**Aim :-**

The Aim of this project is to generate model provided in the dataset using various classification and regression techniques in R. Analyze the results presents in the model and optimize model which is most suitable for this dataset.

**Dataset Description : -** Dataset contains information about information used for bank marketing. Bank Marketing dataset contains 21 variables and 4119 Rows. This Dataset is provided by a Portuguese banking institution based on bank marketing campaigns and based on real world problem.

**Training Dataset : 3530**

**Testing Dataset : 589**

**Link to Dataset :**

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

**Data Cleaning** :- Dataset contains 21 attributes in the training dataset and 20 in the test dataset & one attribute in the target test dataset**.** No N/A values found when summary of all dataset performed . Attribute List as follows :-

Input variables:  
# bank client data:  
1 - age (numeric)  
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','unknown')  
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. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# 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')

Following code used for initial variable analysis :-

**# Setting up default working directory path of the project**

>setwd("D:\\Plan\\DataScience\\project\\Bank project\\bank-additional")

**#Getting working directory**

**> getwd()**

**# Read CSV file from the working directory**

**>** DataBank<-DataBank<-read.csv("bank-additional.csv", sep=";", header=TRUE,quote = "\" ")

**# file: file name**

**#** header**: 1st line as header or not, logical  
 # sep: field separator**

**# quote: quoting characters**

**# Determine structure of CSV file**

str(DataBank)

'data.frame': 4119 obs. of 21 variables:

$ age : int 30 39 25 38 47 32 32 41 31 35 ...

$ job : Factor w/ 12 levels "admin.","blue-collar",..: 2 8 8 8 1 8 1 3 8 2 ...

$ marital : Factor w/ 4 levels "divorced","married",..: 2 3 2 2 2 3 3 2 1 2 ...

$ education : Factor w/ 8 levels "basic.4y","basic.6y",..: 3 4 4 3 7 7 7 7 6 3 ...

$ default : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 1 1 1 2 1 2 ...

$ housing : Factor w/ 3 levels "no","unknown",..: 3 1 3 2 3 1 3 3 1 1 ...

$ loan : Factor w/ 3 levels "no","unknown",..: 1 1 1 2 1 1 1 1 1 1 ...

$ contact : Factor w/ 2 levels "cellular","telephone": 1 2 2 2 1 1 1 1 1 2 ...

$ month : Factor w/ 10 levels "apr","aug","dec",..: 7 7 5 5 8 10 10 8 8 7 ...

$ day\_of\_week : Factor w/ 5 levels "fri","mon","thu",..: 1 1 5 1 2 3 2 2 4 3 ...

$ duration : int 487 346 227 17 58 128 290 44 68 170 ...

$ campaign : int 2 4 1 3 1 3 4 2 1 1 ...

$ pdays : int 999 999 999 999 999 999 999 999 999 999 ...

$ previous : int 0 0 0 0 0 2 0 0 1 0 ...

$ poutcome : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 1 2 2 1 2 ...

$ emp.var.rate : num -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...

$ cons.price.idx: num 92.9 94 94.5 94.5 93.2 ...

$ cons.conf.idx : num -46.2 -36.4 -41.8 -41.8 -42 -37.5 -37.5 -42 -42 -36.4 ...

$ euribor3m : num 1.31 4.86 4.96 4.96 4.19 ...

$ nr.employed : num 5099 5191 5228 5228 5196 ...

$ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

**# Determine Number of rows in the Dataset**

**> nrow(DataBank)**

**[1] 4119**

**# Number of the columns in the dataset**

**> ncol(DataBank)**

**[1] 21**

**# Determine if Dataset has any missing or not applicable value**

**> sum(is.na(DataBank))**

**[1] 0**

**Models:**

**1. Logistic Regression**

Logistic Regression is a type of classification Model.in this model we attempt to predict outcome of dependent variable using one or more independent variables. independent variables can be categorical or numerical . logistic regression is based on logistic function it always takes values between 0 or 1.

The Algorithm starts with splitting dataset into training and testing dataset

# import library caTools for splitting dataset

> library(caTools)

# splitting dataset into two part

# a 90% of training dataset

# b 10% testing dataset

> split<-sample.split(DataBank,SplitRatio = 0.9)

> training<-subset(DataBank,split==1)

> testing<-subset(DataBank,split==0)

> nrow(testing)

[1] 589

> nrow(training)

[1] 3530

# prepare model with training dataset

# is display type of the model in R family i.e. "Binomial Type"

# y is a response variable whose probability we need to Determine

> model1<-glm(y~.,training,family = "binomial")

> summary(model1)

Call:

glm(formula = y ~ ., family = "binomial", data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.8304 -0.2891 -0.1788 -0.1146 2.8093

Coefficients: (1 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.225e+02 1.348e+02 -0.909 0.3634

age 9.504e-03 8.846e-03 1.074 0.2826

jobblue-collar -2.042e-01 2.935e-01 -0.696 0.4866

jobentrepreneur -5.981e-01 5.423e-01 -1.103 0.2701

jobhousemaid 9.073e-02 5.053e-01 0.180 0.8575

jobmanagement -3.903e-01 3.118e-01 -1.252 0.2106

jobretired -1.191e-01 3.788e-01 -0.314 0.7532

jobself-employed -8.566e-01 4.594e-01 -1.865 0.0622 .

jobservices 2.047e-01 2.985e-01 0.686 0.4928

jobstudent 1.274e-01 4.244e-01 0.300 0.7640

jobtechnician 1.678e-01 2.413e-01 0.695 0.4870

jobunemployed 3.456e-01 4.205e-01 0.822 0.4112

jobunknown -6.361e-01 8.560e-01 -0.743 0.4574

maritalmarried 3.476e-01 2.741e-01 1.268 0.2048

maritalsingle 3.765e-01 3.092e-01 1.218 0.2234

maritalunknown 3.339e-01 1.152e+00 0.290 0.7719

educationbasic.6y 4.171e-01 4.313e-01 0.967 0.3335

educationbasic.9y 2.161e-01 3.440e-01 0.628 0.5298

educationhigh.school 1.920e-01 3.371e-01 0.570 0.5689

educationilliterate -1.140e+01 5.354e+02 -0.021 0.9830

educationprofessional.course 1.358e-01 3.668e-01 0.370 0.7112

educationuniversity.degree 2.280e-01 3.403e-01 0.670 0.5028

educationunknown 2.121e-01 4.241e-01 0.500 0.6170

defaultunknown 6.468e-02 2.300e-01 0.281 0.7785

defaultyes -9.022e+00 5.354e+02 -0.017 0.9866

housingunknown -6.946e-01 6.123e-01 -1.134 0.2566

housingyes -6.293e-02 1.476e-01 -0.426 0.6699

loanunknown NA NA NA NA

loanyes -3.261e-03 1.992e-01 -0.016 0.9869

contacttelephone -1.238e+00 3.170e-01 -3.905 9.40e-05 \*\*\*

monthaug 2.987e-01 4.517e-01 0.661 0.5083

monthdec 1.046e+00 7.345e-01 1.425 0.1542

monthjul -1.902e-01 3.860e-01 -0.493 0.6222

monthjun 5.973e-01 4.688e-01 1.274 0.2026

monthmar 2.707e+00 5.841e-01 4.636 3.56e-06 \*\*\*

monthmay -4.043e-01 3.175e-01 -1.273 0.2029

monthnov -5.822e-01 4.536e-01 -1.284 0.1993

monthoct 2.878e-01 5.831e-01 0.493 0.6217

monthsep -1.638e-01 6.512e-01 -0.252 0.8014

day\_of\_weekmon 2.367e-01 2.283e-01 1.037 0.3000

day\_of\_weekthu 5.837e-02 2.344e-01 0.249 0.8033

day\_of\_weektue 4.351e-02 2.376e-01 0.183 0.8547

day\_of\_weekwed 2.649e-01 2.416e-01 1.097 0.2728

duration 5.118e-03 2.755e-04 18.578 < 2e-16 \*\*\*

campaign -7.156e-02 4.708e-02 -1.520 0.1285

pdays -3.085e-05 7.864e-04 -0.039 0.9687

previous 2.845e-01 1.961e-01 1.451 0.1469

poutcomenonexistent 6.003e-01 3.234e-01 1.856 0.0634 .

poutcomesuccess 1.534e+00 7.758e-01 1.977 0.0481 \*

emp.var.rate -7.544e-01 5.180e-01 -1.457 0.1452

cons.price.idx 1.258e+00 8.930e-01 1.408 0.1590

cons.conf.idx 6.552e-02 2.903e-02 2.257 0.0240 \*

euribor3m -1.004e-01 4.562e-01 -0.220 0.8257

nr.employed 4.856e-04 1.084e-02 0.045 0.9643

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2436.8 on 3529 degrees of freedom

Residual deviance: 1379.2 on 3477 degrees of freedom

AIC: 1485.2

Number of Fisher Scoring iterations: 12

|  |
| --- |
| **Following can be predicted by the model**  Null deviance : How well response variable is predicted by a model that includes  only the intercept (grand mean). Here it is 2436.8 with 3529 degrees of freedom and  on other hand Residual deviance (How well response variable is predicted by a model  that includes the intercept (grand mean and independent variables). is decreased by  1379.2 with 3477 degrees of freedom.    Fisher’s scoring algorithm is a derivative of Newton’s method for solving maximum likelihood problems numerically. For model1 we see that Fisher’s  Scoring Algorithm needed 12 iterations to perform the fit.  The Akaike Information Criterion (AIC) : The Akaike Information Criterion (AIC)  provides a method for assessing the quality of your model through comparison of related  models.Also help to optimize model.  **# For example here in below summary of model**  > **anova(model1, test="Chisq")**    Df Deviance Resid. Df Resid. Dev Pr(>Chi)  NULL 3529 2436.8  age 1 11.02 3528 2425.8 0.0009026 \*\*\*  job 11 52.56 3517 2373.2 2.154e-07 \*\*\*  marital 3 14.42 3514 2358.8 0.0023892 \*\*  education 7 7.83 3507 2350.9 0.3480959  default 2 21.80 3505 2329.1 1.844e-05 \*\*\*  housing 2 0.46 3503 2328.7 0.7940638  loan 1 0.01 3502 2328.7 0.9134790  contact 1 64.43 3501 2264.2 1.001e-15 \*\*\*  month 9 146.93 3492 2117.3 < 2.2e-16 \*\*\*  day\_of\_week 4 0.52 3488 2116.8 0.9713337  duration 1 486.03 3487 1630.8 < 2.2e-16 \*\*\*  campaign 1 9.32 3486 1621.4 0.0022690 \*\*  pdays 1 116.46 3485 1505.0 < 2.2e-16 \*\*\*  previous 1 6.65 3484 1498.3 0.0099230 \*\*  poutcome 2 10.05 3482 1488.3 0.0065845 \*\*  emp.var.rate 1 65.69 3481 1422.6 5.273e-16 \*\*\*  cons.price.idx 1 33.33 3480 1389.2 7.793e-09 \*\*\*  cons.conf.idx 1 10.00 3479 1379.2 0.0015690 \*\*  euribor3m 1 0.07 3478 1379.2 0.7885636  nr.employed 1 0.00 3477 1379.2 0.9642835  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Here "\*" present how much in dependent variable significant in predicting  Dependent variable in the model.  \*\*\* = 99.9% significant  \*\*= 99% significant  \* = 95% significant  . = 90% significant  Probability of education is highter > 0.5 (i.e value is 0.3480959  Fail to reject null hypothesis ) so education can be eliminated from the model  So as described below  Model2<-glm(y~.-education,training,family = "binomial")  Summary(Model2)    Null deviance: 2436.8 on 3529 degrees of freedom  Residual deviance: 1380.5 on 3484 degrees of freedom  AIC: 1472.5  Residual deviance increased by (1380.5-1379.2) and AIC (1485.2-1472.5)  Ideally Residual deviance should not increase and AIC should decrease at he same time to  Optimize the model.  model3<-glm(y~.-nr.employed ,training,family = "binomial")  summary(model3)  Null deviance: 2436.8 on 3529 degrees of freedom  Residual deviance: 1379.2 on 3478 degrees of freedom  AIC: 1483.2  This values(Residual deviance: 1379.2/ AIC: 1483.2 ) are much better the previous  Interpretations for nr.employed can be removed from the model to optimize this model.  **# predicting from the model with training data**  > train\_result<-predict(model1,training,type = "response")  > table(ActualValue=training$y,PredictedValue=train\_result>0.5)  **# below confusion matrix**  # This will help to determine classifier performance  #Diagonal entries are correctly classifiers cases  ALL Rows represents actual values  All columns present predicted values  PredictedValue  ActualValue FALSE TRUE  no 3065 79  yes 217 169  Accuracy of this matrix = (correctly predicted values/total values)  3065+169/3065+169+217+79 = 3234/3530 =91.6 %  Precision = 169/217+169 (sensitivity) = 43.78%  Precision refers to the reproducibility of this result that is you  get the same result every time  Recall = 169/169+79 = 68.14 %  (% of correct items are selected)  Specapacity = 3065/3065+79 = 97.48 %  (% of incorrect items selected)  > testing\_result<-predict(model1,testing,type = "response")  > table(ActualValue=testing$y,PredictedValue=testing\_result>0.5)  PredictedValue  ActualValue FALSE TRUE  no 508 16  yes 35 30  > 3234/3530  [1] 0.9161473 = 91.34 % is the accuracy of this model with testing dataset  > |
|  |

testing\_result<-predict(model1,testing,type = "raw")

|  |  |
| --- | --- |
| |  | | --- | |  | |

**Testing\_result :** This will gives probability of subscribing to a term

deposit.