Problem Statement:

You are hired by a sports analysis agency to understand the selection process of high school football players into college with a full or partial scholarship. You are provided details of 6215 high school graduates who have been inducted into 4-year degree colleges with either full or partial scholarships. You have to help the agency in predicting whether a high school graduate will win a full scholarship on the basis of the information given in the data set. Also, find out the important factors which are instrumental in winning a full scholarship in colleges.

Data Dictionary for Market Segmentation:

1. Scholarship: Won a college scholarship: Full / Partial
2. Academic Score: High school academic performance of a candidate
3. Score on Plays Made: A composite score based on the achievements on the field
4. Missed Play Score: A composite score based on the failures on the field
5. Injury Propensity: This has 4 ordinal levels: High, Moderate, Normal and Low. It has been calculated based on what proportion of time a candidate had an injury problem
6. School Type: 4 types of schools based on their location
7. School Score: A composite score based on the overall achievement of the candidates' school, based on the school's academic, sports and community service performance
8. Overall Score: A composite score based on a candidate's family financial state, school performance, psychosocial attitude etc.
9. Region: Region of the country where the school is located

The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, especially identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have a symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. For this is a classification problem, the dependence of the response on the predictors needs to be investigated.
Two different classification techniques are to be applied. However, the EDA part remains the same for both of them.
For easier interpretation of the models, later on, it may be better to code Full = 1 and Partial = 0. You may assume the opposite, but then you have to be very careful about the interpretation of the logistic model coefficients later.

Overview on the data

	Academic_Score	Score_on_Plays_Made	Missed_Play_Score	Injury_Propensity	School_Type	School_Score	Overall_Score	Region	Scholarship
0	7.0	0.27	0.36	High	A	0.45	8.8	Eastern	Partial
1	6.3	0.30	0.34	Low	C	0.49	9.5	Eastern	Partial
2	8.1	0.28	0.40	Moderate	C	0.44	10.1	Eastern	Partial
3	7.2	0.23	0.32	Moderate	C	0.40	9.9	Eastern	Partial
4	7.2	0.23	0.32	Moderate	C	0.40	9.9	Eastern	Partial

Inference from the data overview

- The given dataset has 6215rows and 9 columns.
- The variables data type is 5 (float,64) and 4 (object) data types
- There are no null values
- There are 947 duplicated values
- Visible outliers are present

Describing all D-types Variables

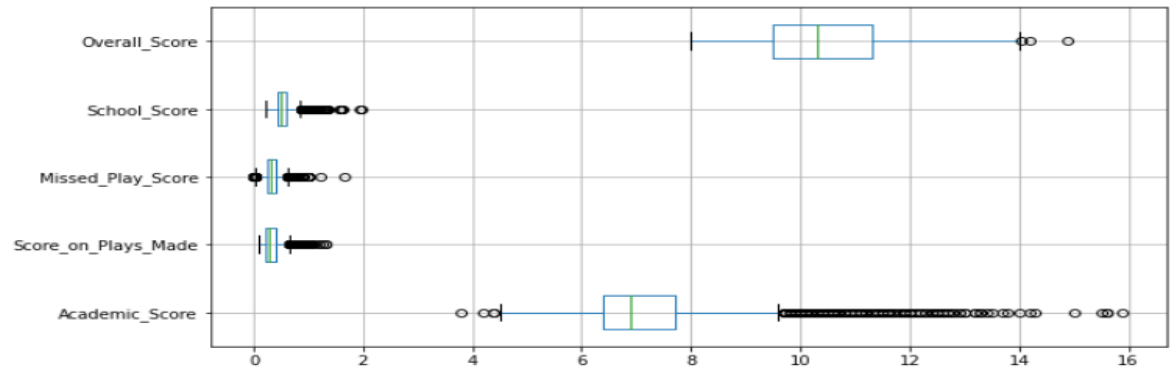
	Academic_Score	Score_on_Plays_Made	Missed_Play_Score	School_Score	Overall_Score
count	6215.000000	6215.000000	6215.000000	6215.000000	6215.000000
mean	7.219252	0.337338	0.319537	0.531448	10.456771
std	1.292237	0.160122	0.145153	0.147245	1.172504
min	3.800000	0.080000	0.000000	0.220000	8.000000
25%	6.400000	0.230000	0.250000	0.430000	9.500000
50%	7.000000	0.290000	0.310000	0.510000	10.200000
75%	7.700000	0.400000	0.390000	0.600000	11.300000
max	15.900000	1.330000	1.660000	1.980000	14.900000

Inference from data describe:

- As we can see from the standard deviation of Academic_score variable we can observe that the difference between the min and max values so there is outlier of extreme values and it is right skewed
- We can observe that all the variables are having same kind of units so we no need to o scaling before modelling
- Overall_Score variable is slightly right skewed and having a outlier as the min , median and mean are floating around the values between 8 and 10.5 but the max value indicate the value as 14

Univariate Analysis

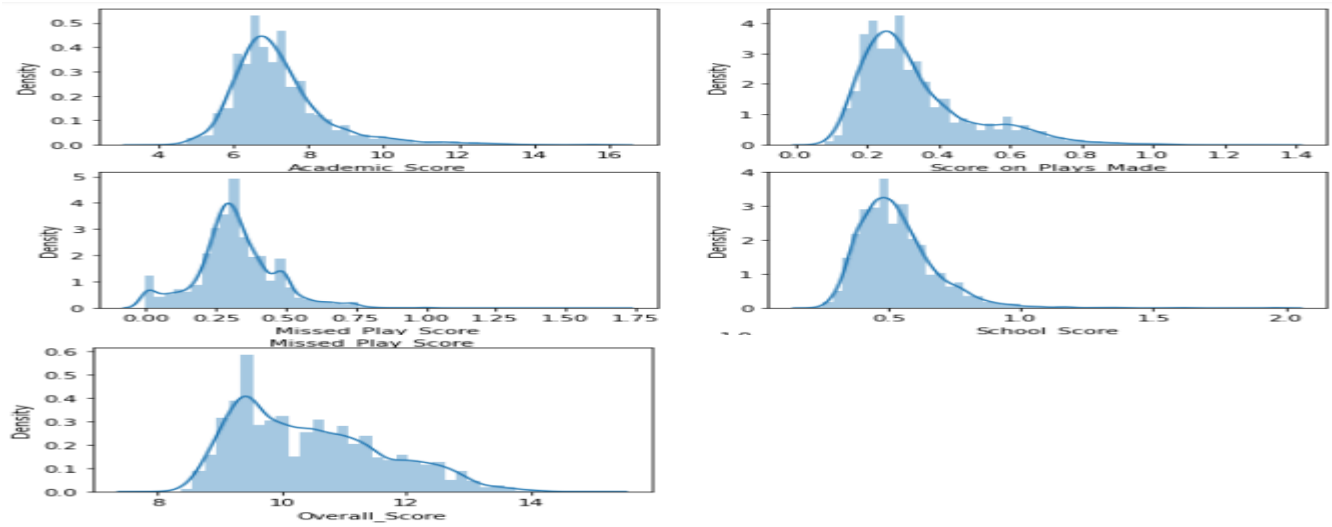
Boxplot



Inference from Boxplot:

The box plots indicate that all the continuous variables have outliers.
The box plot of 'Academic_Score has an extreme value of above 15.

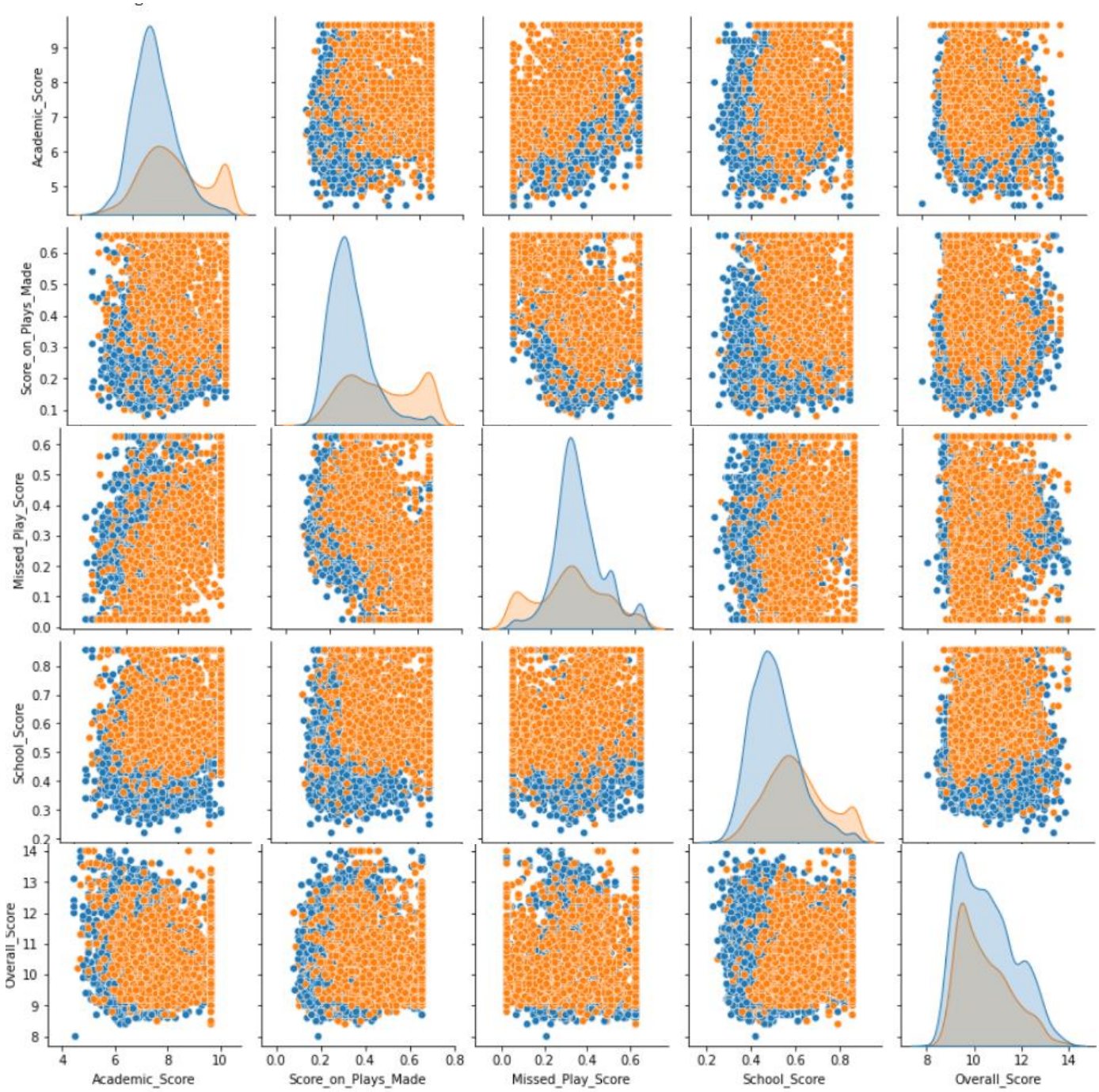
Histogram



Inference from Histogram:

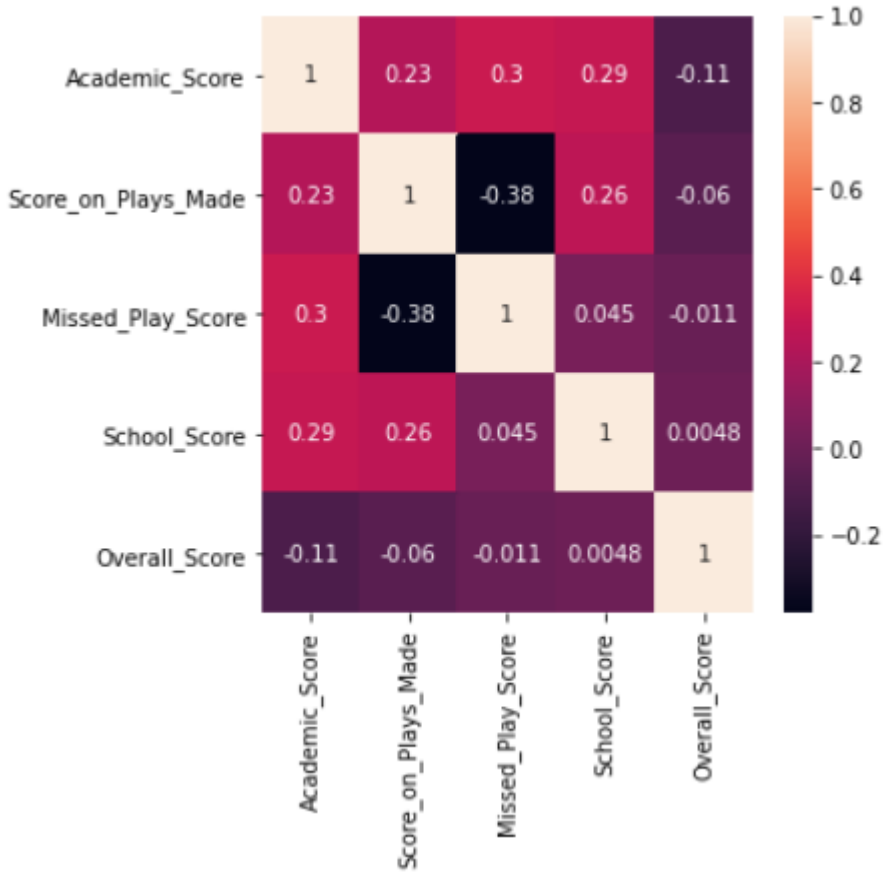
The Histogram indicate that all the continuous variables have Right skewed distribution.
So we need to handle outlier for uniform distribution.

Multivariate Analysis



Inference from pair plot:

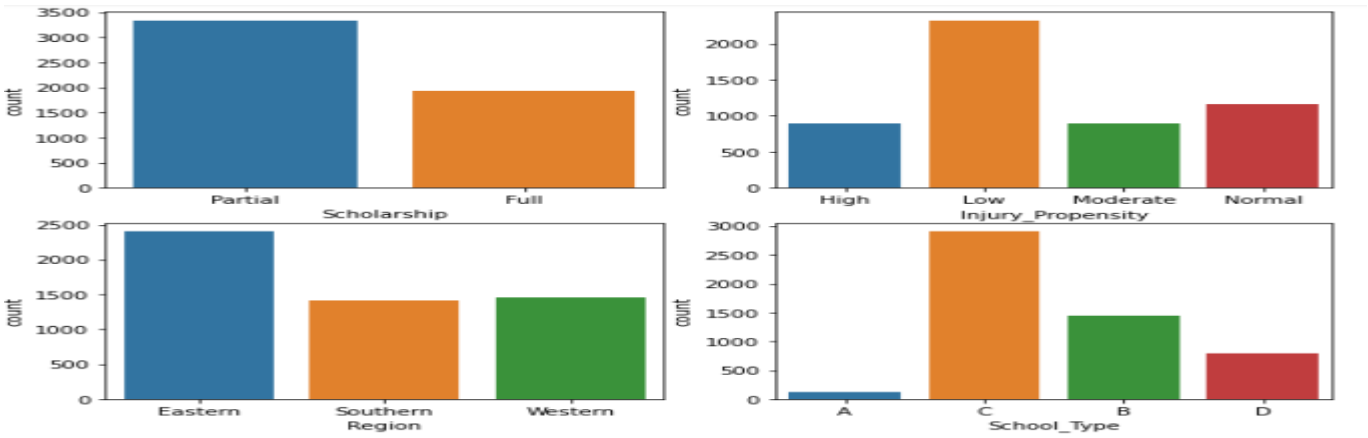
- The class imbalance problem is evident in the plots above.
- We can find there is no linear relationship between the variables



Inference from Heat map:

- There is some amount of multi collinearity between
  Score_on_plays_made - Missed_play_score,Academic_score and School_Score variables
- There is some amount of correlation between **Academic score** – Overall Score and School Score

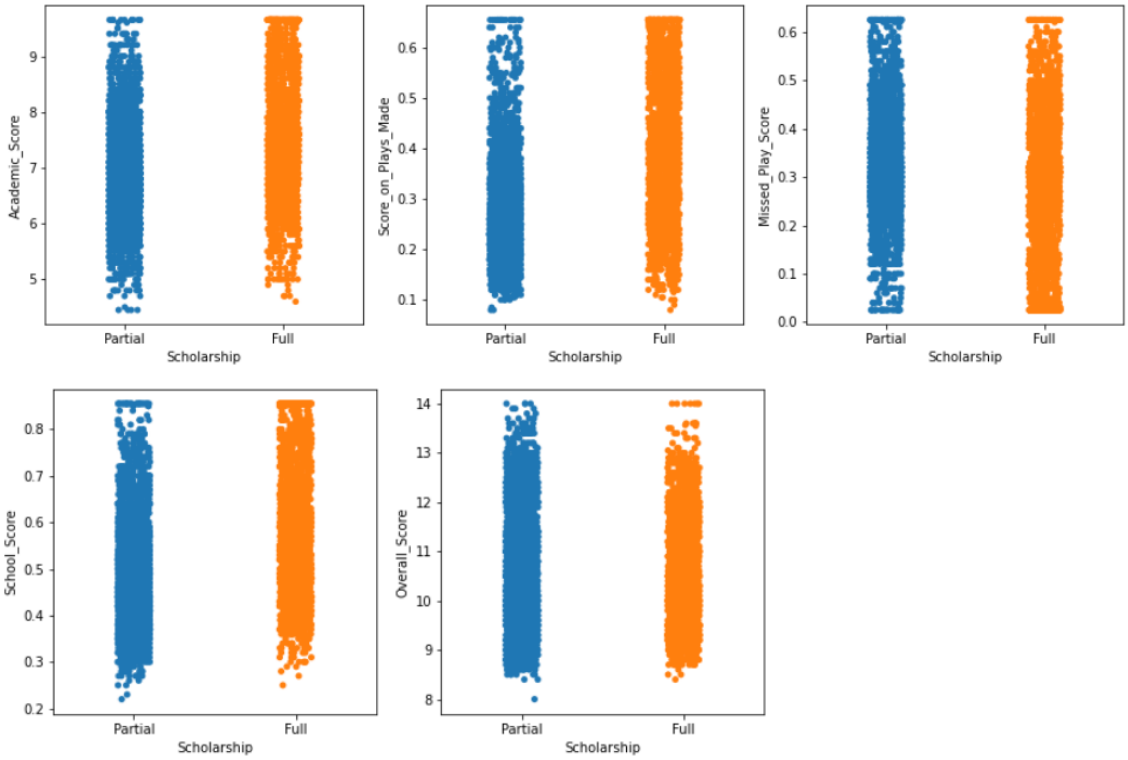
CountPlot:



Inference form Count Plot:

- The Scholarship Column (Dependent Variable) has a class imbalance as we can see partial class is more than 60% of the full class
- The School type C sample is more collected from the population
- Low injury students count are high compared to others
- Eastern region schools play an important role in this data

Strip Plot:



Inference form Strip Plot:

- When School Score is low then the partial scholarship is given to the student
- When Missed play Score is low the full scholarship is given to the students
- When Academic Score is low then the partial scholarship is given to the student

Data Pre-processing

```
def remove_outlier(col):
    sorted(col)
    Q1,Q3=np.percentile(col,[25,75])
    IQR=Q3-Q1
    lower_range= Q1-(1.5 * IQR)
    upper_range= Q3+(1.5 * IQR)
    return lower_range, upper_range
```

```
lr,ur=remove_outlier(data["Academic_Score"])
data["Academic_Score"]=np.where(data["Academic_Score"]>ur,ur,data["Academic_Score"])
data["Academic_Score"]=np.where(data["Academic_Score"]<lr,lr,data["Academic_Score"])
lr,ur=remove_outlier(data["Score_on_Plays_Made"])
data["Score_on_Plays_Made"]=np.where(data["Score_on_Plays_Made"]>ur,ur,data["Score_on_Plays_Made"])
data["Score_on_Plays_Made"]=np.where(data["Score_on_Plays_Made"]<lr,lr,data["Score_on_Plays_Made"])
lr,ur=remove_outlier(data["Missed_Play_Score"])
data["Missed_Play_Score"]=np.where(data["Missed_Play_Score"]>ur,ur,data["Missed_Play_Score"])
data["Missed_Play_Score"]=np.where(data["Missed_Play_Score"]<lr,lr,data["Missed_Play_Score"])
lr,ur=remove_outlier(data["School_Score"])
data["School_Score"]=np.where(data["School_Score"]>ur,ur,data["School_Score"])
data["School_Score"]=np.where(data["School_Score"]<lr,lr,data["School_Score"])
lr,ur=remove_outlier(data["Overall_Score"])
data["Overall_Score"]=np.where(data["Overall_Score"]>ur,ur,data["Overall_Score"])
data["Overall_Score"]=np.where(data["Overall_Score"]<lr,lr,data["Overall_Score"])
```

Inferences:

As from the Univariate analysis we found Outliers in all the Continuous variables so from the above code we have treated the Outliers, by capping the less than lower range values(Q1-(1.5 * IQR)) as lower range itself and greater than upper range values(Q3+(1.5 * IQR)) as upper range. So instead of deleting the outlier we are saving the information by this method.

Inferential Analysis:

Chi2 Test – Feature Importance

```
(array([20.3510347 , 24.2346512 ,  0.21811514]), array([6.44582240e-06, 8.52840144e-07, 6.40479855e-01]))
School_Type          6.404799e-01
Region               8.528401e-07
Injury_Propensity    6.445822e-06
dtype: float64
```

Inference from Chi2 test – Feature Importance

From the test we can find that the **'Region'** variable is the most important categorical variable compared with the dependent variable **'Scholarship'** as the Region variable has lowest P value and highest F score compared to other variables

Chi2 Test – Test of proportions

```
from scipy.stats import chisquare
School_Type=chisquare(data['School_Type'].value_counts())
Region=chisquare(data['Region'].value_counts())
Injury_Propensity=chisquare(data['Injury_Propensity'].value_counts())


print('The test of  proportions P value  for each column respectively are:\n',School_Type[1],Region[1],Injury_Propensity[1])
```

```
The test of  proportions P value  for each column respectively are:
 0.0 1.5983140044235471e-78 7.68115806851711e-230
```

Inference from Chi2 Test of Proportions

We can find that the value is < 0.05 and there is significant change in the proportions of the classes in the categorical variables

Feature Engineering

Categorical Variables Value Count

```
# Converting categorical variables into COntinuous variables
for column in data.columns:
    if data[column].dtype =='object':
        print(column.upper(),':  ',data[column].nunique())
```

```
INJURY_PROPENSITY :  4
SCHOOL_TYPE :  4
REGION :  3
SCHOLARSHIP :  2
```

Conversion of Object type data into integers:

Data mining algorithms in Python can take only numerical columns. It cannot take string / object types. All columns with data type object will need to be changed for both Independent Variable (IV) and Dependent Variable (DV) into integer.

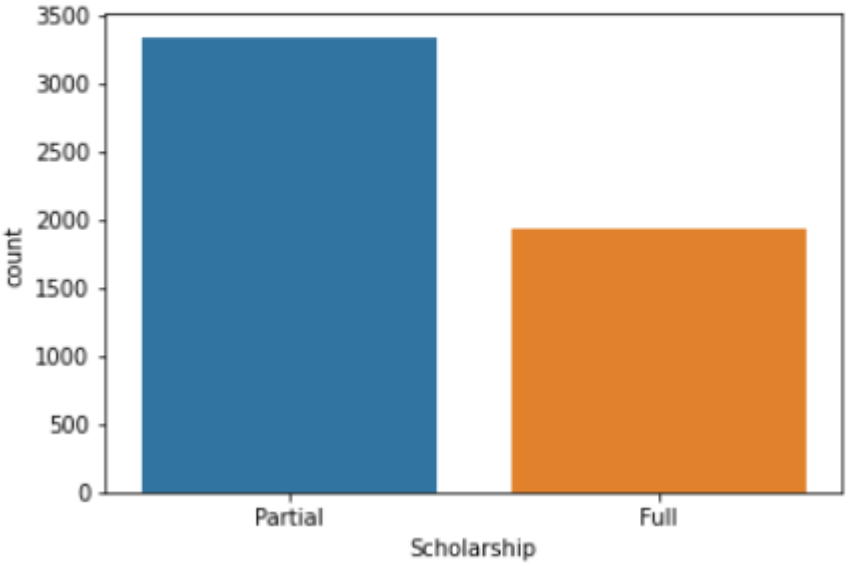
```
# dropping  the Scholarship feature as it is a target variable
cat1=['Injury_Propensity','School_Type','Region']
```

```
#We are applying the One-Hot method  for converting the columns
data =pd.get_dummies(data, columns=cat1,drop_first=True)
data['Scholarship']=np.where(data['Scholarship']=='Partial',0,1)
data.head(2)
```

Scholarship Column is encoded with **Label Encoding** method
Other Categorical variables are handled using **One Hot Encoding** method

Logistic Regression (MLE): Data Split: Split the data into test (30% of the data) and train (70% of the data), build classification model CART, Random Forest, Artificial Neural Network

```
Partial    63.344723
Full       36.655277
Name: Scholarship, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7ff75e1c5b50>
```



As per the problem statement 'SCHOLARSHIP' feature is our dependent variable. From the above Bar Graph we can find that the proportion of the two classes in the claimed feature, so there is small amount of class imbalance. For handling the imbalance we can implement some techniques like SMOTE, etc. But here we are not treating the the class imbalance and we are proceeding further for the model building part.

MODEL 1 and Its VIF Score

Logit Regression Results

Dep. Variable:	Scholarship	**No. Observations:**	5268
Model:	Logit	**Df Residuals:**	5254
Method:	MLE	**Df Model:**	13
Date:	Mon, 24 May 2021	**Pseudo R-squ.:**	0.3243
Time:	14:45:41	**Log-Likelihood:**	-2339.0
converged:	True	**LL-Null:**	-3461.6
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-8.5568	0.568	-15.063	0.000	-9.670	-7.443
Academic_Score	0.4301	0.044	9.765	0.000	0.344	0.516
Score_on_Plays_Made	5.1126	0.323	15.833	0.000	4.480	5.745
Missed_Play_Score	-1.4614	0.345	-4.233	0.000	-2.138	-0.785
School_Score	2.8541	0.313	9.123	0.000	2.241	3.467
Overall_Score	0.2004	0.045	4.488	0.000	0.113	0.288
Injury_Propensity_Low	1.6779	0.150	11.165	0.000	1.383	1.972
Injury_Propensity_Moderate	0.5662	0.157	3.616	0.000	0.259	0.873
Injury_Propensity_Normal	1.2244	0.154	7.941	0.000	0.922	1.527
School_Type_B	-2.1160	0.332	-6.378	0.000	-2.766	-1.466
School_Type_C	-0.8168	0.291	-2.811	0.005	-1.386	-0.247
School_Type_D	0.2617	0.279	0.940	0.347	-0.284	0.808
Region_Southern	-0.4816	0.091	-5.273	0.000	-0.661	-0.302

```
Academic_Score  VIF =  1.71
Score_on_Plays_Made  VIF =  1.61
Missed_Play_Score  VIF =  1.53
School_Score  VIF =  1.3
Overall_Score  VIF =  2.01
Injury_Propensity_Low  VIF =  3.55
Injury_Propensity_Moderate  VIF =  2.08
Injury_Propensity_Normal  VIF =  2.6
School_Type_B  VIF =  13.89
School_Type_C  VIF =  12.67
School_Type_D  VIF =  6.12
Region_Southern  VIF =  1.23
Region_Western  VIF =  1.26
```

Inference:

As we can find that form the VIF score > 5 are already well explaied by other variables , so here we are elimminating the variaables School Type B,C and D, also School_Type D has higher P value also

MODEL 2 and Its VIF Score

Logit Regression Results

Dep. Variable:	Scholarship	**No. Observations:**	5268
Model:	Logit	**Df Residuals:**	5257
Method:	MLE	**Df Model:**	10
Date:	Mon, 24 May 2021	**Pseudo R-squ.:**	0.3030
Time:	14:46:53	**Log-Likelihood:**	-2412.6
converged:	True	**LL-Null:**	-3461.6
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.7234	0.470	-16.442	0.000	-8.644	-6.803
Academic_Score	0.6203	0.040	15.513	0.000	0.542	0.699
Score_on_Plays_Made	5.9718	0.313	19.057	0.000	5.358	6.586
Missed_Play_Score	-1.4204	0.337	-4.219	0.000	-2.080	-0.761
School_Score	3.9417	0.294	13.388	0.000	3.365	4.519
Overall_Score	-0.1305	0.034	-3.865	0.000	-0.197	-0.064
Injury_Propensity_Low	0.8665	0.116	7.493	0.000	0.640	1.093
Injury_Propensity_Moderate	0.1421	0.134	1.064	0.287	-0.120	0.404
Injury_Propensity_Normal	0.6051	0.130	4.653	0.000	0.350	0.860
Region_Southern	-0.4851	0.090	-5.404	0.000	-0.661	-0.309
Region_Western	-0.0316	0.089	-0.353	0.724	-0.207	0.144

```
Academic_Score  VIF =  1.4
Score_on_Plays_Made  VIF =  1.51
Missed_Play_Score  VIF =  1.49
School_Score  VIF =  1.19
Overall_Score  VIF =  1.24
Injury_Propensity_Low  VIF =  2.46
Injury_Propensity_Moderate  VIF =  1.72
Injury_Propensity_Normal  VIF =  2.18
Region_Southern  VIF =  1.23
Region_Western  VIF =  1.24
```

Inference:

As we can find that from the VIF score > 5 are already well explaied by other variables , so here we are having all the variaables less than 5 , but we has higher P value for variables Region western, Injurity propensity Moerate so we are eliminationg that twoo variables for nect model

MODEL 3 and Its VIF Score

Logit Regression Results

Dep. Variable:	Scholarship	**No. Observations:**	5268
Model:	Logit	**Df Residuals:**	5259
Method:	MLE	**Df Model:**	8
Date:	Mon, 24 May 2021	**Pseudo R-squ.:**	0.3028
Time:	14:46:54	**Log-Likelihood:**	-2413.2
converged:	True	**LL-Null:**	-3461.6
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.7088	0.469	-16.427	0.000	-8.629	-6.789
Academic_Score	0.6207	0.040	15.529	0.000	0.542	0.699
Score_on_Plays_Made	5.9557	0.309	19.255	0.000	5.349	6.562
Missed_Play_Score	-1.4410	0.336	-4.289	0.000	-2.099	-0.783
School_Score	3.9358	0.294	13.382	0.000	3.359	4.512
Overall_Score	-0.1246	0.033	-3.746	0.000	-0.190	-0.059
Injury_Propensity_Normal	0.5247	0.107	4.901	0.000	0.315	0.735
Injury_Propensity_Low	0.7882	0.090	8.744	0.000	0.612	0.965
Region_Southern	-0.4746	0.084	-5.644	0.000	-0.639	-0.310

```
Academic_Score  VIF =  1.4
Score_on_Plays_Made  VIF =  1.47
Missed_Play_Score  VIF =  1.49
School_Score  VIF =  1.19
Overall_Score  VIF =  1.21
Injury_Propensity_Low  VIF =  1.55
Injury_Propensity_Normal  VIF =  1.52
Region_Southern  VIF =  1.09
```

MODEL 3

Inference

As from the VIF score and P value everything is perfect so there is no need of further models

Splitting the data into Training Set and Test Set

```
from sklearn.model_selection import train_test_split
```

```
Train,Test = train_test_split(data,test_size=0.3,random_state=1,stratify=data['Scholarship'])
```

```
print(Train.shape)
print(Test.shape)
```

```
(3687, 14)
(1581, 14)
```

```
print(Train['Scholarship'].value_counts(1))
print(Test['Scholarship'].value_counts(1))
```

```
0    0.633577
1    0.366423
Name: Scholarship, dtype: float64
0    0.633144
1    0.366856
Name: Scholarship, dtype: float64
```

For model building we need to split the data into Train and Test.Before splitting the data into Train and Test we need to separate Dependent variable and Independent variable and store in y and x data frames respectively, these steps remain common for all three models

For our model we are splitting the data into 70 % Train and 30 %Test.by using Stratify sampling method so that same proportions in the train and test data will be maintained. Random state provides you the freedom to work on different system while using the same sample data from the population. The Random state value chosen as 10

Logistic Regression Model

```
from sklearn.linear_model import LogisticRegression
LR = LogisticRegression(solver='newton-cg',penalty='none')
```

We Choose Logistic regression classifier as our first model , actually logistic regression is a type of linear model which is used in the classification problem because it brings the non-linearity in the model with the help of sigmoid curve transformation . Mainly logistic regression can be easily interpretable same like linear regression.

By using Logistic regression we have created three models as same as in the MLE method . so please find below the comparison of performance metrics

LDA Model

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
LDA = LinearDiscriminantAnalysis()
```

We used the model LDA classifier as our second model, as Linear discriminate analysis is use for multiclass classification and used to rank the independent variables using Z score (this Z score is different from Standard scalar) and also LDA is used for dimension reduction internally based on the classes.

LDA mainly concentrate on between group variance for separation of the classes to be maximum and the main assumption is the equality of variance in the classes of DV and the IV are independent of each other (it can be tested using Chi2 test of independence)

Here we used same three models which is used already in the MLE

Performance metrics: used for calculating the model performance in classification problems are

Accuracy Score- Companies use machine learning models to make practical business decisions.

Confusion Matrix - Confusion matrices make it easy to tell how accurate a model's outcomes are.

ROC (**Receiver Operating Characteristic Curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

AUC score (Area under the curve): The AUC value lies between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier.

RECALL = TP/TP+FN

PRECISION = TP/TP+FP

F1 Score = 2(SENSITIVITY * PRECISION)/(SENSITIVITY + PRECISION).

.

Logistic Regression Model:

Accuracy Scores:

Models	Accuracy Scores
Model 1	0.8407919717927854
Model 2	0.8294005966910768
Model 3	0.8299430431244914

Classification Report:

Model 1

	precision	recall	f1-score	support
0	0.83	0.94	0.88	1001
1	0.87	0.66	0.75	580
accuracy			0.84	1581
macro avg	0.85	0.80	0.81	1581
weighted avg	0.84	0.84	0.83	1581

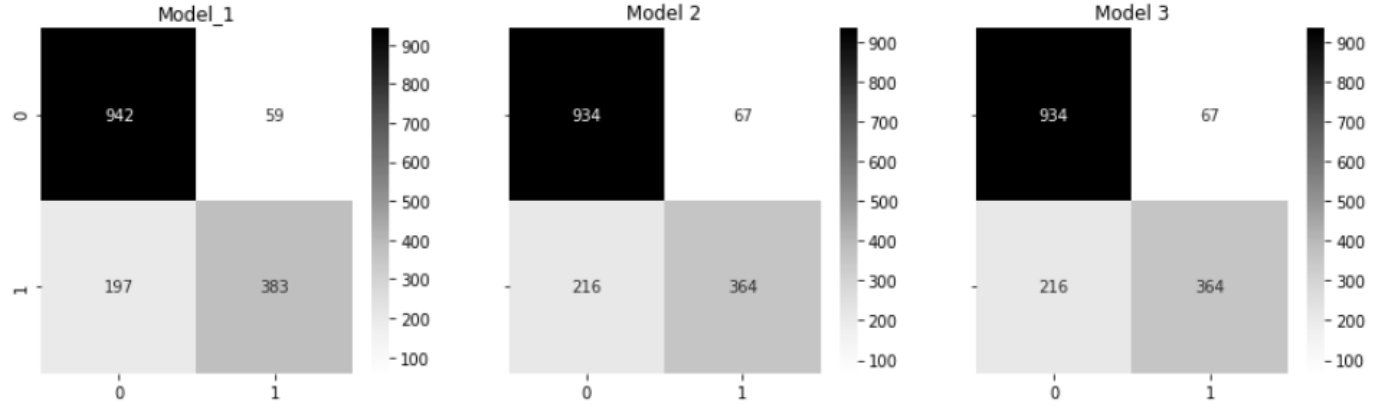
Model 2

	precision	recall	f1-score	support
0	0.81	0.93	0.87	1001
1	0.84	0.63	0.72	580
accuracy			0.82	1581
macro avg	0.83	0.78	0.79	1581
weighted avg	0.82	0.82	0.81	1581

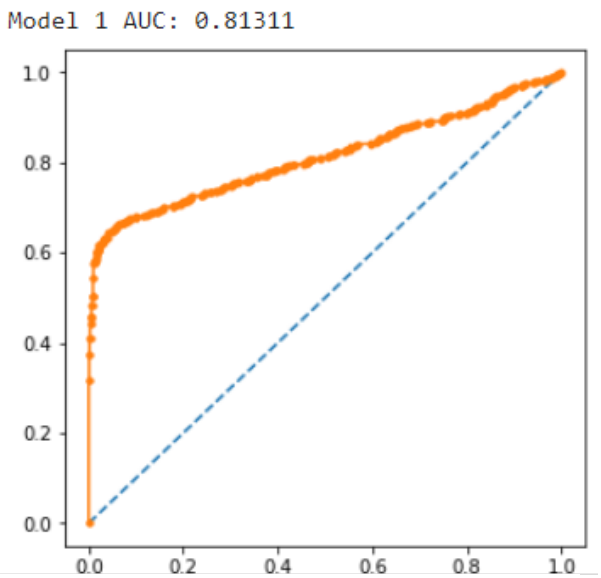
Model 3

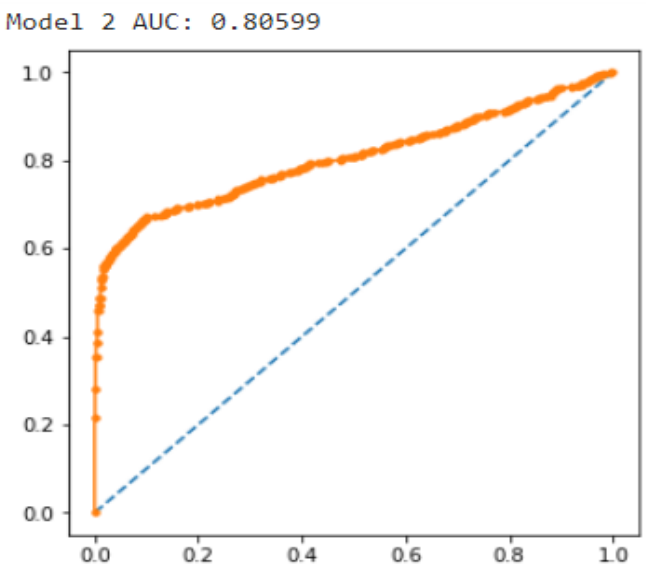
	precision	recall	f1-score	support
0	0.81	0.93	0.87	1001
1	0.84	0.63	0.72	580
accuracy			0.82	1581
macro avg	0.83	0.78	0.79	1581
weighted avg	0.82	0.82	0.81	1581

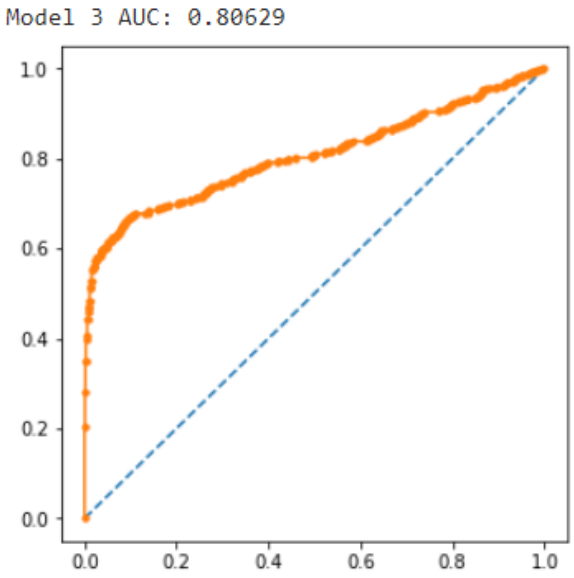
Confusion Matrix



AUC and ROC







Inference:

In the initial logistic model model the Accuracy scores was good and after feature elimination the accuracy score goes down , also from the classification report and confusion matrix model1 plays a good role in classifying the classes efficiently compared to other models.

LDA Model:

Accuracy Score

Models	Accuracy Scores
Model 1	0.8424193110930296
Model 2	0.8310279359913209
Model 3	0.8315703824247356

Classification Report

LDA Model 1

	precision	recall	f1-score	support
0	0.83	0.95	0.88	1001
1	0.87	0.66	0.75	580
accuracy			0.84	1581
macro avg	0.85	0.80	0.81	1581
weighted avg	0.84	0.84	0.83	1581

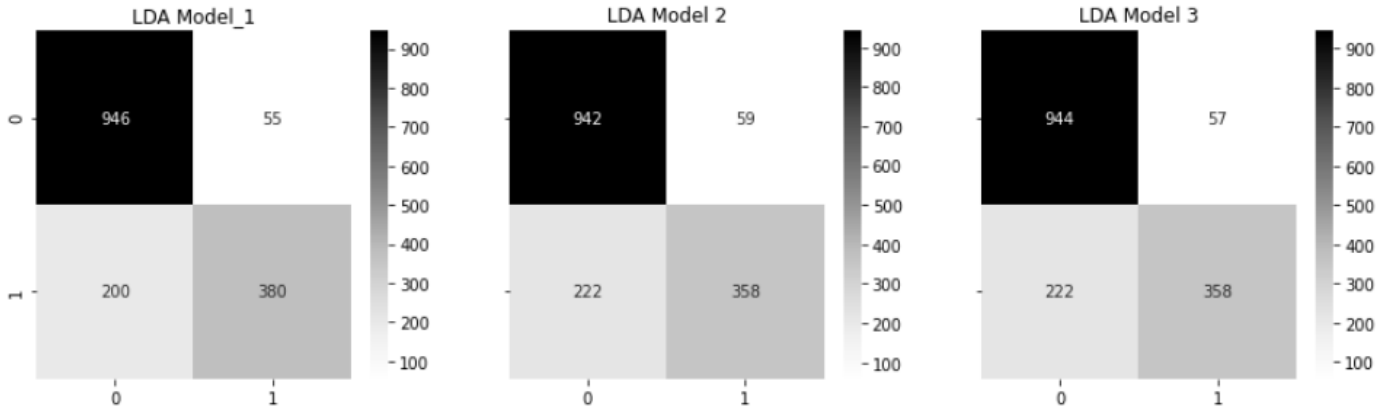
LDA Model 2

	precision	recall	f1-score	support
0	0.81	0.94	0.87	1001
1	0.86	0.62	0.72	580
accuracy			0.82	1581
macro avg	0.83	0.78	0.79	1581
weighted avg	0.83	0.82	0.81	1581

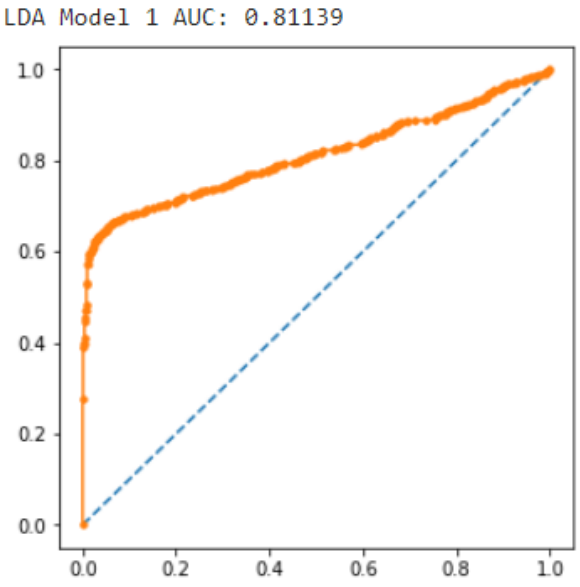
LDA Model 3

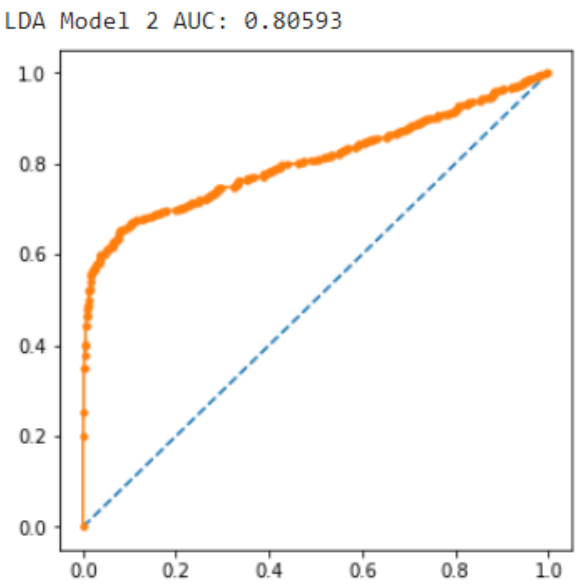
	precision	recall	f1-score	support
0	0.81	0.94	0.87	1001
1	0.86	0.62	0.72	580
accuracy			0.82	1581
macro avg	0.84	0.78	0.80	1581
weighted avg	0.83	0.82	0.82	1581

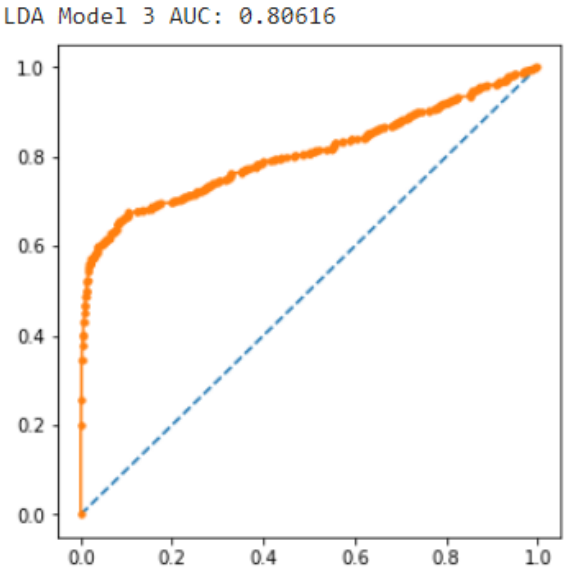
Confusion Matrix



AUC and ROC







Inference:

In the initial LDA model the Accuracy scores was good and after feature elimination the accuracy score goes down , also from the classification report and confusion matrix model1 plays a good role in classifying the classes efficiently compared to other models.

Inference on the comparison of model performance metrics:

a. LDA model classify the classes effectively than logistic regression model
b. Also in LDA the model1 has 1% more than the recall percentage as compared to the Logistic regression model1 recall
c. So from the above inferences we came to a conclusion for the best model as Linear discriminant model

Inference: Basis on these predictions, what are the business insights and recommendations

a. we can find that students having good Academic score and Overall Score is from schooltype C got a Scholarship and having low injuries
b. we can find that students having good Academic score and Overall Score is from schooltype D got a Scholarship and have normal injuries
c. we can find that students having good Academic score and Overall Score is again from schooltype C got a Scholarship and have moderate injuries
d. we can find that students having good Academic score and Overall Score is from schooltype B got a Scholarship and have low injuries

So from this we can get an inference that students from School type C can be given more importance during selection process and next importance can be given to school type B.

Also this inference can change if we collect more samples as naturally in this datatset we have more number of datatpoints belongs to SchoolType C

.

e. From the built prediction model (LDA) using the historical data provided by the company we can able to predict around 84% of the Scholarship Status corectly. So we can use this model for predicting the scholarship full or partial
f. Also, if we can able to discuss more with concern domain person of the company we can able to collect more insights on the same problem statement and build a better accurate model for predicting the future Scholarship.