL PROJECT ON TELE-MARKETING CAMPAIGNS OF EUROPEAN BANKING INSTITUTIONS"

The data is about telemarketing campaigns of a European banking institution. The European bank wants to predict which clients will secure a term deposit based on a set of information on client and purchase of term deposit. The marketing is usually based on phone calls. Often, a client need to be persuaded multiple times in order to assess if the product (bank term deposit) would be or not subscribed. Predictive modelling approach will help the bank to manage their telemarketing campaign efficiently.

So basically, here we will start our project with having a brief outline of the project i.e. by adopting certain methodology to proceed further. Here we will use CTQR consulting framework and DER analytics framework. From which we will decide which modelling technique to be applied and the technology to be used is also decided. After deciding these we will proceed with the project and the first part here is we will check for the missing values and the outliers. And after that imputation of done. And then we will scale the data so that the model developed won't be biased. We will use R to build the model , the models include Logistic Regression , Decision Tree , SVM , Naïve Bayes, Random Forest and KNN.

In further report we will discuss in detail about these models and the results which we have obtained using these models.

2. INTRODUCTION

What is CTQR Consulting Framework?

We use CTQR framework to thoroughly understand the data first before applying the descriptive methods on it. So, it is very important to understand what data is about and what are its main objectives.

Context: The data is about telemarketing campaigns of a European banking institution.

Trigger: European bank wants to know which clients will secure a term deposit based on a set of information on client and purchase of term deposit.

Question: Is it possible to predict the clients which will secure the term deposit?

Response: Yes, it is possible with the help of DER Analytics framework.

Decision: Here in this process depending on the type of the data we will decide which of the following techniques to be applied on the data in order to predict the clients. As the dependent variable is categorical we will be making the decision (Yes/No).

The techniques lead to the type of technology which we want to use like R, Python, SAS etc.

Estimation: Here we don't use estimation technique as it is used while forecasting something.

Rank: Ranking is also not used as it is not an optimization problem.

3. LITERATURE SURVEY

1) The review of the literature survey has revealed very interesting facts about the Success of Bank Telemarketing . Many people have worked on this topic by adopting very different approaches.

2) S´ergio Moro et.al , during the year 2014 used different Data Mining models to predict the success of telemarketing . A Portuguese retail bank was addressed, with data collected from 2008 to 2013, thus including the effects of the recent financial crisis. Here they predicted the results using different predictive models like Decision Trees(DT), Logistic regression , Neural Network(NN). They used two metrics to compare the results AUC and ALIFT . And found out that Neural Network is giving the best results.

3) Leo Breiman et.al , during the year of 1984 , used the data of Household food security (HFS) represents the guiding principle underlying many rural development projects. Here they are using CART (classification and regression techniques) . Thus they were able to predict how can they use the existing information to learn what variables would provide us with an indication of which households are most likely to be food insecure?

4) David Arnott and Graham Pervan , during the year 2008 published a papeon decision support system(DSS). Decision support systems (DSS) is the area of the information systems (IS) discipline that is focused on supporting and improving managerial decision-making. In terms of contemporary professional practice, DSS includes personal decision support systems, group support systems, executive information systems, online analytical processing systems, data warehousing, and business intelligence.

4. DESCRIPTIVE ANALYSIS

The Descriptive analysis includes the following :

1) Data load and cleanup
2) Feature selection
3) Applying the algorithm

1. Data load and cleanup:

In this step we perform the following tasks:

1) Load the train and test dataset.
2) Check the train dataset for missing values and the outliers.
3) Delete the unique columns.

2. Feature selection:

As our dataset has more number of variables feature selection becomes a difficult task. So we directly build the model by overfitting and then get the summary of the model , find the significant variables and then build one more model with only significant variables .

3. Applying the Algorithms:

The different types of algorithms which we will be using in this model are as follows:-

1) Logistic Regression
2) Decision Tree
3) SVM
4) Naïve Bayes
5) Knn
6) Random Forest

5. METHODOLOGY

1) **Logistic Regression**: Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

   Assumptions of Logistic regression are :-

   1) The dependent variable should be categorical.
   2) There shouldn't be any multi-collinearity between independent variables.
   3) There should be relationship between dependent and independent variables.

   Advantages of Logistic regression :-

   1) It is more robust: the independent variables don't have to be normally distributed, or have equal variance in each group.

   2) It does not assume a linear relationship between the IV and DV .

   3) It may handle nonlinear effects .

   4) The DV need not be normally distributed .

   5) There is no homogeneity of variance assumption .

   6) Normally distributed error terms are not assumed .

   7) It does not require that the independents be interval .

   Dis-advantages of Logistic regression :-

1) Logistic regression attempts to predict outcomes based on a set of independent variables, but if we include the wrong independent variables, the model will have little to no predictive value.
2) Logistic regression works well with the categorical variables but it cannot be used to predict the continuous dependent variable.

3) Logistic regression requires that each data point be independent of all other data points. If observations are related to one another, then the model will tend to overweight the significance of those observations.

4) Logistic regression attempts to predict outcomes based on a set of independent variables, but logit models are vulnerable to overconfidence. That is, the models can appear to have more predictive power than they actually do as a result of sampling bias.

Results from the actual codes :-

```
model <-glm(y~ . , family = binomial(link = 'logit'), data = train_set)

summary(model)

model1<-
glm(y~job+education+contact+month+day_of_week+duration+campaign+poutcome+emp.var.rate+cons.price.idx+cons.conf.idx , family = binomial(link = 'logit') , data = train_set)
summary(model1)
```

```
Call:
glm(formula = y ~ job + education + contact + month + day_of_week +
    duration + campaign + poutcome + emp.var.rate + cons.price.idx +
    cons.conf.idx, family = binomial(link = "logit"), data = train_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9797  -0.2990  -0.1837  -0.1310   3.1406

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.27051    0.16456 -19.874  < 2e-16 ***
job2           -0.27287    0.09269  -2.944 0.003239 **
job3           -0.29353    0.15395  -1.907 0.056565 .
job4            0.01177    0.17379   0.068 0.946005
job5           -0.01168    0.10028  -0.117 0.907235
job6            0.13733    0.10756   1.277 0.201665
```

```
job7           -0.19361    0.14327  -1.351 0.176594
job8           -0.15160    0.10318  -1.469 0.141779
job9            0.09522    0.12427   0.766 0.443555
job10          -0.01700    0.08466  -0.201 0.840878
job11           0.11559    0.15042   0.768 0.442228
education2      0.04904    0.14171   0.346 0.729281
education3     -0.00833    0.11200  -0.074 0.940711
education4      0.02710    0.10715   0.253 0.800340
education5      1.12682    0.85397   1.320 0.187001
education6      0.06173    0.11957   0.516 0.605683
education7      0.22483    0.10377   2.167 0.030259 *
contact2       -0.51334    0.08299  -6.186 6.17e-10 ***
month2          0.68675    0.12328   5.571 2.54e-08 ***
month3          0.38562    0.23443   1.645 0.099981 .
month4          0.25204    0.11227   2.245 0.024778 *
month5          0.01063    0.10745   0.099 0.921170
month6          1.80673    0.13932  12.968  < 2e-16 ***
month7         -0.53282    0.09051  -5.887 3.94e-09 ***
month8         -0.24807    0.11530  -2.151 0.031441 *
month9          0.25157    0.14418   1.745 0.081009 .
month10         0.20485    0.15452   1.326 0.184930
day_of_week2   -0.21880    0.08029  -2.725 0.006430 **
day_of_week3    0.06326    0.07692   0.822 0.410841
day_of_week4    0.07372    0.07919   0.931 0.351876
day_of_week5    0.21933    0.07842   2.797 0.005159 **
duration        1.22730    0.02319  52.935  < 2e-16 ***
campaign       -0.04561    0.01413  -3.229 0.001243 **
poutcome2       0.49834    0.07509   6.637 3.21e-11 ***
poutcome3       1.84489    0.10194  18.097  < 2e-16 ***
emp.var.rate   -0.99638    0.02885 -34.533  < 2e-16 ***
cons.price.idx  0.75396    0.03936  19.157  < 2e-16 ***
cons.conf.idx   0.10896    0.02827   3.854 0.000116 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be
1)

    Null deviance: 20299  on 28831  degrees of freedom
Residual deviance: 11898  on 28794  degrees of freedom
AIC: 11974
```

```
# prediction
> y_pred<-predict(model1 , newdata = val_set[-20])
> y_pred = ifelse(y_pred>=0.5,1,0)
> table(y_pred)
y_pred
    0     1
11745   611
> write.csv(y_pred,file ="Final prediction")
> # confusion matrix
> cm<-table(y_pred, val_set$y)
> cm

y_pred     0     1
     0 10777   968
     1   187   424
>
> # accuracy
> acc<- function(cm){
+   tp<-cm[2,2]
+   fp<-cm[2,1]
+   acc<-tp/(tp+fp)
+   acc
+ }
> acc(cm) #
[1] 0.6939444
```

2) **Decision Tree** :-A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. The exact point of split of the node is found using the log worth values.

Advantages of decision trees :

1) Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
2) Allow the addition of new possible scenarios.
3) Help determine worst, best and expected values for different scenarios.
4) Can be combined with other decision techniques.
5) No imputation is required as the decision tree algorithm handles the missing values and the outliers.

Disadvantages of decision trees:

1) They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
2) They are often relatively inaccurate. Many other predictors perform better with similar data. This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to interpret as a single decision tree.
3) For data including categorical variables with different number of levels, information gain in decision trees is biased in favor of those attributes with more levels.

Results of Decision tree :-

```
classifier <- rpart(formula = y~., data = train_set)
>
> # prediction
>  y_pred2<-predict(classifier  ,  newdata  =  val_set[-20]  ,
```

```
type = 'class')
>
>
> # confusion matrix
> cm2 = table(y_pred2 , val_set$ y)
> cm2

y_pred2     0     1
      0 10495   631
      1   469   761
>
> # accuracy
> acc(cm2)
[1] 0.618699
```

3) **SVM** : A support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection . Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Advantages of svm :

1) The main advantage of the svm is that it can actually consider all the non linear data points by using the kernels.

Disadvantages of svm :

1) It is computationally very complex and takes lot of time to compute .

Results of SVM :-

```
library(e1071)
> set.seed(123)
> classifier1<-svm(formula = y~.,data = train_set, type
= 'C-classification' ,kernel = 'radial')
> classifier2<-svm(formula = y~.,data = train_set, type
= 'C-classification' ,kernel = 'sigmoid')
```

```
> classifier3<-svm(formula = y~.,data = train_set, type
= 'C-classification' ,kernel = 'polynomial')
> y_pred3<-predict(classifier1, newdata = val_set[-20])
> y_pred4<-predict(classifier2, newdata = val_set[-20])
> y_pred5<-predict(classifier3, newdata = val_set[-20])
> cm3<-table(y_pred3,val_set$y)
> cm3

y_pred3     0     1
      0 10734   904
      1   230   488
>
> cm4<-table(y_pred4,val_set$y)
> cm4

y_pred4     0     1
      0 10318   853
      1   646   539
>
> cm5<-table(y_pred5,val_set$y)
> cm5

y_pred5     0     1
      0 10830  1023
      1   134   369
> acc(cm3) # accuracy is 67.88
[1] 0.6796657
> acc(cm4) # accuracy 45.48
[1] 0.4548523
> acc(cm5)
[1] 0.7335984
```

4) **Naïve Bayes**: In machine learning, *naive Bayes classifiers* are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independent assumptions between the features.

Advantages of naïve bayes :

1) Easy to implement .
2) Requires a small amount of dataset to estimate the parameters .

3) Good results are obtained in most of the cases.

Disadvantages of naïve bayes :

1) Dependencies exist among the variables.
2) Dependencies among the variables cannot be modelled by naïve bayes classifier.

Results of Naïve bayes:

```
set.seed(5465)
> classifier4<-naiveBayes(y~.,data = train_set)
>
> # prediction
> y_pred6<-predict(classifier4,newdata = val_set[-20])
>
> #confusion matrix
> cm6<-table(y_pred6,val_set$y)
> cm6

y_pred6    0    1
      0 9328  433
      1 1636  959
>
> # accuracy
> acc(cm6) #accuracy = 36.95%
[1] 0.3695568
```

5)**KNN Classifier** : In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression . In both cases, the input consists of the *k* closest training examples in the feature space.In *k-NN classification*, the output is a class membership.

An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

Advantages of knn :

1) Effective if the training data is large .

Disadvantages of knn :

1) Need to determine the value of parameter k (number of nearest neighbors).
2) Distance based learning is not clear which type of distance to use and which attribute to use to produce the best results.
3) Computation time is quite high as the distance between each dataset has to be calculated in order to classify them.

Results of Knn as follows :

```
library(class)
>    y_pred7    <-knn(train    =    train_set,test    =
val_set,cl=train_set$y,k=300,prob = TRUE)
>
> #confusion matrix
> cm7<-table(y_pred7,val_set$y)
> cm7

y_pred7     0     1
      0 10956  1329
      1     8    63
>
> #accuracy
> acc(cm7) [1] 0.8873239
```

6) **Random Forest** : Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The advantages of random forest are:

1) It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
2) It runs efficiently on large databases.
3) It can handle thousands of input variables without variable deletion.

4) It gives estimates of what variables are important in the classification.

5) It generates an internal unbiased estimate of the generalization error as the forest building progresses.

6) It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

7) It has methods for balancing error in class population unbalanced data sets.

8) Generated forests can be saved for future use on other data.

9) Prototypes are computed that give information about the relation between the variables and the classification.

10) It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.

11) The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.

12) It offers an experimental method for detecting variable interactions

Dis-advantages of Random forest :

1) Random forests have been observed to overfit for some datasets with noisy classification/regression tasks.
2) For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

Results of the Random Forest :-

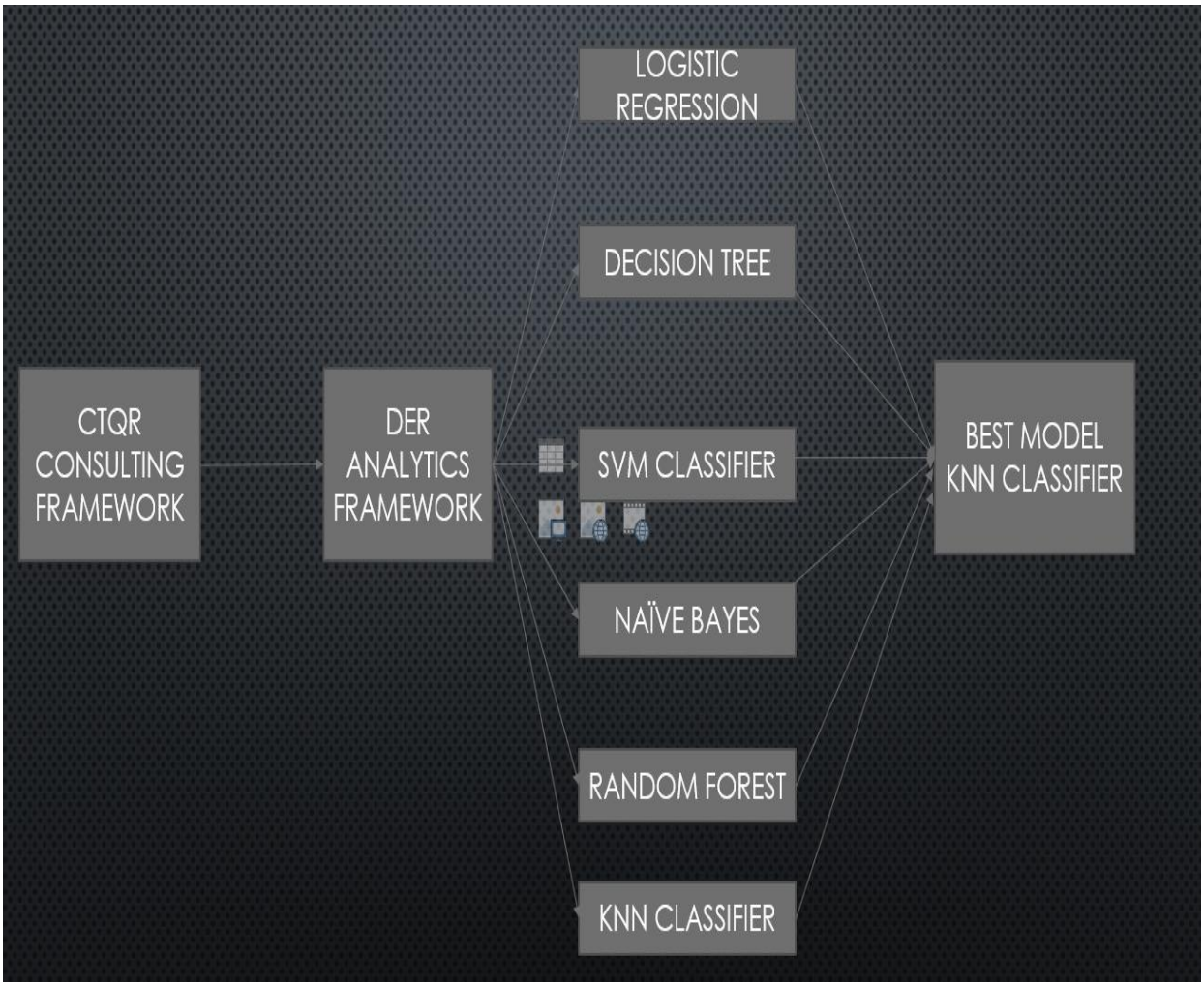
```
classifier4 = randomForest(x=train_set[-20], y = train_set$y, ntree = 200)
>
> # prediction
```

```
> y_pred8<-predict(classifier4,newdata = val_set[-20] )
>
> # confusion matrix
> cm8<- table(y_pred8,val_set$y)
> cm8

y_pred8     0     1
      0 10549   669
      1   415   723
>
> # accuracy
> acc(cm8)  # accuracy = 63.53%
[1] 0.6353251
```

6. CHAMPION CHALLENGER MODEL

After looking at the various results of the models which we have build, we came to the conclusion that knn gives the best accuracy of 88% . So the knn classifier is used to predict the outcome from the test dataset.



7. PREDICTION ON TEST DATASET

Results of the prediction using KNN :

```
# prediction on test data
>
>   result   <-knn(train   =   train_set[-20],test   =
test_set,cl=train_set$y,k=300,prob = TRUE)
> test_set<-data.frame(test_set,result)
> summary(test_set)
       age.v1               job       marital   education
default  housing  loan
 Min.    :-1.883753    1        :854    1: 348    1: 243
1:3089   1:1402   1:2583
 1st Qu.:-0.803627    10       :573    2:1791    2: 150
2:   1   2:1688   2: 507
 Median :-0.214468   2        :554   3: 951   3: 407
 Mean   : 0.000000   8        :276             4: 728
 3rd Qu.: 0.669271   5        :265             5:   1
 Max.   : 4.793387   7        :126             6: 454
                     (Other):442              7:1107
 contact        month       day_of_week       duration.v1
campaign
 1:2108   7       :981   1:580        Min.   :-0.982137
Min.   : 1.000
 2: 982   4       :514   2:642        1st Qu.:-0.588138
1st Qu.: 1.000
          2       :495   3:630        Median :-0.296427
Median : 2.000
          8       :387   4:613        Mean   : 0.000000
Mean   : 2.509
          5       :365   5:625        3rd Qu.: 0.211225
3rd Qu.: 3.000
          1       :169                Max.   :12.819186
Max.   :35.000
          (Other):179
    previous                   poutcome      emp.var.rate
cons.price.idx.v1
 Min.   :0.0000   1: 360   Min.   :-3.4000   Min.   :-
2.2721540
 1st Qu.:0.0000   2:2602   1st Qu.:-1.8000   1st Qu.:-
0.7803963
 Median :0.0000   3: 128   Median : 1.1000   Median :-
0.1505810
```

```
 Mean   :0.2081             Mean   :-0.0468   Mean   :
0.0000000
 3rd Qu.:0.0000             3rd Qu.: 1.4000   3rd Qu.:
0.7881681
 Max.   :6.0000             Max.   : 1.4000   Max.   :
2.1075373

   cons.conf.idx.V1                        euribor3m.V1
nr.employed.V1    result
 Min.   :-2.1450550   Min.   :-1.6082939   Min.   :-
2.5828077   0:3075
 1st Qu.:-0.4385156   1st Qu.:-1.2252465   1st Qu.:-
0.8111527   1:  15
 Median :-0.2489002   Median : 0.7764304   Median :
0.3904348
 Mean   : 0.0000000   Mean   : 0.0000000   Mean   :
0.0000000
 3rd Qu.: 0.8887928   3rd Qu.: 0.8357519   3rd Qu.:
0.8755152
 Max.   : 2.8902895   Max.   : 0.8832091   Max.   :
0.8755152
```

8. CONCLUSION

1. Which machine learning approach is appropriate to find the solution for the mentioned problem?

We found out that knn is the best machine learning approach for this dataset.

2. Predict the term deposit subscription for the Bank additional dataset and conclude if the telemarketing campaign was a success or not.

After we run summary on the result variable we came to know that there are only 16 people out of 3090 people who secured the term deposit .Thus we can conclude that the telemarketing campaign was an utter failure.

3. What is are the key differentiators between the ones who have subscribed (Yes) and who did not (No).

As we have used knn machine learning algorithm we actually don't know which are the significant variables effecting the prediction process.

9. REFERENCE

1) Moro et al., 2014] 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.

2) Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. Wadsworth & Brooks. Monterey, CA, 1984.

3) David Arnott and Graham Pervan. Eight key issues for the decision support systems discipline. Decis. Support Syst., 44(3):657–672, 2008.