**PREDICTIVE MODELLING BUSINESS REPORT**

***Abhishek Chauhan***

***Mrinalini Kulkarni***

***Pavan Krovi***

***Nivedita Kulkarni***

***Yogesh Bhangale***

***PGPDSA DEC’19 BATCH – Group 6***

***Date: Feb 07, 2021***

**CONTENTS**

1. Problem 1: Linear Regressions
2. Assumptions …………………………………………………………………………………
3. Importing Packages ………………………………………………………………………
4. EDA ……………………………………………………………………………………………….
5. Linear Regression.….…………………………………………………………………….
6. Actionable Insights……………………………………………………………………….
7. Recommendations……………………………………………………………………….
8. Problem 2: Logistic Regression and Linear Discriminant Analysis (LDA)
9. Assumptions …………………………………………………………………………………
10. Importing Packages ………………………………………………………………………
11. EDA ……………………………………………………………………………………………….
12. Logistic Regression……………………………………………………………………….
13. Linear Discriminant Analysis.….…………………………………………………….
14. Actionable Insights……………………………………………………………………….
15. Recommendations……………………………………………………………………….

**PROBLEM 1: LINEAR REGRESSION**

**Exploratory Data Analysis for Problem 1**

The dataset provided to us is stored as “[cubic\_zirconia.csv](https://olympus.greatlearning.in/courses/9046/files/1113623/download?verifier=70bhgMbPWpCatRtDGpCwAZpsDn1nZqB0gWtG9T0k&wrap=1)” which contains data of 26967 customers and 11 variables namely

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describes the cut quality of the cubic zirconia. Quality is in increasing order: Fair, Good, Very Good, Premium, Ideal. |
| Colour | Colour of the cubic zirconia. |
| Clarity | Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3 |
| Depth | The Height of a cubic zirconia piece, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

Table : List of variables

**Solution:**

**Importing the Dataset**

The dataset in question is imported in jupyter notebook using **pd.read\_csv ()** function and will store the dataset in “**claim\_df**”. The top 5 rows of the dataset are viewed using **pd.head ()** function.

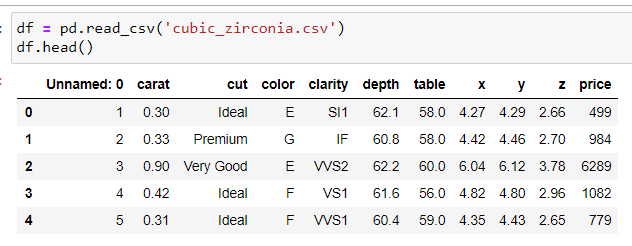


Figure : Head of Table

**Observations from Basic EDA:**

* Total no.of variable are 10.
* We have 3 catagorical and 8 numerical variable available before working with model.
* We have 10

**Dimension of the Dataset**

Let us check the number of rows and the number of columns in the dataframe-



Figure : Dimension of dataset

**Structure of the Dataset**

Structure of the Dataset can be computed using **df.info()** function

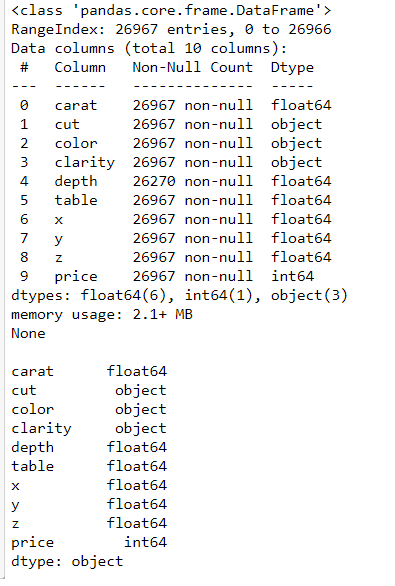
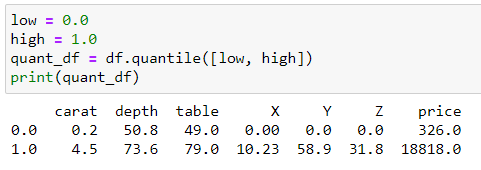


Figure : Structure of DataSet



**Summary of the Dataset**

The summary of the dataset can be computed using pd.describe () function.

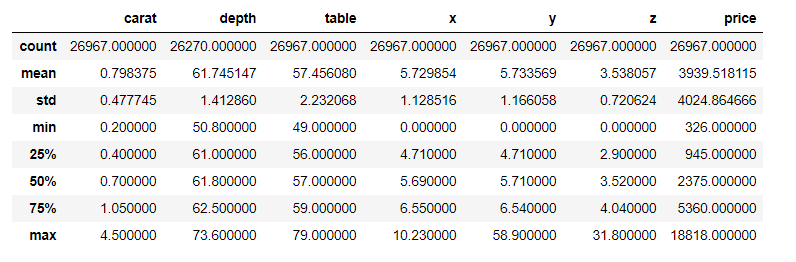


Figure : Summary of Dataset

**Checking for Missing Values**

The missing values or “NA” needs to be checked and dropped from the dataset for the ease of evaluation and null values can give errors or disparities in results. Missing Values can be computed using **.isnull().sum()** function.

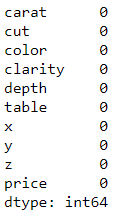


Figure : Missing Values

There are no missing values in the dataset hence we can go ahead with building the model.

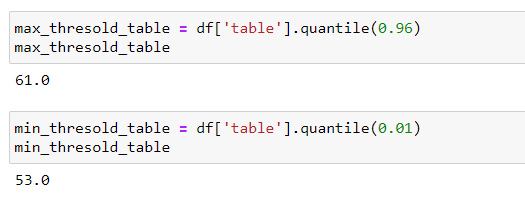
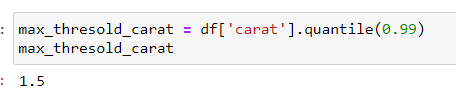
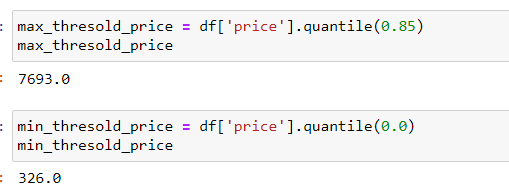
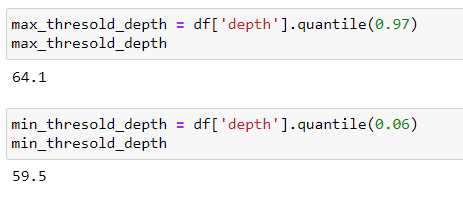


Figure : Threshold calculation

Using Boxplot we can observe Price is having highest

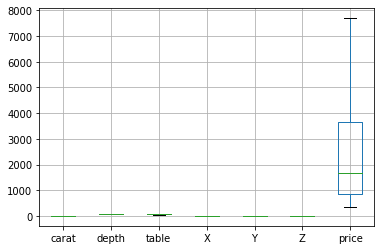
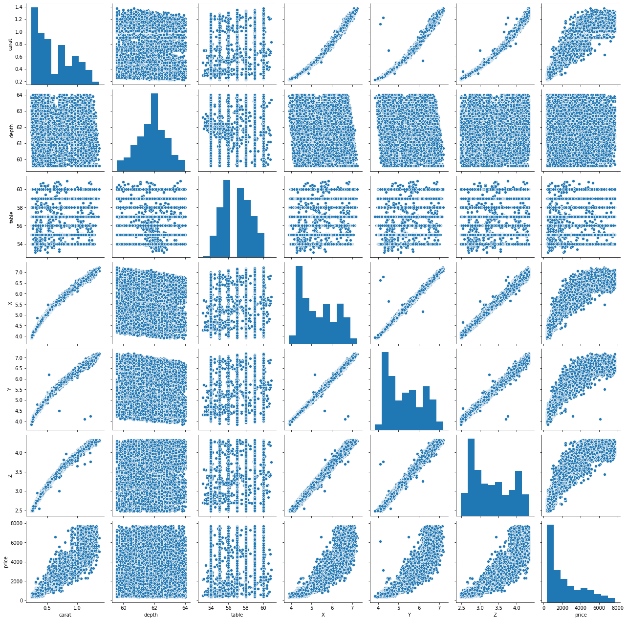
****

Figure : Boxplot to Observe Price value

**** Figure : Pairplot

from Pair plot we can see data is dependent with all the variable.

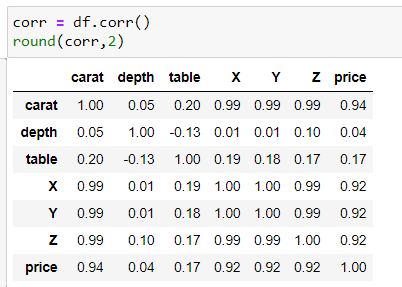
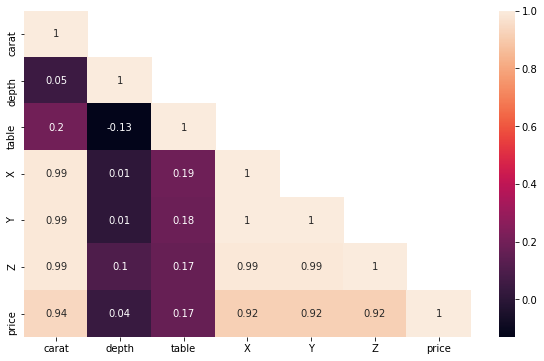


Figure : Corelation

From Heat map we can see X,Y,Z variable are having same 0.99 followed by Price.

**** Figure : Heat Map

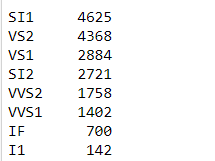
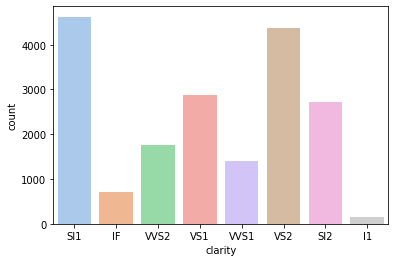
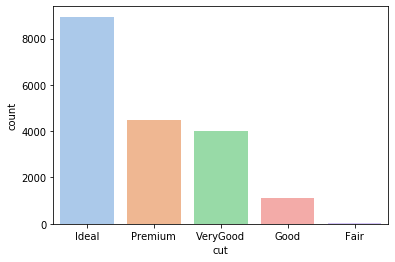
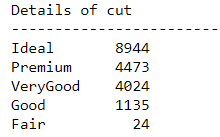
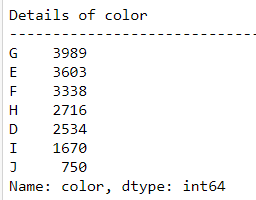
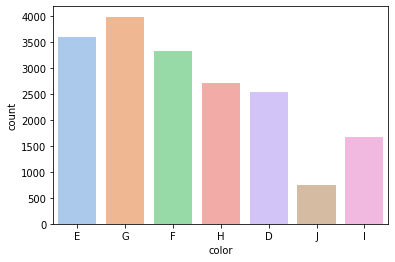


Figure : Bar plot to identify values

**Univariate Exploratory Data Analysis for the continuous variables**

we have used boxplot to check outlier for analysing univariate EDA for continues variable.

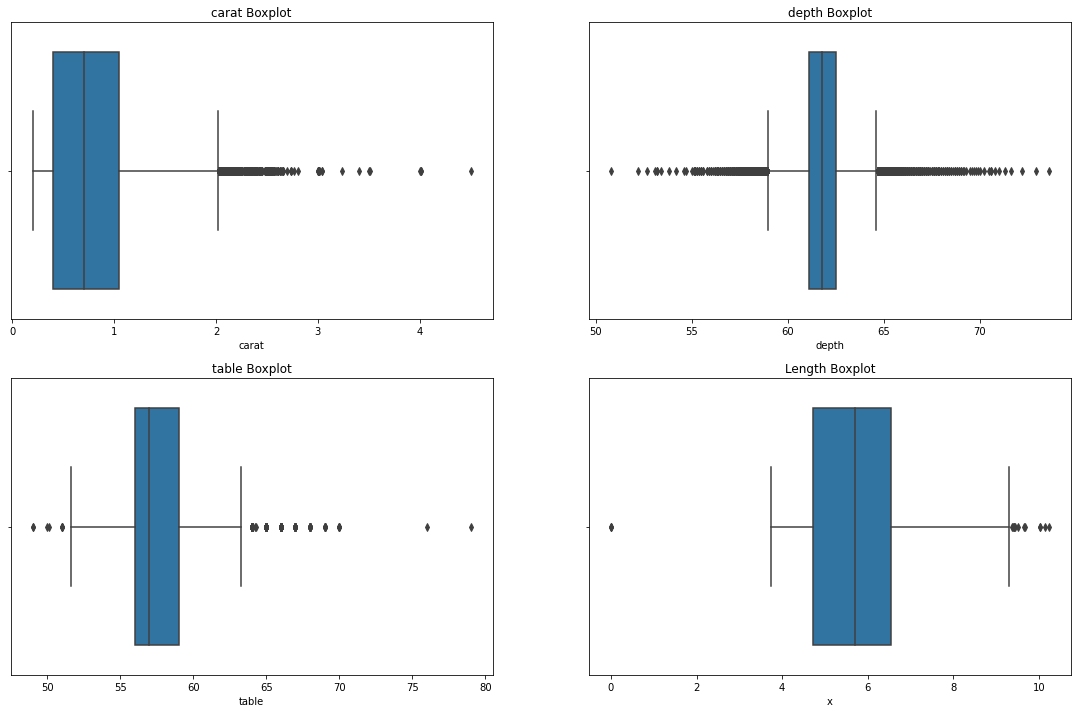


Figure : Univariate EDA for continuous variables

From above image we can see Carat Boxplot is right skewed, depth boxplot having both right and left skewed, table boxplot is right skewed and length boxplot is having centralised data.

# bivariate Exploratory Data Analysis

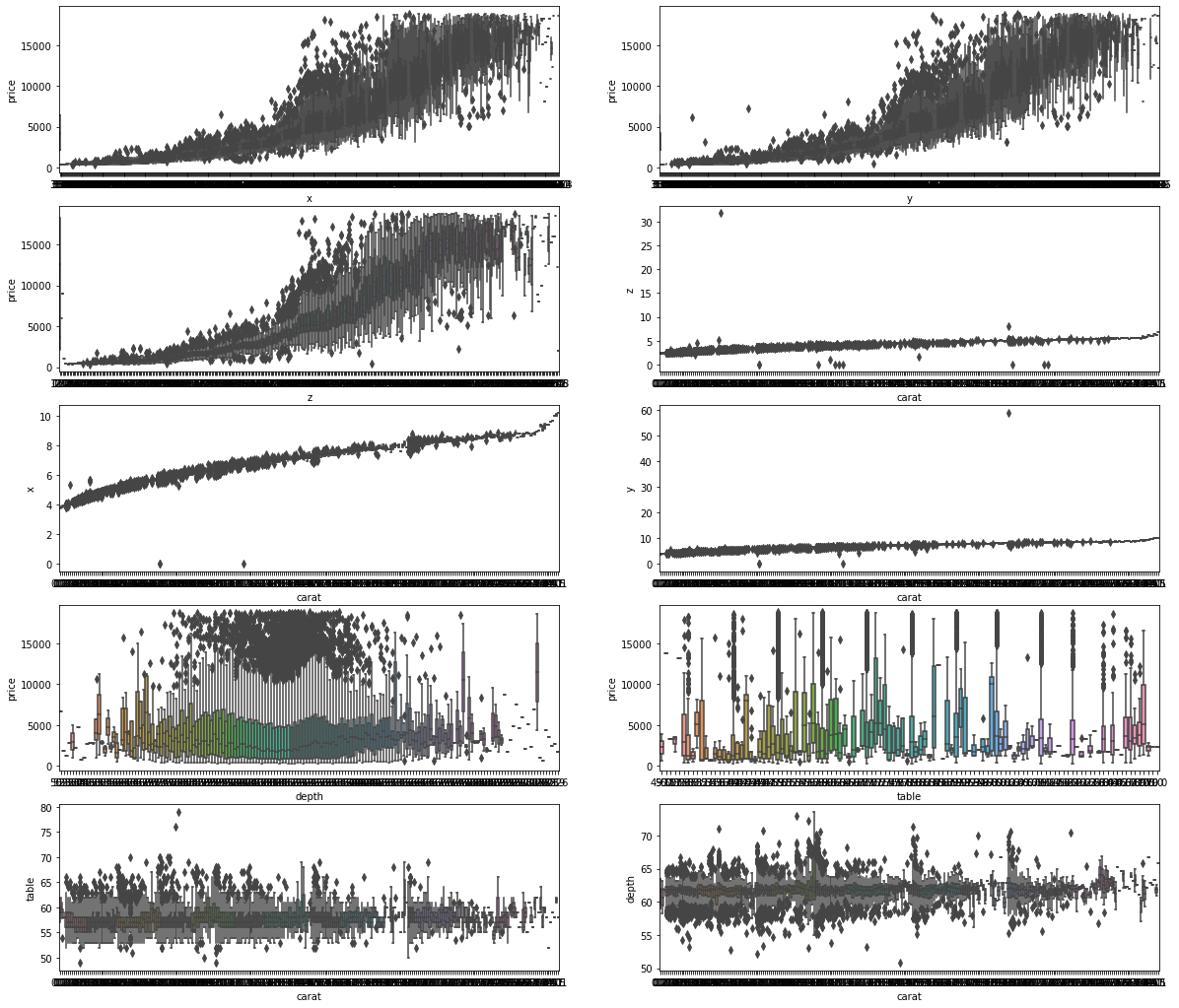


Figure : Bivariate EDA

**Build various iterations of the Linear Regression model using appropriate variable selection techniques for the full data.**

# Model 1: Model using dummy without dropping one level

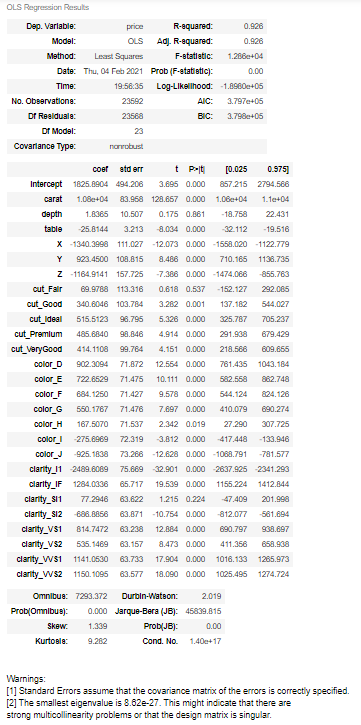


Figure : Model 1

# Model using appropriate number of dummy variable levels.

# 

Figure : Model 2

From above models it shows R-Squared values without dropping one level dummy data and using appropriate no.of dummy variable levels are same : **0.926**

We have checked multicollinearity for full data after checking VIF values and by removing highest VIF values we have created final formula for model by X,Y,Z, cut\_ideal and ,'clarity\_SI1' and cut\_Premium we got adj.R Squared value **0.913.**

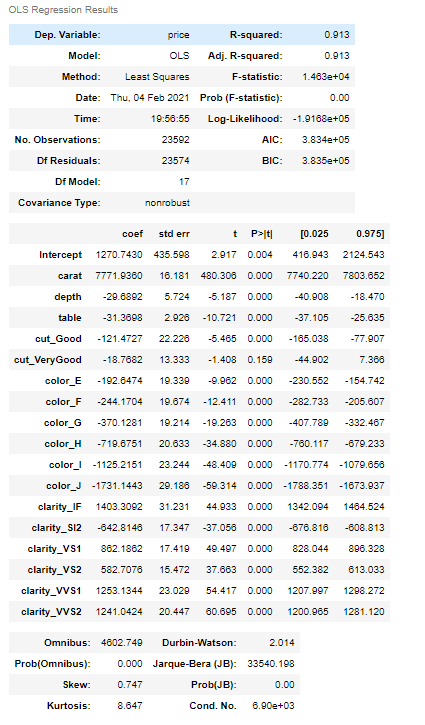


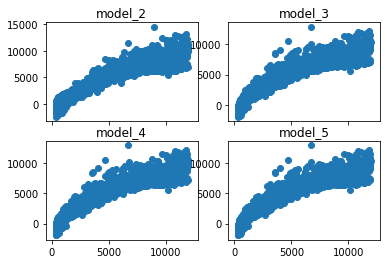
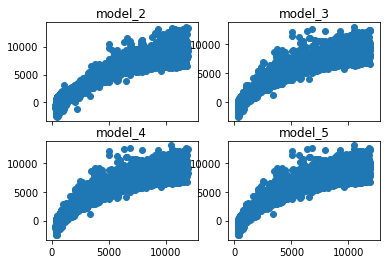
Figure : Final Model building

* + - Linear regression was applied and the following observations were noticed

|  |  |
| --- | --- |
| The coefficient for carat is | 3509.59701 |
| The coefficient for depth is | 34.80358105 |
| The coefficient for table is | 4.011278409 |
| The coefficient for x is | 51.23638972 |
| The coefficient for y is | 359.7823243 |
| The coefficient for z is | 650.6647277 |
| The coefficient for cut\_Good is | 326.5272714 |
| The coefficient for cut\_Ideal is | 497.4164669 |
| The coefficient for cut\_Premium is | 485.9633009 |
| The coefficient for cut\_Very Good is | 431.9485576 |
| The coefficient for color\_E is | -173.9106803 |
| The coefficient for color\_F is | -123.5264989 |
| The coefficient for color\_G is | -218.0401695 |
| The coefficient for color\_H is | -380.8231158 |
| The coefficient for color\_I is | -698.8529482 |

**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**

|  |  |  |
| --- | --- | --- |
|  | Training Data | Test Data |
| RMSE of model\_2 | 566304.033451395 | 579477.5076139265 |
| RMSE of model\_3 | 667178.6286775426 | 669830.8788656069 |
| RMSE of model\_4 | 667257.2140665484 | 669935.4110414303 |
| RMSE of model\_5 | 667290.5082177826 | 670032.7859784868 |



Training data Test Data

## Inference: Predictions derived business insights & recommendations

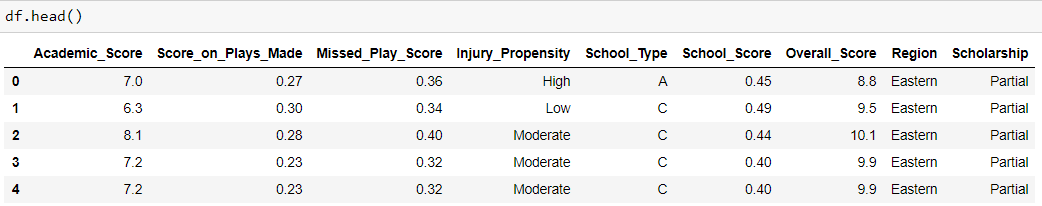
* + - **The five most important attributes for the price of Zirconia are:**
      * Carat is the most important attribute as when it increase the price of Zirconia increases by 3509
      * Clarity is the second most important attribute with its variants VVS2,IF,VVS1,VS1,VS2 and so on so forth which affects the price as they increase.
      * Color is the third most important attribute which affects the price in as they increase the price tends to decrease in the sequence : J,I,H,G,E,F
      * The forth most important attribute is the cut in sequence : Ideal, Premium,
      * Very Good & Good
      * The fifth & sixth important attribute are Z & Y which have positive impact on price. As they increase, the price increases.

**PROBLEM 2: LINEAR REGRESSION**

* + - 1. **EDA Observations:**
* The dataset provided to us as “**Football+Scholarship.csv**” which contains data of 6215 graduate students and 9 variables.
* Dataset does not have any null values.

**Importing the Dataset**

The dataset in question is imported in jupyter notebook using **pd.read\_csv ()** function and will store the dataset in “**claim\_df**”. The top 5 rows of the dataset are viewed using **pd.head ()** function.



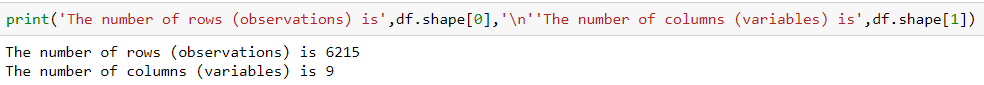
**Observations from Basic EDA:**

-Total Number of variables are 10.

-8 independent variable and one target variable – Scholarship.

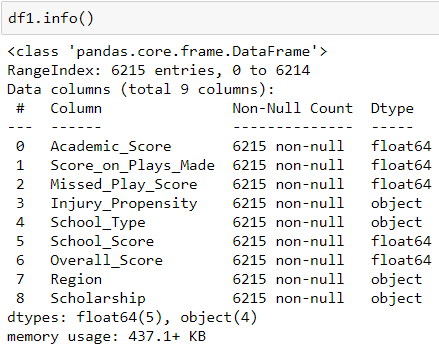
**Dimension of the Dataset**

Let us check the number of rows and the number of columns in the dataframe-



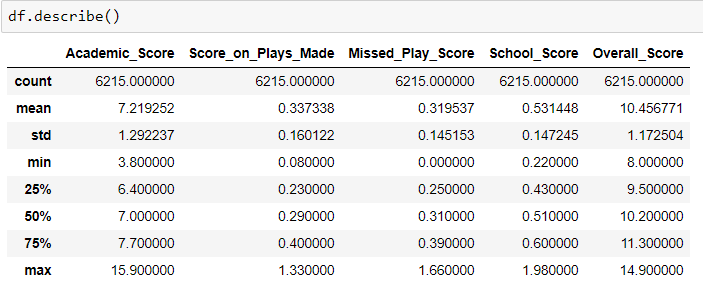
**Structure of the Dataset**

Structure of the Dataset can be computed using **df.info()** function.



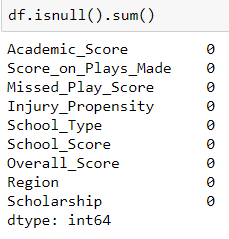
**Summary of the Dataset**

The summary of the dataset can be computed using pd.describe () function.



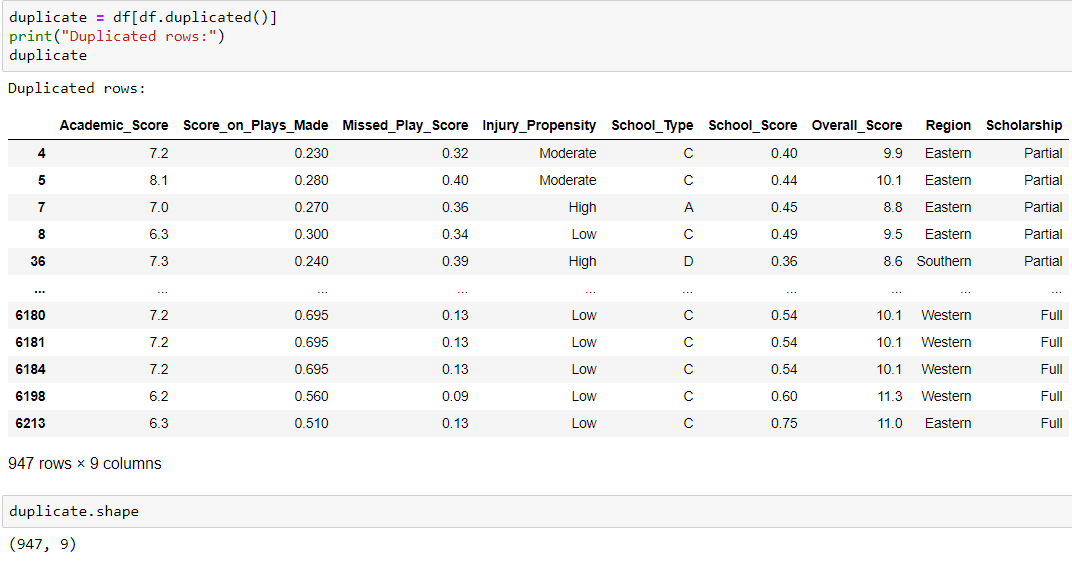
**Checking for Missing Values**

The missing values or “NA” needs to be checked and dropped from the dataset for the ease of evaluation and null values can give errors or disparities in results. Missing Values can be computed using **.isnull().sum()** function.



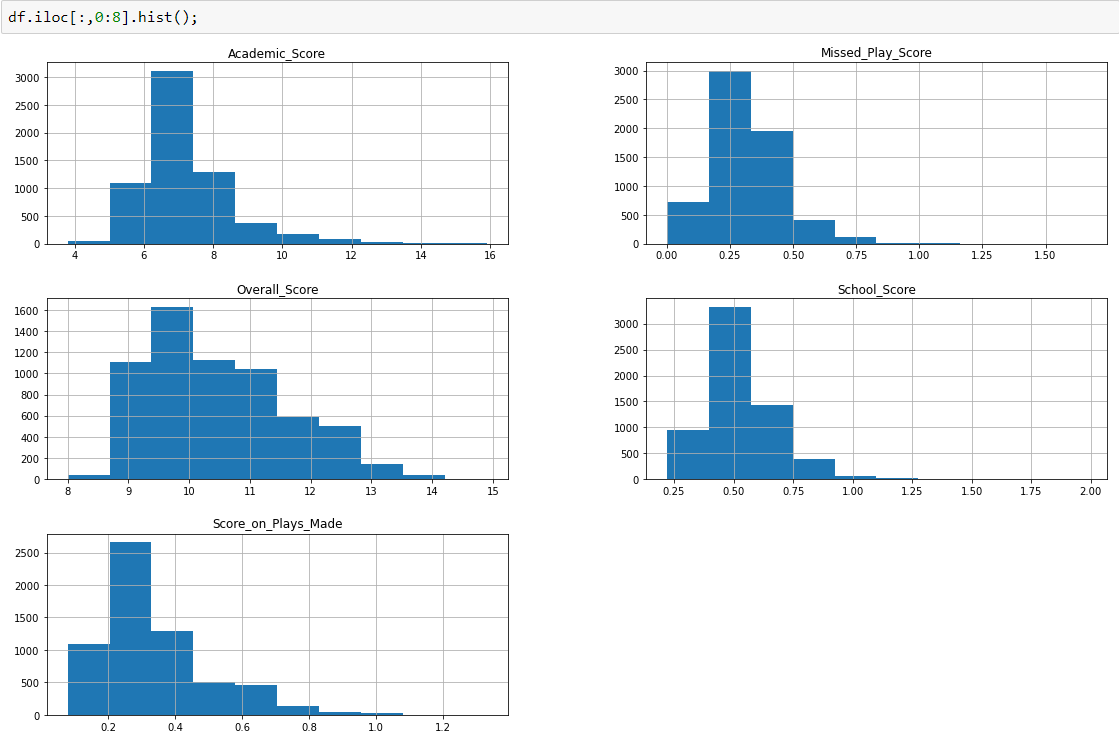
There are no missing values in the dataset hence we can go ahead with building the model.

**Duplicate Records:**

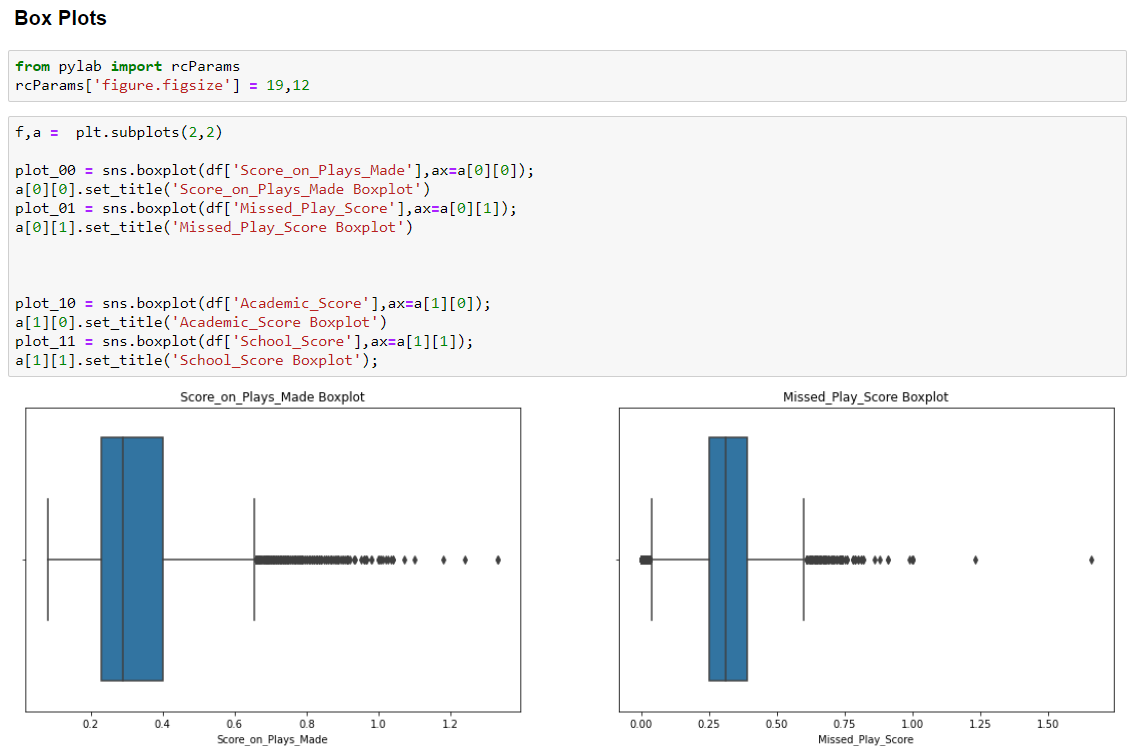


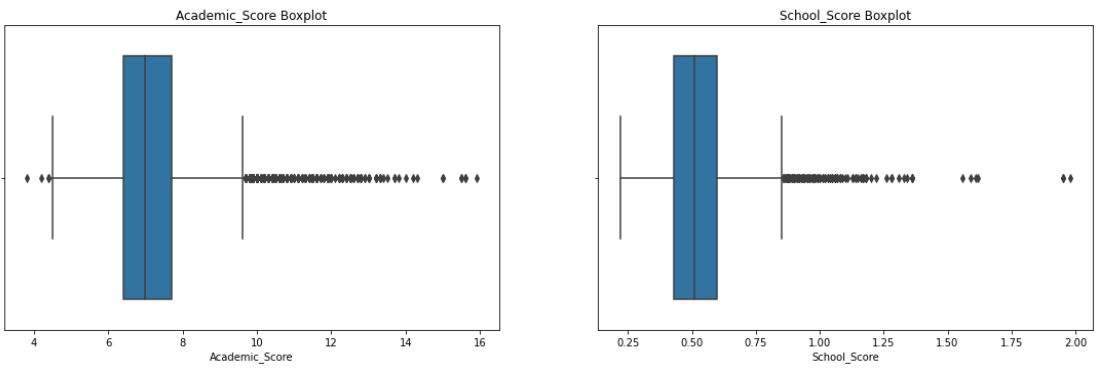
**Univariate Analysis**

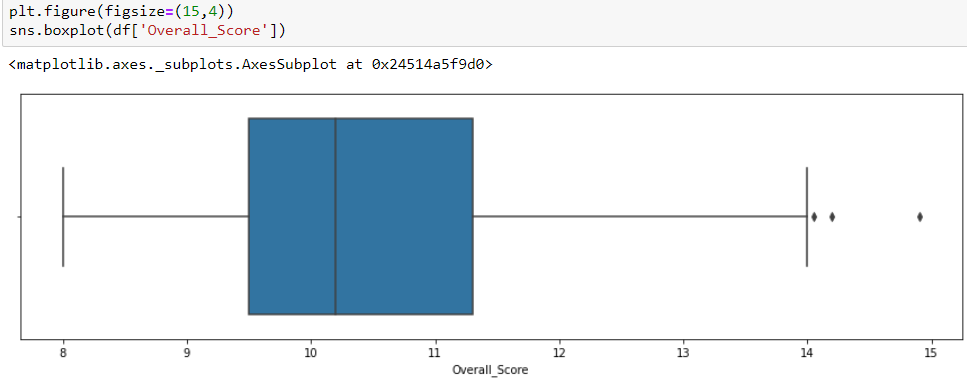
Histograms are plotted for all the numerical variables and it can be seen that We can see that data is rightly skewed.

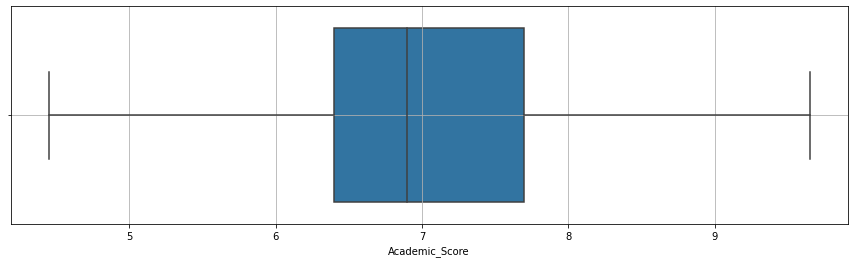


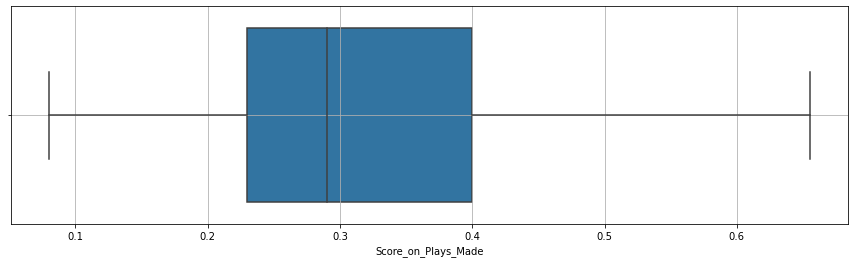
**Boxplots of Variables to check for Outliers**

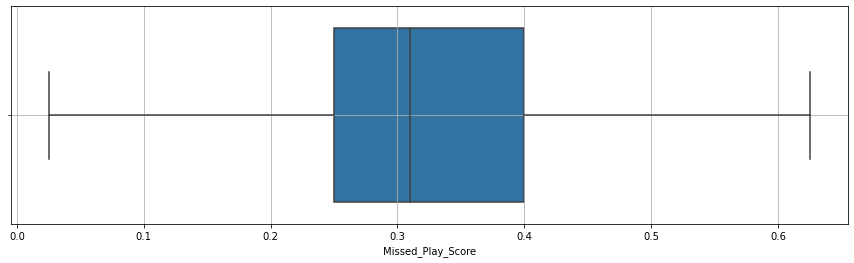


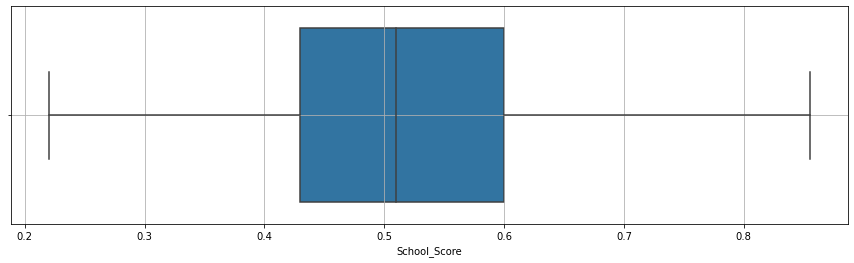


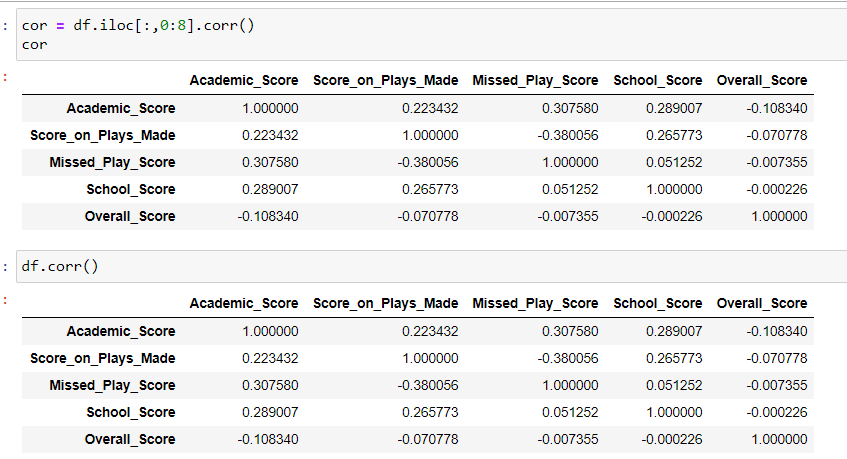












**Bivariate**

**Outliers Treatment**

Combined boxplot for all variables before outlier treatment–

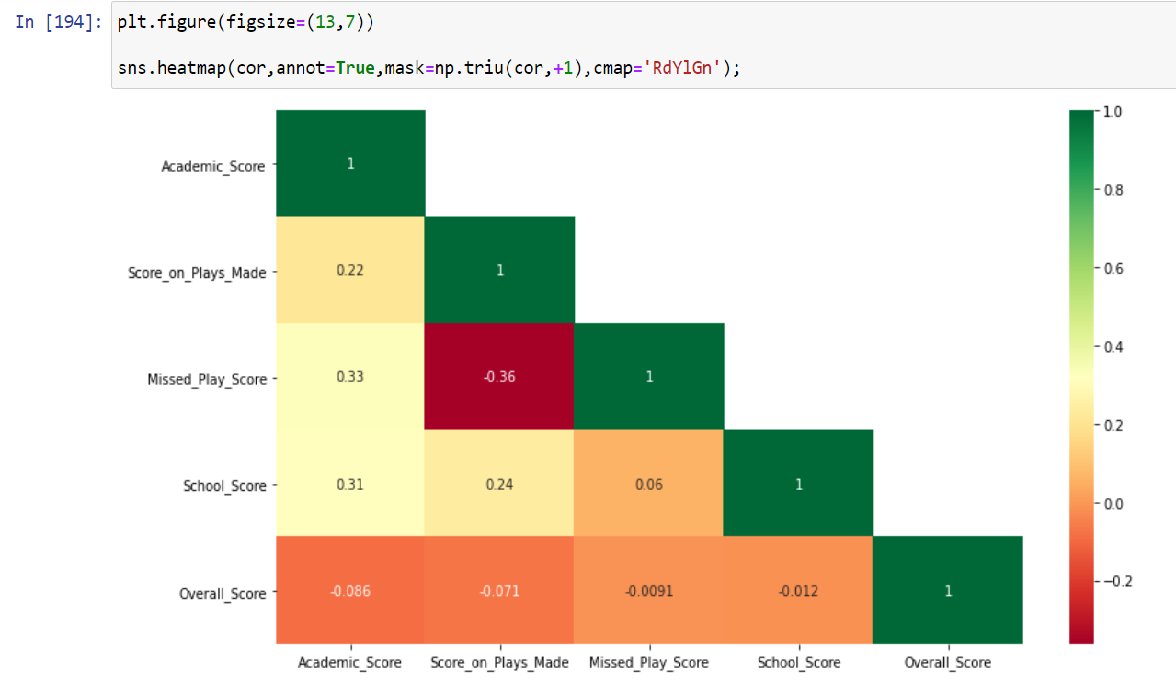
**Inference**: After plotting the Boxplots for all the numerical variables we can conclude that a very high number of outliers are present in the variables namely, **Score\_on\_Plays\_Made, Missed\_Play\_Score, School\_Score, Overall\_Score, Academic\_Score** which means that we need to treat these outlier values so as to proceed further with our model building and analysis as these values can create errors and can deviate from the actual results.

We can conclude from the above graphs that the majority of the customers doing a claim in our data belong to age group of 25-40 with the type of Tour Agency firm being Travel Agency, Channel being Online, Product name being Customised Plan and Destination being Asia.

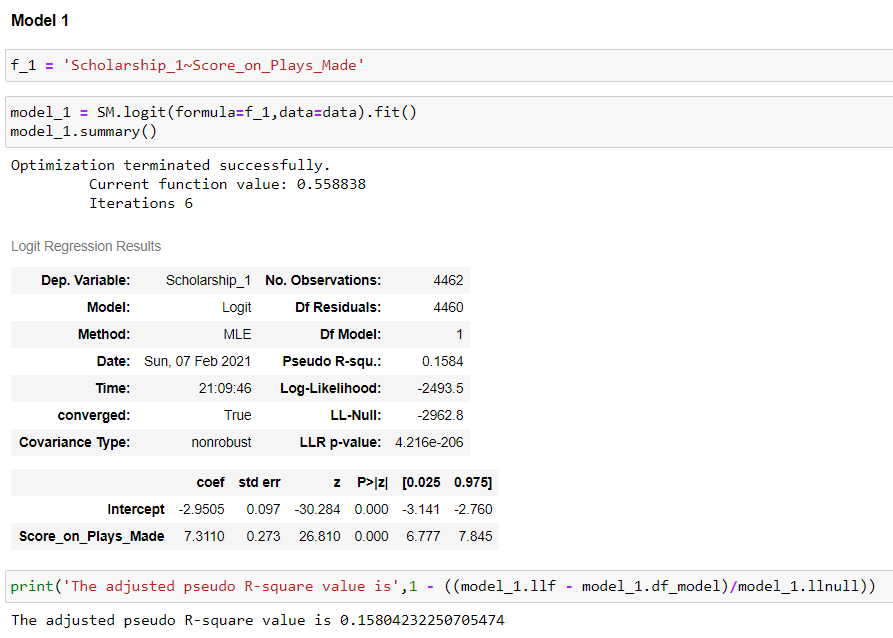
**Multivariate Analysis**

**Heat Map (Relationship Analysis)**

We will now plot a Heat Map or Correlation Matrix to evaluate the relationship between different variables in our dataset. This graph can help us to check for any correlations between different variables.



Interpretation from the above heat map – There is highest corelation between Academic Score and Missed Play Score. And also, Academic Score and School Score.

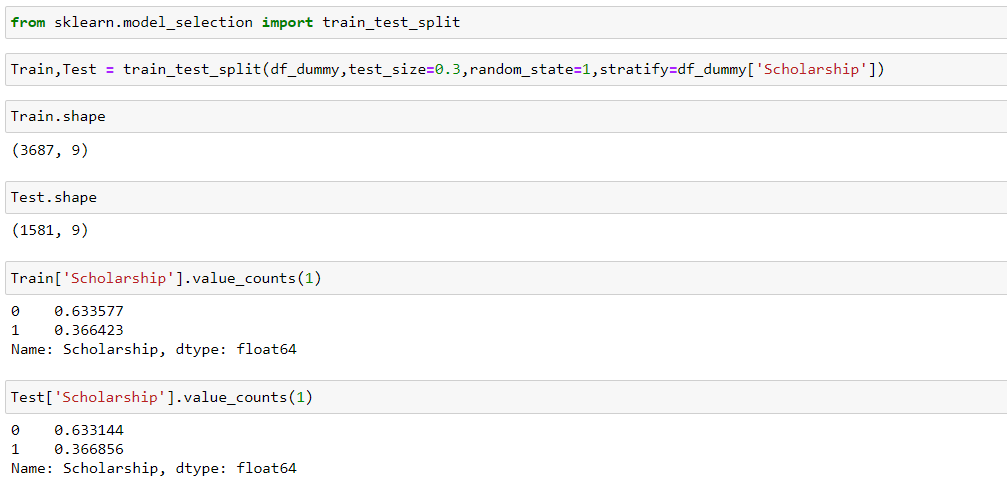
**2.2 Build various iterations of the Logistic Regression model using appropriate variable selection techniques for the full data. Compare values of model selection criteria for proposed models. Compare as many criteria as you feel are suitable.**

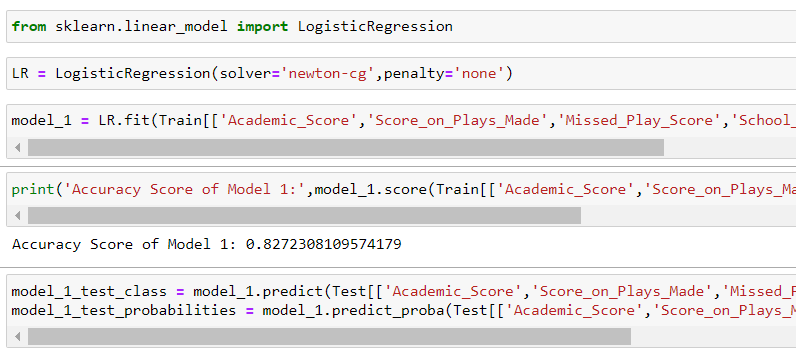
Optimization terminated successfully.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Current function value | Iterations | The adjusted pseudo R-square value is |
| Model\_1 | 0.558838 | 6 | 0.15804232250705474 |
| Model\_2 | 0.518559 | 6 | 0.21836631309626142 |
| Model\_3 | 0.473284 | 6 | 0.2828374003740155 |
| Model\_4 | 0.477047 | 6 | 0.27750807669713196 |
| Model\_5 | 0.484761 | 6 | 0.266227966543902 |
| Model\_6 | 0.484795 | 6 | 0.26651442016778726 |
| Model\_7 | 0.484823 | 6 | 0.266810445108563 |

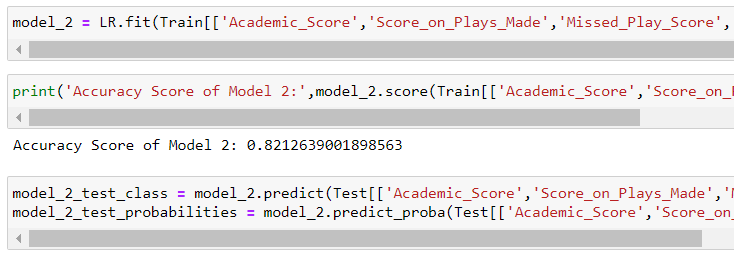
Codes are available in shared Python sheet.

**2.3 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.**

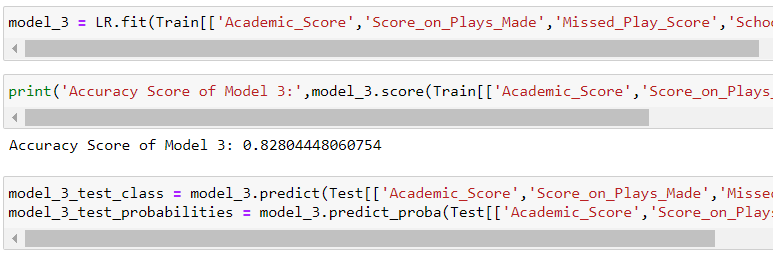
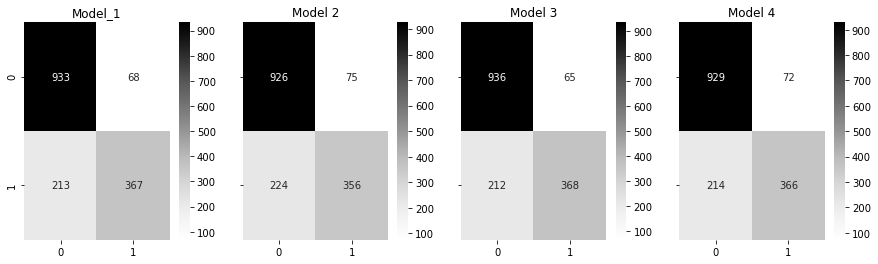


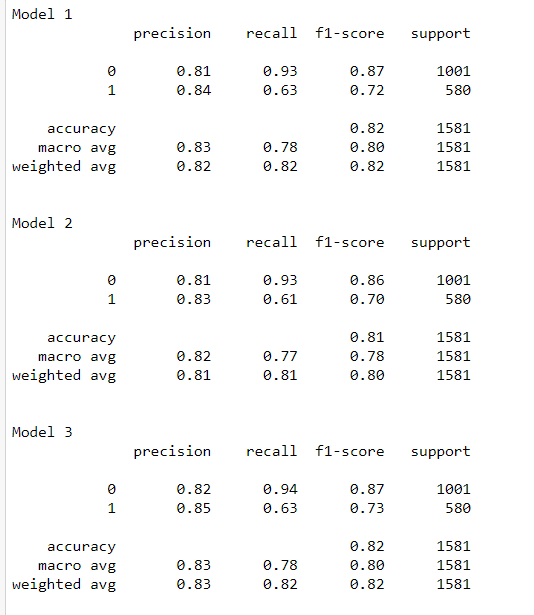
**Model 1:**

Model 2:

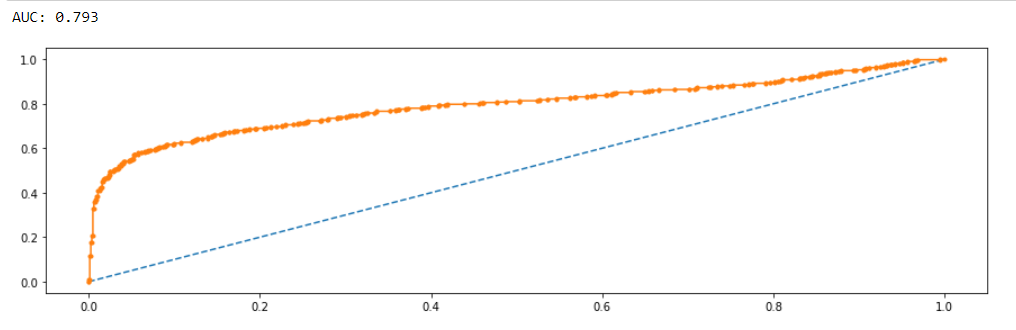


Model 3:

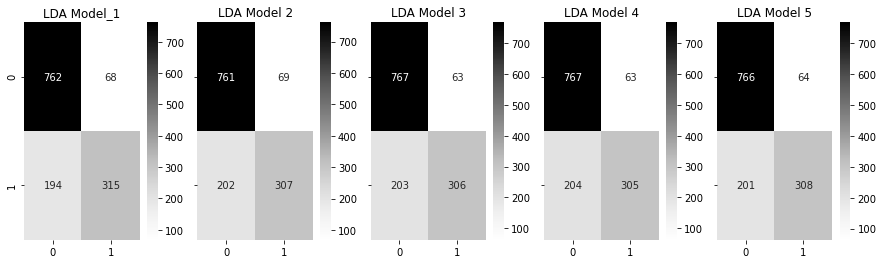


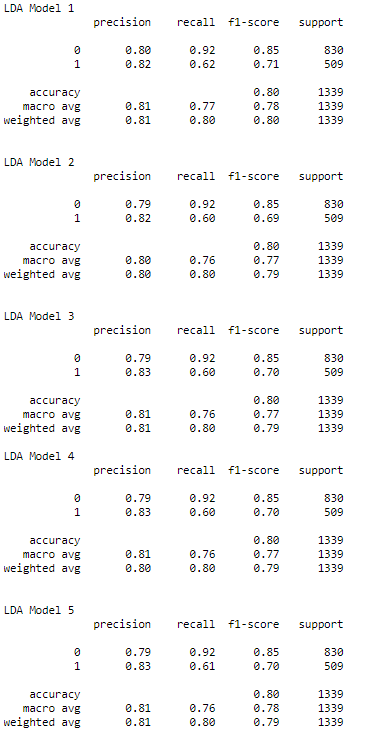


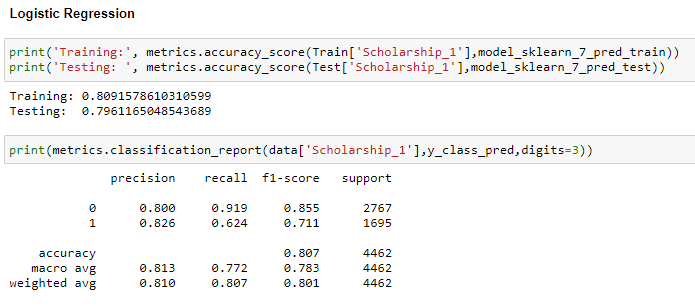
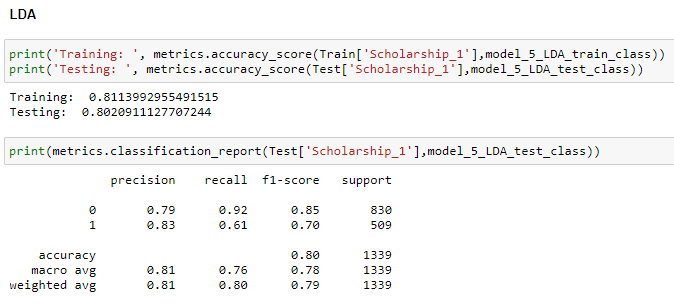
|  |  |  |
| --- | --- | --- |
|  | Training | Testing |
| Accuracy Score - Model 4 | 0.8178033941722702 | 0.7953696788648245 |
| Accuracy Score - Model 5 | 0.8097982708933718 | 0.79462285287528 |
| Accuracy Score - Model 6 | 0.8101184758245277 | 0.79462285287528 |
| Accuracy Score - Model 7 | 0.8091578610310599 | 0.7961165048543689 |



**2.4.** **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.**







**2.5 Business Report documenting the results and stating the actionable insights and the recommendations**

The Importance factors which are instrumental in winning a scholarship in collages are :

1. Score\_on\_Plays\_Made, School\_Score, School\_Type\_D, School\_Type\_C, Academic\_Score

2. In which **Score\_on\_Plays\_Made** having **highest** preference and Academin\_score has least score.

3. **Score\_on\_Plays\_Made** is the most importance attribute as when it increases probability of getting scholarship increases.

4. School\_Score, School\_Type\_D, School\_Type\_C, Academic\_Score which has positive impact on scholarship.