IIMA CONVERT PREDICTION USING LOGISTIC REGRESSION

INTRODUCTION:

MBA is one of the most popular courses in India.There are plethora of colleges which run 2 year or 1 year MBA curriculum.However, students are interested to join only the best colleges.Moreover, intake at top colleges especially at IIM'S are very few and they also ensure that they never compromise on the quality of the students. Hence it is essential for students to understand the selection process which IIMs follow to recruit the best minds.Here, we consider the case of IIM Ahemedabad's approach towards selecting right candidates.

PROBLEM STATEMENT:

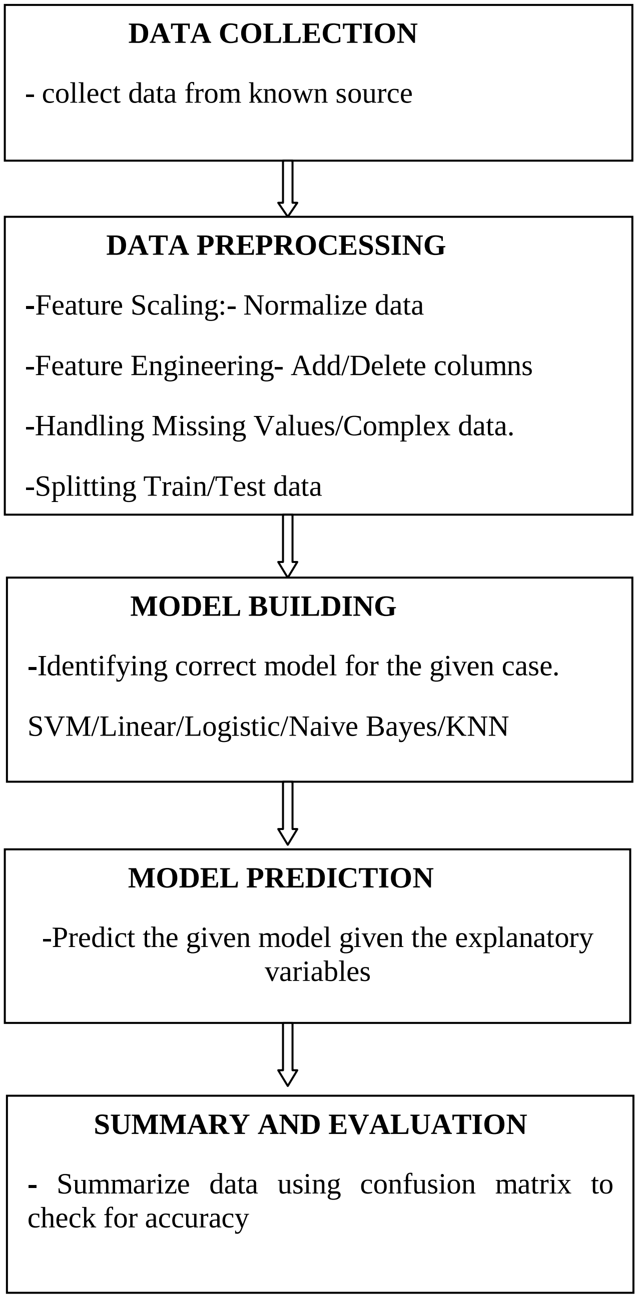
Predict if a student can get into IIM Ahmedabad.

DATA SET:

Variable Name	Description	Type	Example
ID No.	Indentification	String	Chhavi
Caste	BC/General/OBC	Character	G
10th	10th score of candidate	Numeric	92
12th	12th score of candidate	Numeric	91
UG Score	Undergrad Score	Numeric	98
Gender	Male/Female	Character	M
Interview Performance	Quality of interview	String	Good
Prestige Tag	Quality of the college IIT/NIT/BITS/SRCC	Numeric	2
Extra Curriculuars/POR	Quality of Extracurricular/POR and Certifications	String	Excellent
Work Experience	Years of Work Exp	Numeric	3
Call	Whether person gets call	Character	Y
Cultural Diversity	Whether person is Culturally different from the normal peer	Character	Y
UG Degree	Engineer/Non Engineer	String	Engineering
Convert	Person gets the seat or not	Character	Y

METHODLOGY:

Basically any machine learning technique follows 5 basic steps



DATA COLLECTION

- collect data from known source

DATA PREPROCESSING

-Feature Scaling:- Normalize data

-Feature Engineering- Add/Delete columns

-Handling Missing Values/Complex data.

-Splitting Train/Test data

MODEL BUILDING

-Identifying correct model for the given case.

SVM/Linear/Logistic/Naive Bayes/KNN

MODEL PREDICTION

-Predict the given model given the explanatory variables

SUMMARY AND EVALUATION

- Summarize data using confusion matrix to check for accuracy

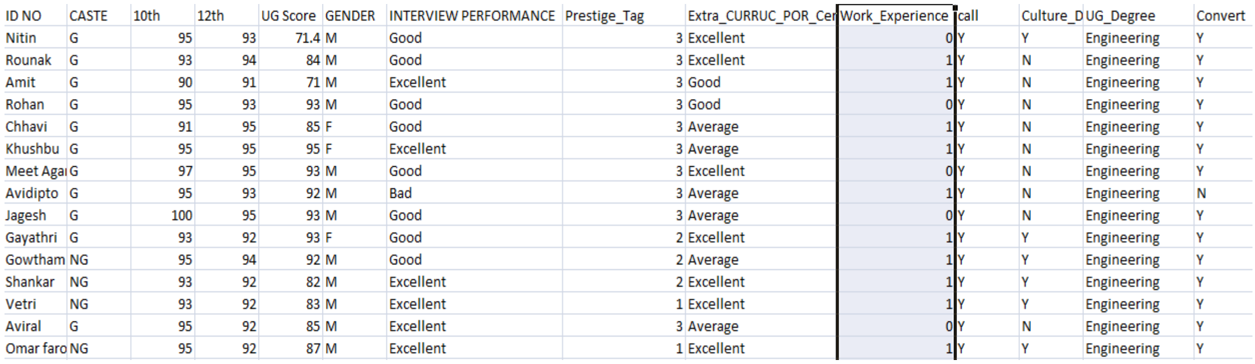
ASSUMPTIONS:

1. Only college being target is IIM Ahmedabad.

2. Logistic Regression is the best for the given model.

DATA COLLECTION:

Here first 16 samples are shown in the figure.There were totally 124 samples for train and test.



ID NO	CASTE	10th	12th	UG Score	GENDER	INTERVIEW PERFORMANCE	Prestige_Tag	Extra_CURRUC_POR_Cer	Work_Experience	call	Culture_D	UG_Degree	Convert
Nitin	G	95	93	71.4	M	Good	3	Excellent	0	Y	Y	Engineering	Y
Rounak	G	93	94	84	M	Good	3	Excellent	1	Y	N	Engineering	Y
Amit	G	90	91	71	M	Excellent	3	Good	1	Y	N	Engineering	Y
Rohan	G	95	93	93	M	Good	3	Good	0	Y	N	Engineering	Y
Chhavi	G	91	95	85	F	Good	3	Average	1	Y	N	Engineering	Y
Khushbu	G	95	95	95	F	Excellent	3	Average	1	Y	N	Engineering	Y
Meet Agai	G	97	95	93	M	Good	3	Excellent	0	Y	N	Engineering	Y
Avidipto	G	95	93	92	M	Bad	3	Average	1	Y	N	Engineering	N
Jagesh	G	100	95	93	M	Good	3	Average	0	Y	N	Engineering	Y
Gayathri	G	93	92	93	F	Good	2	Excellent	1	Y	Y	Engineering	Y
Gowtham	NG	95	94	92	M	Good	2	Average	1	Y	Y	Engineering	Y
Shankar	NG	93	92	82	M	Excellent	2	Excellent	1	Y	Y	Engineering	Y
Vetri	NG	93	92	83	M	Excellent	1	Excellent	1	Y	Y	Engineering	Y
Aviral	G	95	92	85	M	Excellent	3	Average	0	Y	N	Engineering	Y
Omar faro	NG	95	92	87	M	Excellent	1	Excellent	1	Y	Y	Engineering	Y

DATA PREPROCESSING:

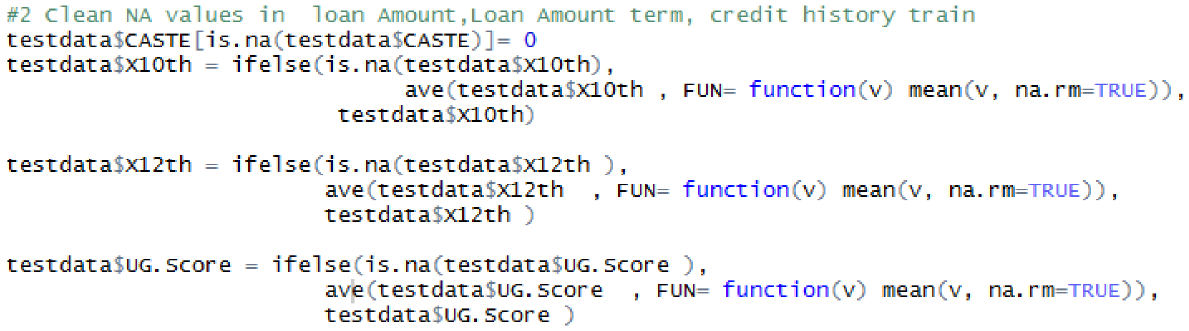
Here, we have data with Missing and NA values.We preprocess the data to make changes.This is done to ensure uniformity in data.Further, we split data into train and test(70:30)

WAYS TO HANDLE MISSING VALUES AND COMPLEX DATA:

1.Replace NAs with either Average or Mode values depending on the circumstance.

2.Another way to handle to handle missing values is to delete them.

SAMPLE CODE:



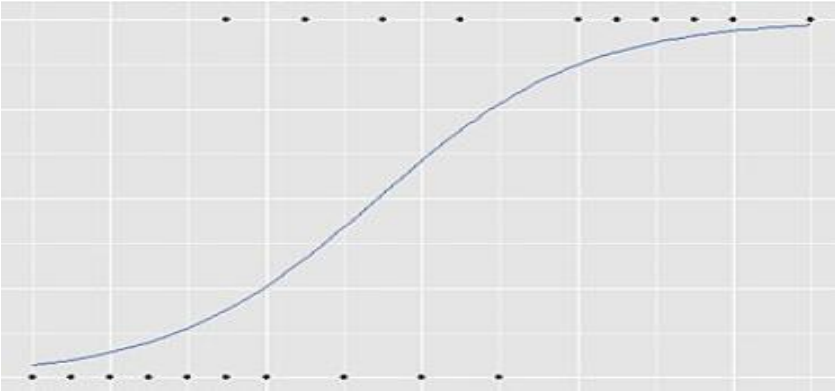
```
#2 Clean NA values in  loan Amount,Loan Amount term, credit history train
testdata$CASTE[is.na(testdata$CASTE)]= 0
testdata$X10th = ifelse(is.na(testdata$X10th),
                              ave(testdata$X10th , FUN= function(v) mean(v, na.rm=TRUE)),
                         testdata$X10th)

testdata$X12th = ifelse(is.na(testdata$X12th ),
                        ave(testdata$X12th  , FUN= function(v) mean(v, na.rm=TRUE)),
                        testdata$X12th )

testdata$UG.Score = ifelse(is.na(testdata$UG.Score ),
                        ave(testdata$UG.Score  , FUN= function(v) mean(v, na.rm=TRUE)),
                        testdata$UG.Score )
```

MODEL BUILDING:

Logistic Regression:

Logistic Regression is a model used to predict a certain class of objects which linear regression fails to achieve.For example in rain prediction, there are only 2 classes- Yes and No.Hence, logistic regression is used in scenarios where classifying the data becomes the sole objective.

Basically logistic regression plots a Sigmoid/logit function curve.It determines the probability in which a certain class of objects lie.



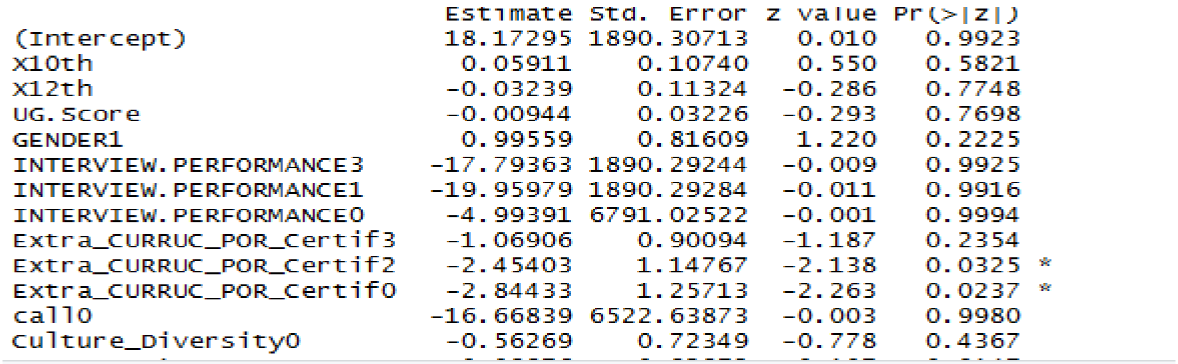
$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}.$$

where ℓ is the log-odds, b is the base of the logarithm, and βs are parameters of the model determined from the train data.

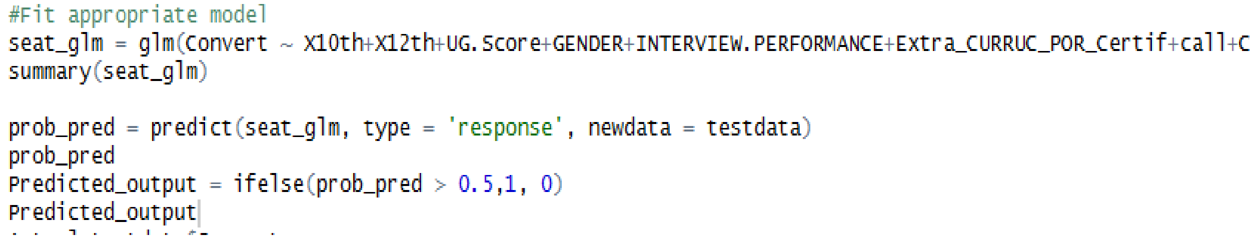
These equations are used to predict the probability of loan prediction.

OBSERVATIONS:



```
                              Estimate Std. Error z value Pr(>|z|)  
(Intercept)                   18.17295 1890.30713   0.010   0.9923  
X10th                          0.05911    0.10740   0.550   0.5821  
X12th                         -0.03239    0.11324  -0.286   0.7748  
UG.Score                      -0.00944    0.03226  -0.293   0.7698  
GENDER1                        0.99559    0.81609   1.220   0.2225  
INTERVIEW.PERFORMANCE3       -17.79363 1890.29244  -0.009   0.9925  
INTERVIEW.PERFORMANCE1       -19.95979 1890.29284  -0.011   0.9916  
INTERVIEW.PERFORMANCE0        -4.99391 6791.02522  -0.001   0.9994  
Extra_CURRUC_POR_Certif3      -1.06906    0.90094  -1.187   0.2354  
Extra_CURRUC_POR_Certif2      -2.45403    1.14767  -2.138   0.0325 *
Extra_CURRUC_POR_Certif0      -2.84433    1.25713  -2.263   0.0237 *
call0                        -16.66839 6522.63873  -0.003   0.9980  
Culture_Diversity0            -0.56269    0.72349  -0.778   0.4367  
```

MODEL CODE:

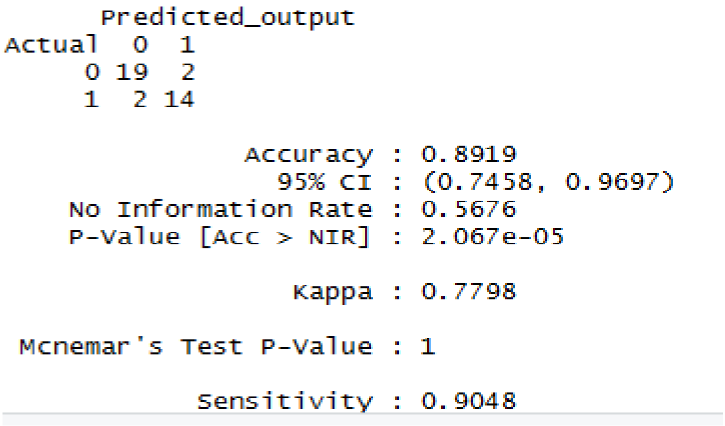


```
#Fit appropriate model
seat_glm = glm(Convert ~ X10th+X12th+UG.Score+GENDER+INTERVIEW.PERFORMANCE+Extra_CURRUC_POR_Certif+call+C
summary(seat_glm)

prob_pred = predict(seat_glm, type = 'response', newdata = testdata)
prob_pred
Predicted_output = ifelse(prob_pred > 0.5,1, 0)
Predicted_output
```

RESULTS:

We achieved an accuracy of 89.19%We can further optimize our model by applying few techniques like Parameter Tuning, Controlling Train/Testsplit and by changing the probability threshold.These techniques can further boost our accuracy.

CONFUSION MATRIX BEFORE OPTIMIZATION:

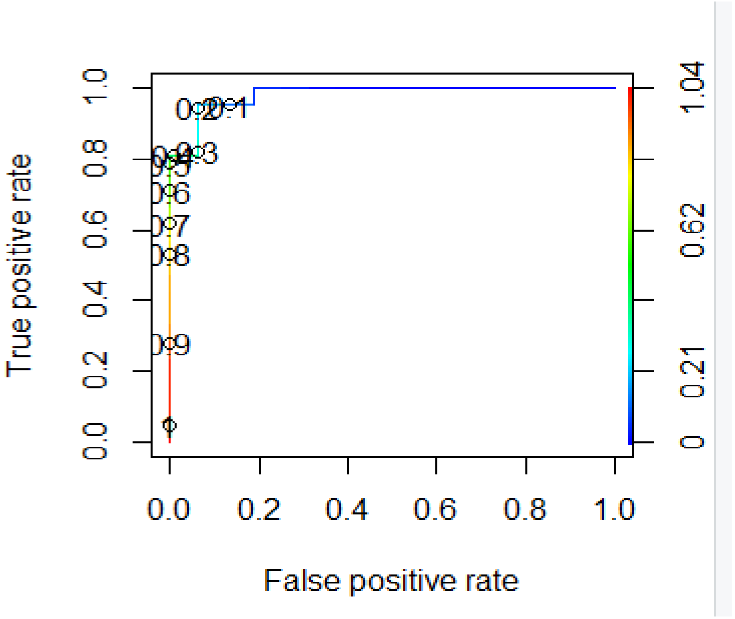


```
       Predicted_output
Actual  0  1
     0 19  2
     1  2 14

               Accuracy : 0.8919
                 95% CI : (0.7458, 0.9697)
    No Information Rate : 0.5676
    P-Value [Acc > NIR] : 2.067e-05

                  Kappa : 0.7798

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9048
```

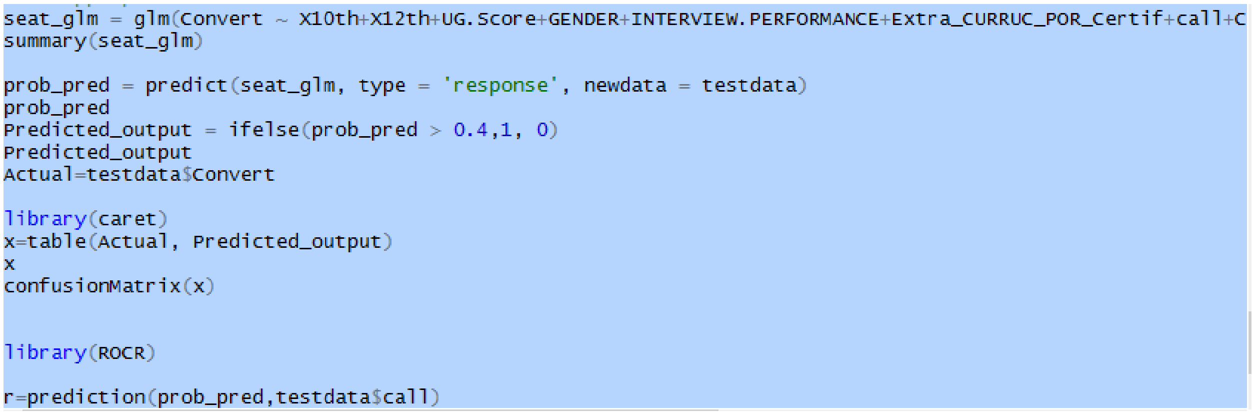
We have achieved accuracy of 89.19.Our model can further be optimized using ROC curve

ROC CURVE:

Optimized model code can be derived from the ROC curve

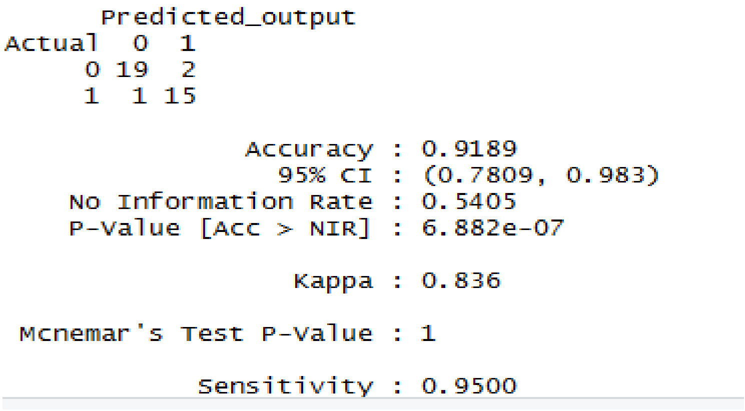


OPTIMIZED MODEL CODE BY TWEAKING THRESHOLD:



```
seat_glm = glm(Convert ~ X10th+X12th+UG.Score+GENDER+INTERVIEW.PERFORMANCE+Extra_CURRUC_POR_Certif+call+C
summary(seat_glm)

prob_pred = predict(seat_glm, type = 'response', newdata = testdata)
prob_pred
Predicted_output = ifelse(prob_pred > 0.4,1, 0)
Predicted_output
Actual=testdata$Convert

library(caret)
x=table(Actual, Predicted_output)
x
confusionMatrix(x)


library(ROCR)

r=prediction(prob_pred,testdata$call)
```

CONFUSION MATRIX AFTER OPTIMIZATION:



```
       Predicted_output
Actual   0  1
     0 19  2
     1  1 15

               Accuracy : 0.9189
                 95% CI : (0.7809, 0.983)
    No Information Rate : 0.5405
    P-Value [Acc > NIR] : 6.882e-07

                  Kappa : 0.836

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9500
```

Our accuracy has further increased from 89.19 to 91.89% by tweaking the threshold.Hence, we are more sure that the right set of candidates convert IIMA call.

SUMMARY:

1. We have basically incorporated Logistic Regression model to predict student's chance of converting IIMA call.We have taken the most important explanatory variables to explain the output variable more efficiently.We infer that extracurriculars,good interview performance are the key parameters to convert IIM Ahmedabad.
2. We have optimized our model and increased its accuracy from **89.19 to 91.89%.**This was achieved because of Optimization. Further, sensitivity of the model is kept minimum. This ensures that th model correctly predicts "Yes" to the students who truly deserve the seat at IIMA.
3. Optimization of the logistic regression can be carried out in two ways.One, by tuning the parameters and other by changing the threshold value.

REFERENCES:

- https://www.youtube.com/results?search_query=logistic+regression+fundamentals
- https://www.researchgate.net/publication/303326261_Machine_Learning_Project
- https://www.andrewng.org/