Predictive Modelling Group Assignment

**Sharmila Sengupta Chowdhry**

[Introduction 3](#_TOC_250026)

[Exploratory Data Analysis 4](#_TOC_250025)

[Univariate and Bivariate Analysis 7](#_TOC_250024)

[Univariate Analysis of Categorical Features with Countplot 10](#_TOC_250023)

[Multivariate Analysis for Continuous Variable with Pairplot 11](#_TOC_250022)

[Multivariate Bivariate Analysis for Continuous Variable with Correlation Matrix and Boxplot 14](#_TOC_250021)

[Bivariate Analysis for Categorical Variables with Countplot 16](#_TOC_250020)

[Categorical & Numerical Analysis with Boxplot 17](#_TOC_250019)

[Outliers Analysis and Treatment 26](#_TOC_250018)

[Logistic Regression 27](#_TOC_250017)

[Model Building 28](#_TOC_250016)

[With all the Features (Part I) 28](#_TOC_250015)

[Without Region\_Western Variable (Part II) 30](#_TOC_250014)

[Interpreting Coefficients with Odds Ratio 30](#_TOC_250013)

[Splitting Data into training (70%) and test (30%) 33](#_TOC_250012)

[Model Evaluation 35](#_TOC_250011)

[Confusion Matrix 35](#_TOC_250010)

[Classification report 35](#_TOC_250009)

[AUC ROC Curve for Logistic Regression 36](#_TOC_250008)

[Recommendation 37](#_TOC_250007)

[Linear Discriminant Analysis (LDA) 38](#_TOC_250006)

[Model Evaluation 39](#_TOC_250005)

[Confusion Matrix 39](#_TOC_250004)

[Classification Report for LDA 39](#_TOC_250003)

[AUC ROC Curves for LDA 40](#_TOC_250002)

[Recommendation 40](#_TOC_250001)

[Combined Matrix Results 41](#_TOC_250000)

**Problem 2:** *Logistic Regression and Linear Discriminant Analysis*

*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.*

# Introduction

Linear and logistic Regression techniques are widely used because of their simplicity and explainability.

Logistic regression is one of the most used statistical procedures in research, as an important procedure in predictive analytics. Unlike linear regression, logistic regression is appropriate for modeling a categorical variable.

To begin our analysis of the data, we would start with data cleaning and preparation and along the way we would be doing Exploratory Data Analysis (EDA). We start EDA by exploring the nature of all the variables, identify the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and how to treat them, and whether the variables have a symmetric or skewed distribution.

We also conduct both univariate and bivariate analyses and pre-processing of data to explore relationships between the predictors as well as the predictor and target variables.. For any regression problem (linear or logistic) the dependence of the response on the predictors needs to be thoroughly investigated.

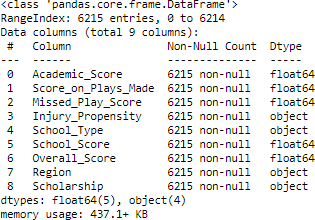
Q1. The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, especially identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing 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.

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.

**Exploratory Data Analysis**

We import all the necessary libraries. Then we read the dataset and conduct basic EDA.



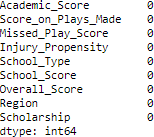
In total, we have 6,215 records across 9 columns in the dataset. The target / dependent / response variable is *Scholarship*, which is an object data type. The other columns are feature / independent / predictor variables.

Of the 8 predictor variables, 5 (*Academic\_Score, Score\_on\_Plays\_Made, Missed\_Play\_Score, School\_Score, Overall\_Score*) are scores of different types and are float datatype. The remaining, 3 (*Injury\_Propensity, School\_Type, Region*) are object datatypes, with 2 or more classes. We will analyse the classes during EDA of categorical variables below.

*Null value identification & treatment*

Missing value produce biased estimates that result in invalid conclusions, thereby compromising the predictive power of the study.

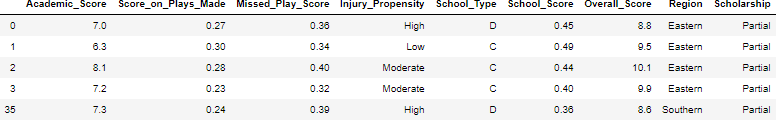
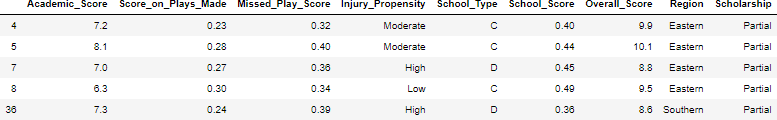
After checking for null values, we found that there are none



There are no null values in the dataset.

*Duplicates*

After checking the duplicates in the data we found that there are 947 duplicates in the data.



From the above two snippets we can see that row number 3 and 4 are duplicates. Similarly, row number 1 and 5 and 0 and 7 are duplicates.

Out of the total 6,215 records, there 947 duplicates which is approximately 0.13% of the total data. Students having the exact same scores (in 2 decimal points in some cases) coming from the same school do not seem like a possibility, therefore we decided to drop the duplicates, hence losing 0.13% of data from the dataset.

Table: Data Information

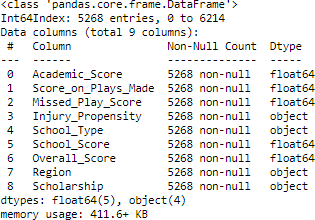
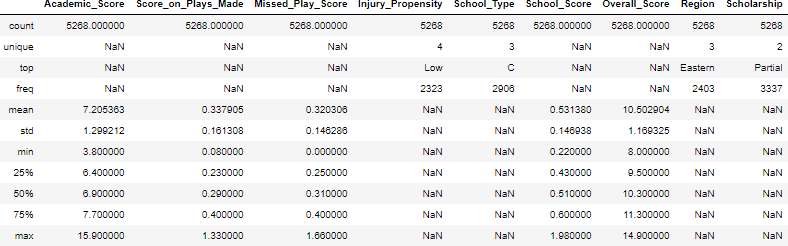


Table: Data Description



As per above, we can see that after dropping duplicates from the dataset we now have a total of 5,268 records. None of the continuous variables look to be normally distributed since their mean is not equal to median (50%).

**Univariate and Bivariate Analysis**

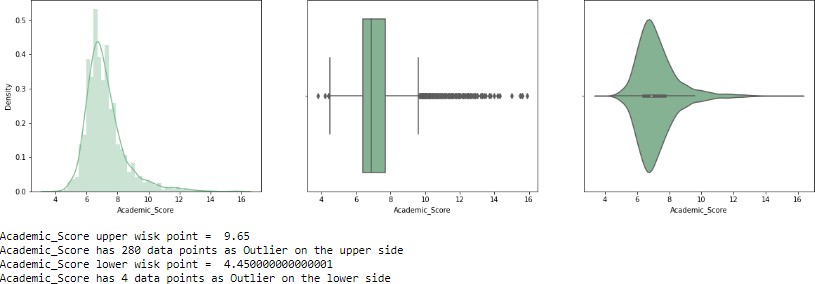
Univariate analysis, refers to the analysis of a single variable. The main purpose of univariate analysis is to summarize and find patterns in the data variable.

*Univariate Analysis for Continuous Variable*

We start univariate analysis with a set of three graphs, a KDE & histogram plot, a boxplot and a violin plot, all to check the distribution of data in the feature, as well as the presence and extent of outliers. Like missing data, outliers substantially affect the model, and thus must be dealt with carefully. We elaborate on this in an upcoming section.

We begin with analysing the continuous variables in the listed order: *Academic\_Score, Score- on\_Plays\_Made, Missed\_Play\_Score, School\_Score*, and *Overall\_Score*.

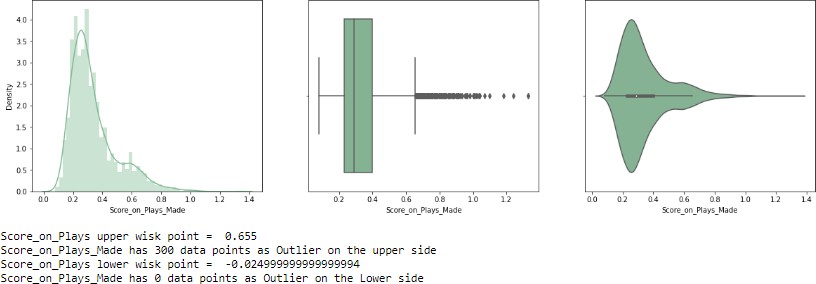
Multiple Charts: Academic\_Score



The univariate analysis of Academic\_Score reveals that most of the observations are in between 5 and around 9.5. There are more outliers on the upper side of distribution, which explains the right skew in the violin plot. We do not know whether these are valid outliers or invalid ones. The treatment of outliers is subjective

 Most of the students in the dataset are scoring between 4 and 10 in academics. There are some students who are scoring between 10 to 16, which is responsible for the skew.

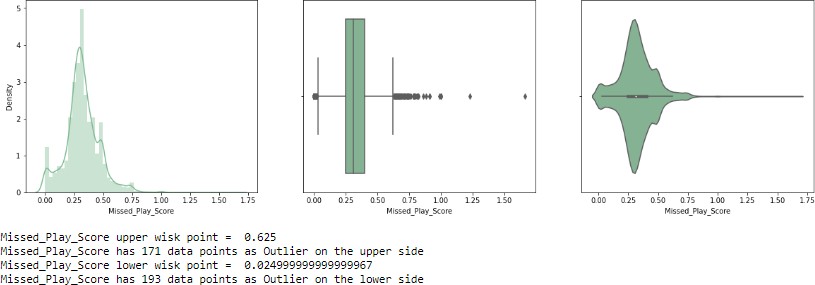
Multiple Charts: Score\_on\_Plays\_Made



Most of the observations of *Score\_on\_Plays* are in between 0.2 and 0.6, but due to the outliers on the higher end of the plot, it has a right skew.

 Most of the students in the dataset are scoring between 0.2 and 0.6 on the field, and a few of them scoring upto 1.4.

Multiple Charts: Missed\_Play\_Score



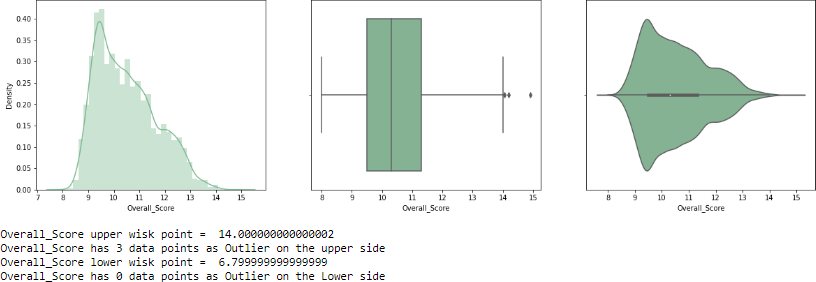
Observations of *Missed\_Play\_Score* are concentrated between 0 and 0.75, with few outliers on the lower side, and many on the upper side, upto 1.75.

*Missed\_Play\_Score* is a composite of misses on the field. As such, a higher number should be inversely proportional to the chance of Full Scholarship, which we can check below with the bivariate analysis (pairplot).

Most of the students in the dataset are scoring between 0.2 and 0.6 on the field, and a few of them are scoring upto 1.4.

*.*

Multiple Charts: Overall\_Score

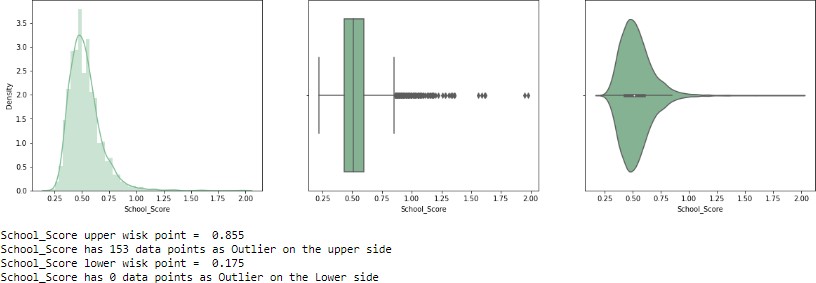


Observations of *Overall\_Score* are concentrated between 8 and 14, with few outliers on the upper side. Half of the students’ overall score is more than 10.25, 3/4ths or 75% students’ overall score is more than 9.5, and a similar number (75%) students have scores between 8 and 11.25 (approximately).

The distribution is not normal. The wider spread of the violin plot denotes higher probability. The highest probability of the data seems to be between 9.25-12.

 Most students have a lower *Overall\_Score* (between 9.25-12), than a higher one.

Multiple Charts: School\_Score



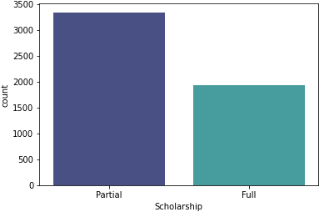
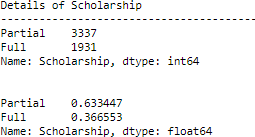
Observations of *School\_Score* are concentrated between 0.3 and 0.75, with outliers on the upper side. almost 3/4ths of the students’ School\_Scores’ are below 0.625. Outliers are between 1 to 2.

The distribution is not normal and has a right skew.

 Excluding the outliers, most students have a School\_Score of 0.25 to 0.87 approximately.

## Univariate Analysis of Categorical Features with Countplot

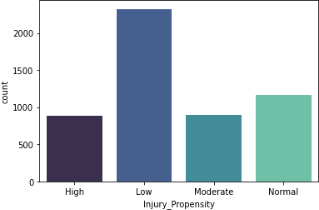
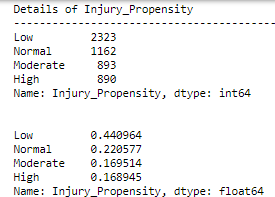
Table & Countplot: Scholarship



The Target Variable Scholarship comprises 63% (3337) students with Partial Scholarship while 37% (1931) students enjoy Full Scholarship. Although the data has an imbalance (⅓ for the class of interest and not ½), in real life datasets we always investigate and analyse the rarer instance, so this is quite expected. Also, the imbalance is not as pronounced, so we expect the model to be able to learn sufficiently from the available data.

In Python, categorical variables are analysed through a function called countplot, pulled from the seaborn library. We analyse the frequency of class distribution in each of the categorical features, Injury\_Propensity, School\_Type, Region and the Target Variable, Scholarship.

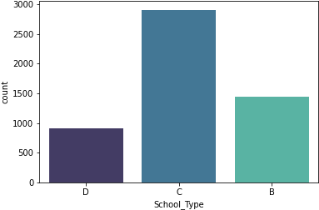
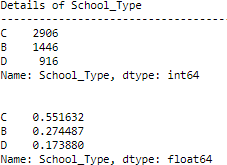
Table & Countplot: Injury\_Propensity



Injury Propensity constitutes of 44% (2323) Low injury, 22% (1162) Normal and 17% (893) for Moderate and 17% (890) High Injury. It seems logical that most students have low

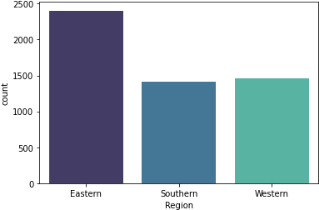
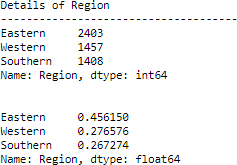
propensity for injury, considering that this is a group of student-athletes and they would have adequate training and fitness over the years.

Table & Countplot: School\_Type



School C constitutes 55% (2906) of the entire data,School B contributes 27% (1446) while School D contributes 18% (916) in the dataset. We have no qualitative information about the schools.

Table & Countplot: Region



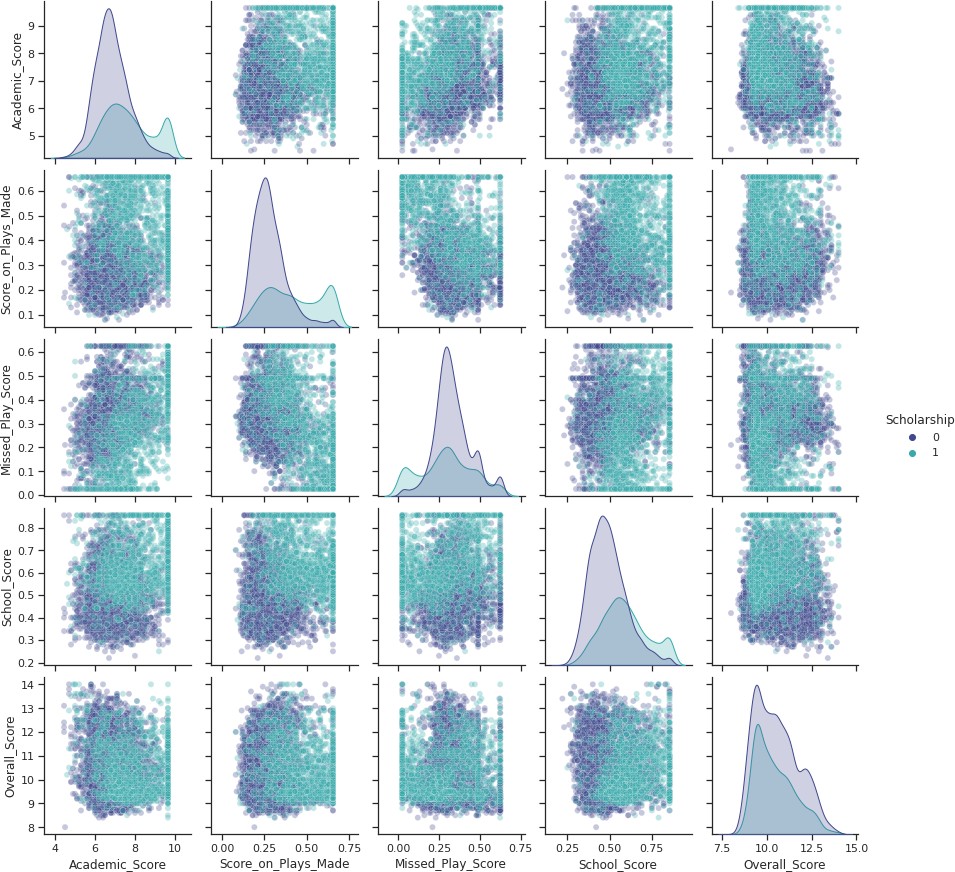
Eastern region account for 45% (2403) of the students in the dataset, whereas Western clocks 28% (1408 students) and 27% (1457) students belong to schools from the Southern region.

## Multivariate Analysis for Continuous Variable with Pairplot

We start the multivariate analysis with the pairplot.

A pairplot allows us to analyse both distribution of single variables and relationships between two variables. Pair plots are a great method to identify trends for further analysis.

Chart: Pairplot



There is a 60-80% overlap between the two classes of the Target Variable, Scholarship, and significant scatter is seen in the bivariate distribution across the dataset. The conclusion we can make is about how the data 'tends' to behave, and how it ‘looks like’ it will act, with little certainty. With that in mind, the following can be seen:

* There isn't much correlation between the independent variables. Multicollinearity does not look like a problem here.
* The KDE (across the diagonal) shows that none of the variables offer good separation of the classes, with *Overall\_Score* showing complete overlap. High degree of separation allows the model to learn distinguishing factors about the dataset, that it applies to the test and unseen data.

*Academic\_Score*

* *Scores\_on\_Plays\_Made*: Higher the *Academic\_Score* and higher the

*Scores\_on\_Plays\_Made*, is likely to result in a Full *Scholarship*, and vice-versa.

* *Missed\_Play\_Score:* Overall, data has a lot of overlaps on both the classes, but it looks like Full *Scholarship* is more likely a result of lower *Missed\_Play\_Score* and higher *Academic\_Score.* That said, higher *Academic\_Score* and higher *Missed\_Play\_Score* also *has* many data points with Full *Scholarship*. (Note: the Boxplot below throws more light on *Missed\_Play\_Score)*
* *School\_Score*: Even if a student has higher *Academic\_Score*, a lower *School\_Score* is more likely to result in a Partial *Scholarship.* The best chance of getting a Full *Scholarship* seems to be average to high *School\_Score* and average to high *Academic\_Score*.
* *Overall\_Score:* This composite score bears little relationship with other variables. It bears no relationship with other variables (escaping multicollinearity, but we need University's help to understand this better).

*Scores\_on\_Plays\_Made*

* *Missed\_Play\_Score:* It seems that Full *Scholarship* can be obtained, irrespective of the

*Missed\_Play\_Score* as long as the *Scores\_on\_Plays\_Made* is average to high.

* *School\_Score:* For a Full *Scholarship*, *School\_Score* needs to be average to high with an average to high *Scores\_on\_Plays\_Made*.
* *Overall\_Score: Overall\_Score* and *Scores\_on\_Plays\_Made* could be high or low for a Full *Scholarship*. Data is scattered and offers no insight. At all levels of *Overall\_Score*, a higher *Scores\_on\_Plays\_Made* will result in Full Scholarship.

*Missed\_Play\_Score*

* School\_Score: A low *School\_Score* will most likely result in a Partial *Scholarship.*
* Overall\_Score: *Overall\_Score* could be high or low, but *Missed\_Play\_Score* needs to be high for a Full *Scholarship*.

*Overall\_Score*

Overall\_Score: For all values of Overall\_Score, low values of *School\_Score* only Partial

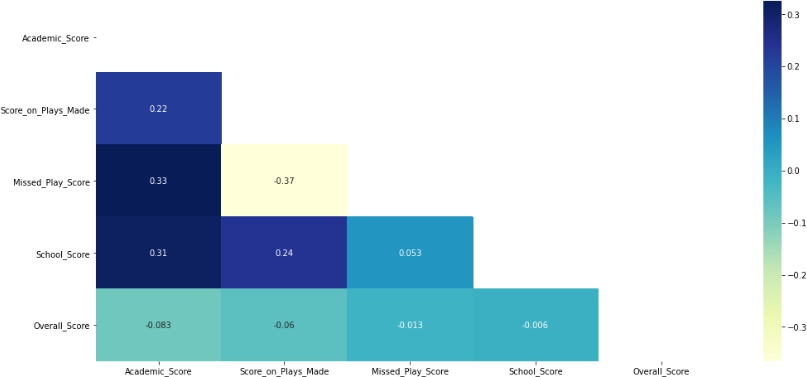
*Scholarship* is seen.

## Multivariate Bivariate Analysis for Continuous Variable with Correlation Matrix and Boxplot

Table: Correlation Matrix



Chart: Boxplot



* *Overall\_Score* has a very low negative correlation with all the 4 variables, depicting that any increase in *Academic\_Score*, *Score\_on\_Plays\_Made, Missed\_Play\_Score*, *School\_Score* will reduce *Overall\_Score* very minimally. Perhaps this can be explained by the computation of *Overall\_Score* itself: it consists of the candidate’s family financial state, school performance, psychosocial attitude, etc. These are not captured elsewhere in the dataset, so we can assume that Overall\_Score is a standalone metric, and therefore, correlationally not strong with the other numerical variables.

Of all the variables with *Overall\_Score*, *Academic\_Score* (-0.08) is the most negatively correlated.

* *School\_Score* is positively correlated with *Score\_on\_Plays\_Made* (0.24) and good *Academic\_Score* (0.31), showing perhaps there are better sports facilities and trainers in a higher scoring school supporting better scores on the field and perhaps an academically bright student can be found in a school with better score. Between the two, a better *School\_Score* is more likely to accompany *Academic\_Score*, than *Score\_on\_Plays\_Made*.
* *Missed\_Play\_Score* is negatively correlated with *Score\_on\_Plays\_Made (-0.37)*, understandably because a better performance on field is oppositely correlated with failures on the field. A positive correlation between *Academic\_Score* and *Missed\_Play\_Score* (0.33) tells us that if a student is academically oriented then she or he might be technically imperfect in sports, leading to errors on the field.

*Missed\_Play\_Score* has little to do with School\_Score (0.05), telling us that even if a candidate is from a ‘better’ school, the failures on the field does not reduce, but minimally increases.

* *Score\_on\_Plays\_Made* is positively correlated with *Academic\_Score* (0.22), showing performance in academics and sports are important for this group of students.

There is an interesting interplay of factors that we want to throw light on.

Theoretically, *Missed\_Play\_Score* and *Score\_on\_Plays\_Made* are opposites of each other, therefore the negative correlation of -0.37. A good sportsman has less failures on the field. However, the relationship with *Academic\_Score* is less straightforward.

*Academic\_Scores* is positively correlated with *Score\_on\_Plays\_Made* (0.22) as well as *Missed\_Play\_Score* (0.33), although the two latter variables themselves are negatively correlated.

The data suggests that even though a student with a higher *Academic\_Scores* is somewhat likely to have a higher *Score\_on\_Plays\_Made* (0.22), but she or he is more likely to have *Missed\_Play\_Score* (0.33).

To put it simply, if a student is scoring high academically, she or he is more likely to be less effective as a sportsman.

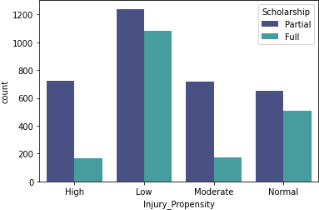
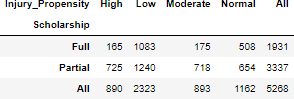
* *Academic\_Score* is positively correlated with *School\_Score*. At 0.31, it is the highest correlation after *Academic\_Score* and *Missed\_Play\_Score*, which is at 0.33. This can be explained by a higher scoring school (a better school, lets say) will have more academic rigor as well as facilities, processes and faculty which leads to better academic performance.

## Bivariate Analysis for Categorical Variables with Countplot

Categorical variables can be analysed using their class frequencies. We have 4 categorical variables: *Injury\_Propensity*, *School\_Type, Region* and target variable, *Scholarship.*

We analyse each of the three against Scholarship to see the prior probabilities.

Chart: Injury\_Propensity vs Scholarship



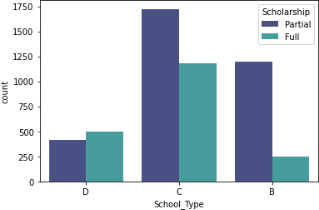
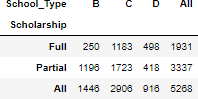
* Majority of the Full *Scholarship* are provided to the students who belongs to Low and Normal *Injury\_Propensity* categories
* Students with High and Moderate Injury propensity seem to have a similar distribution in *Scholarship* (both Partial and Full)
* Students with Low *Injury\_Propensity* will a have higher chance of getting the

*Scholarship,* both Partial and Full.

* High and Moderate *Injury\_Propensity* students are most likely to receive Partial

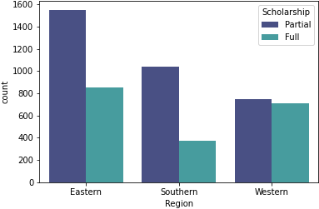
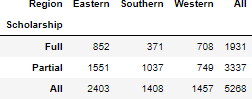
*Scholarship*,than Full. .

Table & Chart: Frequency of School\_Type vs Scholarship



* School D more has students with Full *Scholarship* than Partial *Scholarship*.
* School C has the highest number of students and thus contributes maximum to the dataset. Full *Scholarship* has 40 % of all students in School C.
* In School B, only 17% students enjoy Full *Scholarship.*

Table & Chart: Frequency of Region vs Scholarship



* It can be seen that Western *Region* has students with highest percentage of Full

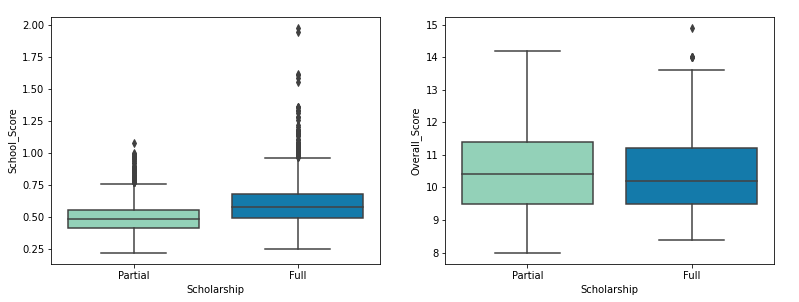
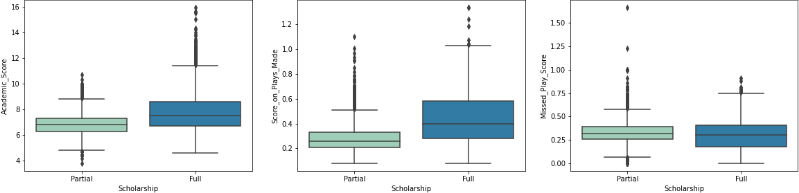
*Scholarship* (49%).

* Even though Eastern *Region* has maximum number of students in the dataset, only 35% of them enjoy Full *Scholarship*.
* Southern *Region* and Western *Region* have similar numbers of students with *Scholarship*, but Southern *Region* has only 26% of students enjoying Full *Scholarship,* compared to 49% of Western *Region*.

### Categorical & Numerical Analysis with Boxplot

In this section, we analyse the spread of how each numerical variable stack up against categorical variables and their classes, starting with the target variable.

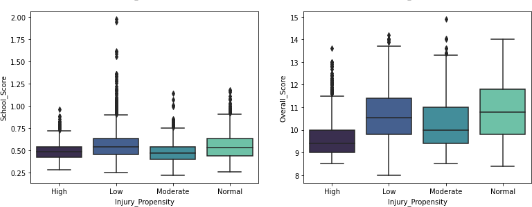
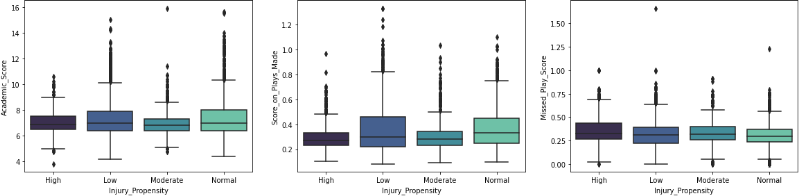
Boxplot: Continuous Variables and Scholarship



The boxplot corroborates the findings of the pairplot above. Additionally, it shows the presence of significant outliers in almost all the variables. Individually:

* Students with higher *Academic\_Score* are more likely to get a *Scholarship*. Students with a score more than 12 are likely to get a Full *Scholarship*.
* *Score-on\_ Plays\_Made* clearly shows the Full *Scholarship* being visibly higher scorer on the field than the recipient of Partial *Scholarship.*
* The boxplot of Full *Scholarship* is much wider in *Missed\_Play\_Score* boxplots, showing that students have higher spread of score, but the median of both the classes are close to each other. This depicts median scores are not very different for Partial and Full *Scholarship*. Partial *Scholarship* does have a large number of outliers on the higher side, which means that the group that is more likely to make higher failures on the field are perhaps more likely to receive Partial *Scholarship*.
* Students with higher *School\_Score* are likely to get a Full *Scholarship*.
* *Overall\_Score* distribution seems similar for both Full and Partial Scholarship.

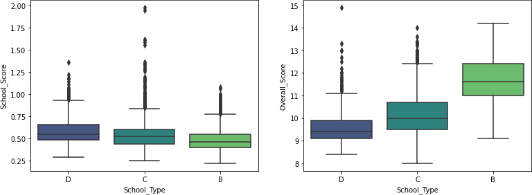
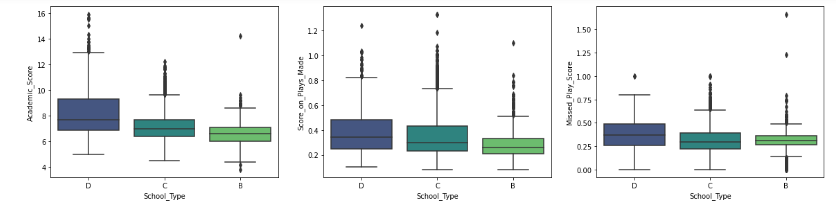
Boxplot: Continuous Variables vs Injury\_Propensity



* The median for all types of Injury\_Propensity seems to be near each other for *Academic\_Score* meaning, the whether *Injury\_Propensity* is High, Low, Moderate or Normal, equal number of candidates score more than the median (approximately 7), as less than the median.
* In *School\_Score*, the most outliers lie in Low *Injury\_Propensity*
* A High Injury\_Propensity is associated with low *Overall\_Score*. A Normal

*Injury\_Propensity* is associated with a higher *Overall\_Score.*

Boxplot: Continuous Variables vs School\_Type

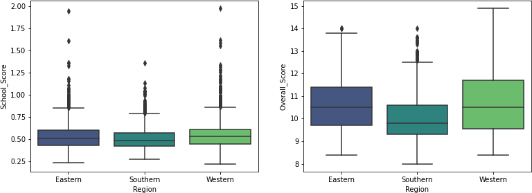
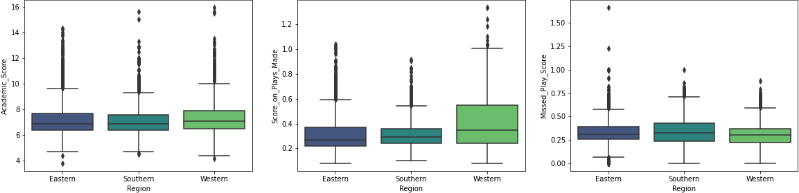


* *School\_Type* D students score the highest *Academic\_Score*. While the lowest scorers belong to *School\_Type* B. All have outliers.
* *School\_Type* D and *School\_Type* C have competing scores when it comes to making points on the field i.e., *Scores\_On\_Play\_Made*. Schools B has least *Scores\_On\_Play\_Made*. All have outliers.
* *School\_Type* D students are making more mistakes on the field compared to *School\_Type* C and *School\_Type* B, which have more outliers than *School\_Type* D. However, *School\_Type* B students are making the least mistakes among all three schools.
* *School\_Type* D has schools with slightly higher *School\_Score*, followed by

*School\_Type* C and then B. All have Outliers.

* The *Overall\_Score* is the highest for *School\_Type* B, followed by *School\_Type* C then *School\_Type* D. Although few outliers on *School\_Type* D have higher values than the other two.

Boxplot: Continuous Variables vs Region



* *Academic\_Score* is comparable for all regions.
* Western *Region* is the highest scorer on the field with a clear lead, as depicted by the

*Score\_on\_Plays\_Made* boxplot.

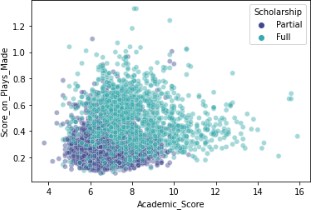
* Eastern and Western *Regions* are similar in mistakes on the field (*Missed\_Play\_Score*), Southern *Region* has little more spread in the data.
* *School\_Scores* are similar for all three Regions.
* *Overall\_Scores* is the highest for Western, closely followed by Eastern, and Southern trailing the two.

**Multivariate Analysis**

Multivariate statistical methods are used to analyze the joint behavior of more than one random variable.

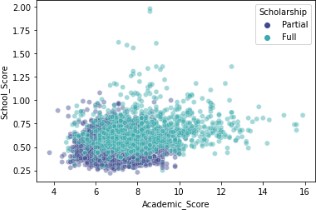
*Scatter Plots*

Scatterplot: Academic\_Score vs Score\_on\_Plays\_Made



* With increase in *Academic\_Score* and *Score\_on\_Play\_Made* chances of getting a Full *Scholarship* will increase. At score more than 11 the student have higher probability of getting a Full *Scholarship.*
* The correlation is positive

Scatterplot: Academic\_Score vs School\_Score



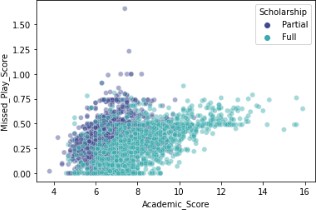
At higher School\_Score (>1.25), the student is more likely to get a Full *Scholarship.* For lower values we can see a overlap of datapoints.

There is somewhat of a positive correlation.

Students with a high *School\_Score* are most likely to get Full *Scholarship*. Students who belong to higher performing Schools and average academic scores AND Students who have high Academic Scores from average schools receive Full *Scholarship.*

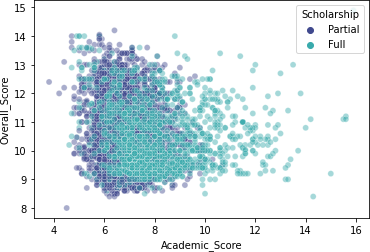
If the *School\_Score* falls below 0.5, likelihood of both possibilities, Full and Partial, coexist. There are few students with low *School\_Score* and low *Academic\_Score* who have got the Full *Scholarship* and this can be investigated further.

Scatterplot: Academic\_Score vs Score\_on\_Plays\_Made



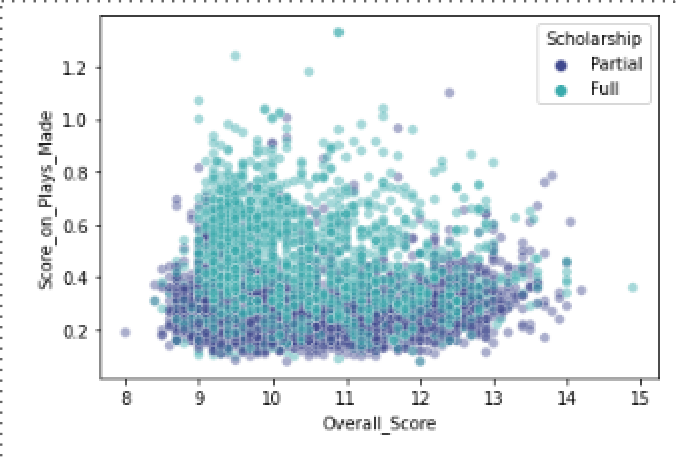
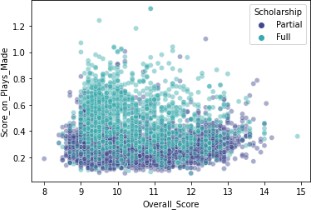
* For higher Academic\_Scores (> 10.5) there is more certainty of Full *Scholarship.*
* Even at the lower *Academic\_Score* and 0 *Missed\_Play\_Scores* students have secured Full *Scholarships*. It means if a student is able to cut down the on field mistakes he might end up getting a Full *Scholarship*, despite a low academic score.

Scatterplot: Academic\_Score vs Overall\_Score



* *At Academic\_Scores* < 6, very few incidents of Full *Scholarship* are present. Perhaps they are only present due to the influence of other variables.
* At *Academic\_Score* > 11, Full Scholarship is almost certain at varying *Overall\_Scores*.
* There is an "outlier" at *Academic\_Score* of 16 and *Overall\_Score* of 15. In the absence of any documentation, it seems like a high performing case, and not an outlier that disturbs the model. We explore Outliers in the next section.

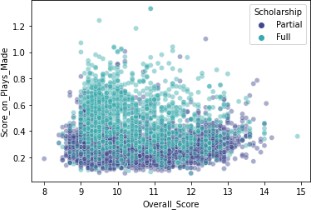
Scatterplot: Academic\_Score vs Overall\_Score



*Overall\_Score* is scattered throughout, there is no significant trend that can be seen here.

For all values of *Overall\_Scores, Scores\_on\_Plays\_Made* > 0.5 has shows likelihood of a Full *Scholarship*. This is in line with our earlier observations in the Pairplot and Boxplots.

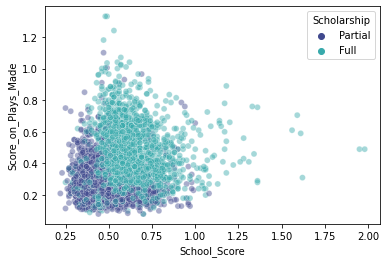
Scatterplot: Academic\_Score vs Overall\_Score



*Overall\_Score* is scattered throughout, there is no significant trend that can be seen here.

For all values of *Overall\_Scores, Scores\_on\_Plays\_Made* > 0.5 has shows likelihood of a Full *Scholarship*. This is in line with our earlier observations in the Pairplot and Boxplots.

Scatterplot: School\_Score vs Score\_on\_Plays\_Made



* Scoring high in *School\_Score* will ensure Full *Scholarship* even if their *Score\_on Play\_Made* is not much.
* Some students with lower *School\_Score* and high *Score\_on\_Play\_Made* have gotten Full *Scholarship*. This can be investigated further.

**Outliers Analysis and Treatment**

*Observations on Outliers Detected:*

1. For *Academic\_Score*, the lower outliers have values 3.8 to 4.4 and higher outliers have values between 9.7 to 15.9. The data dictionary, does not mention the scorecard with the range.. Logically, a10,12 or even 15 marks do not seem unrealistic, or invalid.
2. If we see the lower outliers for variable *Missed\_Play\_Score*, there are many records with value 0.0 to 0.02. and values higher outliers are between 0.63 to 1, except two records which are 1.23 and 1.66. Again, the value does not seem unrealistic.
3. School\_Score and *Overall\_Score*.that on which scale *School\_Score/Overall\_Score* has been measured.

We did not want to drop the outliers initially, given that the points are not invalid. After identification of outliers, we need to analyse them from the validity point, and from the point of view of industry or a business expert. At this point, we would have consulted with the

experts and gotten their view on whether the outliers are possible / valid ones or impossible / invalid.

However, keeping in mind the performance of the Logistic Model, we will be treating the outliers by flooring and capping them.

Q2. Use Full Data to develop a logistic regression model to identify significant predictors. Check whether the proposed model is free of multicollinearity.

Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?

Compare values of model selection criteria for proposed models. Compare as many criteria as you feel are suitable.

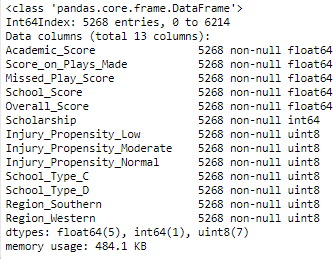
**Logistic Regression**

Logistic regression is the appropriate regression (prediction) analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Encoding follows a lexicographic ordering, so in the “Football” dataset F of Full in Scholarship variable (Dependent variable) comes first, so it will be coded as 0 and Partial will be coded as 1. Hence, we manually coded it because we wanted Full to be encoded as 1 to assist model interpretation.

For other independent variables in the dataset we have used n-1 encoding technique to convert all the categorical variables into binary.

Data information



As per above snippet, post encoding, we can see that there are 5,268 rows and 13 variables in total with Scholarship is the Dependent Variable and rest are Independent Variables.

## Model Building

#### With all the Features (Part I)



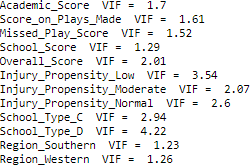
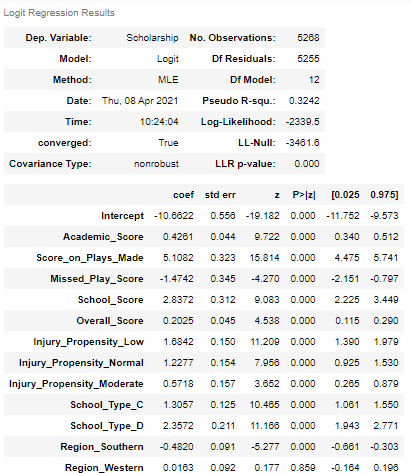
*Interpretation:*

The method of estimation follows an iterative process called Fisher Scoring Algorithm. It is important that the algorithm converges. In the current situation convergence took place in 6 iterations.



*Checking Multicollinearity in the predictor variables Using VIF*

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. In general, VIF greater than 5 have high/severe multicollinearity between variables. Since, no variable has high VIF value thus we will check the p-value from the snippet of logistic regression above.



What is p-value? How do we interpret it? why do we need it?

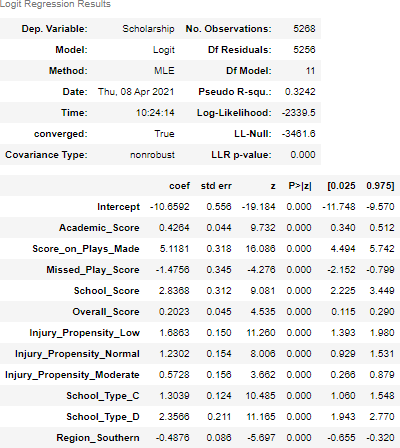
p-Value means the probability of finding something by pure chance in a dataset.

H0 or null hypothesis in this case is that any changes in predictor variable does not change the target variable. The target and predictor are independent of each other.

At a threshold confidence level of 0.05 (or 5%) if we have a p-value of 0.07, lets say, we will accept the null-hypothesis that the predictor variable has no role in ‘predicting’ the target variable because changes in the predictor is not associated with a corresponding change in the target variable.

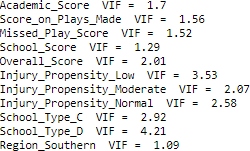
p-Values help in determining whether the variable is significant to determining the predictor variable or not. Since p-value of the Region\_Western variable is 85.9% which is greater than 5% hence, the variable is not significant. Thus, we decided to drop the Region\_Western variable from the data and again build the Logistic Regression results.

#### Without Region\_Western Variable (Part II)



*Checking Multicollinearity in the predictor variables Using VIF*

Again, if we check the VIF values of all the variables, none of them have high VIF factor and p-values are also 0 now for all the variables.



## Interpreting Coefficients with Odds Ratio

Unlike a Linear Regression, where coefficients directly denote the degree of relationship between the Dependent and Independent Variable (because they are linearly related), in Logistic Regression, the relationship is that of the variable increase or decrease leading to the odds of the dependent variable changing by an exponent of the coefficient.

Checking the coefficient values, the positive coefficients indicate predictors for which an increase in the value of the predictor is associated with an increase in the probability of Full *Scholarship*, the target class of interest, now denoted by 1. Similarly, negative coefficients indicate predictors associated with reducing the probability of *Full Scholarship*.

Unit increases in *Academic\_Score, Score\_on\_Plays\_Made, School\_Score and Overall\_Score*

variables are associated with an increase in the probability of *Full Scholarship*.

Higher *Missed\_Play\_Score* reduces the probability of *Full Scholarship*. Similarly,

*Region\_Southern* also reduces the probability of *Full Scholarship*.

Let’s look at the extent of relationship between the target and predictor variables through Odds Ratio.

**Odds Ratio**

|  |  |
| --- | --- |
| Intercept | 0.000023 |
| Academic\_Score | 1.531682 |
| Score\_on\_Plays\_Made | 167.013704 |
| Missed\_Play\_Score | 0.228634 |
| School\_Score | 17.061191 |
| Overall\_Score | 1.224262 |
| Injury\_Propensity\_Low | 5.399585 |
| Injury\_Propensity\_Normal | 3.421877 |
| Injury\_Propensity\_Moderate | 1.773297 |
| School\_Type\_C | 3.683764 |
| School\_Type\_D | 10.554614 |
| Region\_Southern | 0.614073 |

Lets take a continuous variables example.

For every unit of increase in *School\_Score*, the log odds of receiving Full *Scholarship* increases by 2.8372, provided every other variable is constant. To reverse the effect of log, we calculate the exponent of 2.8372, which is 17.06.

For every unit increase in *School\_Score*, the odds of Full *Scholarship* increase by 17.06 times or by 1601% .

Similarly,

For every unit of increase in *Academic\_Score*, the odds of receiving Full Scholarship increases by 1.53 times, or 53% .

For every unit of increase in *Missed\_Play\_Score*, the odds of receiving Full Scholarship reduces by 0.22 times, or 78% .

For every unit of increase in *Overall\_Score*, the odds of receiving Full Scholarship increases by 1.22 times, or 22% .

For every unit of increase in *Score\_on\_Plays\_Made*, the odds of receiving Full Scholarship increases by 167 times, or by 6701%.

For categorical variables,

If the student has *Injury\_Propensity\_Low*, the odds of receiving Full Scholarship increases by

5.4 times, or 439%.

If the student has *Injury\_Propensity\_Normal*, the odds of receiving Full Scholarship increases by 3.42 times, or 242%.

If the student has *Injury\_Propensity\_Moderate*, the odds of receiving Full Scholarship increases by 1.77 times, or 77%.

If a student has *School\_Type\_C*, the odds of receiving Full Scholarship increases by 3.68 times, or 268% .

If a student has *School\_Type\_D*, the odds of receiving Full Scholarship increases by 10.55 times, or 955% .

If a student has *Southern Region*, the odds of receiving Full Scholarship decreases by 0.61 times, or 39% .

From the EDA, we could arrive at a similar point. Full *Scholarships* are higher when:

* *School\_Score* are higher
* *Score\_on\_Play\_Made* are higher
* Higher *Academic\_Score* and higher the *Score\_on\_Plays\_Made*
* Higher *Academic\_Score* and higher the *School\_Score*
* At 54.3%, *School\_D* more has students with Full *Scholarship*. Best chance of getting a Full Scholarship is in School type D.
* *Injury\_Propensity* categories are Low and Normal

**How to read Odds Ratio?**

If the Odds Ratio is above 1, the outcome has a positive relationship with the target.

Odds Ratio for a continuous predictor variable is the number of times the odds change the target variable, for a single unit increase in the continuous predictor variable, given other variables are constant.

To calculate the percentage change, we deduct 1 from the ratio (10 if the number is in tens, 100 if the number is in hundreds, and so on) and then multiply the ratio by 100 (for %). E.g. odds of 1.25 will become 25%,

2.5 will become 150%

21.34 will become 1134%

167.2 will become 6720%

If the Odds Ratio is below 1, the predictor has a negative relationship with the target. We see this in the case of *Southern\_Region* and *Missed\_Play\_Score.*

We deduct the number from 100, and arrive at the % less likely.

E.g. If the odds ratio for *Missed\_Play\_Scores* is 0.22, it means that each unit increase in this variable will reduce the likelihood of the chosen target variable (Full Scholarship) by

0.22 times. It can also be interpreted as: with each unit increase in the predictor variable, the probability of target reduces by 78% (1-0.22).

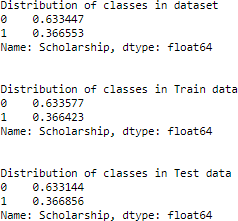
Q3. Split the data into training (70%) and test (30%). Build the various iterations of the Linear Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics.

If prediction accuracy of the full scholarship is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (II) to compare accuracy in training and test sets. Compare the final model of Part (II) and the proposed one in Part (III). Which model provides the most accurate prediction? If the model found in Part (II) is different from the proposed model in Part (III), give an explanation.

## Splitting Data into training (70%) and test (30%)

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm.

Target Variable classes are uniformly distributed in Train and Test:



With the help from the strategy function, we ensure that the Train and Test groups have the same composition of Full and Partial as the sample dataset. This helps maintain class balance during validation.

We are splitting the data and comparing between train and test for the selected model. We are interested in good model performance, as well as explainability.

Our first model (Part I) was using all the variables. All VIF scores were within 5, suggesting the correlation between variables are not alarming and would not affect model performance. Having taken care of the VIF, we assess the significance of variables.

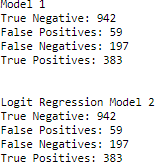
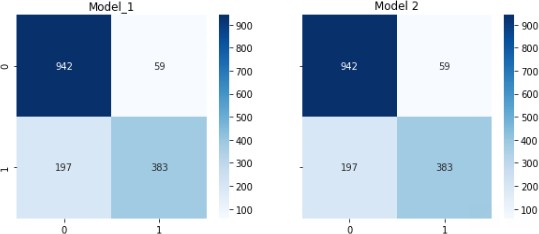
We found that for one variable *Region\_Western,* the p-value was lower than 0.05, thereby suggesting its insignificance in influencing the target variable.

In the second model (Part II), we dropped *Region\_Western.*

We made predictions on Part I and Part II Test set and are comparing their performance below:

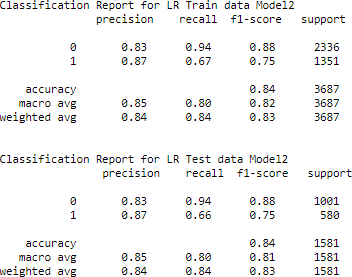
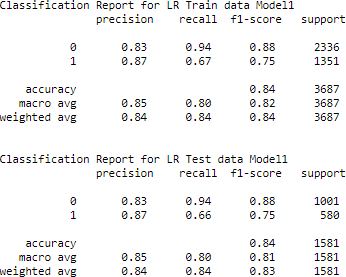
## Model Evaluation

#### Confusion Matrix



#### Classification report

Model 1 Model 2



From the above 2 models matrix it can be observed that even after dropping the “*Region\_Western*” variable there is no change in the prediction and thus the model performs in exactly the same way as Model 1.

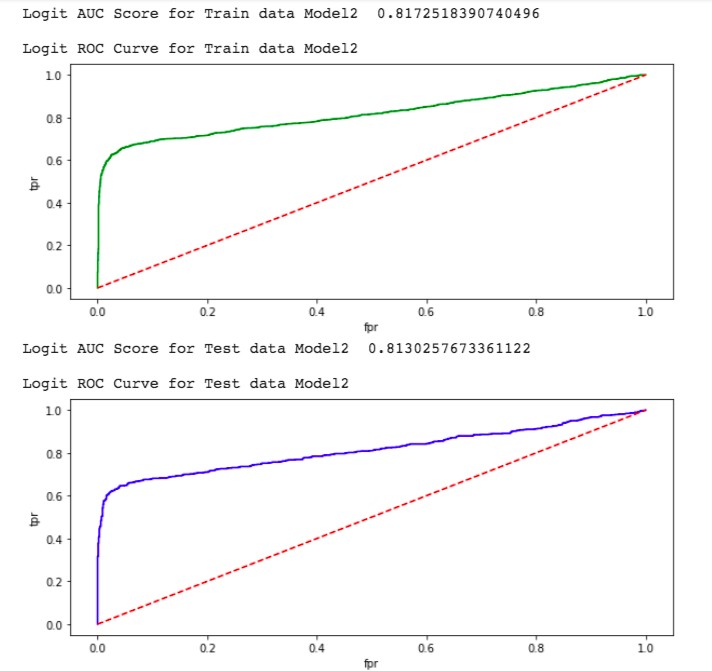
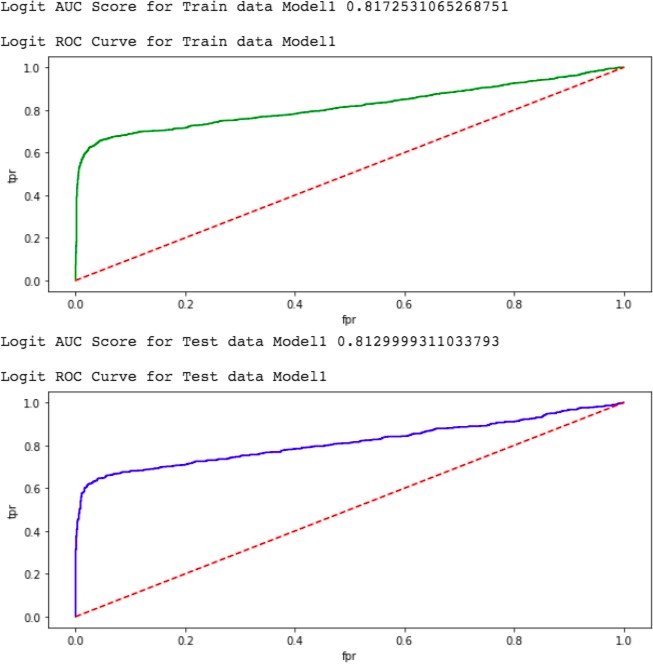
Classification metrics that we will be considering closely are : Accuracy, Recall / Sensitivity, Precision, F1-Score.

Accuracy is one metric for evaluating classification models.

Precision measures the proportion of positive identifications that are *actually* correct. Recall attempts to measure the proportion of *actual positives* that are identified correctly. F1 Score is the harmonic mean between the precision and the recall, and balances the two.

#### AUC ROC Curve for Logistic Regression

Model 1 Model 2



ROC curve makes it easy to identify the best threshold for making a decision, that which threshold the model performs better. From the above ROC curve, we can see that both the models perform equally good, with Part II 2 resulting in slightly better prediction than Part I .

AUC represents the model’s degree of separability between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. AUC Score is also very close in both the models. We will pick Model 2 (Part II) since it has all significant variables and the AUC Score is marginally higher.

---

In an imbalance dataset, accuracy is not considered reliable and other metrics are more relied upon.

*The question is - is it more expensive or undesirable to have a false negative or a false positive?*

A False negative is a case where a student has been predicted as not eligible for Full

*Scholarship,* yet in reality she or he is eligible.

A False positive is a case where a student has been predicted as eligible for Full *Scholarship*

by the model*,* yet in reality she or he is not eligible.

If the university is new and is trying to promote itself, it will be very careful not to miss the best athletes applying. It doesn’t mind if a few non-meritorious students are selected by the model. In this case, we can live with a higher false positive and Precision Score is more important.

However, for example, if the university is well established and has a reputation to protect, it will be very selective about identifying only the best. In this case, we can opt for higher false negative and Recall score is more important.

## Recommendation

Model 2 (Part II) shows no overfitting or underfitting. Performance-wise:

* it is able to predict with 84% accuracy (Accuracy Score)
* 87% of class 1 predictions (Full *Scholarship*) are actually correct (Precision)
* 67% of the actual Full *Scholarship* records have been actually identified correctly by the model (Recall)
* the F1 Score is 75% (Precision & Recall)

In Logistic Regression, we recommend Model 2 (Part II).

Q4. Use the same training-test data split in Part (III) to develop a suitable Linear Discriminant Analysis (LDA) model. Use the same to on the test data. Compare the final output from the logistic regression model and LDA.

**Linear Discriminant Analysis (LDA)**

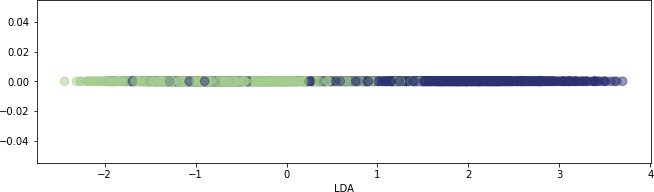
Linear Discriminant Analysis (LDA) is a dimensionality reduction technique. As the name implies dimensionality reduction techniques reduce the number of dimensions (i.e. variables) in a dataset while retaining as much information as possible.

LDA assumes that all classes are linearly separable through hyperplanes in the feature space created to distinguish the classes. If there are three classes then the LDA draws two hyperplanes and projects the data onto these hyperplanes in a way that maximize the distance between the means of two classes and minimizing the variation between each category. The hyperplanes represent the newly formed dimensions which are linear combinations / transformation of dimensions in the dataset.

Although normality is an assumption for LDA, even without that assumption being satisfied, LDA has been seen to perform quite satisfactorily.

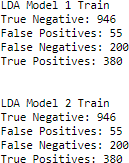
We built the Models in python. Model 2 has been able to create the some kind of separability between classes (Full *Scholarship* and Partial *Scholarship*) as shown below.

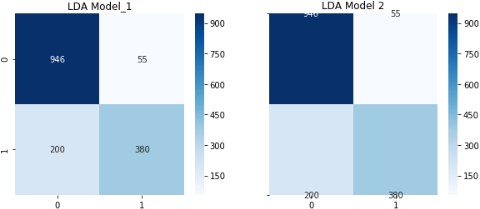
Scatterplot: Linear Segregation of Data by LDA algorithm



## Model Evaluation

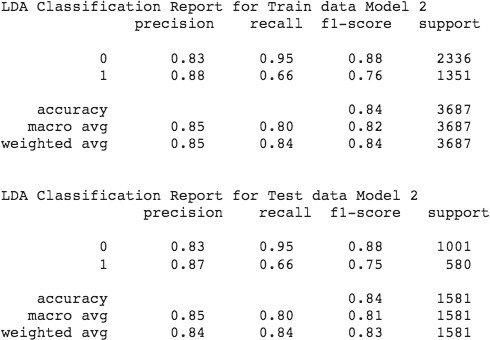
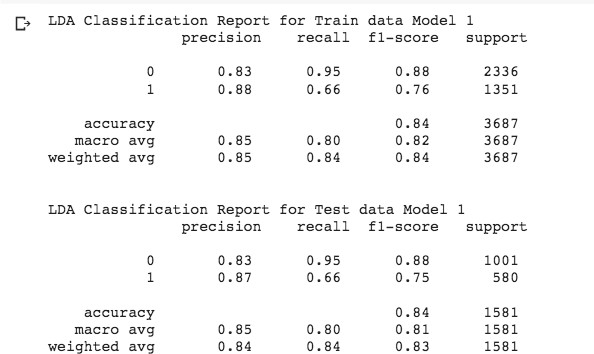
#### Confusion Matrix





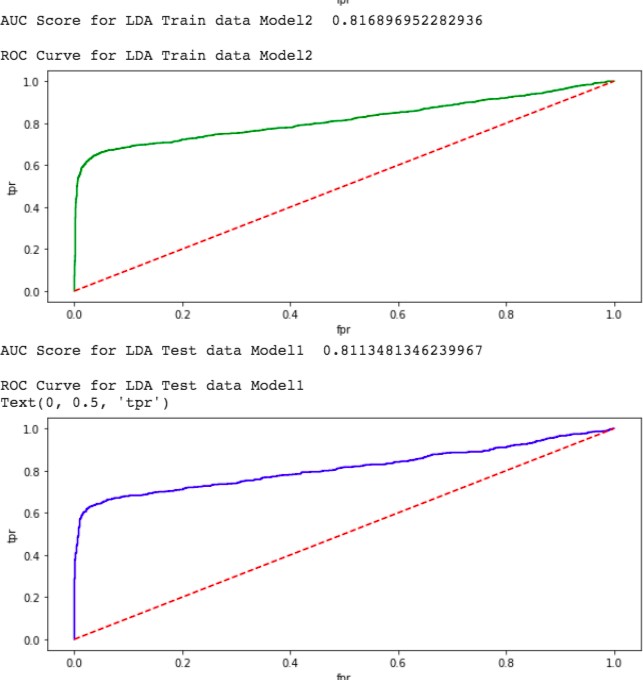
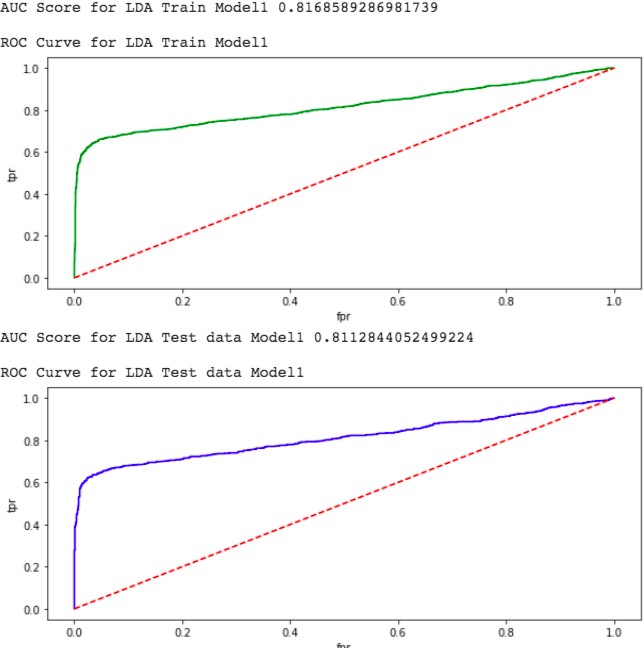
#### Classification Report for LDA

Model 1 Model 2



#### AUC ROC Curves for LDA

Model 1 Model 2



## Recommendation

LDA Model 2 (Part II) shows no overfitting or underfitting. Performance-wise:

* it is able to predict with 84% accuracy (Accuracy Score)
* 87% of class 1 predictions (Full *Scholarship*) are actually correct (Precision)
* 66% of the actual Full *Scholarship* records have been actually identified correctly by the model (Recall)
* the F1 Score is 75% (Precision & Recall)

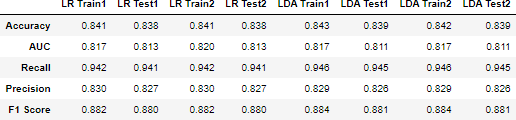
In LDA, we recommend Model 2 (Part II) since it has only significant variables.

Q5.Business Report documenting the results and stating the actionable insights and the recommendations.

**Combined Matrix Results**

We have performed 2 predictive modelling techniques, Logistic Regression and Linear Discriminant Analysis, with 2 models. Here are the summary performance for each:

Table: Summary performance of various models



While the LDA Model 2 performs exactly the same as LDA Model 1 and both perform slightly better than the Linear Regression Models. For predictive power, LDA Model 2 is better, however, for explainability, Logistic Regression Model 2 is better. Although the two Logistic Regression Models offer the same performance, the second model only has significant variables, so that is most preferable.

The most preferable instances for Full Scholarship are:

* A higher *School\_Score*
* A higher *Score\_on\_Play\_Made*
* Higher *Academic\_Score* and *Score\_on\_Plays\_Made*
* Higher *Academic\_Score* and *School\_Score*
* Being from *Type School\_D*
* *Injury\_Propensity* categories with Low and Normal