

Contents

Contents	1
Problem Statement and Objective:	3
Understanding the dataset.....	3
1. Data Dictionary and data information:	3
2. EDA – Descriptive analysis.....	4
Datatypes:	4
Response variable distribution	5
Data distribution:	6
Data discrepancy:	6
Duplicate entries:	6
Missing data:	6
Skewness:.....	6
Outliers:7	
Histograms and Boxplots for predictors:	9
Bar plot analysis - object type variables.....	14
Multivariate analysis	15
3. Data Pre-processing	18
Outliers Treatment	18
4. Logistic regression with full data	19

4.1	Model 1 :	19
4.2	Model 2 :	21
4.3	Model 3 :	23
4.4	Model 4 :	25
4.5	Model 5 :	27
4.6	Model 6 :	29
4.7	Model Comparison :	31
5.	Logistic regression with train-test data.....	32
5.7	Model Comparison:	37
5.8	Comparison of best models between full dataset and train – test dataset:	37
6.	Linear Discriminant Analysis (LDA)	39
6.4	Model Comparison:	42
6.5	Comparison of best models from Logistic regression and LDA.....	43
7.	Conclusion	44

Problem Statement and Objective:

The goal is to understand the selection process of high school football players into college with a full or partial scholarship. We are provided details of 6215 high school graduates who have been inducted into 4-year degree colleges with either full or partial scholarships.

The objective is to:

- predict whether a high school graduate will win a full scholarship and,
- find out the important factors which are instrumental in winning a full scholarship.

Understanding the dataset

1. Data Dictionary and data information:

Data has **6215 rows and 9 columns** as given below including target variable:

Variable Name	Description	Variable Type	Variable role
Scholarship	It explains the type with values as 'Partial and 'Full'.	object	Response
Academic_Score	This is the high school academic performance of candidate	float	Predictor
Score_on_Plays_Made	This variable represents the score of a candidate as per achievements on the field.	float	Predictor
Missed_Play_Score	This represents the score of the candidate as per the failures on the field.	float	Predictor
Injury_Propensity	This column is calculated as based on how much time the candidate was injured. It has 4 values: High, Moderate, Normal, Low.	object	Predictor
School_Type	This represents type of school based on the location of school. There are 4 unique values : A, B,C,	object	Predictor

	D.		
School_Score	This variable represents a composite score based on the overall achievement of the candidates' school, based on the schools academic, sports and community service performance.	float	Predictor
Overall_Score	This represents a composite score based on a candidate's family financial state, school performance, psychosocial attitude etc	float	Predictor
Region	Length of the cubic zirconia in mm.	object	Predictor

2. EDA – Descriptive analysis.

Datatypes:

- There are **5 float and 4 object** datatypes in the given dataset.
- From describe () method, we may notice that all **the variables are on same scale**.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Academic_Score	6215	NaN	NaN	NaN	7.21925	1.29224	3.8	6.4	7	7.7	15.9
Score_on_Plays_Made	6215	NaN	NaN	NaN	0.337338	0.160122	0.08	0.23	0.29	0.4	1.33
Missed_Play_Score	6215	NaN	NaN	NaN	0.319537	0.145153	0	0.25	0.31	0.39	1.66
Injury_Propensity	6215	4	Low	2650	NaN	NaN	NaN	NaN	NaN	NaN	NaN
School_Type	6215	4	C	3384	NaN	NaN	NaN	NaN	NaN	NaN	NaN
School_Score	6215	NaN	NaN	NaN	0.531448	0.147245	0.22	0.43	0.51	0.6	1.98
Overall_Score	6215	NaN	NaN	NaN	10.4568	1.1725	8	9.5	10.2	11.3	14.9
Region	6215	3	Eastern	2835	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Scholarship	6215	2	Partial	4028	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- Only Missed_play_score starts from zero** which is understandable as it is the score based on failures on field which could be zero for a sportsman.
- In object types, we have three predictors with following unique records. There seem to be **no redundancy in object categorization**.

Column Injury_Propensity has 4 unique values:

Low	2650
Normal	1319
High	1181
Moderate	1065

Name: Injury_Propensity, dtype: int64

Column School_Type has 4 unique values:

C	3384
B	1620
D	1042
A	169

Name: School_Type, dtype: int64

Column Region has 3 unique values:

Eastern	2835
Western	1724
Southern	1656

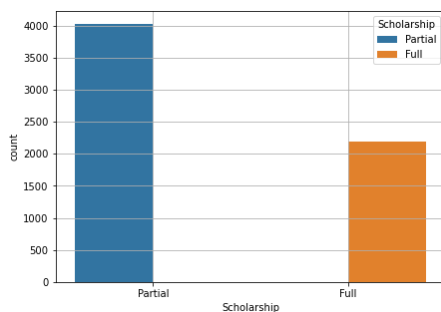
Name: Region, dtype: int64

Response variable distribution

- Proportion of classes in the data is sufficiently divided into both classes though the Full Scholarship is represented less and Partial Scholarship is represented more.
- Approximately, 65% data which includes 4028 records belong to Partial Scholarship class, and approx. 35% of data which includes 2187 records belongs to from Full Scholarship class

```
print(df['Scholarship'].value_counts())  
print(df['Scholarship'].value_counts(normalize=True))
```

```
Partial    4028  
Full       2187  
Name: Scholarship, dtype: int64  
Partial    0.648109  
Full       0.351891  
Name: Scholarship, dtype: float64
```



Data distribution:

- All variables are on same scale or range.
- Data is more or less normal as mean and median values for all variables are very close.
- Minimum and maximum values for all variables are well within limits possible values, hence the values in the data are valid entries.
- For Injury_Propensity, 'low' category is top frequency value, for School_Type top category is 'C' and for Region top frequented category is 'Eastern'.

Data discrepancy:

There is no discrepancy observed in the records so far.

Duplicate entries:

There are 947 identical records in the dataset, but since there is no unique identification column for each record, it could be different students with identical scores. Hence, we will consider these records as separate entries.

Missing data:

#	Column	Non-Null Count	Dtype
0	Academic_Score	6215 non-null	float64
1	Score_on_Plays_Made	6215 non-null	float64
2	Missed_Play_Score	6215 non-null	float64
3	Injury_Propensity	6215 non-null	object
4	School_Type	6215 non-null	object
5	School_Score	6215 non-null	float64
6	Overall_Score	6215 non-null	float64
7	Region	6215 non-null	object
8	Scholarship	6215 non-null	object

There are no null values in the dataset.

Skewness:

All variables are skewed to right.

Academic_Score	1.750576
Score_on_Plays_Made	1.405692
Missed_Play_Score	0.492118
School_Score	1.733181
Overall_Score	0.601023

- *Strongly skewed* – ‘Academic_Score’, ‘School_Score’ are strongly skewed with skewness value is greater than 1.7 and ‘Score_on_Plays_Made’ is also strongly skewed with skewness more than 1.
- *Moderately skewed* – ‘Overall_Score’ is moderately skewed with skewness of 0.60.
- *Weakly skewed / Symmetric* – ‘Missed_Play_Score’ has skewness of more than 0.49. We may consider it as weakly skewed as the value greater than 0.5 is considered moderately skewed and here the value is close to 0.5.

Outliers:

We may observe that all continuous variable columns have outliers.

Variable Name	No. of Outliers
Academic_Score	336
Injury_Propensity	0
Missed_Play_Score	481
Overall_Score	3
Region	0
Scholarship	0
School_Score	179
School_Type	0
Score_on_Plays_Made	339

While analyzing the outliers, we need to keep in mind following points:

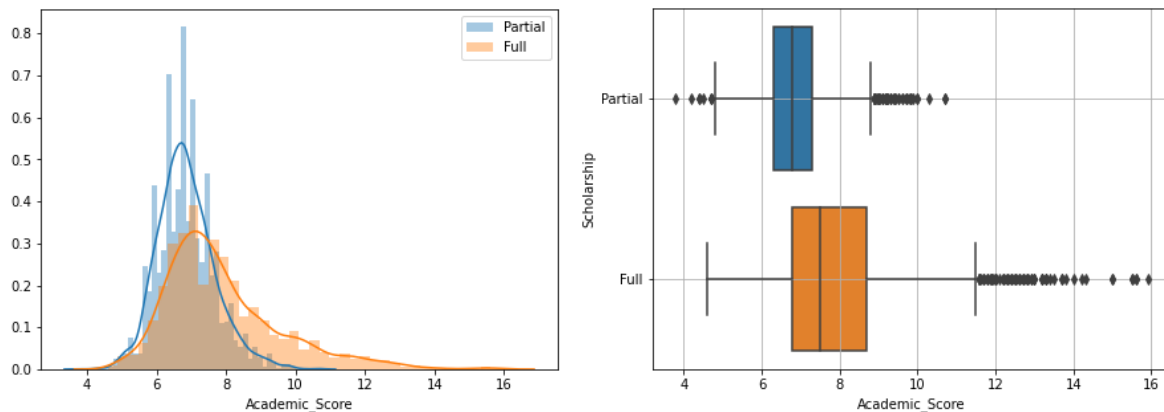
- Whether the outlier is valid or a discrepancy – this will be explored in our visual EDA in next section.
- If the outlier is valid, then how much does it impact the whole data – it can be noted that the *percentage of the outliers for all variables in overall*

*data is very small. The **highest number of outliers are for column Missed_Play_Score is 481 which is less than 7.739% of the whole data.***

EDA – Visual Analysis

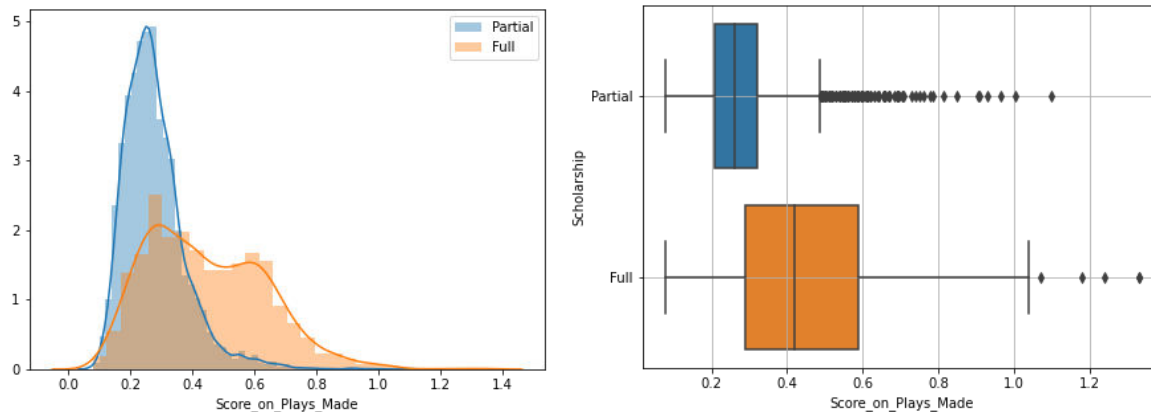
Histograms and Boxplots for predictors:

Academic_Score



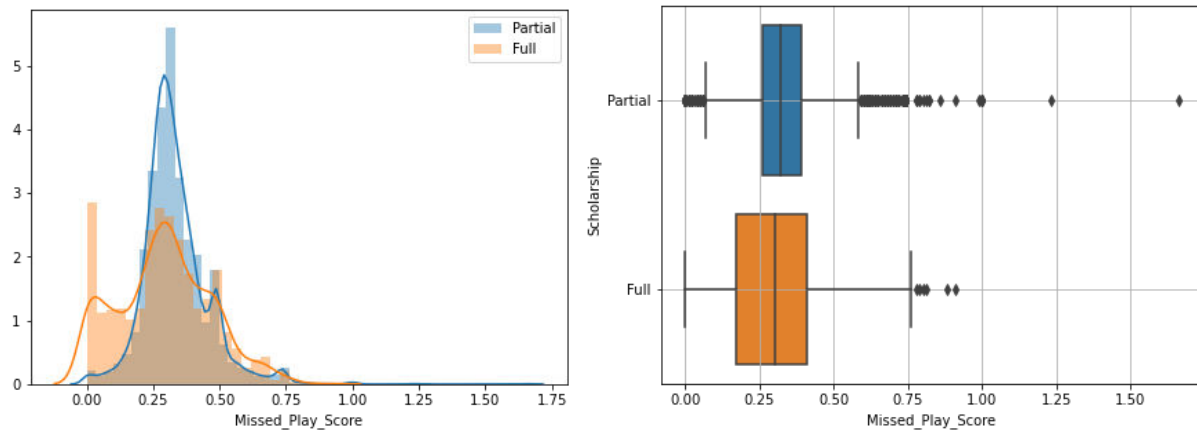
- From graphs, we may see that the *'Academic_Score' is skewed strongly to the right.*
- Class Partial has higher kurtosis and class Full has higher skewness.
- However, if we look into the distribution of two classes, *Partial class seems more normally distributed than Full.*
- **The column has 336 outliers but these are valid outliers** as the range maximum and minimum values are possible values for Academic_Scores.
- Partial class has outliers on both sides while Full class has only upper outliers.
- From the maximum and minimum values in both classes, we understand that the outliers are valid outliers.
- Span of distribution for Full class is on higher score values as compared to Partial class.

Score_on_Plays_Made:



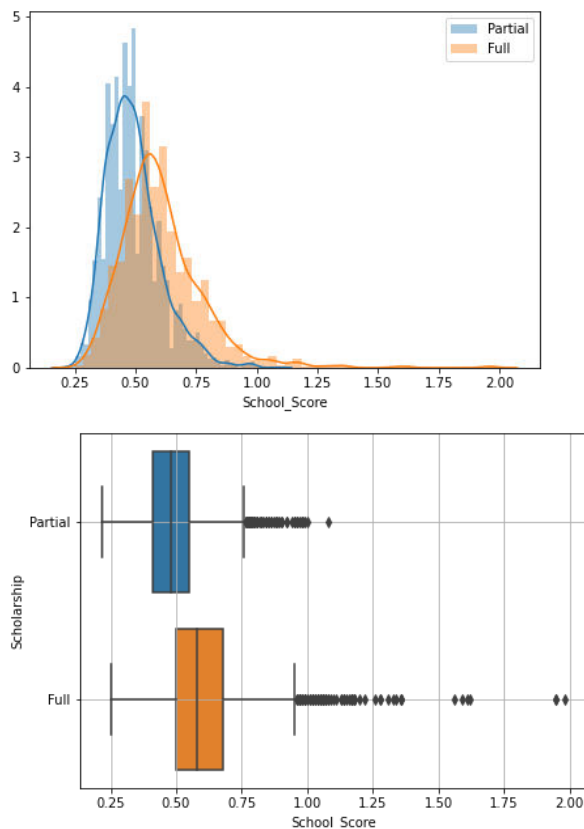
- Distribution of Score_on_Plays_Made is strongly skewed to right.
- We may see that even in this case class Partial class seems more normally distribute than class Full.
- Class Partial has higher kurtosis and class Full has higher skewness.
- There are **339 outliers** in the column on upper side, however, Full class contributes very less to the number of outliers.
- From the maximum and minimum values in both classes, we understand that the outliers are valid outliers.
- We may also observe that the class Full has distribution more spanned on higher side of the score value while class Partial is spanned more on the lower value of score.

Missed_Play_Score:



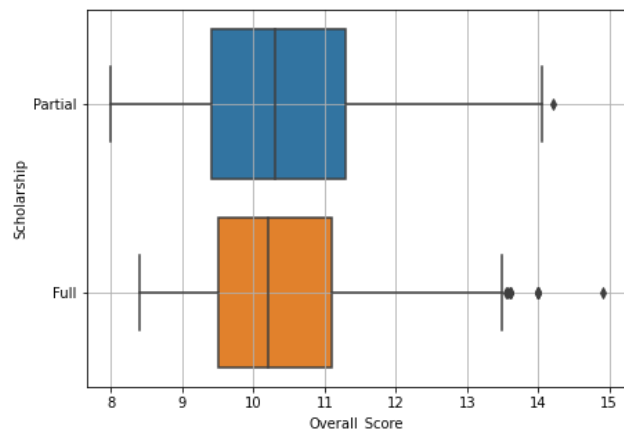
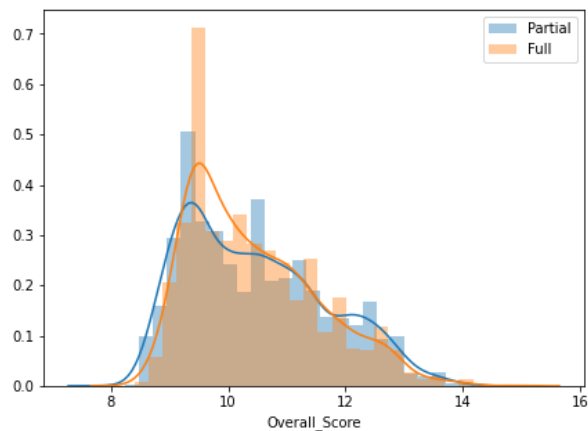
- Distribution of Missed_Plays_Score is more or less symmetrical. We may observe a very slight skewness to right.
- Even in this column, class Partial has higher kurtosis and class Full has higher skewness.
- There are **481 outliers** in the column. Class Full outliers only on upper side while class Partial has outlier on both sides.
- From the maximum and minimum values in both classes, we understand that the outliers are valid outliers.
- The span of distribution for class Full is wider than class Partial but the distribution is more towards lower value of Missed_Play_Score. This does make sense as Missed_Play_Score is not a good credit for a Sportsman.

School_Score:



- Distribution of School_Score is more skewed to right specially for class Full.
- Even in this column, class Partial has higher kurtosis and class Full has higher skewness but as compared to other variables, kurtosis of Full in this column is sufficiently higher and the two peaks are comparable.
- There are **179 outliers** in the column. Both classes have outliers only on upper side.
- From the maximum and minimum values in both classes, we understand that the outliers are valid outliers.
- The span of distribution for class Full is closer to class Partial in this column but still, class Full distribution is on higher scores as compared to class Partial.

Overall_Score

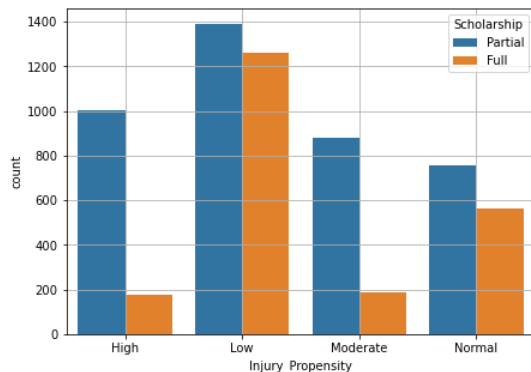


- Distribution of School_Score is more or less overlapping for two classes hence both distributions are in the same span of values.
- Column is moderately skewed to the right for both classes.
- Both classes have similar skewness and kurtosis.
- There are **only 3 outliers** overall in the column. Both classes have outliers only on upper side.
- Since, overall score is derived from other score values, we may consider the outliers in this column are valid outliers as drivers of overall score are well within possible range.
- The span of distribution for both classes is more or less in the same range of score values. This column may not be a good contributor for predicting classes as the distribution for both classes for Overall_Score column is mostly overlapping.

Bar plot analysis - object type variables

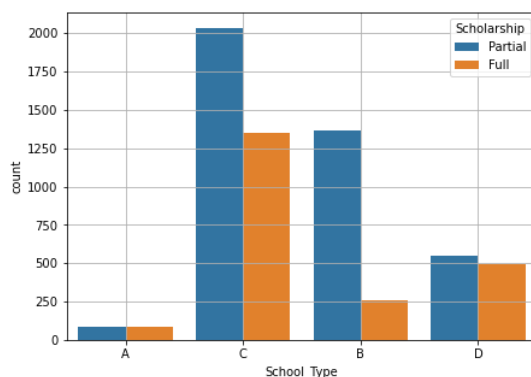
We have used bar plots to check if object variables have any particular pattern in determining the classes for the Scholarship variable.

Injury_Propensity



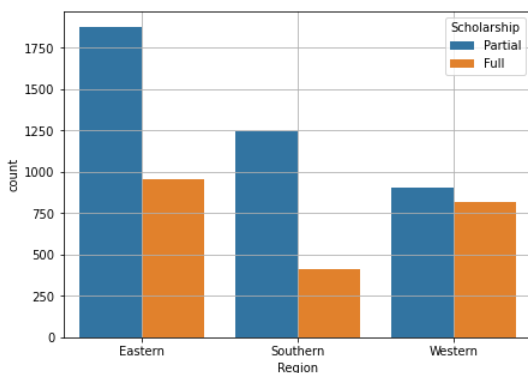
- Highest number of students belong to the 'low' Injury_Propensity, followed by Normal, High and Moderate.
- Proportion of the two classes in Low and Normal category is almost equal with number of students with Full are only slightly lower than Partial. Hence, these features may or may not be very helpful for classification.
- In other two categories, proportion of Full is much lower than Partial.

School_Type:



- Highest number of students belong to the 'C' School_Type, followed by B, D and A.
- Proportion of the two classes in A and D category is almost equal with number of students with Full are only slightly lower than Partial. Hence, these features may or may not be very helpful for classification.
- In other two categories i.e. C and B, proportion of Full is lower than Partial.
- 'C' School_Type has highest student for both Full and Partial classes.

Region:

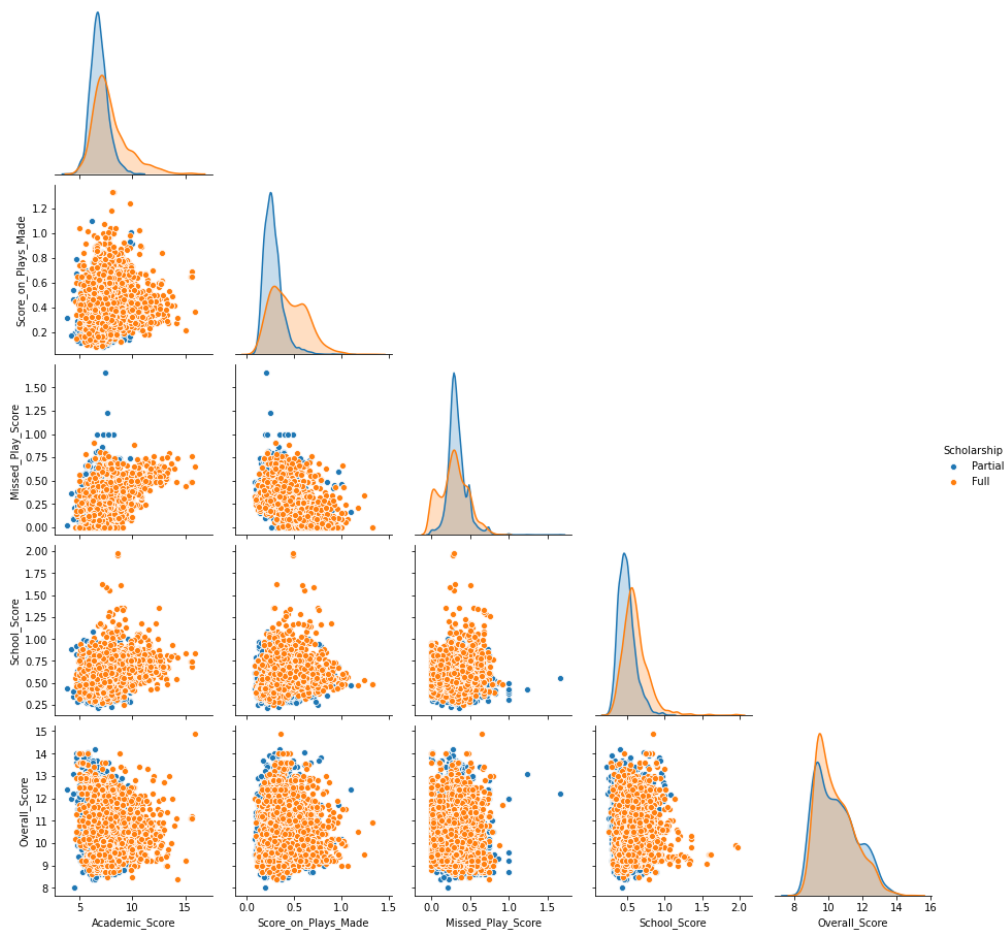


- Eastern Region has highest number of students overall and for both classes individually followed by Southern and Western.
- Proportion of the two classes in Western category is almost equal with number of students with Full are only slightly lower than Partial. Hence, this feature may or may not be very helpful for classification.
- In other two categories i.e., Eastern and Southern, proportion of Full is considerably lower than Partial.

Multivariate analysis

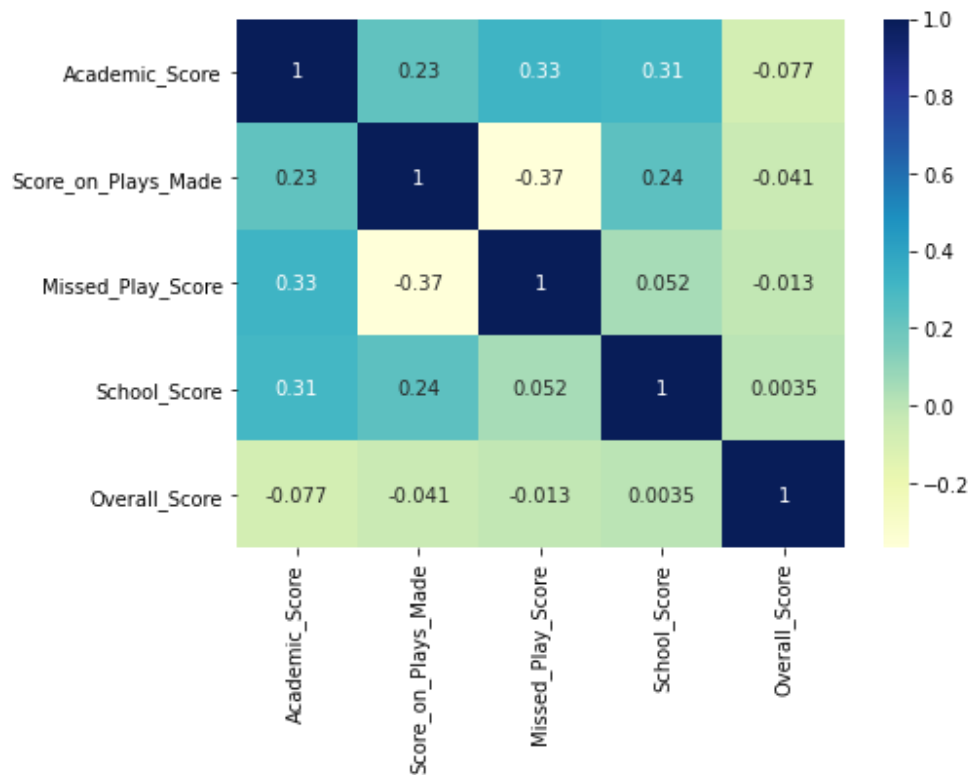
Pair plot

From pair-plot we may check the linear relationship between the variables.



- As we may see the datapoints are distributed all over the scatter plots in various combinations, we conclude there is very little or no correlation between different types of scores.
- The pattern holds true for both classes i.e., Full and Partial.

Heat Map



- Our observation from pair plot is confirmed from the heat map.
- It is clear that highest correlation coefficient value is -0.37 between Missed_Play_Score and Score_on_Plays_Made.
- Inverse relation between two can be understandable as one is count of failures and other is count of achievements on the field.

3. Data Pre-processing

Outliers Treatment

In point 4. Outliers, we discussed the basis of analyzing outliers on following two points:

Whether the outlier is valid or a discrepancy?

From our detailed analysis in boxplots, we have concluded that outliers in all the features are valid. The validation has been made on industry standards or logical interpretation of available data.

Since, all the outliers are valid, we must now investigate the second condition

What is the impact of outliers on data?

If the outliers are valid and not significant in numbers, the effect of such outliers would be negligible and will not impact the slope too much. In present data, we can see the *percentage of the outliers for all variables in overall data is significant enough to be not dropped.*

Outliers before scaling	Outlier after scaling
Academic_Score 336	Academic_Score : 130
Injury_Propensity 0	Score_on_Plays_Made : 82
Missed_Play_Score 481	Missed_Play_Score : 27
Overall_Score 3	School_Score : 71
Region 0	Overall_Score : 12
Scholarship 0	Injury_Propensity_Low : 0
School_Score 179	Injury_Propensity_Moderate : 0
School_Type 0	Injury_Propensity_Normal : 0
Score_on_Plays_Made 339	School_Type_B : 0
dtype: int64	School_Type_C : 0
	School_Type_D : 0
	Region_Southern : 0
	Region_Western : 0

To retain the outlier and not miss significant amount of data, we have applied scaling (zscore) on the data which reduced the number of outliers which we capped to the upper and lower bound respectively.

4. Logistic regression with full data

We have built the logistic regression model with full data set without any train-test split and built multiple iterations as follows:

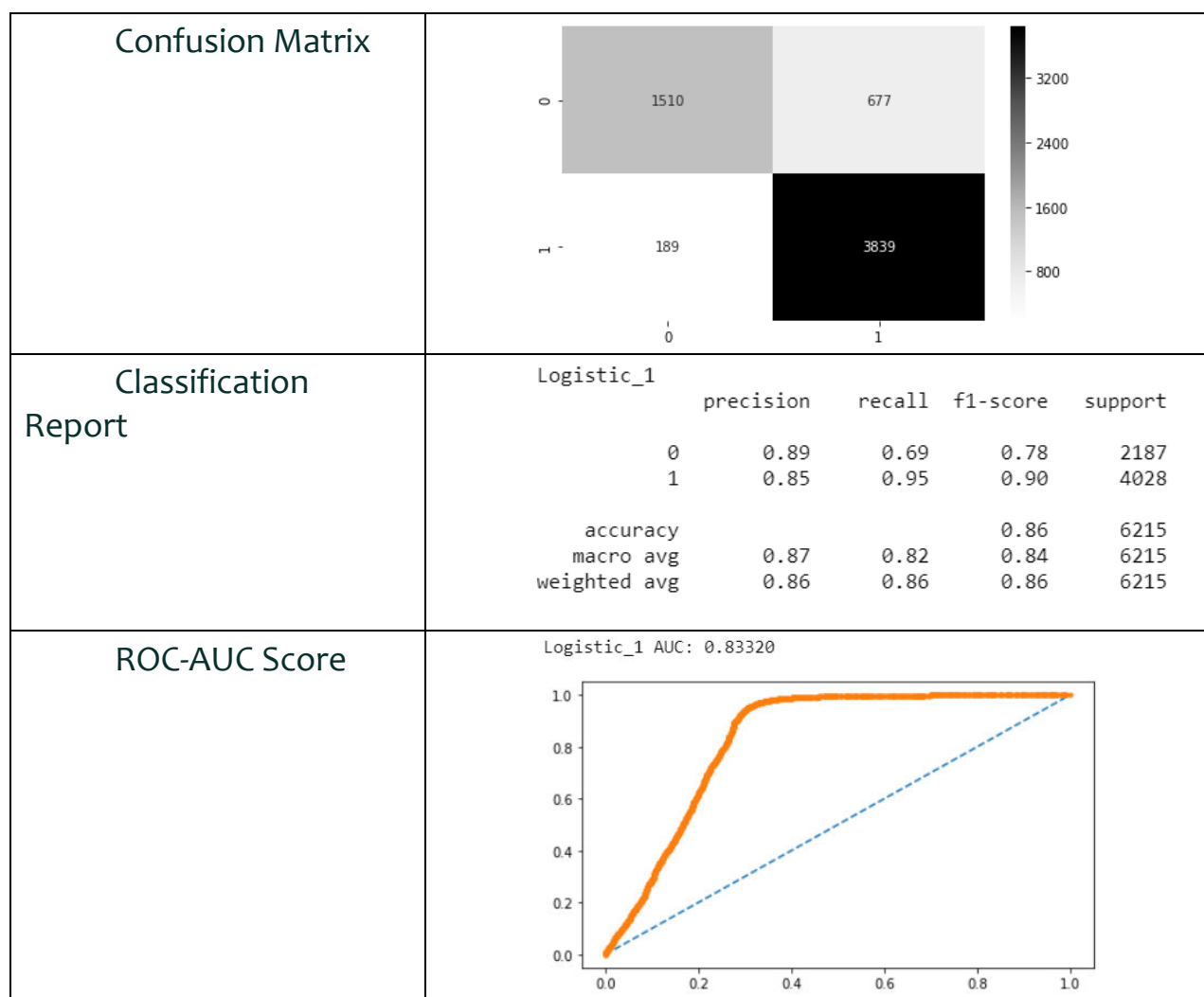
4.1 Model 1 :

This is the very first iteration where we have considered all the variables to build the model and checked various evaluation parameters. At the end we have checked the VIF (variance inflation factor) to check for the multicollinearity amongst the variables based on which we will proceed to the next iteration. Following are the results of this iteration.

Logit Regression Results

Dep. Variable:	Scholarship	No. Observations:	6215
Model:	Logit	Df Residuals:	6201
Method:	MLE	Df Model:	13
Date:	Mon, 01 Feb 2021	Pseudo R-squ.:	0.3643
Time:	12:36:48	Log-Likelihood:	-2562.4
converged:	True	LL-Null:	-4031.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.7276	0.036	20.288	0.000	0.657	0.798
Academic_Score	-0.5839	0.053	-11.108	0.000	-0.687	-0.481
Score_on_Plays_Made	-0.8546	0.048	-17.870	0.000	-0.948	-0.761
Missed_Play_Score	0.2090	0.046	4.533	0.000	0.119	0.299
School_Score	-0.4350	0.042	-10.314	0.000	-0.518	-0.352
Overall_Score	-0.2572	0.051	-5.059	0.000	-0.357	-0.158
Injury_Propensity_Low	-0.9148	0.072	-12.637	0.000	-1.057	-0.773
Injury_Propensity_Moderate	-0.2358	0.057	-4.137	0.000	-0.348	-0.124
Injury_Propensity_Normal	-0.5334	0.061	-8.679	0.000	-0.654	-0.413
School_Type_B	0.8678	0.145	5.978	0.000	0.583	1.152
School_Type_C	0.3091	0.144	2.149	0.032	0.027	0.591
School_Type_D	-0.1262	0.102	-1.242	0.214	-0.325	0.073
Region_Southern	0.1966	0.039	5.049	0.000	0.120	0.273
Region_Western	-0.0152	0.039	-0.386	0.699	-0.092	0.062



VIF is also check for all the above variables used to build this model and

```

Academic_Score VIF = 1.84
Score_on_Plays_Made VIF = 1.59
Missed_Play_Score VIF = 1.54
School_Score VIF = 1.29
Overall_Score VIF = 2.08
Injury_Propensity_Low VIF = 3.58
Injury_Propensity_Moderate VIF = 2.06
Injury_Propensity_Normal VIF = 2.56
School_Type_B VIF = 13.32
School_Type_C VIF = 12.15
School_Type_D VIF = 6.17
Region_Southern VIF = 1.23
Region_Western VIF = 1.25

```

their values are as follows:

We can see that we get high VIF values for 'School_Type_B' , School_Type_C' and School_Type_D'. But the p-values of School_Type_B' , School_Type_C' still indicates them to be the significant variable, so we select 'School_Type_D' to be dropped for the next iteration thus to remove the multicollinearity and use only significant variables to build the model.

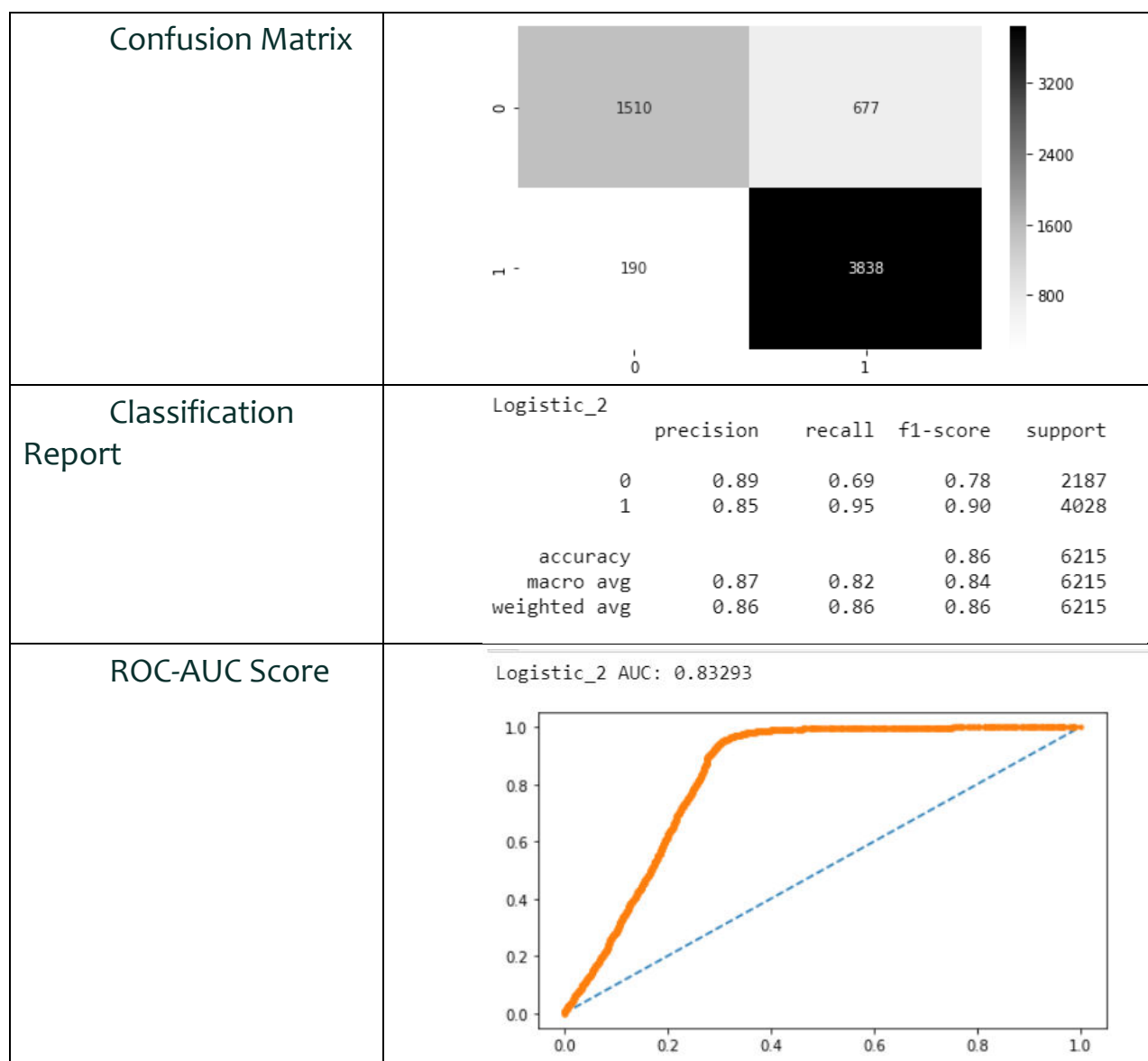
4.2 Model 2 :

As concluded in the previous iteration, we will drop 'School_Type_D' and proceed with model building and check the results as follows:

Logit Regression Results

Dep. Variable:	Scholarship	No. Observations:	6215
Model:	Logit	Df Residuals:	6202
Method:	MLE	Df Model:	12
Date:	Tue, 02 Feb 2021	Pseudo R-squ.:	0.3642
Time:	23:49:13	Log-Likelihood:	-2563.2
converged:	True	LL-Null:	-4031.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.7271	0.036	20.273	0.000	0.657	0.797
Academic_Score	-0.5751	0.052	-11.039	0.000	-0.677	-0.473
Score_on_Plays_Made	-0.8536	0.048	-17.843	0.000	-0.947	-0.760
Missed_Play_Score	0.2104	0.046	4.560	0.000	0.120	0.301
School_Score	-0.4312	0.042	-10.257	0.000	-0.514	-0.349
Overall_Score	-0.2613	0.051	-5.150	0.000	-0.361	-0.162
Injury_Propensity_Low	-0.9216	0.072	-12.745	0.000	-1.063	-0.780
Injury_Propensity_Moderate	-0.2402	0.057	-4.216	0.000	-0.352	-0.129
Injury_Propensity_Normal	-0.5376	0.061	-8.747	0.000	-0.658	-0.417
School_Type_B	1.0093	0.090	11.193	0.000	0.833	1.186
Region_Southern	0.1965	0.039	5.047	0.000	0.120	0.273
Region_Western	-0.0156	0.039	-0.396	0.692	-0.093	0.062
School_Type_C	0.4639	0.072	6.444	0.000	0.323	0.605



VIF is also check for all the above variables used to build this model and

```

Academic_Score VIF = 1.81
Score_on_Plays_Made VIF = 1.59
Missed_Play_Score VIF = 1.54
School_Score VIF = 1.29
Overall_Score VIF = 2.07
Injury_Propensity_Low VIF = 3.56
Injury_Propensity_Moderate VIF = 2.06
Injury_Propensity_Normal VIF = 2.56
School_Type_B VIF = 5.78
School_Type_C VIF = 3.15
Region_Southern VIF = 1.23
Region_Western VIF = 1.25
          
```

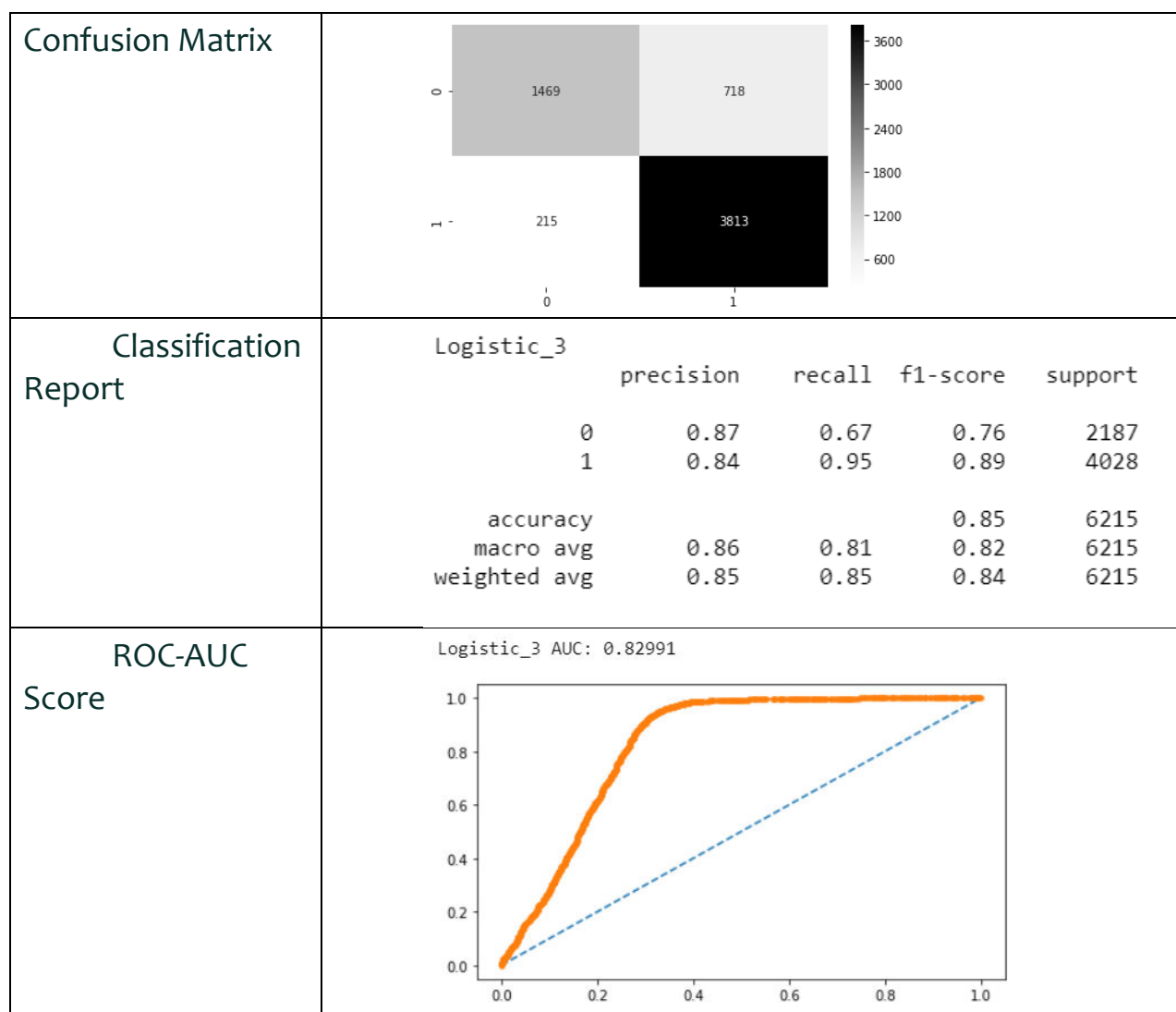
their values are as follows:

We can see that we get high VIF values for 'School_Type_B' , which indicates the multicollinearity in the data, so this variable is selected to be removed in the next iteration.

4.3 Model 3 :

As discussed in the previous iteration, we will drop 'School_Type_B' along with 'School_Type_D' and proceed with model building and check the results as follows:

Logit Regression Results						
Dep. Variable:	Scholarship	No. Observations:	6215			
Model:	Logit	Df Residuals:	6203			
Method:	MLE	Df Model:	11			
Date:	Tue, 02 Feb 2021	Pseudo R-squ.:	0.3475			
Time:	23:49:15	Log-Likelihood:	-2630.2			
converged:	True	LL-Null:	-4031.1			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.7251	0.035	20.586	0.000	0.656	0.794
Academic_Score	-0.8050	0.047	-17.086	0.000	-0.897	-0.713
Score_on_Plays_Made	-0.9678	0.047	-20.696	0.000	-1.059	-0.876
Missed_Play_Score	0.1959	0.046	4.297	0.000	0.107	0.285
School_Score	-0.5671	0.040	-14.184	0.000	-0.645	-0.489
Overall_Score	0.0751	0.040	1.864	0.062	-0.004	0.154
Injury_Propensity_Low	-0.4812	0.055	-8.672	0.000	-0.590	-0.372
Injury_Propensity_Moderate	-0.0261	0.050	-0.519	0.604	-0.124	0.072
Injury_Propensity_Normal	-0.2483	0.052	-4.810	0.000	-0.349	-0.147
Region_Western	-0.0112	0.039	-0.290	0.772	-0.087	0.065
Region_Southern	0.2031	0.038	5.292	0.000	0.128	0.278
School_Type_C	-0.1839	0.039	-4.699	0.000	-0.261	-0.107



VIF is also check for all the above variables used to build this model and their values are as follows:

```

Academic_Score VIF = 1.45
Score_on_Plays_Made VIF = 1.51
Missed_Play_Score VIF = 1.54
School_Score VIF = 1.2
Overall_Score VIF = 1.37
Injury_Propensity_Low VIF = 2.42
Injury_Propensity_Moderate VIF = 1.84
Injury_Propensity_Normal VIF = 2.12
School_Type_C VIF = 1.24
Region_Southern VIF = 1.23
Region_Western VIF = 1.25

```


We can check that we get relatively good VIF values for all the variables used, which indicates there is no more multicollinearity in the data, so now we will drop the variable on the basis of its significance(p-value). Now we select 'Region_Western' to be dropped since it has highest p-value and then build the next iteration of model.

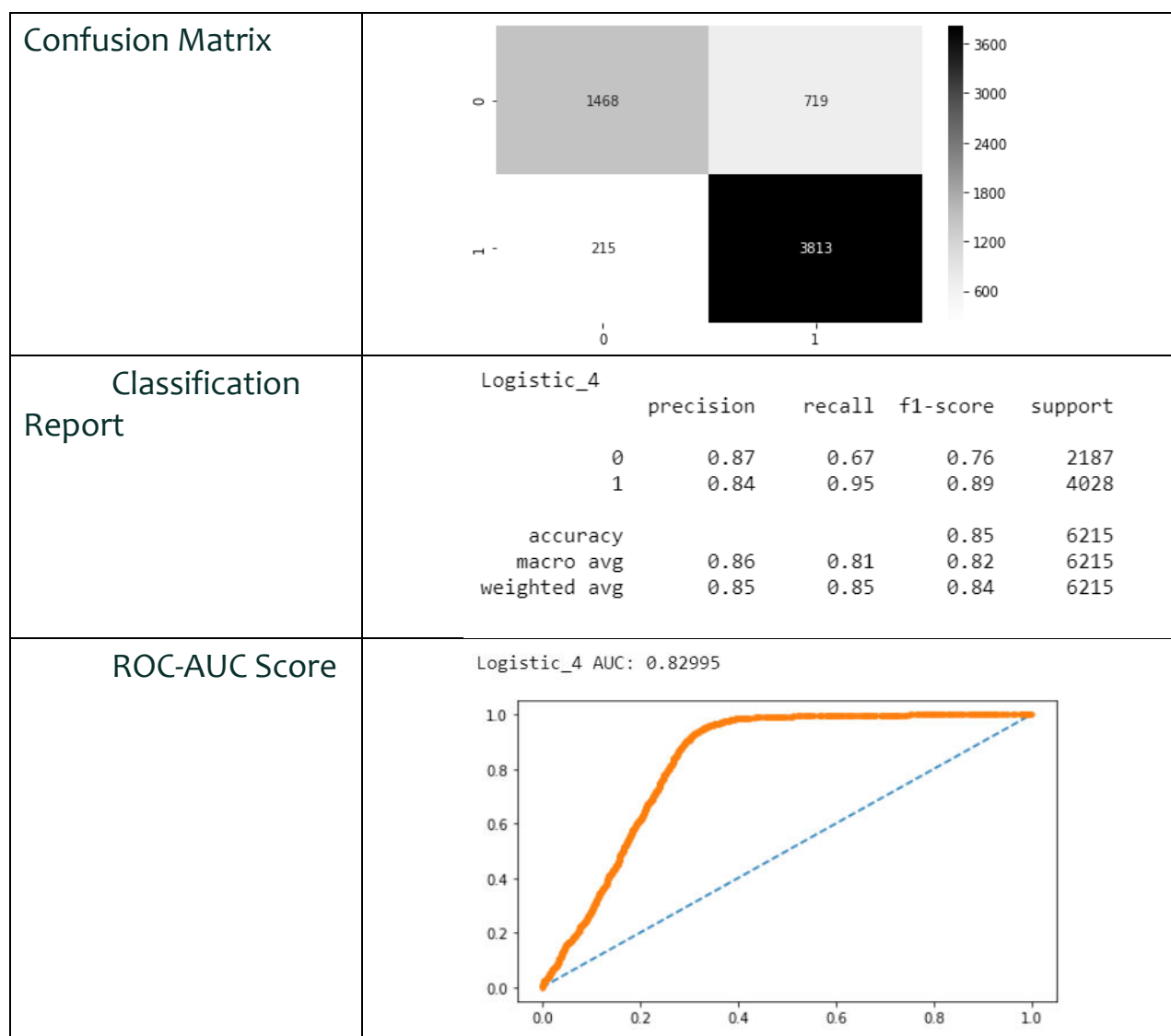
4.4 Model 4 :

As discussed in the previous iteration, we will drop 'Region_Western' along with 'School_Type_B','School_Type_D' and proceed with model building and check the results as follows:

Logit Regression Results

Dep. Variable:	Scholarship	No. Observations:	6215
Model:	Logit	Df Residuals:	6204
Method:	MLE	Df Model:	10
Date:	Tue, 02 Feb 2021	Pseudo R-squ.:	0.3475
Time:	23:49:16	Log-Likelihood:	-2630.3
converged:	True	LL-Null:	-4031.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.7251	0.035	20.589	0.000	0.656	0.794
Academic_Score	-0.8052	0.047	-17.090	0.000	-0.898	-0.713
Score_on_Plays_Made	-0.9700	0.046	-21.012	0.000	-1.060	-0.880
Missed_Play_Score	0.1963	0.046	4.307	0.000	0.107	0.286
School_Score	-0.5671	0.040	-14.183	0.000	-0.645	-0.489
Overall_Score	0.0753	0.040	1.868	0.062	-0.004	0.154
Injury_Propensity_Low	-0.4828	0.055	-8.738	0.000	-0.591	-0.374
Injury_Propensity_Moderate	-0.0267	0.050	-0.532	0.595	-0.125	0.072
Injury_Propensity_Normal	-0.2498	0.051	-4.863	0.000	-0.350	-0.149
Region_Southern	0.2070	0.036	5.755	0.000	0.136	0.277
School_Type_C	-0.1826	0.039	-4.697	0.000	-0.259	-0.106



VIF is also check for all the above variables used to build this model and their values are as follows:

```
Academic_Score VIF = 1.45
Score_on_Plays_Made VIF = 1.46
Missed_Play_Score VIF = 1.53
School_Score VIF = 1.2
Overall_Score VIF = 1.37
Injury_Propensity_Low VIF = 2.4
Injury_Propensity_Moderate VIF = 1.84
Injury_Propensity_Normal VIF = 2.1
School_Type_C VIF = 1.23
Region_Southern VIF = 1.09
```

We discussed in previous iteration we get relatively good VIF values for all the variables used, which indicates there is no more multicollinearity in the data, so now we will drop the variable on the basis of its significance(p-value). Now we select 'Injury_Propensity_Moderate' to be dropped since it has highest p-value and then build the next iteration of model.

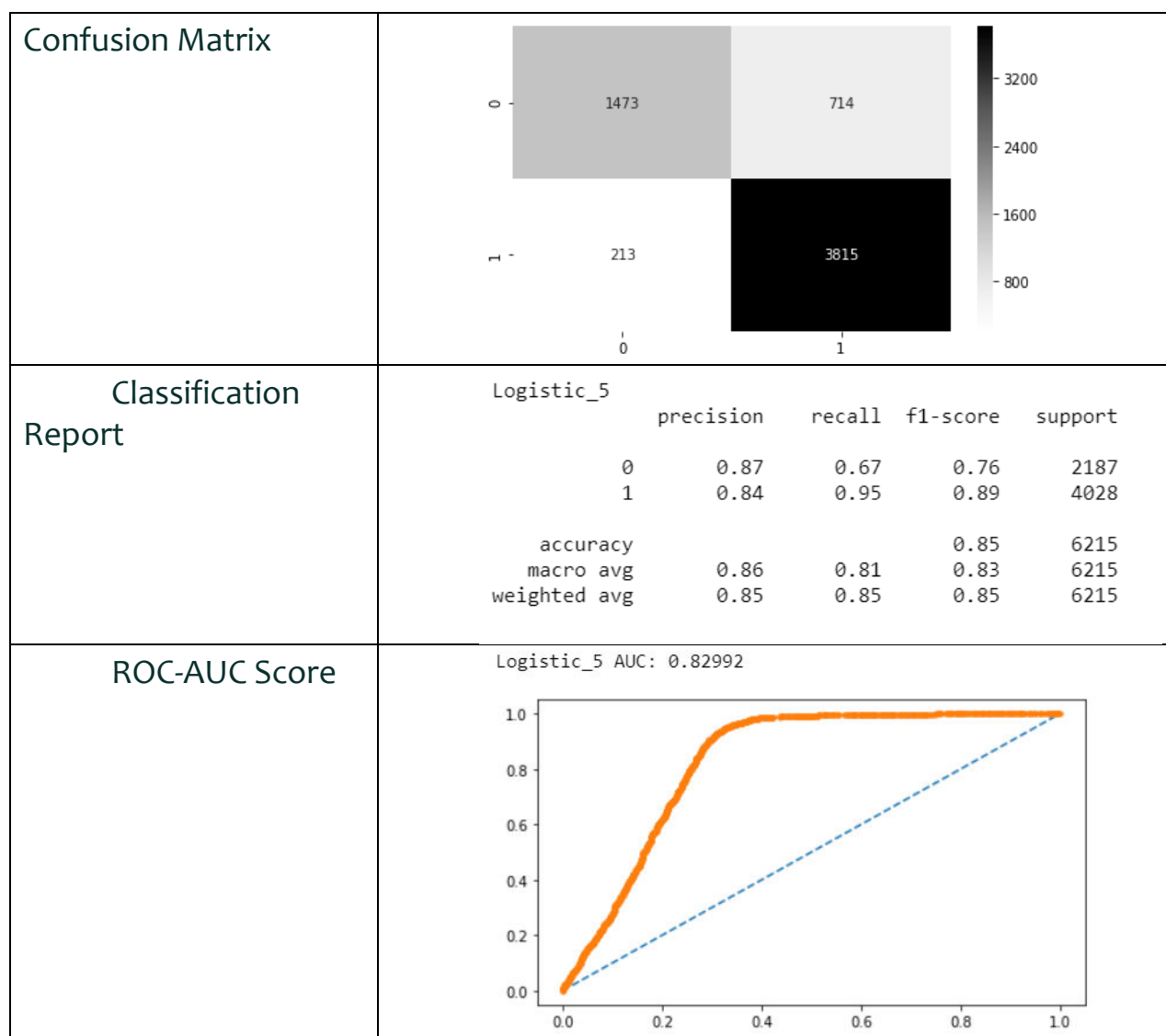
4.5 Model 5 :

Now we will drop 'Injury_Propensity_Moderate' along with 'Region_Western', 'School_Type_B', 'School_Type_D' and proceed with model building and check the results as follows:

Logit Regression Results

Dep. Variable:	Scholarship	No. Observations:	6215
Model:	Logit	Df Residuals:	6205
Method:	MLE	Df Model:	9
Date:	Tue, 02 Feb 2021	Pseudo R-squ.:	0.3475
Time:	23:49:18	Log-Likelihood:	-2630.4
converged:	True	LL-Null:	-4031.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.7253	0.035	20.600	0.000	0.656	0.794
Academic_Score	-0.8055	0.047	-17.102	0.000	-0.898	-0.713
Score_on_Plays_Made	-0.9702	0.046	-21.021	0.000	-1.061	-0.880
Missed_Play_Score	0.1970	0.046	4.325	0.000	0.108	0.286
School_Score	-0.5662	0.040	-14.177	0.000	-0.644	-0.488
Overall_Score	0.0699	0.039	1.794	0.073	-0.006	0.146
Injury_Propensity_Low	-0.4642	0.043	-10.885	0.000	-0.548	-0.381
Injury_Propensity_Normal	-0.2342	0.042	-5.569	0.000	-0.317	-0.152
Region_Southern	0.2071	0.036	5.760	0.000	0.137	0.278
School_Type_C	-0.1887	0.037	-5.084	0.000	-0.261	-0.116



VIF is also check for all the above variables used to build this model and their values are as follows:

```

Academic_Score VIF = 1.45
Score_on_Plays_Made VIF = 1.46
Missed_Play_Score VIF = 1.53
School_Score VIF = 1.2
Overall_Score VIF = 1.27
Injury_Propensity_Low VIF = 1.55
Injury_Propensity_Normal VIF = 1.51
School_Type_C VIF = 1.1
Region_Southern VIF = 1.09
          
```

We get good VIF values for all the variables used, which indicates there is no more multicollinearity in the data, so now we will drop the variable on the basis of its significance(p-value). Now we select 'Overall_Score' to be dropped since it has highest p-value and then build the next iteration of model.

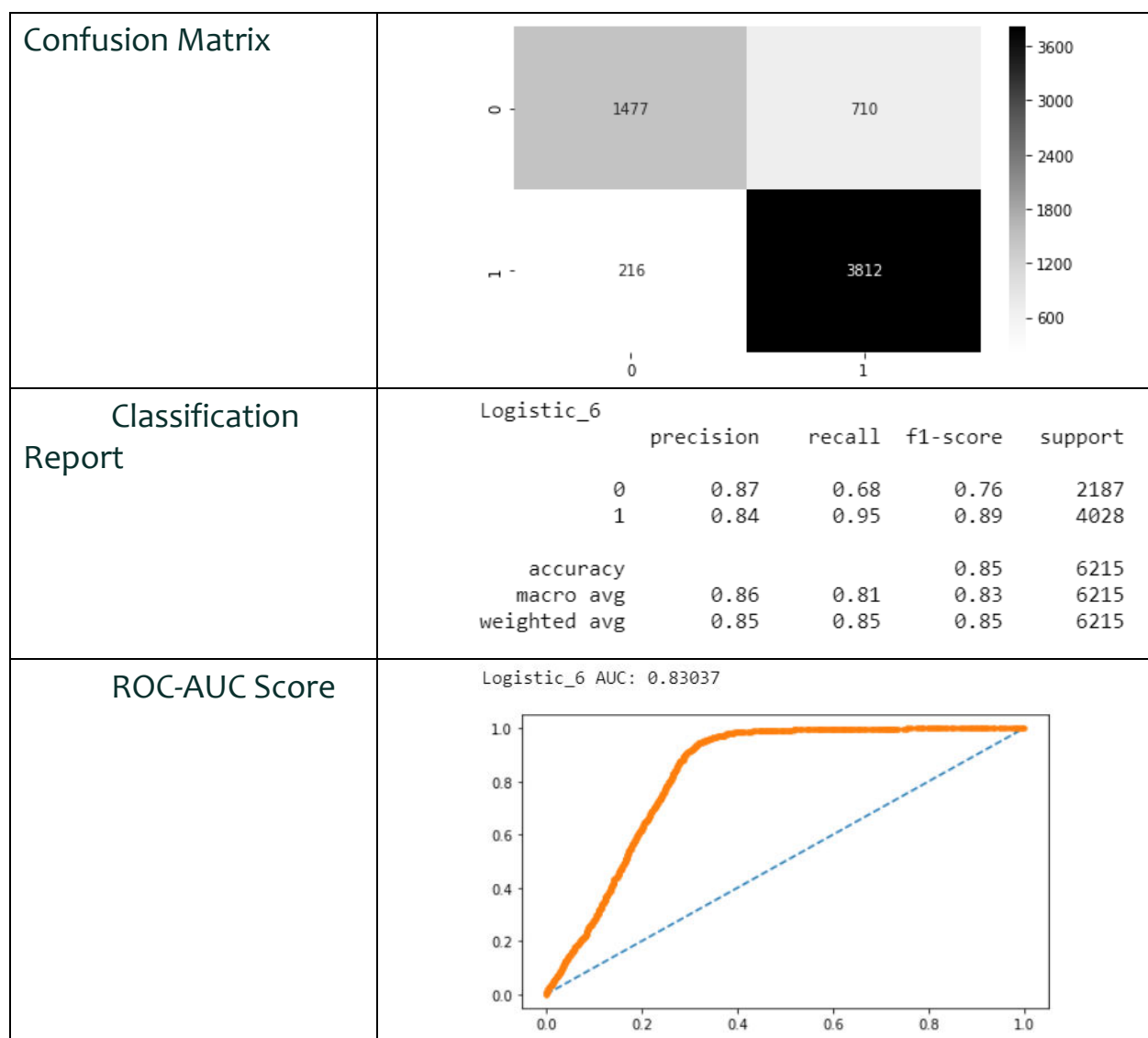
4.6 Model 6 :

We select 'Overall_Score' to drop along with 'Injury_Propensity_Moderate', 'Region_Western', 'School_Type_B', 'School_Type_D' and proceed with model building and check the results as follows:

Logit Regression Results

Dep. Variable:	Scholarship	No. Observations:	6215
Model:	Logit	Df Residuals:	6206
Method:	MLE	Df Model:	8
Date:	Tue, 02 Feb 2021	Pseudo R-squ.:	0.3471
Time:	23:49:19	Log-Likelihood:	-2632.0
converged:	True	LL-Null:	-4031.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.7293	0.035	20.763	0.000	0.660	0.798
Academic_Score	-0.8160	0.047	-17.460	0.000	-0.908	-0.724
Score_on_Plays_Made	-0.9710	0.046	-21.057	0.000	-1.061	-0.881
Missed_Play_Score	0.2005	0.046	4.393	0.000	0.111	0.290
School_Score	-0.5652	0.040	-14.173	0.000	-0.643	-0.487
Injury_Propensity_Low	-0.4437	0.041	-10.806	0.000	-0.524	-0.363
Injury_Propensity_Normal	-0.2132	0.040	-5.280	0.000	-0.292	-0.134
Region_Southern	0.1973	0.036	5.533	0.000	0.127	0.267
School_Type_C	-0.2066	0.036	-5.776	0.000	-0.277	-0.137



VIF is also check for all the above variables used to build this model and their values are as follows:

```
Academic_Score VIF = 1.42
Score_on_Plays_Made VIF = 1.46
Missed_Play_Score VIF = 1.53
School_Score VIF = 1.2
Injury_Propensity_Low VIF = 1.41
Injury_Propensity_Normal VIF = 1.38
School_Type_C VIF = 1.06
Region_Southern VIF = 1.05
```

With this, we have reached to the final iteration of model where get good VIF values for all the variables used, which indicates there is no more multicollinearity in the data, and also we have all the variables to be significant in this model.

4.7 Model Comparison :

So, to summarize we have built 6 iterations of logistic model with the complete data set dropping one variable at each iteration to reach to the final model. We also checked different evaluation parameters which will now be compared as follows:

Models	Variable dropped	Accuracy	Recall (partial=1)	Recall (Full=0)	ROC-AUC
Iteration 1	None	0.86	0.95	0.69	0.833
Iteration 2	School_Type_D	0.86	0.95	0.69	0.832
Iteration 3	School_Type_B	0.85	0.95	0.67	0.829
Iteration 4	Region_Western	0.85	0.95	0.67	0.829
Iteration 5	Injury_Propensity_Moderate	0.85	0.95	0.67	0.829
Iteration 6	Overall_Score	0.85	0.95	0.68	0.830

It is observed that there is not much difference in the scores amongst all the models. But according to the problem statement the class of importance for us is 'Full', so we will check the recall value of 'Full' first.

According to the table, Iteration 1 and 2 have the best recall value of 'Full', but since in those 2 models we see high VIF of variables, those models cannot be considered free of multicollinearity.

The **best model is then Iteration 6** since it is free of multicollinearity and using only the significant variables. The accuracy is also decent plus the ROC-AUC score is better than Iteration 3,4 and 5. Hence, the Model 6 is the optimal model.

5. Logistic regression with train-test data

Now for this question we have split the data set into training set(70%) and testing set (30%). We have followed the same iterations built in previous question but this time with less data (70% - training) and did prediction for both test and train data set to check the following evaluation parameters.

For the iterations we have performed with train-test data set, the order in which variables are dropped is same as of previous question since they show the similar order of importance.

The evaluation parameters for all the 6 models are checked as follows.

5.1 Model 1

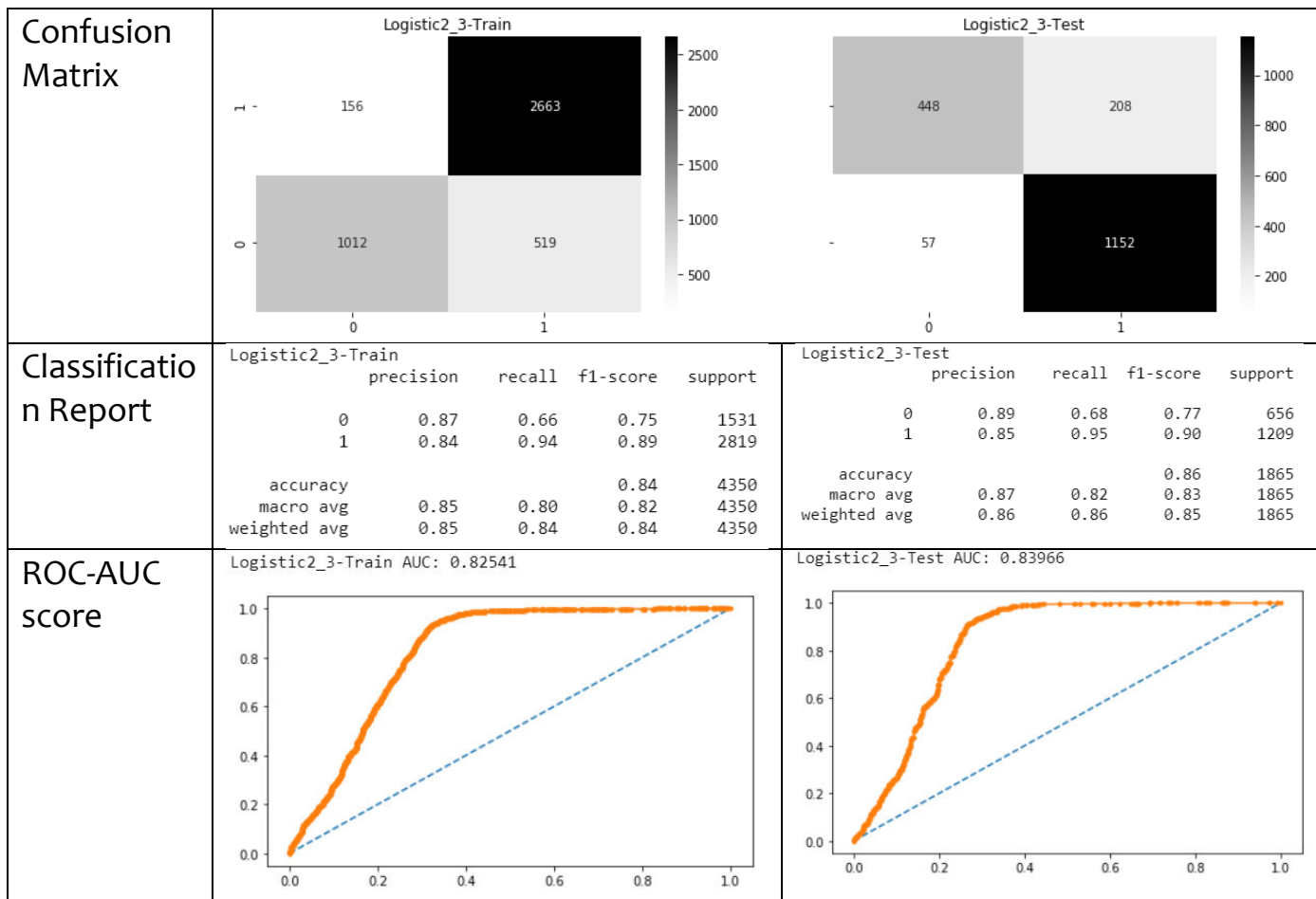
Metrics	Train Data	Test Data																																																												
Confusion Matrix	<div><p>Logistic2_1-Train</p><table border="1"><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>1044</td><td>142</td></tr><tr><th>1</th><td>487</td><td>2677</td></tr></tbody></table></div>		0	1	0	1044	142	1	487	2677	<div><p>Logistic2_1-Test</p><table border="1"><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>460</td><td>48</td></tr><tr><th>1</th><td>196</td><td>1161</td></tr></tbody></table></div>		0	1	0	460	48	1	196	1161																																										
	0	1																																																												
0	1044	142																																																												
1	487	2677																																																												
	0	1																																																												
0	460	48																																																												
1	196	1161																																																												
Classification Report	<div><p>Logistic2_1-Train</p><table><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr><tr><td>0</td><td>0.88</td><td>0.68</td><td>0.77</td><td>1531</td></tr><tr><td>1</td><td>0.85</td><td>0.95</td><td>0.89</td><td>2819</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>4350</td></tr><tr><td>macro avg</td><td>0.86</td><td>0.82</td><td>0.83</td><td>4350</td></tr><tr><td>weighted avg</td><td>0.86</td><td>0.86</td><td>0.85</td><td>4350</td></tr></table></div>		precision	recall	f1-score	support	0	0.88	0.68	0.77	1531	1	0.85	0.95	0.89	2819	accuracy			0.86	4350	macro avg	0.86	0.82	0.83	4350	weighted avg	0.86	0.86	0.85	4350	<div><p>Logistic2_1-Test</p><table><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr><tr><td>0</td><td>0.91</td><td>0.70</td><td>0.79</td><td>656</td></tr><tr><td>1</td><td>0.86</td><td>0.96</td><td>0.90</td><td>1209</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>1865</td></tr><tr><td>macro avg</td><td>0.88</td><td>0.83</td><td>0.85</td><td>1865</td></tr><tr><td>weighted avg</td><td>0.87</td><td>0.87</td><td>0.86</td><td>1865</td></tr></table></div>		precision	recall	f1-score	support	0	0.91	0.70	0.79	656	1	0.86	0.96	0.90	1209	accuracy			0.87	1865	macro avg	0.88	0.83	0.85	1865	weighted avg	0.87	0.87	0.86	1865
	precision	recall	f1-score	support																																																										
0	0.88	0.68	0.77	1531																																																										
1	0.85	0.95	0.89	2819																																																										
accuracy			0.86	4350																																																										
macro avg	0.86	0.82	0.83	4350																																																										
weighted avg	0.86	0.86	0.85	4350																																																										
	precision	recall	f1-score	support																																																										
0	0.91	0.70	0.79	656																																																										
1	0.86	0.96	0.90	1209																																																										
accuracy			0.87	1865																																																										
macro avg	0.88	0.83	0.85	1865																																																										
weighted avg	0.87	0.87	0.86	1865																																																										
ROC-AUC score	<div><p>Logistic2_1-Train AUC: 0.82775</p><table border="1"><thead><tr><th>fpr</th><th>tpr</th></tr></thead><tbody><tr><td>0.0</td><td>0.0</td></tr><tr><td>0.1</td><td>0.3</td></tr><tr><td>0.2</td><td>0.6</td></tr><tr><td>0.3</td><td>0.9</td></tr><tr><td>0.4</td><td>0.98</td></tr><tr><td>0.5</td><td>1.0</td></tr><tr><td>1.0</td><td>1.0</td></tr></tbody></table></div>	fpr	tpr	0.0	0.0	0.1	0.3	0.2	0.6	0.3	0.9	0.4	0.98	0.5	1.0	1.0	1.0	<div><p>Logistic2_1-Test AUC: 0.84612</p><table border="1"><thead><tr><th>fpr</th><th>tpr</th></tr></thead><tbody><tr><td>0.0</td><td>0.0</td></tr><tr><td>0.1</td><td>0.3</td></tr><tr><td>0.2</td><td>0.6</td></tr><tr><td>0.3</td><td>0.9</td></tr><tr><td>0.4</td><td>0.98</td></tr><tr><td>0.5</td><td>1.0</td></tr><tr><td>1.0</td><td>1.0</td></tr></tbody></table></div>	fpr	tpr	0.0	0.0	0.1	0.3	0.2	0.6	0.3	0.9	0.4	0.98	0.5	1.0	1.0	1.0																												
fpr	tpr																																																													
0.0	0.0																																																													
0.1	0.3																																																													
0.2	0.6																																																													
0.3	0.9																																																													
0.4	0.98																																																													
0.5	1.0																																																													
1.0	1.0																																																													
fpr	tpr																																																													
0.0	0.0																																																													
0.1	0.3																																																													
0.2	0.6																																																													
0.3	0.9																																																													
0.4	0.98																																																													
0.5	1.0																																																													
1.0	1.0																																																													

5.2 Model 2

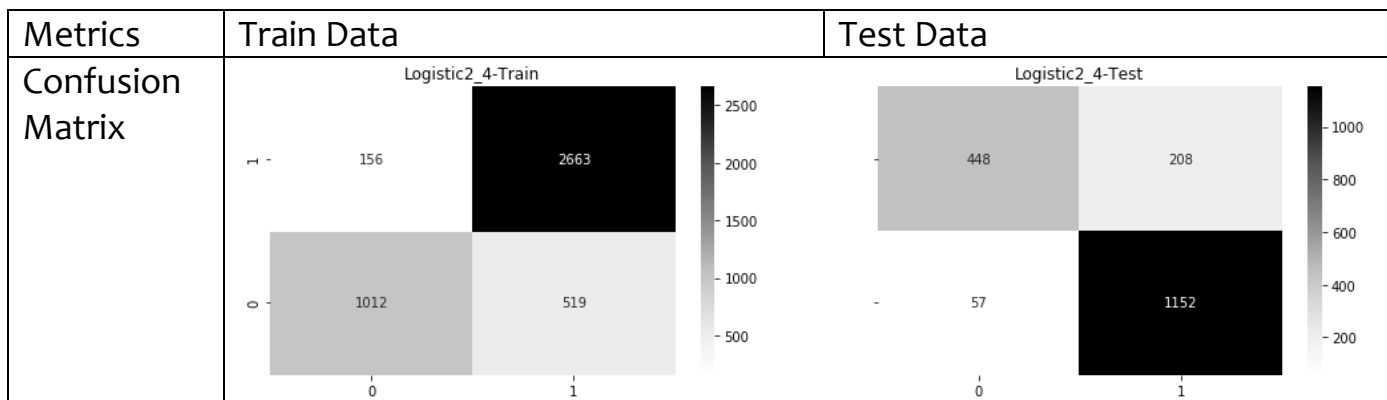
Metrics	Train Data	Test Data																																																												
Confusion Matrix	<div><p>Logistic2_2-Train</p><table><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>1044</td><td>487</td></tr><tr><th>1</th><td>140</td><td>2679</td></tr></tbody></table></div>		0	1	0	1044	487	1	140	2679	<div><p>Logistic2_2-Test</p><table><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>460</td><td>196</td></tr><tr><th>1</th><td>48</td><td>1161</td></tr></tbody></table></div>		0	1	0	460	196	1	48	1161																																										
	0	1																																																												
0	1044	487																																																												
1	140	2679																																																												
	0	1																																																												
0	460	196																																																												
1	48	1161																																																												
Classification Report	<div><p>Logistic2_2-Train</p><table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.88</td><td>0.68</td><td>0.77</td><td>1531</td></tr><tr><td>1</td><td>0.85</td><td>0.95</td><td>0.90</td><td>2819</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>4350</td></tr><tr><td>macro avg</td><td>0.86</td><td>0.82</td><td>0.83</td><td>4350</td></tr><tr><td>weighted avg</td><td>0.86</td><td>0.86</td><td>0.85</td><td>4350</td></tr></tbody></table></div>		precision	recall	f1-score	support	0	0.88	0.68	0.77	1531	1	0.85	0.95	0.90	2819	accuracy			0.86	4350	macro avg	0.86	0.82	0.83	4350	weighted avg	0.86	0.86	0.85	4350	<div><p>Logistic2_2-Test</p><table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.91</td><td>0.70</td><td>0.79</td><td>656</td></tr><tr><td>1</td><td>0.86</td><td>0.96</td><td>0.90</td><td>1209</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>1865</td></tr><tr><td>macro avg</td><td>0.88</td><td>0.83</td><td>0.85</td><td>1865</td></tr><tr><td>weighted avg</td><td>0.87</td><td>0.87</td><td>0.86</td><td>1865</td></tr></tbody></table></div>		precision	recall	f1-score	support	0	0.91	0.70	0.79	656	1	0.86	0.96	0.90	1209	accuracy			0.87	1865	macro avg	0.88	0.83	0.85	1865	weighted avg	0.87	0.87	0.86	1865
	precision	recall	f1-score	support																																																										
0	0.88	0.68	0.77	1531																																																										
1	0.85	0.95	0.90	2819																																																										
accuracy			0.86	4350																																																										
macro avg	0.86	0.82	0.83	4350																																																										
weighted avg	0.86	0.86	0.85	4350																																																										
	precision	recall	f1-score	support																																																										
0	0.91	0.70	0.79	656																																																										
1	0.86	0.96	0.90	1209																																																										
accuracy			0.87	1865																																																										
macro avg	0.88	0.83	0.85	1865																																																										
weighted avg	0.87	0.87	0.86	1865																																																										
ROC-AUC score	<div><p>Logistic2_2-Train AUC: 0.82706</p></div>	<div><p>Logistic2_2-Test AUC: 0.84612</p></div>																																																												

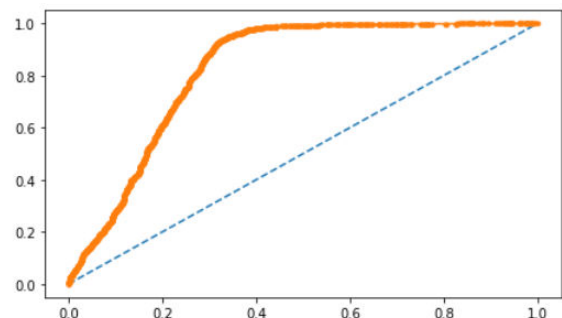
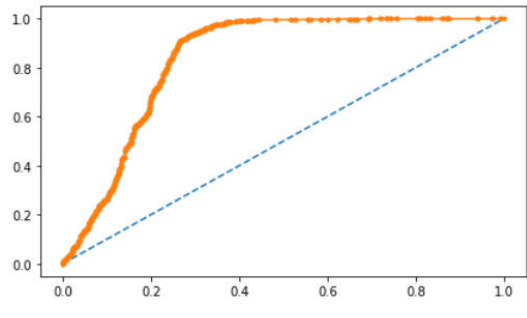
5.3 Model 3

Metrics	Train Data	Test Data
---------	------------	-----------

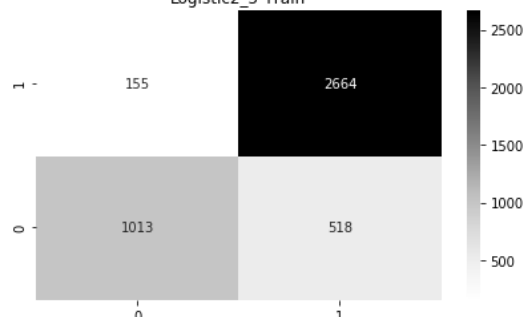
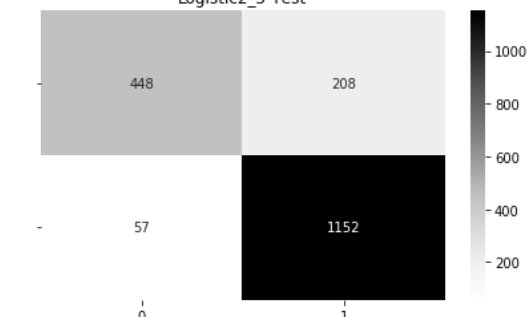


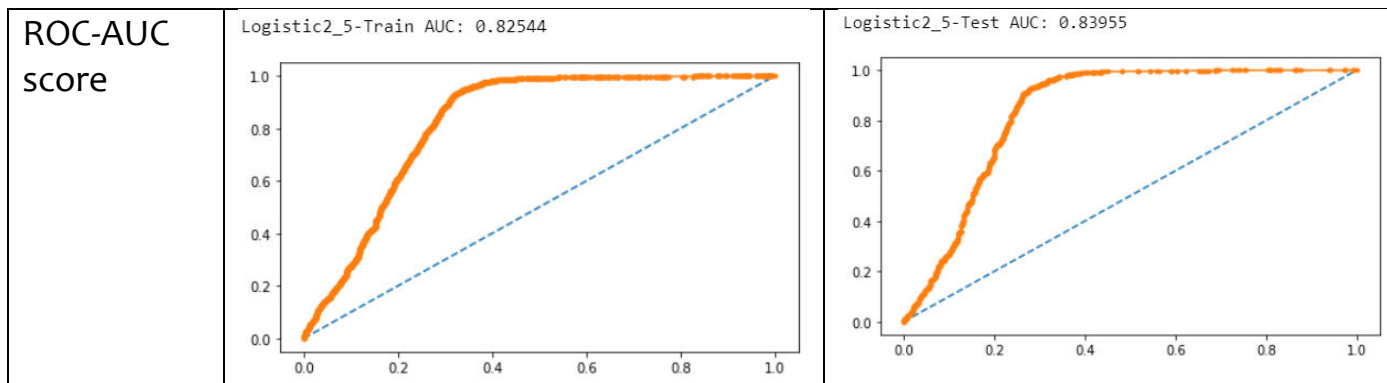
5.4 Model 4



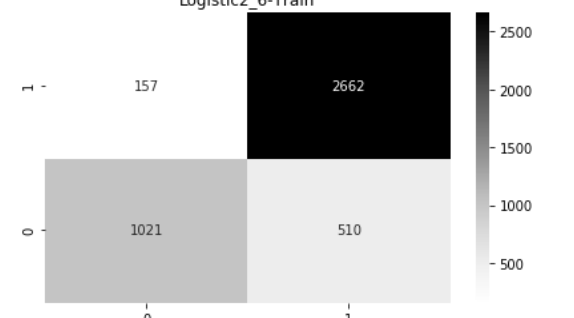
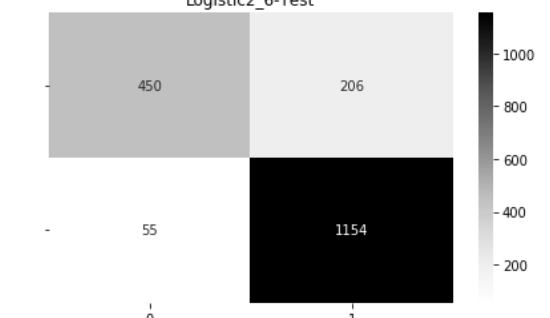
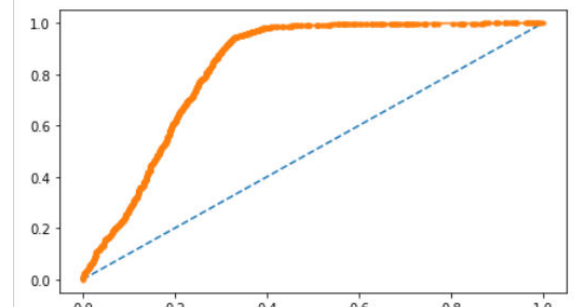
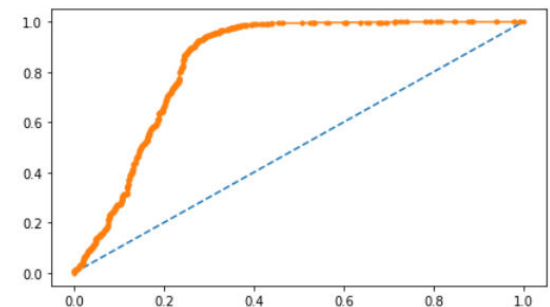
Classification Report	Logistic2_4-Train <pre> precision recall f1-score support 0 0.87 0.66 0.75 1531 1 0.84 0.94 0.89 2819 accuracy 0.84 4350 macro avg 0.85 0.80 0.82 4350 weighted avg 0.85 0.84 0.84 4350 </pre>	Logistic2_4-Test <pre> precision recall f1-score support 0 0.89 0.68 0.77 656 1 0.85 0.95 0.90 1209 accuracy 0.86 1865 macro avg 0.87 0.82 0.83 1865 weighted avg 0.86 0.86 0.85 1865 </pre>
	ROC-AUC score Logistic2_4-Train AUC: 0.82540 	ROC-AUC score Logistic2_4-Test AUC: 0.83967 

5.5 Model 5

Metrics	Train Data	Test Data
Confusion Matrix	Logistic2_5-Train 	Logistic2_5-Test 
Classification Report	Logistic2_5-Train <pre> precision recall f1-score support 0 0.87 0.66 0.75 1531 1 0.84 0.95 0.89 2819 accuracy 0.85 4350 macro avg 0.85 0.80 0.82 4350 weighted avg 0.85 0.84 0.84 4350 </pre>	Logistic2_5-Test <pre> precision recall f1-score support 0 0.89 0.68 0.77 656 1 0.85 0.95 0.90 1209 accuracy 0.86 1865 macro avg 0.87 0.82 0.83 1865 weighted avg 0.86 0.86 0.85 1865 </pre>



5.6 Model 6

Metrics	Train Data	Test Data																																																												
Confusion Matrix	<div><p>Logistic2_6-Train</p></div>	<div><p>Logistic2_6-Test</p></div>																																																												
Classification Report	<div><p>Logistic2_6-Train</p><table><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr><tr><td>0</td><td>0.87</td><td>0.67</td><td>0.75</td><td>1531</td></tr><tr><td>1</td><td>0.84</td><td>0.94</td><td>0.89</td><td>2819</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.85</td><td>4350</td></tr><tr><td>macro avg</td><td>0.85</td><td>0.81</td><td>0.82</td><td>4350</td></tr><tr><td>weighted avg</td><td>0.85</td><td>0.85</td><td>0.84</td><td>4350</td></tr></table></div>		precision	recall	f1-score	support	0	0.87	0.67	0.75	1531	1	0.84	0.94	0.89	2819	accuracy			0.85	4350	macro avg	0.85	0.81	0.82	4350	weighted avg	0.85	0.85	0.84	4350	<div><p>Logistic2_6-Test</p><table><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr><tr><td>0</td><td>0.89</td><td>0.69</td><td>0.78</td><td>656</td></tr><tr><td>1</td><td>0.85</td><td>0.95</td><td>0.90</td><td>1209</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>1865</td></tr><tr><td>macro avg</td><td>0.87</td><td>0.82</td><td>0.84</td><td>1865</td></tr><tr><td>weighted avg</td><td>0.86</td><td>0.86</td><td>0.86</td><td>1865</td></tr></table></div>		precision	recall	f1-score	support	0	0.89	0.69	0.78	656	1	0.85	0.95	0.90	1209	accuracy			0.86	1865	macro avg	0.87	0.82	0.84	1865	weighted avg	0.86	0.86	0.86	1865
	precision	recall	f1-score	support																																																										
0	0.87	0.67	0.75	1531																																																										
1	0.84	0.94	0.89	2819																																																										
accuracy			0.85	4350																																																										
macro avg	0.85	0.81	0.82	4350																																																										
weighted avg	0.85	0.85	0.84	4350																																																										
	precision	recall	f1-score	support																																																										
0	0.89	0.69	0.78	656																																																										
1	0.85	0.95	0.90	1209																																																										
accuracy			0.86	1865																																																										
macro avg	0.87	0.82	0.84	1865																																																										
weighted avg	0.86	0.86	0.86	1865																																																										
ROC-AUC score	<div><p>Logistic2_6-Train AUC: 0.82587</p></div>	<div><p>Logistic2_6-Test AUC: 0.84042</p></div>																																																												

5.7 Model Comparison:

So, to summarize we have built 6 iterations of logistic model with the train-test data set dropping one variable at each iteration to reach to the final model. We also checked different evaluation parameters which will now be compared as follows:

Models	Variable dropped	Dataset used	Accuracy	Recall (Partial=1)	Recall (Full=0)	ROC-AUC
Iteration 1	None	Train	0.86	0.95	0.69	0.827
		Test	0.87	0.96	0.70	0.846
Iteration 2	School_Type_D	Train	0.86	0.95	0.68	0.827
		Test	0.87	0.96	0.70	0.846
Iteration 3	School_Type_B	Train	0.84	0.94	0.66	0.825
		Test	0.86	0.95	0.68	0.839
Iteration 4	Region_Western	Train	0.84	0.95	0.66	0.825
		Test	0.86	0.94	0.68	0.839
Iteration 5	Injury_Propensity_Moderate	Train	0.85	0.95	0.66	0.825
		Test	0.86	0.95	0.68	0.839
Iteration 6	Overall_Score	Train	0.85	0.94	0.67	0.825
		Test	0.86	0.95	0.69	0.840

Comparing all the above 6 iterations, we find **iteration 6 is the best** because it has all the significant variables with no multicollinearity in the model. Also, we observe in iteration 6 dropping 5 variables compared to iteration 1 doesn't create significant impact on the model scores.

We get almost same score with the difference of 0.01 in test and train accuracy when compared to model 1. Recall of class 'Full' is decent with the value of 0.67 for train and 0.69 for test. ROC-AUC score is also same as that of model 1 up to 2 decimals both for train and test.

5.8 Comparison of best models between full dataset and train – test dataset:

Metrics	Full data	Train-test data	
		Train data	Test data
Accuracy	0.85	0.85	0.86
Recall (Partial=1)	0.95	0.94	0.95

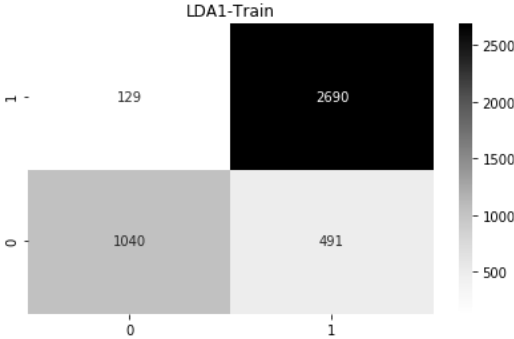
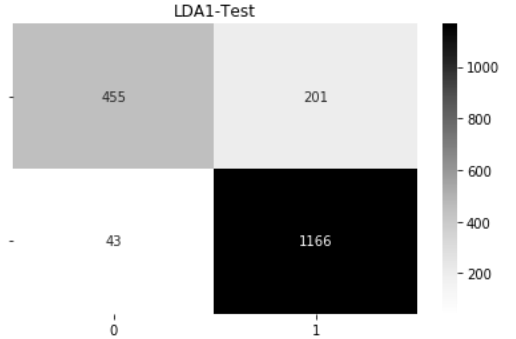
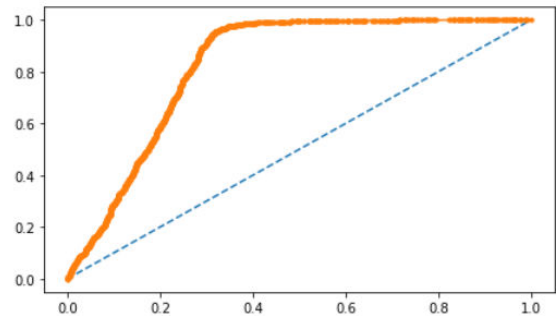
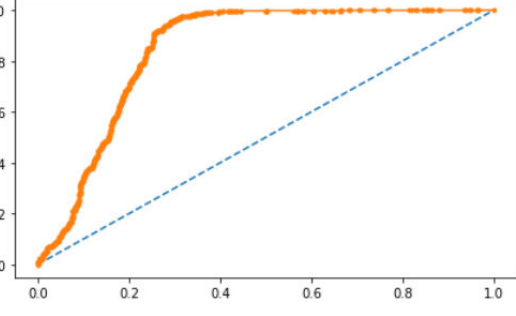
Recall (Full=0)	0.68	0.67	0.69
ROC-AUC	0.830	0.825	0.840

We do not see much of the difference between the 2 models may be because the data variation is same. There is only a difference of 0.01 amongst all the metrics which could be understood as the amount of data used is different in all models. Both the models provide prediction of similar accuracy.

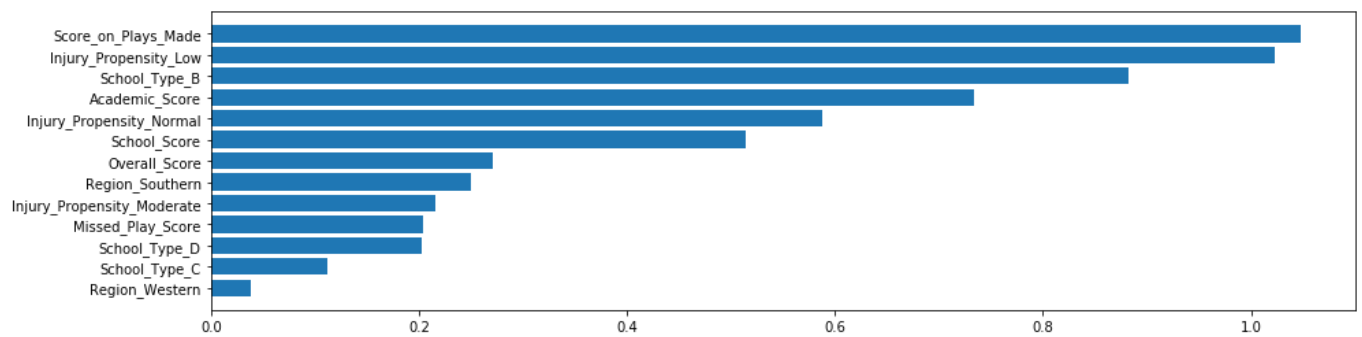
6. Linear Discriminant Analysis (LDA)

In this question, we used the same train-test split created in the previous question to build the multiple iterations of Linear Discriminant Analysis. Here, we have used `feature_importance()` to decide on which variable to be dropped in the next iteration. The evaluation metrics of all the models are checked as follows:

6.1 Model 1

Metrics	Train Data	Test Data																																																												
Confusion Matrix	<div>LDA1-Train</div>	<div>LDA1-Test</div>																																																												
Classification Report	<div>LDA1-Train<table><tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td></tr><tr><td>0</td><td>0.89</td><td>0.68</td><td>0.77</td><td>1531</td></tr><tr><td>1</td><td>0.85</td><td>0.95</td><td>0.90</td><td>2819</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>4350</td></tr><tr><td>macro avg</td><td>0.87</td><td>0.82</td><td>0.83</td><td>4350</td></tr><tr><td>weighted avg</td><td>0.86</td><td>0.86</td><td>0.85</td><td>4350</td></tr></table></div>		precision	recall	f1-score	support	0	0.89	0.68	0.77	1531	1	0.85	0.95	0.90	2819	accuracy			0.86	4350	macro avg	0.87	0.82	0.83	4350	weighted avg	0.86	0.86	0.85	4350	<div>LDA1-Test<table><tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td></tr><tr><td>0</td><td>0.91</td><td>0.69</td><td>0.79</td><td>656</td></tr><tr><td>1</td><td>0.85</td><td>0.96</td><td>0.91</td><td>1209</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>1865</td></tr><tr><td>macro avg</td><td>0.88</td><td>0.83</td><td>0.85</td><td>1865</td></tr><tr><td>weighted avg</td><td>0.87</td><td>0.87</td><td>0.86</td><td>1865</td></tr></table></div>		precision	recall	f1-score	support	0	0.91	0.69	0.79	656	1	0.85	0.96	0.91	1209	accuracy			0.87	1865	macro avg	0.88	0.83	0.85	1865	weighted avg	0.87	0.87	0.86	1865
	precision	recall	f1-score	support																																																										
0	0.89	0.68	0.77	1531																																																										
1	0.85	0.95	0.90	2819																																																										
accuracy			0.86	4350																																																										
macro avg	0.87	0.82	0.83	4350																																																										
weighted avg	0.86	0.86	0.85	4350																																																										
	precision	recall	f1-score	support																																																										
0	0.91	0.69	0.79	656																																																										
1	0.85	0.96	0.91	1209																																																										
accuracy			0.87	1865																																																										
macro avg	0.88	0.83	0.85	1865																																																										
weighted avg	0.87	0.87	0.86	1865																																																										
ROC-AUC score	<div>LDA1-Train AUC: 0.82738</div>	<div>LDA1-Test AUC: 0.84648</div>																																																												

This model is the first iteration which is built using all the variables and their importance in the model is observed as follows:

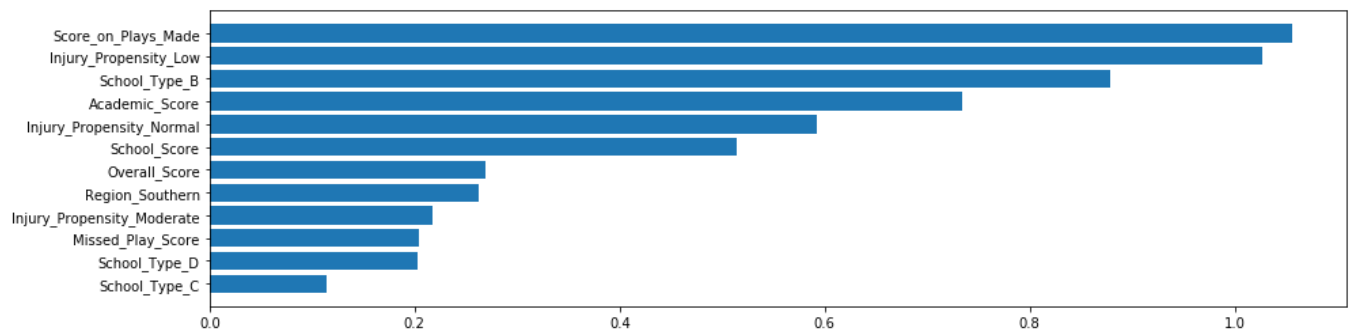


From the above figure it is observed that 'Region_Western' is the least important of all the variables, so it will be dropped for the next iteration.

6.2 Model 2

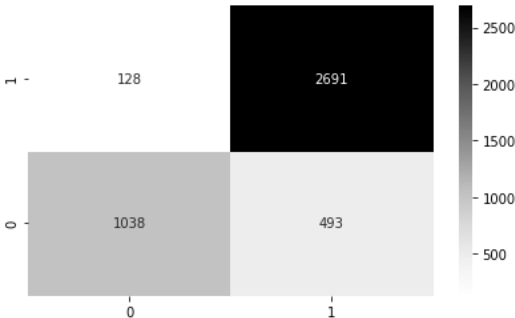
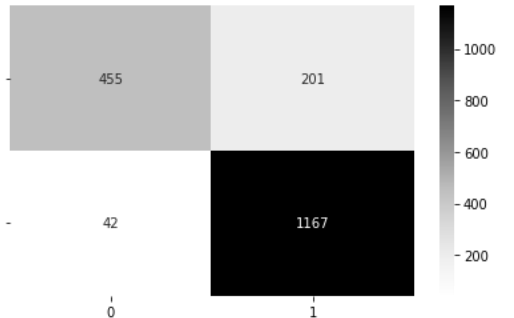
Metrics	Train Data	Test Data																																																												
Confusion Matrix	<div><p>LDA2-Train</p><table><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>1</th><td>129</td><td>2690</td></tr><tr><th>0</th><td>1039</td><td>492</td></tr></tbody></table></div>		0	1	1	129	2690	0	1039	492	<div><p>LDA2-Test</p><table><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>1</th><td>455</td><td>201</td></tr><tr><th>0</th><td>42</td><td>1167</td></tr></tbody></table></div>		0	1	1	455	201	0	42	1167																																										
	0	1																																																												
1	129	2690																																																												
0	1039	492																																																												
	0	1																																																												
1	455	201																																																												
0	42	1167																																																												
Classification Report	<div><p>LDA2-Train</p><table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.89</td><td>0.68</td><td>0.77</td><td>1531</td></tr><tr><td>1</td><td>0.85</td><td>0.95</td><td>0.90</td><td>2819</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>4350</td></tr><tr><td>macro avg</td><td>0.87</td><td>0.82</td><td>0.83</td><td>4350</td></tr><tr><td>weighted avg</td><td>0.86</td><td>0.86</td><td>0.85</td><td>4350</td></tr></tbody></table></div>		precision	recall	f1-score	support	0	0.89	0.68	0.77	1531	1	0.85	0.95	0.90	2819	accuracy			0.86	4350	macro avg	0.87	0.82	0.83	4350	weighted avg	0.86	0.86	0.85	4350	<div><p>LDA2-Test</p><table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.92</td><td>0.69</td><td>0.79</td><td>656</td></tr><tr><td>1</td><td>0.85</td><td>0.97</td><td>0.91</td><td>1209</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>1865</td></tr><tr><td>macro avg</td><td>0.88</td><td>0.83</td><td>0.85</td><td>1865</td></tr><tr><td>weighted avg</td><td>0.88</td><td>0.87</td><td>0.86</td><td>1865</td></tr></tbody></table></div>		precision	recall	f1-score	support	0	0.92	0.69	0.79	656	1	0.85	0.97	0.91	1209	accuracy			0.87	1865	macro avg	0.88	0.83	0.85	1865	weighted avg	0.88	0.87	0.86	1865
	precision	recall	f1-score	support																																																										
0	0.89	0.68	0.77	1531																																																										
1	0.85	0.95	0.90	2819																																																										
accuracy			0.86	4350																																																										
macro avg	0.87	0.82	0.83	4350																																																										
weighted avg	0.86	0.86	0.85	4350																																																										
	precision	recall	f1-score	support																																																										
0	0.92	0.69	0.79	656																																																										
1	0.85	0.97	0.91	1209																																																										
accuracy			0.87	1865																																																										
macro avg	0.88	0.83	0.85	1865																																																										
weighted avg	0.88	0.87	0.86	1865																																																										
ROC-AUC score	<div><p>LDA2-Train AUC: 0.82751</p><p>The ROC curve for LDA2-Train shows a strong performance, with the orange curve well above the diagonal blue line. The AUC is 0.82751.</p></div>	<div><p>LDA2-Test AUC: 0.84643</p><p>The ROC curve for LDA2-Test shows a strong performance, with the orange curve well above the diagonal blue line. The AUC is 0.84643.</p></div>																																																												

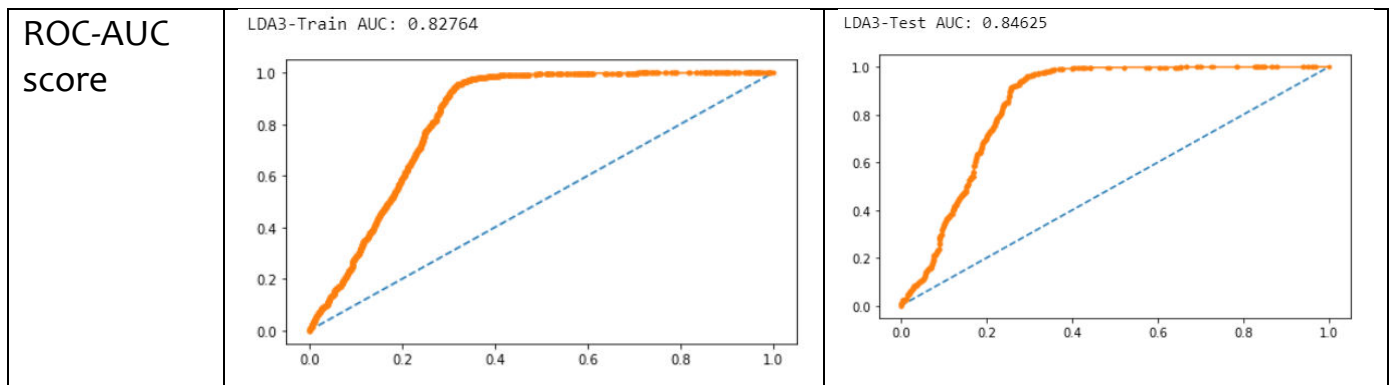
This model is the second iteration which is built by dropping variable 'Region_Western' and importance of all other variables in the model is observed as follows:



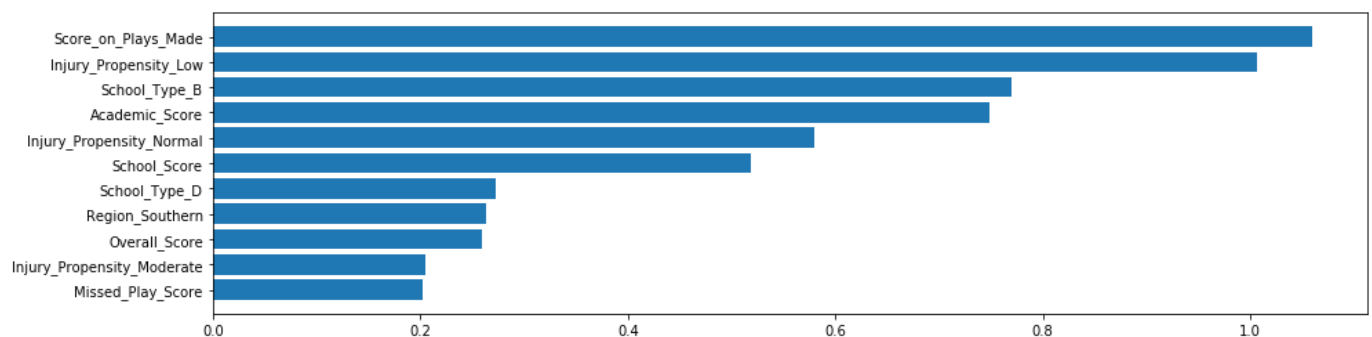
After 'Region_Western', 'School Type C' is the least important of all the variables, so it will be dropped for the next iteration.

6.3 Model 3

Metrics	Train Data					Test Data				
Confusion Matrix										
	LDA3-Train					LDA3-Test				
Classification Report	LDA3-Train					LDA3-Test				
		precision	recall	f1-score	support		precision	recall	f1-score	support
	0	0.89	0.68	0.77	1531	0	0.92	0.69	0.79	656
	1	0.85	0.95	0.90	2819	1	0.85	0.97	0.91	1209
	accuracy			0.86	4350	accuracy			0.87	1865
	macro avg	0.87	0.82	0.83	4350	macro avg	0.88	0.83	0.85	1865
	weighted avg	0.86	0.86	0.85	4350	weighted avg	0.88	0.87	0.86	1865



This model is the final iteration which is built by dropping variable 'Region_Western' and 'School_Type_C' then the importance of all other variables in the model is observed as follows:



Now, the remaining variables seems to be important enough to be kept in the model.

6.4 Model Comparison:

So, to summarize we have built 3 iterations of Linear Discriminant Analysis with the train-test data set dropping one variable at each iteration to reach to the final model. We also checked different evaluation parameters which will now be compared as follows:

Models	Variable dropped	Dataset used	Accuracy	Recall (Partial=1)	Recall (Full=0)	ROC-AUC
Iteration 1	None	Train	0.86	0.95	0.68	0.827
		Test	0.87	0.96	0.69	0.846
Iteration	Region_Western	Train	0.86	0.95	0.68	0.827

2		Test	0.87	0.97	0.69	0.846
Iteration	School_Type_C	Train	0.86	0.95	0.68	0.827
3		Test	0.87	0.97	0.69	0.846

Comparing all the above 3 iterations, we find **iteration 3 is the best** because it has all the significant variables in the model. Also, we observe in iteration 3 dropping 2 variables compared to iteration 1 doesn't create significant impact on the model scores.

We get same score in terms of test and train accuracy for all the models. Recall of class 'Full' is decent with the value of 0.68 for train and 0.69 for test. ROC-AUC score is also same as that of model 1 for both for train and test.

6.5 Comparison of best models from Logistic regression and LDA

Metrics	Logistic Regression		LDA	
	Train data	Test data	Train data	Test data
Accuracy	0.85	0.86	0.86	0.87
Recall (Partial=1)	0.94	0.95	0.95	0.97
Recall (Full=0)	0.67	0.69	0.68	0.69
ROC-AUC	0.825	0.840	0.827	0.846

From the above table it is observed that LDA is performing a bit better as both the train and test accuracy is increased by 0.01. Recall of 'Full' is same for the test data but slightly (by 0.01) higher for train data. There is greater increase in the ROC-AUC score for both the train and test data.

The almost similar scores could be observed by the 3rd iteration of Logistic Regression but the variables which are dropped are different along with their order of dropping.

In summary there is not significant difference between the results of both the models thus both can be recommended.

7. Conclusion

1. Based on the study of feature importance from Logistic regression and LDA, the 2 most important variables for model prediction are: 'Score_on_Plays_Made' and 'Injury_Propensity_Low'. Besides them, 'School_Type_B' and 'Academic_Score' also have potential contribution in prediction. Hence, from business perspective, **higher the score in these features for a student, higher is his / her chance to obtain Full Scholarship.** It is *recommended for an individual student to main these 2 scores high.*
2. The next important factor in winning the scholarship of any type is high 'School_Score'. Thus, schools are recommended to take actions for improving their score.
3. Next two important predictors are 'Injury_Propensity' and 'Missed_Play_Score' and the students are advised *to maintain both these as minimum as they would reduce the chances for acquiring a Full Scholarship.*
4. Overall, we may see that the chances of Full Scholarship are high with good health and good sports ship maintained by the student.