LOGISTIC REGRESSION REPORT

Prepared by:

Team Data Insighters

Sreekar Bethu

Meghna Amin

Balacoumarane Vetrivel

Mohammed Ismail Khan

CONTENTS

1. Introduction 1
   1.1 Objective 1
   1.2 Data Source 1

2 Exploratory Data Analysis 1
   2.1 Data Cleaning 1
   2.2 Summary of the population 1
   2.3 Numeric Distribution 2
   2.4 Target vs Numeric predictors 2
   2.5 Target vs Categorical predictors 3

3 Sampling 3
   3.1 Sampling Techniques analyzed before building Model 3
      3.1.1 Numeric variable distribution in Sampled dataset 4
      3.1.2 Target VS Numeric predictors in Sampled dataset 4
   3.2 Correlation between numeric variables 5

4 Assumptions 5

5 Building the model 6
   5.1 Train and Test 6
   5.2 Regression 7
      5.2.1 Variable Significance 7
      5.2.2 Confusion Matrix and ROC Curve for Train and Test Data 8

6 Goodness of Fit 8
   6.1 Loglikelihood test 8
   6.2 HL test 9

7 Conclusion 9

1 Introduction:

Direct marketing is a form of advertising where organizations communicate directly to customers through a variety of media including phone calls, text messaging, emails etc. In direct marketing companies provide physical marketing materials to consumers to communicate information about a product or service. Banks engage in direct marketing to sell and provide services. A Portuguese Banking Institution has provided data related to direct marketing campaigns. The campaigns were based on phone calls to their customers to offer term deposit subscriptions.

1.1 Objective:

To build a model to identify the factors that influence client's decision to subscribe to term deposit using the predictive modelling method of Logistic Regression.

1.2 Data Source:

The dataset is about direct marketing campaigns of the Portuguese Bank Marketing and is obtained from the University of California, Irvine (UCI) Machine Learning Repository. During the marketing campaign multiple phone calls were made to the customers during the period between May 2008 to November 2010. The client responses and predictor variable information are used to assess whether the client will subscribe to the bank term deposit or not.

There are 21 variables and 41188 observations in the dataset. Amongst the 21 variables, 10 are continuous and other 10 are categorical. The target variable y is the binary response indicating whether the client has or has not subscribed to a term deposit.

2 Exploratory Data Analysis

Removed duplicates from the whole population. Added a new variable '*ynum*' representing 'y' (target variable) as numeric, 1 <- 'yes' & 0 <- 'no'.

2.1 Data Cleaning:

Missing Values: On checking missing values for the dataset in R, no missing values were found. This can be seen while summarizing the population in R.

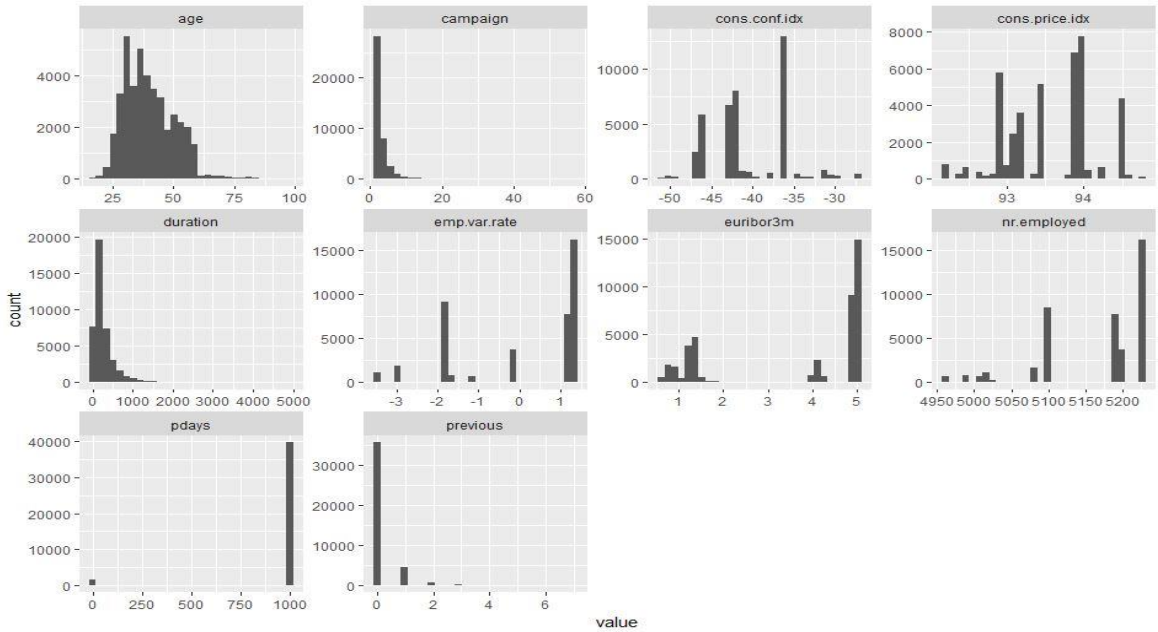
2.2 Summary of the population:

```
> summary(bank_population)
      age                job            marital                     education        default
 Min.   :17.00   admin.     :10419   divorced: 4611   university.degree  :12164   no     :32577
 1st Qu.:32.00   blue-collar: 9253   married :24921   high.school        : 9512   unknown: 8596
 Median :38.00   technician : 6739   single  :11564   basic.9y           : 6045   yes    :    3
 Mean   :40.02   services   : 3967   unknown :   80   professional.course: 5240
 3rd Qu.:47.00   management : 2924                    basic.4y           : 4176
 Max.   :98.00   retired    : 1718                    basic.6y           : 2291
                 (other)    : 6156                    (other)            : 1748
    housing           loan            contact          month       day_of_week    duration
 no     :18615   no     :33938   cellular :26135   may    :13767   fri:7826    Min.   :   0.0
 unknown:  990   unknown:  990   telephone:15041   jul    : 7169   mon:8512    1st Qu.: 102.0
 yes    :21571   yes    : 6248                     aug    : 6176   thu:8618    Median : 180.0
                                                    jun    : 5318   tue:8086    Mean   : 258.3
                                                    nov    : 4100   wed:8134    3rd Qu.: 319.0
                                                    apr    : 2631               Max.   :4918.0
                                                    (other): 2015
   campaign          pdays          previous          poutcome      emp.var.rate      cons.price.idx
 Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252   Min.   :-3.40000   Min.   :92.20
 1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35551   1st Qu.:-1.80000   1st Qu.:93.08
 Median : 2.000   Median :999.0   Median :0.000   success    : 1373   Median : 1.10000   Median :93.75
 Mean   : 2.568   Mean   :962.5   Mean   :0.173                       Mean   : 0.08192   Mean   :93.58
 3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000                       3rd Qu.: 1.40000   3rd Qu.:93.99
 Max.   :56.000   Max.   :999.0   Max.   :7.000                       Max.   : 1.40000   Max.   :94.77

 cons.conf.idx     euribor3m      nr.employed      y              ynum
 Min.   :-50.8   Min.   :0.634   Min.   :4964   no :36537   Min.   :0.0000
 1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099   yes: 4639   1st Qu.:0.0000
 Median :-41.8   Median :4.857   Median :5191               Median :0.0000
 Mean   :-40.5   Mean   :3.621   Mean   :5167               Mean   :0.1127
 3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228               3rd Qu.:0.0000
 Max.   :-26.9   Max.   :5.045   Max.   :5228               Max.   :1.0000
```

Target Variable distribution: The population data is having almost 90% of responses as 'no' and remaining 10% is 'yes'. Building a model with the whole population will lead to high type2 error, which should be avoided. Sampling can be done to get the responses in proper proportions.

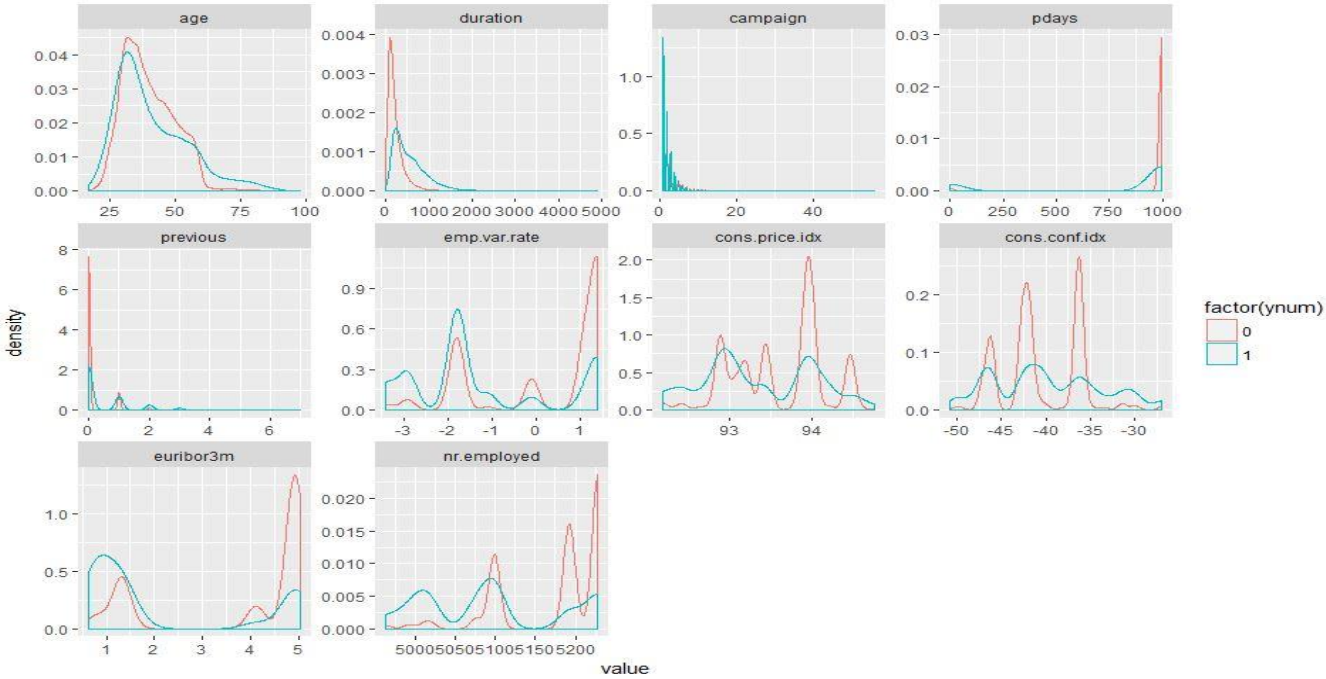
2.3 Numeric Distribution



From this graphical representation '*pdays'* and '*previous'* can be categorized as below:

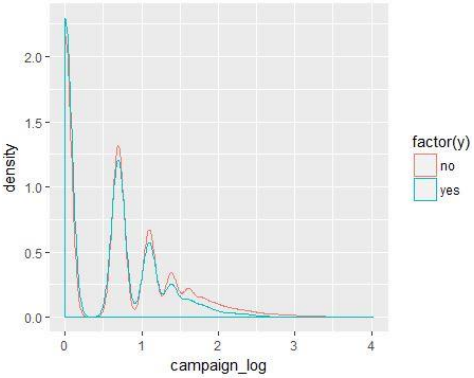
1. pdays: It represents the number of days passed since the customer was last contacted from the previous campaign.
   a) if pdays=999 then 'Not contacted previous campaign'
   b) if pdays<>999 then 'Contacted during previous campaign'
2. previous: it represents the number of times the customer was contacted before this campaign.
   a) if previous=0 then 'Never contacted before'
   b) if previous <> 0 then 'Contacted before'
3. Distribution of other numeric variables doesn't show any logical information to categorize.

2.4 Target VS Numeric predictors:



From the above graphs it can be inferred that:

a) Distribution of '*Age*' vs '*y*' is similar for both the responses 'yes' or 'no'. Therefore, the variable '*Age*' in predictors may not lead to an accurate model.
b) Distribution of '*Campaign'* vs y is highly skewed; therefore, log transformation of the '*campaign'* variable was plotted, and the result is seen as below.

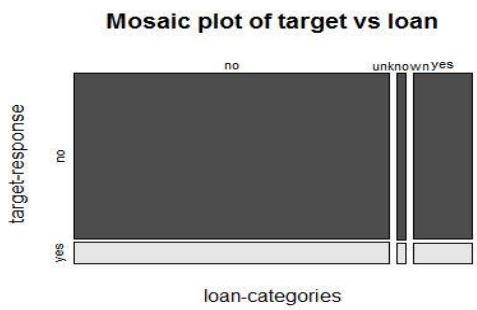


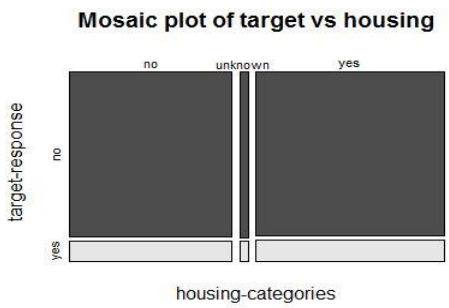
The log distribution of '*campaign'* vs '*y'* variable is similar for the responses 'Yes' and 'No'. Thus, the variable '*campaign'* as a predictor will not lead to an accurate model.

2.5 Target Vs Categorical Variables

On plotting the distributions for all categorical variables, it was observed that for the variables '*loan'* and '*housing*', distribution of the target response is the same for the categories 'Yes' and 'No'. Thus, the variables '*loan'* and '*housing*' as a predictor will not lead to an accurate model.

Following are the graphs for '*loan'* vs '*y*' and '*housing'* vs '*y*':





3 Sampling:

3.1 Sampling Techniques analyzed before building Model:

Sampling is done to get the responses in proper proportions. Different Sampling Techniques were implemented before finalizing the best Sampling method to be employed to build the model.

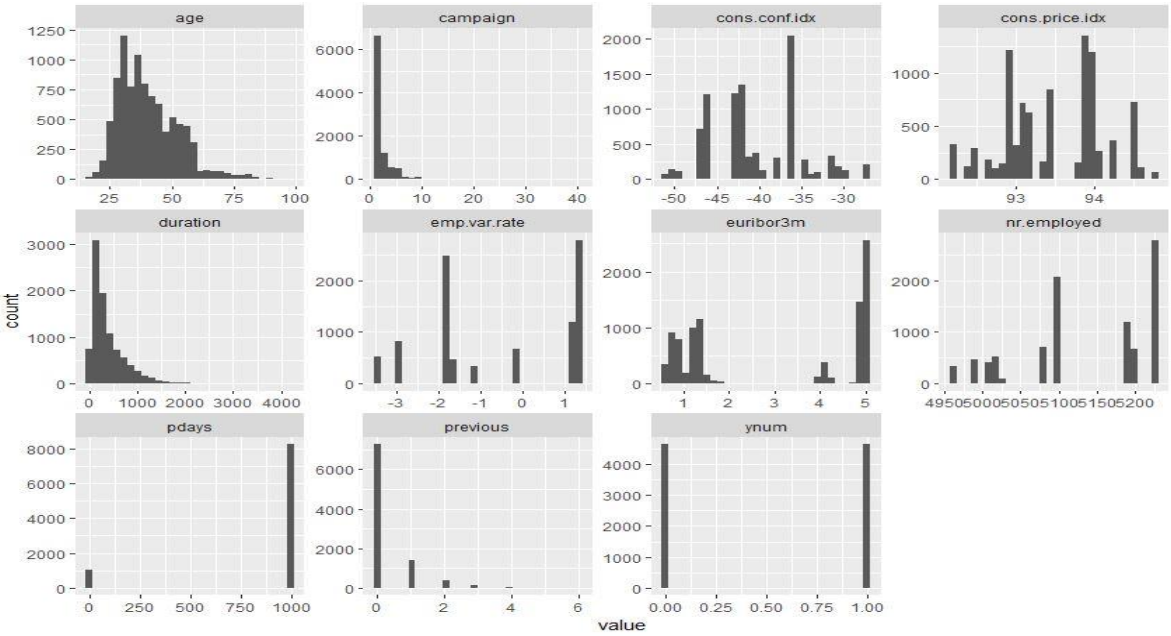
Sampling	Data	No. of records	AIC	Accuracy	Sensitivity	Specificity	ROC value	Significance (HL)
Over	Train	43844	54926	74.03	62.68	85.37	0.795	1.34E-13
	Test	29230		74.45	63.59	85.31	0.7959	
Under	Train	5564	5987.7	74.52	62.38	86.67	0.796	0.34
	Test	3712		73.6	62.02	85.18	0.7842	
SMOTE	Train	10495	12906	74.45	73.14	75.93	0.8133	1.665E-15
	Test	6997		74.18	72.26	76.35	0.8067	

Based on above table, it can be inferred that the model built using under sampling technique passes the Statistical Significance Test (HL Test p-value>0.05).

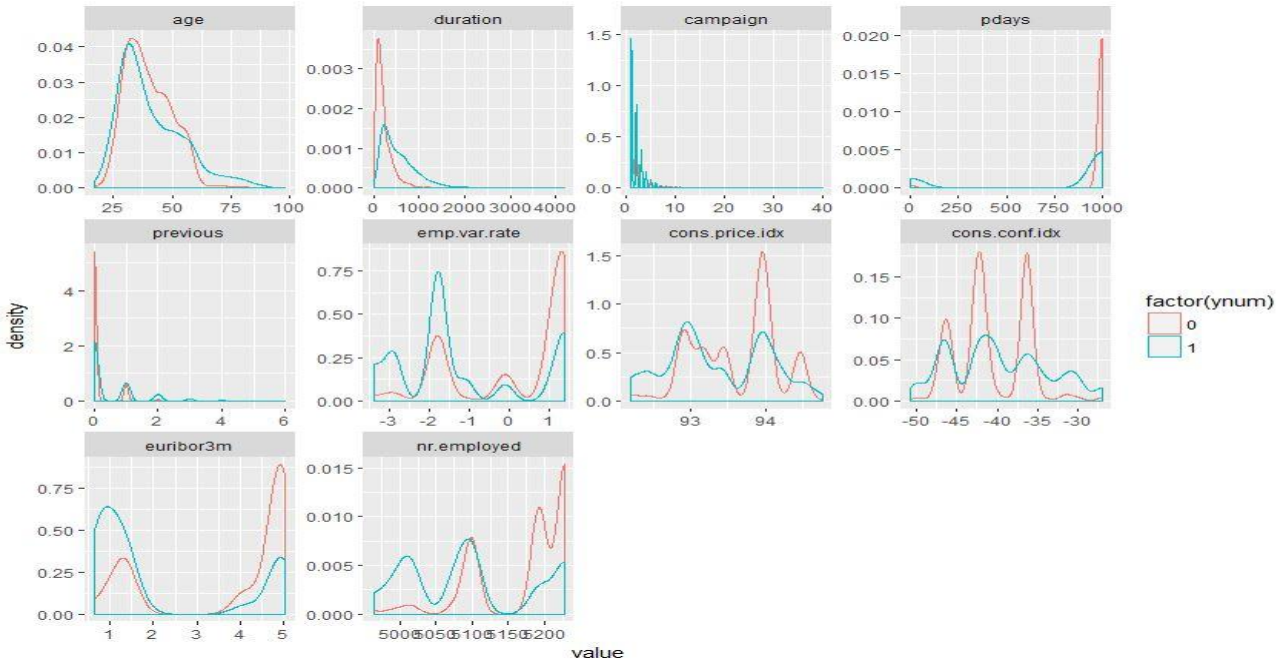
Under-sampling reduces the number of observations from majority class to make the data set balanced. In this case, number of observations with response 'no' is decreased to match the number of observations with response 'yes'.

Below are the plots of sampled data set to show that there isn't much deviation in distribution of sampling w.r.t whole population.

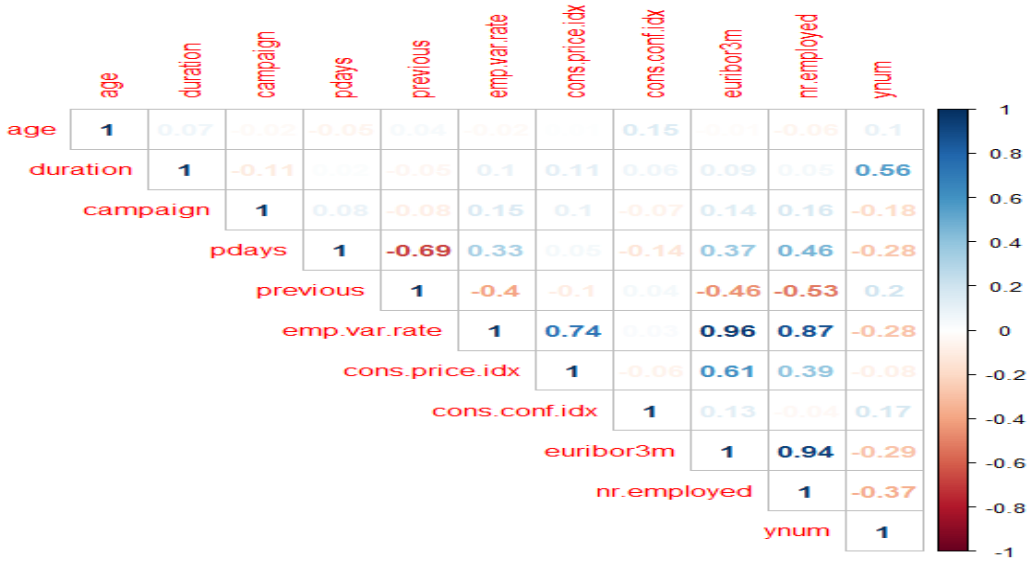
3.1.1 Numeric variable distribution in Sampled dataset:



3.1.2 Target VS Numeric predictors in Sampled dataset:



3.2 Correlation between numeric variables:



As per the above plot, there can be a relation between *pdays* & *previous*, *emp.var.rate* & *cons.price.idx*, *emp.var.rate*&*euribor3m*, *emp.var.rate* & *nr.employed*.

a) *Pdays* & *previous* are categorized as per the initial analysis. These numeric variables will not be considered in the model.
b) *emp.var.rate*, *cons.price.idx*, *euribor3m* are correlated as per the numbers but logically they are indicators based on time frame. As the time frames are different for the three variables, they cannot be considered as correlated.

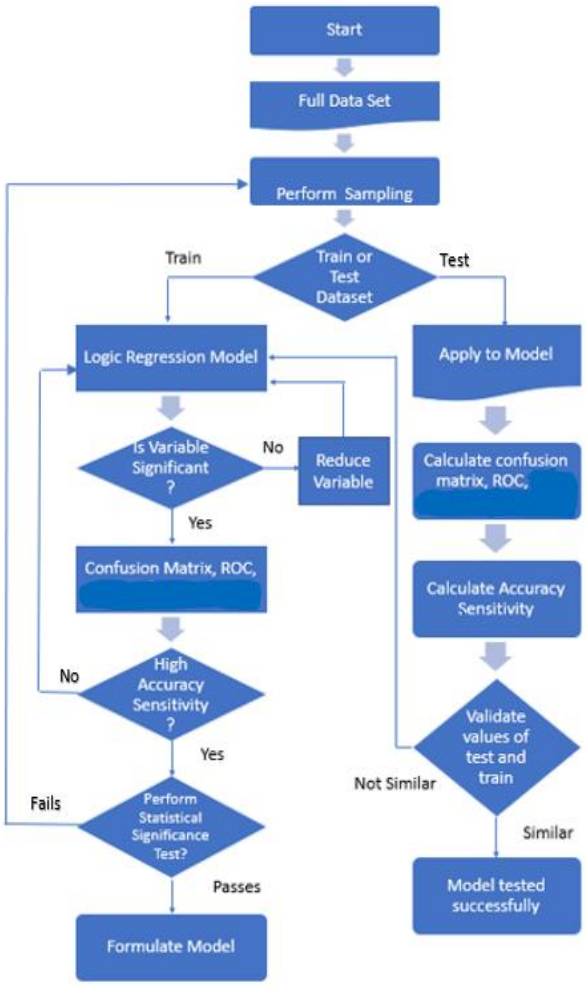
4 Assumptions:

The following predictors variable will not have much impact on the target response:

1) *Age* doesn't affect the distribution of the responses of 'y'
2) *Marital* status is a demographic variable
3) In *Default*, the no. of observations with the value 'yes' are only 3, therefore the variable is very skewed. Including this variable may not result in an accurate model
4) *Housing* and loan have similar distribution against the target response 'y'
5) *Duration* shouldn't be considered to predict a realistic model
6) *Campaign* doesn't affect the distribution of the responses of 'y'
7) The variable *nr.employed* is highly correlated with *emp.var.rate*. As per trial and error method, it can be inferred that the model performs better without *nr.employed*

5 Building the Model

The flowchart below represents the steps performed for building the model.



5.1 Train and Test:

The dataset is split into train and test in the ratio 60:40.

Justification: Using the trial and error method, the accuracy and sensitivity of the model is measured under three different scenarios starting with 80:20 then 70:30 and lastly 60:40 as the train data should be greater than the test data. The results are mentioned below:

Split ratio	Data	No. of records	AIC	Accuracy	Sensitivity	Specificity	ROC value	Significance (HL)
80:20	Train	7422	8004.9	74.09	61.6	86.58	0.7926	0.126
	Test	1856		73.44	61.53	85.34	0.7884	
70:30	Train	6494	6994	74.28	61.84	86.73	0.7947	0.39
	Test	2784		73.35	60.92	85.78	0.785	
60:40	Train	5566	5987.7	74.52	62.38	86.67	0.796	0.34
	Test	3712		73.6	62.02	85.18	0.7842	

Inference drawn from the table is that the Accuracy and Sensitivity values of 60:40 split is higher and there is more similarity between the Train and Test data result. Therefore the 60:40 split is chosen to build the model.

5.2 Regression

The flow chart above represents the entire process of Logistic Regression.

- Built the model without considering the variables tagged as insignificant in EDA
- Checked for significant variables through various iterations
- Generated the confusion matrix, ROC and Lift curve to check the efficiency of model
- Tested the model with Test Data
- Performed HL and LR test to check for the statistical significance of the model

5.2.1 Variable Significance

All the variables are considered during trial and error methods of building the model, based on the exploratory data analysis and inferential analysis, the insignificant variables obtained are as follows:

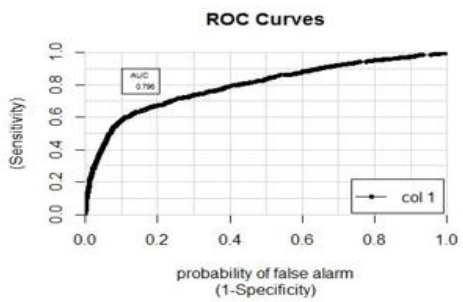
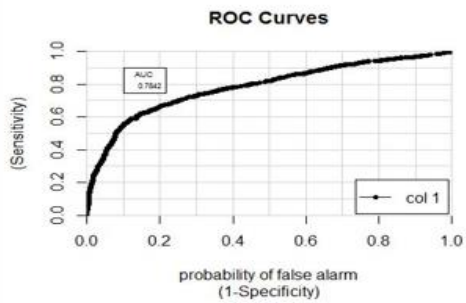
Insignificant Variable	Reason
Age	As per the distributions mentioned in EDA
Default	Highly skewed distribution of variable towards 'no'
Duration	To predict realistic model
nr.employed	High correlation with *emp.var.rate* variable and model is deteriorated when this variable is used
Loan	In population dataset, distribution of target variable is similar between loan categories 'yes' or 'no'
Housing	In population dataset, distribution of target variable is similar between loan categories 'yes' or 'no'
Marital	Demographic variables don't cause much impact in predictive analysis
Campaign	As per the distributions mentioned in EDA
cons.conf.idx	Insignificant in the iterations. Tried various combinations of the variables in building the model and they didn't show any insignificance.
day_of_week	
Education	
previous_cat	

List of significant variables obtained that impact the target variable is as follows:

Significant Variable	Reason
Job	Distribution of categories in job doesn't follow any pattern w.r.t. target
Contact	Person contacted through cellular phone have higher chances of subscribing to term deposit
Month	Person contacted in yearend have higher chances of subscribing to term deposit
Poutcome	People with 'success' as *poutcome* have higher chances of subscribing to term deposit
pdays_cat	People have been contacted in the previous campaign have higher chances of subscribing to term deposit
emp.var.rate	Person has higher chances of subscribing to term deposit when *emp.var.rate* is less than -1
cons.price.idx	Person has higher chances of not subscribing to term deposit when *cons.price.idx* is around 93 to 94
euribor3m	Person have higher chances of subscribing to term deposit when *euribor3m* is less than 2

5.2.2. Confusion Matrix and ROC Curve for Train and Test Data

The following confusion matrix and ROC was generated for the model

	Train Data	Test Data
Confusion Matrix	Confusion Matrix and Statistics Reference Prediction 0 1 0 2412 1047 1 371 1736 Accuracy : 0.7452 95% CI : (0.7336, 0.7566) No Information Rate : 0.5 P-Value [Acc > NIR] : < 2.2e-16 Kappa : 0.4905 Mcnemar's Test P-Value : < 2.2e-16 Sensitivity : 0.6238 Specificity : 0.8667 Pos Pred Value : 0.8239 Neg Pred Value : 0.6973 Prevalence : 0.5000 Detection Rate : 0.3119 Detection Prevalence : 0.3785 Balanced Accuracy : 0.7452 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction 0 1 0 1581 705 1 275 1151 Accuracy : 0.736 95% CI : (0.7215, 0.7501) No Information Rate : 0.5 P-Value [Acc > NIR] : < 2.2e-16 Kappa : 0.472 Mcnemar's Test P-Value : < 2.2e-16 Sensitivity : 0.6202 Specificity : 0.8518 Pos Pred Value : 0.8072 Neg Pred Value : 0.6916 Prevalence : 0.5000 Detection Rate : 0.3101 Detection Prevalence : 0.3842 Balanced Accuracy : 0.7360 'Positive' Class : 1
ROC Curve		

Comparing the confusion matrix and ROC curve results for Train and Test data, it can be inferred that the accuracy and sensitivity of the model is nearly the same for both the cases.

6 Goodness of Fit:

6.1 Loglikelihood test

In this test the null hypothesis is that the reduced model is true, and the null hypothesis can be rejected if the p-value is low.

The result of the Loglikelihood test is as below:

```
Likelihood ratio test

Model 1: train$ynum ~ (age + job + marital + education + default + housing +
    loan + contact + month + day_of_week + duration + campaign +
    pdays + previous + poutcome + emp.var.rate + cons.price.idx +
    cons.conf.idx + euribor3m + nr.employed + y + previous_cat +
    pdays_cat) - duration - pdays_cat - previous_cat - y
Model 2: train$ynum ~ (age + job + marital + education + default + housing +
    loan + contact + month + day_of_week + duration + campaign +
    pdays + previous + poutcome + emp.var.rate + cons.price.idx +
    cons.conf.idx + euribor3m + nr.employed + y + previous_cat +
    pdays_cat) - previous_cat - education - day_of_week - cons.conf.idx -
    campaign - marital - housing - loan - nr.employed - duration -
    default - age - pdays - previous - y
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  52 -2949.8
2  28 -2965.9 -24 32.061     0.1255
```

The output gives an acceptable p-value ($p > 0.05$), hence the reduced model has a better fit than the full model.

6.2 HL Test

The HL test gives the goodness of fit of the model, the results is seen as follows:

```
$C

        Hosmer-Lemeshow C statistic

data:  fitted(model_sample) and train$ynum
X-squared = 9.0268, df = 8, p-value = 0.34


$H

        Hosmer-Lemeshow H statistic

data:  fitted(model_sample) and train$ynum
X-squared = 12.833, df = 8, p-value = 0.1177
```

The output gives an acceptable p-value ($p > 0.05$), hence the model has a good fit.

7 Conclusion:

The AUC for the curve is around 79%, so the model is accurate in its prediction.

The final model consists of the significant variables JOB, CONTACT, MONTH, PDAYS, POUTCOME, EMPOLYMENT VARIATION RATE, CONSUMER PRICE INDEX AND EURIBOR3M.

These variables are considered with their respective coefficients to build the final logit regression equation as mentioned below.

```
train$ynum ~ (age + job + marital + education + default + housing +
    loan + contact + month + day_of_week + duration + campaign +
    pdays + previous + poutcome + emp.var.rate + cons.price.idx +
    cons.conf.idx + euribor3m + nr.employed + y + previous_cat +
    pdays_cat) - previous_cat - education - day_of_week - cons.conf.idx -
    campaign - marital - housing - loan - nr.employed - duration -
    default - age - pdays - previous - y
```

Below are the significant variables and their coefficients:

```
Coefficients:
                          Estimate Std. Error z value     Pr(>|z|)
(Intercept)             -155.58609   17.76882  -8.756      < 2e-16 ***
jobblue-collar            -0.32718    0.09474  -3.454     0.000553 ***
jobentrepreneur           -0.15508    0.17722  -0.875     0.381555
jobhousemaid              -0.31716    0.22241  -1.426     0.153854
jobmanagement             -0.06497    0.13288  -0.489     0.624890
jobretired                 0.07649    0.15156   0.505     0.613777
jobself-employed          -0.20303    0.17392  -1.167     0.243067
jobservices               -0.33612    0.12448  -2.700     0.006930 **
jobstudent                 0.56177    0.22517   2.495     0.012600 *
jobtechnician             -0.12597    0.10032  -1.256     0.209221
jobunemployed             -0.14067    0.21370  -0.658     0.510362
jobunknown                -0.04578    0.37358  -0.123     0.902461
contacttelephone          -0.63478    0.11990  -5.294 0.0000001196 ***
monthaug                   0.76664    0.16388   4.678 0.0000028968 ***
monthdec                   0.55262    0.39541   1.398     0.162242
monthjul                   0.20158    0.15164   1.329     0.183729
monthjun                  -0.42133    0.16642  -2.532     0.011351 *
monthmar                   1.53033    0.27303   5.605 0.0000000208 ***
monthmay                  -0.40886    0.12152  -3.365     0.000767 ***
monthnov                  -0.45479    0.17627  -2.580     0.009877 **
monthoct                  -0.01812    0.24570  -0.074     0.941223
monthsep                   0.76163    0.29921   2.545     0.010913 *
poutcomenonexistent        0.46877    0.10902   4.300 0.0000170895 ***
poutcomesuccess           -0.22729    0.75215  -0.302     0.762509
emp.var.rate              -1.52553    0.17920  -8.513      < 2e-16 ***
cons.price.idx             1.65745    0.18601   8.911      < 2e-16 ***
euribor3m                  0.58452    0.13361   4.375 0.0000121524 ***
pdays_catnever contacted  -1.97625    0.73949  -2.672     0.007530 **
```