

**Project Report of**

**Predictive Modelling Module**

**Project Report**

Submitted to



**Submitted By**

**Group No. 5 Batch: 2021** **Location: Pune**

**Group Members**

1. **Niranjan Dhavan**
2. **Sagar Belagali**
3. **Nimit Kumar**
4. **Pravin Kumar**
5. **Takshay Sheetigar**
6. **Praveen Kulandia Arasu**
7. **Abhinav Pathak**

**Problem 1: Linear Regression**

**Problem Statement –**

**You are hired by a company named Gemstone Co Ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of approximately 27,000 pieces of cubic zirconia (which is an inexpensive synthesized diamond alternative with similar qualities of a diamond).**

**Objective**

**Objective is 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.**

**cubic\_zirconia Data**

The data dictionary is given below.

1. Carat - Carat weight of the cubic zirconia

2. Cut - Describes the cut quality of the cubic zirconia. Quality is in increasing order: Fair, Good, Very Good, Premium, Ideal.

3. Colour - Colour of the cubic zirconia.

4. 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

5. Depth - The Height of a cubic zirconia piece, measured from the Culet to the table, divided by its average Girdle Diameter.

6. Table - The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.

7. Price - Price of the cubic zirconia.

8. X - Length of the cubic zirconia in mm.

9. Y - Width of the cubic zirconia in mm.

10. Z - Height of the cubic zirconia in mm.

**Performing exploratory data analysis on the dataset. Showcasing some charts & graphs.**

1. Loading the data set- We will be loading the “cubic\_zirconia.csv” file using pandas library in python. For this, we will be using read\_csv file.

2. The head function will tell us the top head records in the data set. By default, python shows you only the top 5 record.



3. The tail function will tell us the last entries records in the data set. By default, python shows you only the last 5 records. Let’s check tail for the totals/subtotals if any. The cubic zirconia data dataset doesn’t contain any total/subtotals. Further basis inspection it was identified that Unnamed: 0 is useless column lets drop the Unnamed: 0 column.



4. The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The data set has 26967 observations and 10 variables in the data set.

5. info() is used to check the Information about the data and the datatypes of each respective attribute.



Figure 1

* Looking at the data in the head function and in info, we come to know that the variables comprise of float, object and integer data types. sklearn in Python does not take the input of object data types while building Linear Regression models. So, we need to convert these variables into some numerical form.
* Basis Figure 1 we can see that there are three object type variables (cut, colour & clarity) which has the object data types which we need to convert into numerical form. Since the variable cut & clarity are in ordinal range, we are replacing the categorical variables with the numbers. Further we shall perform one hot encoding for colour variable. The conversion is done before running the linear regression model.
* Further basis above figure 1 we can see that there are 697 null values in the depth variable. Let’s go ahead and drop the null values from the dataset. Post Dropping the null values we can see that we have lost the 2.58% of data loss. (i.e. Left with 26270 entries).
* Also, it was identified that there were 34 duplicated rows in the dataset. Will go ahead and drop the duplicated values from the dataset.

6. The described method will help to see how data has been spread for numerical values. We can clearly see the minimum value, mean values, different percentile values, and maximum values for the Income data set.



Figure 2

* Basis Figure 2 We can see that min value as 0 in x, y & z. Where x, y, & z means the Length, width & height as 0. Where it cannot be zero.
* We can see there are 8 rows with Dimensions 'Zero'. We will Drop them as it seems better choice instead of filling them with any of Mean or Median.
* Post Cleaning the data we can see that we have lost the 2.74% of data. (i.e. Left with 26228 entries)

**Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. We performed uni-variate and bi-variate analysis to get a better overview and to find outliers in our dataset. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it doesn’t affect the performance of our model.

**Uni-Variate Analysis.**

***Target / Predictor Variable Analysis - price.***

Chart

Description automatically generated

Figure 3

* The Price Variable distribution seems to be Highly Left-skewed.
* Basis skewness value we can see that distribution is highly skewed.
* Basis box plot we can see that there are outliers in the variable. Let’s treat the outliers in the further process.

***Response / Dependent Variable Analysis***

*1).* ***Analyzing Feature: cut***

*Chart, histogram

Description automatically generated*

Figure 4

* Basis above figure 4 we can see that ideal cut diamond have comparably huge demand in the market followed by the premium cut.
* Basis above figure 4 we can also infer that quality of cut changes the demand for the product goes down.

*2).* ***Analyzing Feature: color***

***Chart, bar chart

Description automatically generated***

Figure 5

* Basis above figure we can see that G Colour diamond have comparably huge demand in the market as it might be lowly priced as compare to Colourless cibic zirconia.
* Basis above figure we can also infer that colourless cubic zirconia are high priced in market as compared to near colourless cubic zirconia.

*3).* ***Analyzing Feature: clarity***

***Chart, histogram

Description automatically generated***

Figure 6

* Basis above figure we can see that SI1 Clarity cubic zirconia have comparably huge demand in the market followed by the VS2 as they might be prices lower as compare to flawless cubix zirconia.
* Basis above figure we can also infer that IF cubic zirconia Has lower demand in the market due to higher price bracket.

*4).* ***Analyzing Feature: depth***

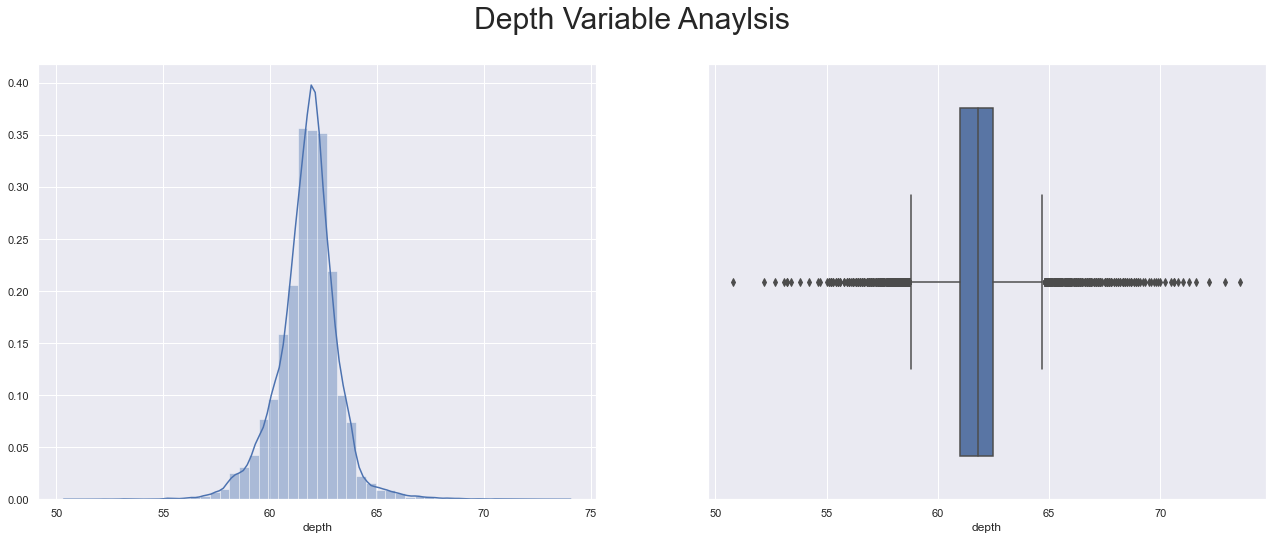
****

Figure 7

* Basis above figure we can see that Data distribution for 'depth' variable is slightly left-skewed.
* Basis skewness value we can see that distribution is approximately symmetric.
* Basis box plot we can see that there are outliers in the variable. Let’s treat the outliers in the further process.

*5).* ***Analyzing Feature: x***

Chart, histogram

Description automatically generated

Figure 8

* Basis above figure we can see that Data distribution for 'x' variable is Right-skewed.
* Basis skewness value we can see that distribution is approximately symmetric.
* Basis box plot we can see that there are outliers in the variable. Let’s treat the outliers in the further process.

*5).* ***Analyzing Feature: y***

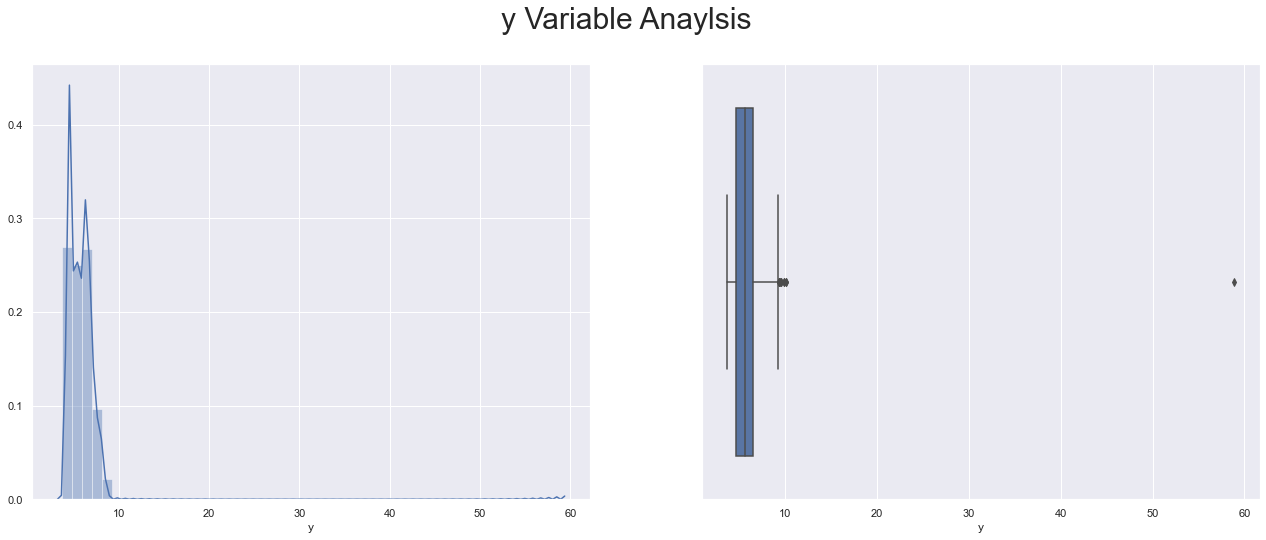
******

Figure 9

* Basis above figure we can see that Data distribution for 'y' variable is highly left-skewed.
* Basis skewness value we can see that distribution is highly skewed.
* Basis box plot we can see that there are couple of outliers in the variable. Let’s treat the outliers in the further process.

*6).* ***Analyzing Feature: z***

***Chart

Description automatically generated***

Figure 10

* Basis above figure we can see that Data distribution for 'z' variable is Right-skewed.
* Basis skewness value we can see that distribution is approximately symmetric.
* Basis box plot we can see that there are couple of outliers in the variable. Let’s treat the outliers in the further process.

**Bi-Variate Analysis.**

Factors influencing price of the property*.*

***Analyzing Feature: Price Vs carat***

******

Figure 11

* *Basis above figure we can see that as the carat cubic zirconia increases the prices increases.*

***Analyzing Feature: Price Vs x (Length)***



Figure 12

* Basis above figure we can see that as the cubic zirconia Length of the cubic zirconia in mm increases the prices increases.

***Analyzing Feature: Price Vs z (Height)***

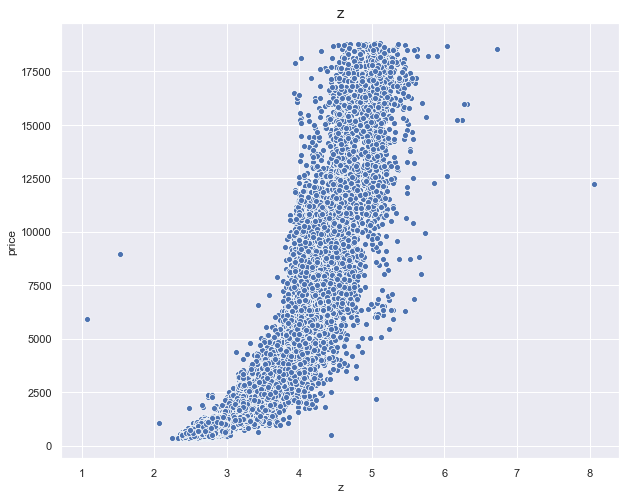


Figure 13

* Basis above figure we can see that as the cubic zirconia Height of the cubic zirconia in mm. increases the prices increases
* ***Analyzing Feature: Price Vs cut***

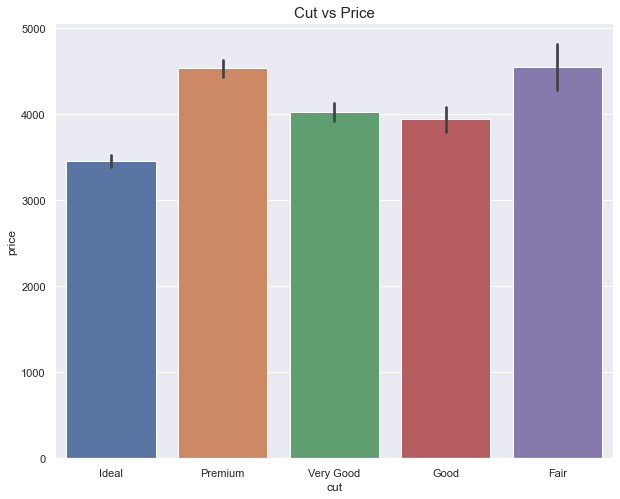


Figure 14

* Basis above figure we can see that fair cut is being highly priced in the market.
* Basis above figure we can see that premium cut has second highly price cubic zirconia in the market.

***Analyzing Feature: Price Vs color***

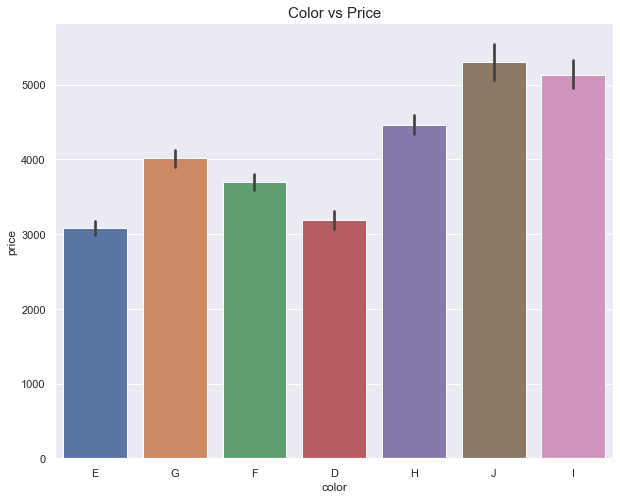
******

Figure 15

* Basis above figure we can see that J Colour is being highly priced in the market.
* Basis above figure we can see that Near colourless cubic zirconia has high prices in the market.

***Analyzing Feature: Price Vs clarity***

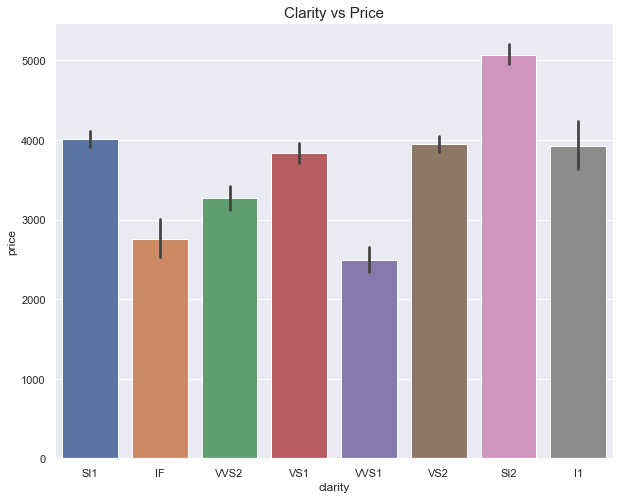


Figure 16

* Basis above figure we can see that SI1 Colour is being highly priced in the market. Followed by SI1
* Basis above figure we can see that Small inclusions cubic zirconia, very small inclusions have higher prices as compared to the Flawless cubic zirconia.

**Multi-Variate Analysis.**

***Analyzing Feature: Price Vs Carat***



Figure 17

* Basis above figure we can see that ideal cut cubic zirconia are Highly priced in the market irrespective of the carat.
* Basis above figure we can see that Fair cut cubic zirconia are lowly priced in the market as compared to the other cuts.
* Basis above figure we can see that premium & very good cut cubic zirconia are averagely priced in the market.

***Analyzing Feature: Price Vs Carat***

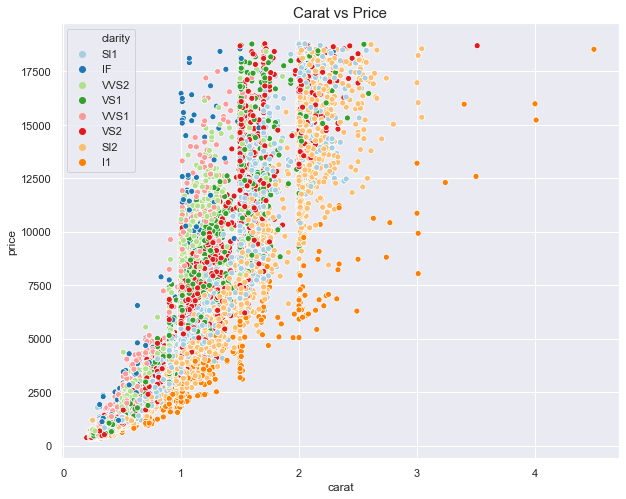
******

Figure 18

* Basis above figure we can see that I1 clarity cubic zirconia are lower priced in the market irrespective of the carat.
* Basis above figure we can see that SI2 clarity cubic zirconia are priced Above I1 clarity cubic zirconia in the market irrespective of the carat.
* Basis above figure we can see that IF cut cubic zirconia are Highly priced in the market as compared to the other cuts despite of being the lower carats (i.e. Below 2carat).

***Analyzing Feature: Price Vs Carat***

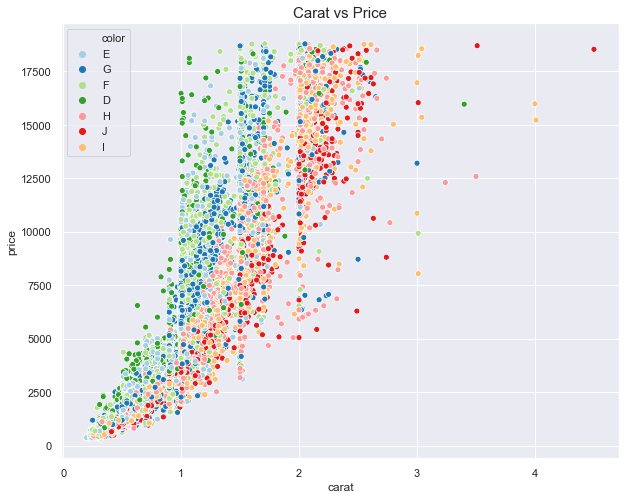


Figure 19

* Basis above figure we can see that H & I Colour cubic zirconia are lower priced in the market irrespective of the carat.
* Basis above figure we can see that D, E & F Colour cubic zirconia are Highley priced in the market.

**Data Pre-Processing**

* Basis analysis of above we can see that there are outliers in the variables. Let’s go ahead and treat the outliers.
* As mentioned above couple of variables are categorical variables, we shall Replace the categorical variables with the numbers as its ordinal range. (i.e. For Variable cut & clarity).
* sklearn in Python does not take the input of object data types when building linear regression model. So, we need to convert these variables into some numerical form. We shall perform one hot encoding for them (i.e. colour variable). Post Data Pre-processing the head of data looks like. 
* We are not scaling the dataset. And proceeded with dataset as it is.

**Linear Regression**

Linear regression is a way to identify a relationship between two or more variables. We use this relationship to predict the values for one variable for a given set of value(s) of the other variable(s). The variable, which is used in prediction is termed as independent/explanatory/regressor variable where the predicted variable is termed as dependent/target/response variable. Linear regression assumes that the dependent variable is linearly related to the estimated parameter(s).

𝑦 = 𝑐+𝑚𝑥

In machine learning and regression literature the above equation is used in the form:

𝑦 = 𝑤0+𝑤1𝑥

Where w0 is intercept on y-axis, w1 is slope of line, x is an explanatory variable and y is the response variable.

1. **Descriptive Linear Regression**
2. **Predictive Linear Regression**

**Descriptive Linear Regression**

* Descriptive Linear Regression – Main Objective of Descriptive Linear Regression is to understand the relation between Features / Variables.
* In Descriptive Linear Regression we need to look after assumptions i.e. VIF Values (for Multicollinearity). And we select variables basis significance of p value. Whereas for descriptive type assumptions and metric both stand important.
* For Descriptive type we don’t divide the dataset into train and test. Also, We Use statmodel for as python library for Descriptive Linear Regression.

**Firstly, we will use descriptive linear regression to understand which all variables are significant variables that impact the price variable.**

**Descriptive linear regression Models**

**Model 1 (Using All the variables)**

* Firstly, we will import statsmodels.formula.api as SM model
* In first model will run the model using all the variables (i.e. price~carat+cut+clarity+depth+table+x+y+z+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J).



* Basis above model we can that adjusted r square is 93.3%. Which seems to be good. Also Checking the p values as the p values seems to significant variables. But the condition number seems to higher side which in turn shows that there is multi-collinearity.
* Let’s check the VIF values for the all the variables.



**Basis above figure we can see that there are many variables having high VIF which shows that there is multicollinearity in the independent variables. Basis VIF scores above as there is multicollinearity Lets consider the threshold of 5 and start dropping the variables of VIF above 5 one by one basis the variable which are less co-related with price variable.**

**Model 2 (Model with Dropping high infinity VIF values (i.e. clarity Variable).**

* In second model will run the model using all the variables (i.e. price~carat+cut+depth+table+x+y+z+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J).



* Basis above model we can that adjusted r square is 90.1%. Which seems to be good but dropping compared to above model 1. Also Checking the p values al the p values seems to significant variables. Except variable depth & color\_F. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to higher side which in turn shows that there is multi-collinearity.
* Let’s check the VIF values.



* Still we see that there are variables with VIF value above 5. Let’s keep on dropping the variable above VIF 5.

**Model 3 (Model with Dropping high infinity VIF values (i.e. clarity & cut)).**

* In third model will run the model using all the variables (i.e. price~carat+depth+table+x+y+z+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J).



* Basis above model we can that adjusted r square is 89.99%. Which seems to be good but dropping compared to above model 2. Also Checking the p values al the p values seems to significant variables. Except variable color\_F. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to higher side which in turn shows that there is multi-collinearity.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
|  | **variables** | **VIF** |
| **0** | carat | 123.445387 |
| **1** | table | 717.346755 |
| **2** | depth | 995.217448 |
| **3** | x | 10258.65012 |
| **4** | y | 9220.487106 |
| **5** | z | 2970.85611 |
| **6** | color\_E | 2.467627 |
| **7** | color\_F | 2.414572 |
| **8** | color\_G | 2.70249 |
| **9** | color\_H | 2.26844 |
| **10** | color\_I | 1.893794 |
| **11** | color\_J | 1.49442 |

* Still we see that there are variables with VIF value above 5. Let’s keep on dropping the variable above VIF 5.

## Model 4 (Model with Dropping high infinity VIF values (i.e. clarity, cut & Depth).

* In fourth model will run the model using all the variables (i.e. price~carat+table+x+y+z+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J).



* Basis above model we can that adjusted r square is 89.99%. Which seems to be good but dropping compared to above model 3. Also Checking the p values al the p values seems to significant variables. Except variable color\_F. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to higher side which in turn shows that there is multi-collinearity.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
|  | **variables** | **VIF** |
| **0** | carat | 86.28756 |
| **1** | table | 266.136814 |
| **2** | x | 9980.010927 |
| **3** | y | 9194.625567 |
| **4** | z | 1571.724914 |
| **5** | color\_E | 2.46172 |
| **6** | color\_F | 2.407718 |
| **7** | color\_G | 2.687687 |
| **8** | color\_H | 2.254623 |
| **9** | color\_I | 1.87998 |
| **10** | color\_J | 1.489809 |

* Still we see that there are variables with VIF value above 5. Let’s keep on dropping the variable above VIF 5.

## Model 5 (Model with Dropping high infinity VIF values (i.e. clarity, cut, Depth & table)).

* In Fifth model will run the model using the variables (i.e. price~carat+x+y+z+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.7%. Which seems to be dropping down compared to above models. Also Checking the p values al the p values seems to significant variables. Except variable color\_F & color\_G. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to be decreasing compared to above models.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
|  | **variables** | **VIF** |
| **0** | carat | 11.324381 |
| **1** | x | 9570.4914 |
| **2** | y | 9193.516103 |
| **3** | z | 1564.980819 |
| **4** | color\_E | 2.433057 |
| **5** | color\_F | 2.394265 |
| **6** | color\_G | 2.671584 |
| **7** | color\_H | 2.23459 |
| **8** | color\_I | 1.860607 |
| **9** | color\_J | 1.478309 |

* Still we see that there are variables with VIF value above 5. Let’s keep on dropping the variable above VIF 5.

## Model 6 Model with Dropping high infinity VIF values (i.e. clarity, cut, Depth, table & z Variable).

* In Sixth model will run the model using the variables (i.e. price~carat+x+y+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.6%. Which seems to be dropping down compared above models. Also Checking the p values al the p values seems to significant variables. Except variable color\_F & color\_G. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to be decreasing compared to above models.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
|  | **variables** | **VIF** |
| **0** | carat | 11.296452 |
| **1** | x | 9026.82853 |
| **2** | y | 8939.023996 |
| **3** | color\_E | 2.431451 |
| **4** | color\_F | 2.392727 |
| **5** | color\_G | 2.667254 |
| **6** | color\_H | 2.229572 |
| **7** | color\_I | 1.857403 |
| **8** | color\_J | 1.476089 |

* Still we see that there are variables with VIF value above 5. Let’s keep on dropping the variable above VIF 5.

## Model 7 Model with Dropping high infinity VIF values (i.e. clarity, cut, Depth, table, z & x Variable Variable).

* In seventh model will run the model using the variables (i.e. price~carat+y+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.3%. Which seems to be dropping down compared above models. Also Checking the p values all the p values seems to significant variables. Except variable color\_F & color\_G. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to be decreasing compared to above models.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
| **variables** | **VIF** |  |
| **0** | carat | 11.001789 |
| **1** | y | 22.930637 |
| **2** | color\_E | 2.430476 |
| **3** | color\_F | 2.391847 |
| **4** | color\_G | 2.665555 |
| **5** | color\_H | 2.226383 |
| **6** | color\_I | 1.857345 |
| **7** | color\_J | 1.476087 |

* Still we see that there are variables with VIF value above 5. Let’s keep on dropping the variable above VIF 5.

## Model 8 - Model with Dropping high infinity VIF values (i.e. clarity, cut, Depth, table, z & y Variable Variable).

* In eighth model will run the model using the variables (i.e. price~carat+color\_E+color\_F+color\_G+color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.2%. Which seems to be coming down from above models. Also Checking the p values al the p values seems to significant variables. Except variable color\_G. Let’s check p values significance once we drop all the variable with high VIF Values. The condition number seems to be decreasing compared to above models.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
| **variables** | **VIF** |  |
| **0** | carat | 3.405663 |
| **1** | color\_E | 1.318681 |
| **2** | color\_F | 1.377663 |
| **3** | color\_G | 1.50346 |
| **4** | color\_H | 1.499119 |
| **5** | color\_I | 1.429428 |
| **6** | color\_J | 1.277313 |

* Now we see that all the variables are within VIF value of 5. Let’s stop dropping.

**Let’s Check features based on high P Value**

For the 𝑡−𝑠𝑡𝑎𝑡𝑖𝑠𝑡𝑖𝑐 for every co-efficient of the Linear Regression the null and alternate Hypothesis is as follows:

**𝐻0 : The variable is significant.**

**𝐻1: The variable is not significant.**

Lower the p-value for the t-statistic more significant are the variables.

## Model 9 - Model with Dropping insignificant P Values (i.e. clarity, cut, Depth, table, z, x, y & color\_F Variable).

* In Ninth model will run the model using the variables (i.e. price~carat+color\_E+color\_G+color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.2%. Which seems to be dropping down compared above models.
* Let’s check the VIF values for the all the variables except clarity variable.

|  |  |  |
| --- | --- | --- |
| **variables** | **VIF** |  |
| **0** | carat | 2.472058 |
| **1** | color\_E | 1.23132 |
| **2** | color\_G | 1.365445 |
| **3** | color\_H | 1.362294 |
| **4** | color\_I | 1.311707 |
| **5** | color\_J | 1.201292 |

**Model 10- Model with Dropping insignificant P Values (i.e** **clarity, cut, Depth, table, z, x, y, color\_F & color\_G Variable).**

* In Tenth model will run the model using the variables (i.e. price~carat+color\_E+color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.2%. Which seems to be coming down from above models in decimals.
* Let’s check the VIF values.

|  |  |  |
| --- | --- | --- |
| **variables** | **VIF** |  |
| **0** | carat | 1.810442 |
| **1** | color\_E | 1.16941 |
| **2** | color\_H | 1.26533 |
| **3** | color\_I | 1.228283 |
| **4** | color\_J | 1.147419 |

## Model 11- Model with Dropping insignificant P Values (i.e clarity, cut, Depth, table, z, x, y, color\_F & color\_G Variable & color\_E)).

* In Eleventh model will run the model using the variables (i.e. price~carat+ color\_H+color\_I+color\_J)



* Basis above model we can that adjusted r square is 89.2%. Which seems to be coming down from above models in decimals.
* Let’s check the VIF values.



**Model Evaluation.**

|  |  |  |
| --- | --- | --- |
| **model\_name** | **model\_perf** | **Adjusted R square** |
| **0** | All Variables | 0.93322 |
| **1** | Dropping clarity | 0.90105 |
| **2** | Dropping clarity & cut | 0.899229 |
| **3** | Dropping clarity, cut & depth | 0.899196 |
| **4** | Dropping clarity, cut, Depth & table | 0.897227 |
| **5** | Dropping clarity, cut, Depth, table & z | 0.895642 |
| **6** | Dropping clarity, cut, Depth, table, z & x | 0.892909 |
| **7** | Dropping clarity, cut, Depth, table, z, x & y | 0.891572 |
| **8** | Dropping clarity, cut, Depth, table, z, x, y & color\_F | 0.891557 |
| **9** | Dropping clarity, cut, Depth, table, z, x, y, color\_F & color\_G | 0.891553 |
| **10** | Dropping clarity, cut, Depth, table, z, x, y, color\_F & color\_G Variable & color\_E | 0.891533 |

**Inference –** Basis above iterations we can see that model No 10 seems to be giving decent results compared to model 1 (which is including all the variables). Also Model no 11 is the model free from multi-collinearity & following all the assumptions check.

Basis Above descriptive linear regression we can say that below variables are the important variables in predicting the prices of Cubic zirconia.

* carat
* color\_H
* color\_I
* color\_J

* We will use Model 1 and Model 11 to predict and check the model evaluation.
* We Have chosen Model 11 because, it has a high Adjusted R Square, with least number of features. Also, for comparison we have taken model no 1. Which includes all the variables.

**Model 1 Predictions & Model 11 Predictions**

**Chart, scatter chart

Description automatically generated**

Figure 20

**Distplot of Residuals for Model 1 & Model 11**

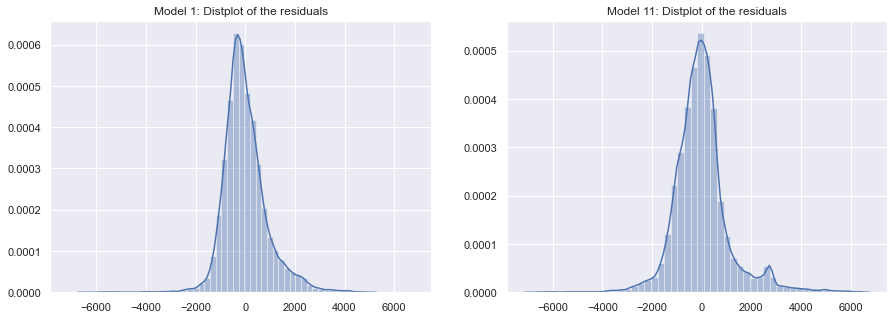
****

Figure 21

**Boxplot of residuals for Model 1 & Model 11.**

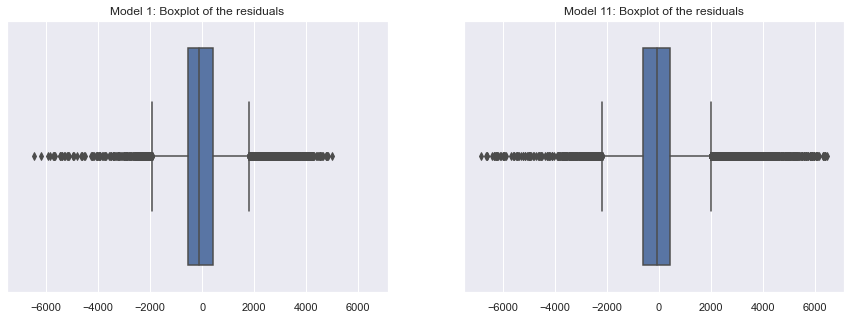
****

Figure 22

**RMSE Scores for Model 1 & 11**

|  |  |
| --- | --- |
|  | **RMSE Score** |
| **Model 1** | **895.0999417** |
| **Model 11** | **1140.985589** |

**Predictive Linear Regression**

* Predictive Linear Regression – Main Objective of Predictive Linear Regression is predictive values for Features / Variables.
* In predictive Linear Regression we no need to look after assumptions. Whereas for predictive type assumptions are not important only metric is important.
* For predictive type we divide the dataset into train and test. Also, We Use sklearn & statmodel for as python library for predictive Linear Regression.

**Predictive Approach using the Models 1, Model 10 & Model 11.**

* from sklearn.linear\_model import LinearRegression
* Splitting the data into the dependent and independent variables.
* from sklearn.model\_selection import train\_test\_split
* Splitting the data into train (70%) and test (30%).
* Using only Model 1 variables to build the model on the training data and predict on the training as well as test data. Results for the model are mentioned below.
* Using only Model 10 variables to build the model on the training data and predict on the training as well as test data. Results for the model are mentioned below.
* Using only Model 11 variables to build the model on the training data and predict on the training as well as test data. Results for the model are mentioned below.

**Model 1, Model 10 & Model 11 Train Prediction Scattered plot.**

Graphical user interface, application

Description automatically generated

Figure 23

**Model 1, Model 10 & Model 11 Test Prediction Scattered plot.**

Graphical user interface

Description automatically generated

Figure 24

**Model Results for Train & Test Dataset.**

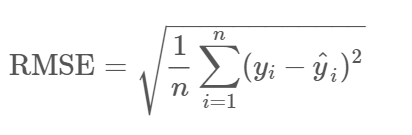
A good fitting model is one where the difference between the actual and observed values or predicted values for the selected model is small and unbiased for train, validation and test data sets.

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model:

* **Mean Squared Error (MSE).**
* **Root Mean Squared Error (RMSE).**
* **Mean Absolute Error (MAE)**

Here in the problem we are using RMSE Score as Metric -

* The most commonly used metric for regression tasks is **RMSE (root-mean-square error)**. This is defined as the square root of the average squared distance between the actual score and the predicted score:

****

* The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.



**Inference** – Basis Model 1 we can see that it has a lowest RMSE Score for Train & test data when compared to the other two models (i.e. Model 10 & Model 11). Lower values of RMSE for Train & test data indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction. The model 1 can be used when the prediction is important irrespective of checking which variables are important.

Whereas Model No 10 & Model 11 are the models which are free from multi-collinearity. Also, in case if client wants to understand which variables in dataset play an important role in changing the prices of cubic zirconia then we can go ahead with model 10 & Model No 11.

**Problem 2 – Logistic Regression & Linear Discriminant Analysis (LDA)**

**Problem Statement -**

**You are hired by a sports analysis agency to understand the selection process of high school football players into college with a full or partial scholarship. You are provided details of 6215 high school graduates who have been inducted into 4-year degree colleges with either full or partial scholarships.**

**Objective**

**Objective is 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.**

[**Football Scholarship**](https://olympus.greatlearning.in/courses/10652/files/2170918/download?verifier=40nQ7DzuxDM9jqwevkbNsFSo3H8BTFig4yvrhFmE&wrap=1) **Data**

The data dictionary is given below.

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

**Performing exploratory data analysis on the dataset. Showcasing some charts & graphs.**

1. Loading the data set- We will be loading the “Football+Scholarship.csv” file using pandas library in python. For this, we will be using read\_csv file.

2. The head function will tell us the top head records in the data set. By default, python shows you only the top 5 record.



3. The tail function will tell us the last entries records in the data set. By default, python shows you only the last 5 records. Let’s check tail for the totals/subtotals if any. The Football+Scholarship data dataset doesn’t contain any total/subtotals.



4. The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The Football+Scholarship data set has 6215 observations and 9 variables in the data set.

5. info() is used to check the Information about the data and the datatypes of each respective attribute.



Figure 25

* Looking at the data in the head function and in info, we come to know that the variables comprise of float and object data types. sklearn in Python does not take the input of object data types while building Logistic Regression models & LDA Models. So, we need to convert these variables into some numerical form.
* Basis Figure 1 we can see that there are three object type variables (Injury\_Propensity, School\_Type, Region & Scholarship) which has the object data types which we need to convert into numerical form. Since the variable Scholarship & Injury\_Propensity are in ordinal range, we are replacing the categorical variables with the numbers. Further we shall perform one hot encoding for School\_Type & Region variable.
* Further basis above figure 1 we can see that there are 0 null values in the dataset.
* Also, it was identified that there were 947 duplicated rows in the dataset. Will go ahead and drop the duplicated values from the dataset. Post Dropping the duplicated values we are left out with 5268 Entries. We have lost approx. 15.24% of data.

6. The described method will help to see how data has been spread for numerical values. We can clearly see the minimum value, mean values, different percentile values, and maximum values for the Income data set.



Figure 26

7). Check Proportion of observations in each of the target classes (Scholarship Variable).

|  |  |  |
| --- | --- | --- |
|  | **Numbers** | **Percentage** |
| **Partial Scholarship (0)** | 3337 | 63% |
| **Full Scholarship (1)** | 1931 | 37% |

* Since 63% & 37% is balanced data. Hence, We Are not changing the threshold value / Probability value. We are going ahead considering it as 0.5 itself.

**Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. We performed uni-variate and bi-variate analysis to get a better overview and to find outliers in our dataset. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it doesn’t affect the performance of our model.

**Uni-Variate Analysis.**

***Target / Predictor Variable Analysis - price.***

#### Analyzing Feature: Scholarship

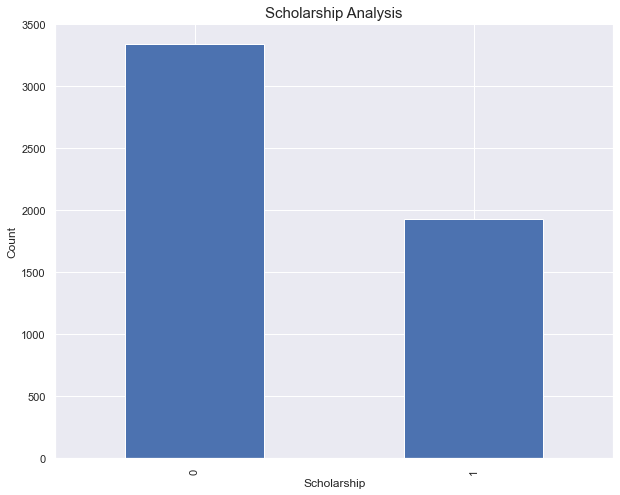


Figure 27

* Basis above figure we can see that 0 (Partial Scholarship) are high in number compared to 1 (Full Scholarship)

***Response / Dependent Variable Analysis***

***Analyzing Feature: Academic\_Score***

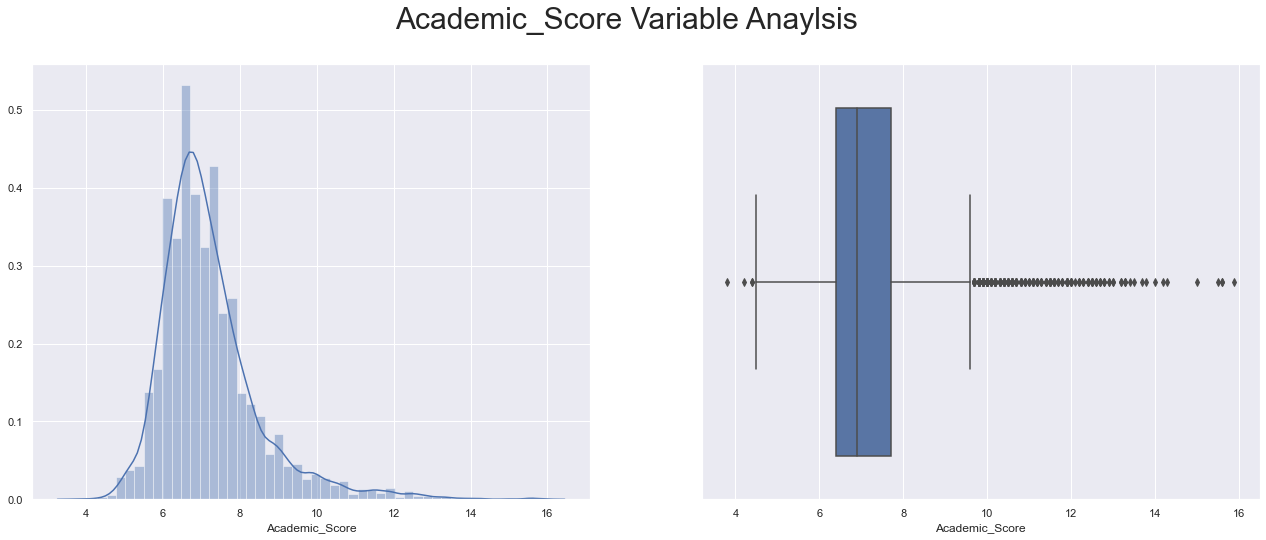


Figure 28

* The Academic\_Score Variable distribution seems to be Right skewed.
* Basis skewness value we can see that distribution is highly skewed.
* Basis box plot we can see that there are outliers in the variable. Let’s treat the outliers in the further process.

***Analyzing Feature: Score\_on\_Plays\_Made***

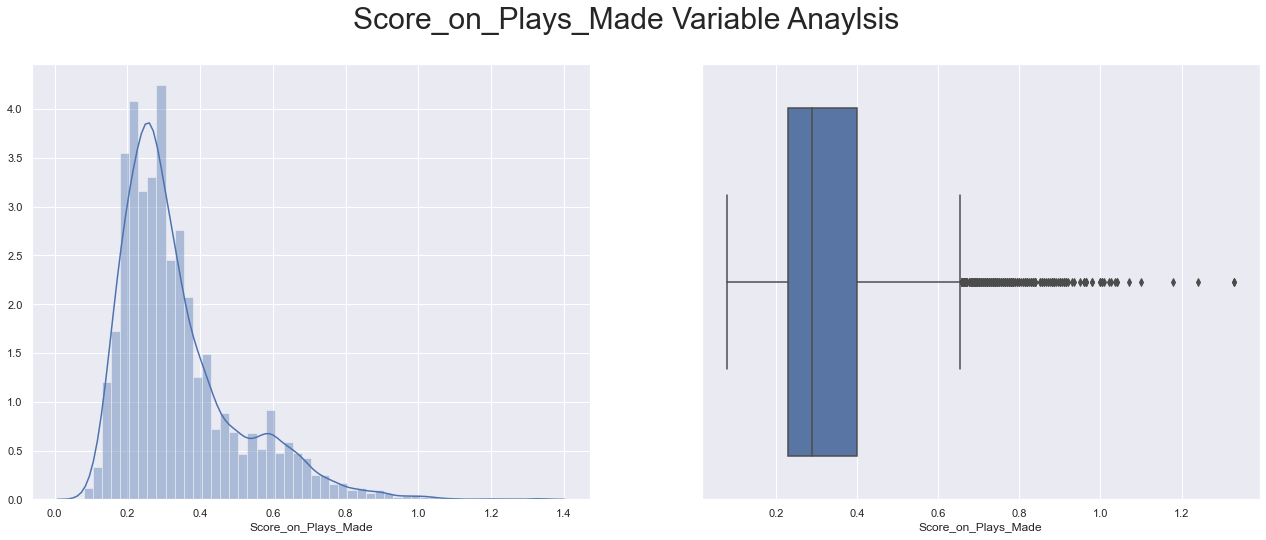


Figure 29

* The Score\_on\_Plays\_Made Variable distribution seems to be Right skewed.
* Basis skewness value we can see that distribution is highly skewed
* Basis box plot we can see that there are outliers in the varaible. Lets treat the outliers in the further process.

***Analyzing Feature: Missed\_Play\_Score***

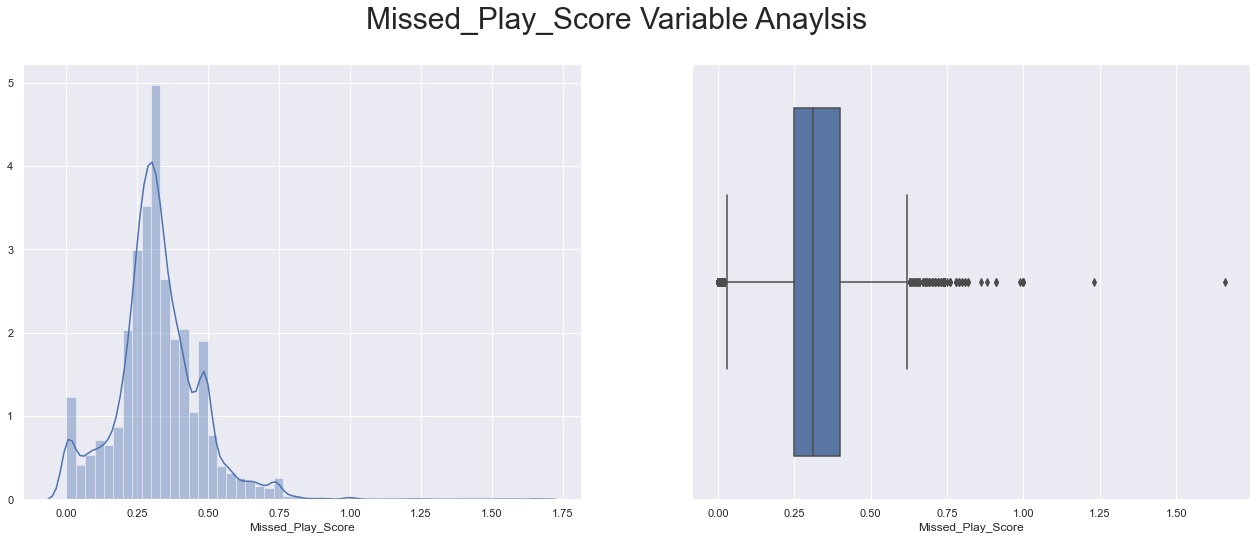


Figure 30

* The Missed\_Play\_Score Variable distribution seems to be Normally skewed.
* Basis skewness value we can see that distribution seems to be approximately symmetric
* Basis box plot we can see that there are outliers in the varaible. Lets treat the outliers in the further process.

***Analyzing Feature: School\_Score***

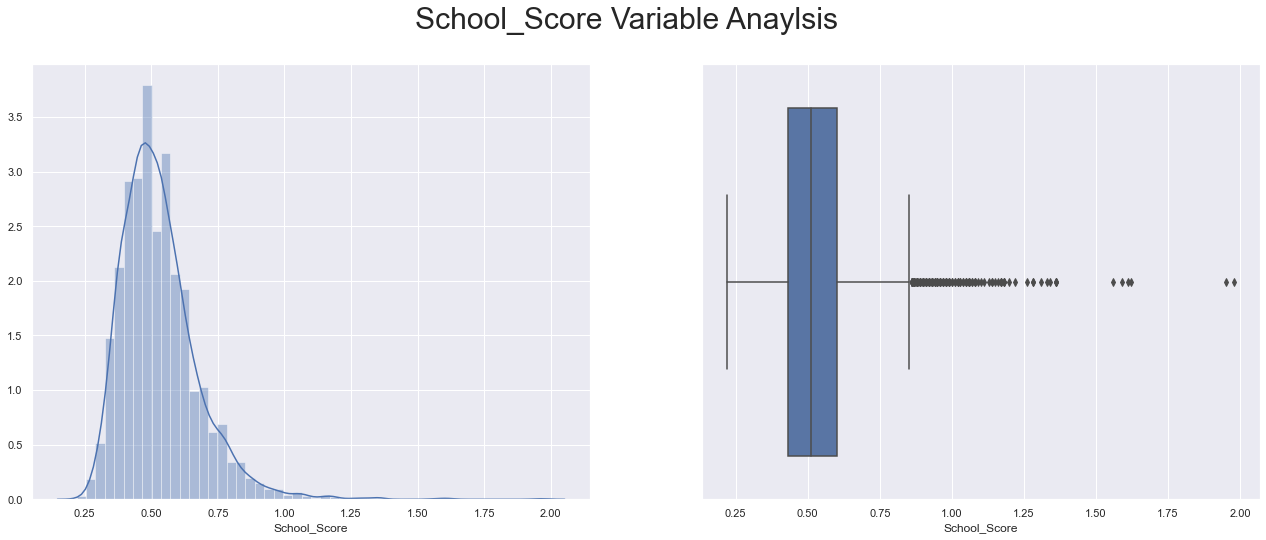
****

Figure 31

* The School\_Score Variable distribution seems to be slightly right skewed.
* Basis skewness value we can see that distribution seems to be approximately symmetric
* Basis box plot we can see that there are outliers in the variable. Let’s treat the outliers in the further process.

***Analyzing Feature: Overall\_Score***

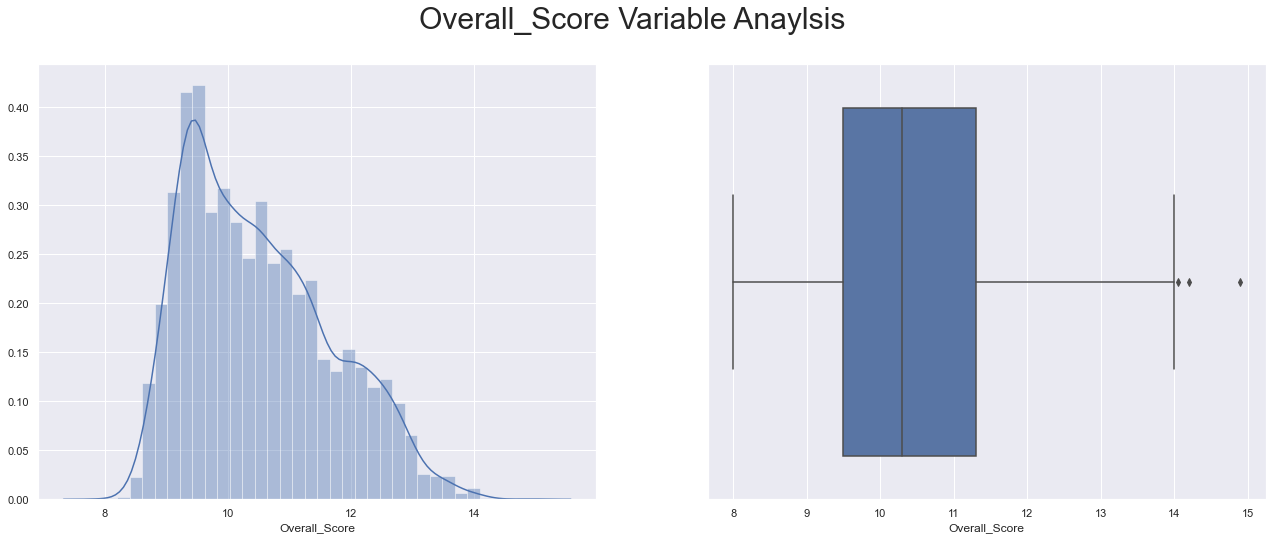
****

Figure 32

* The Overall\_Score Variable distribution seems to be right skewed.
* Basis skewness value we can see that distribution seems to be approximately symmetric
* Basis box plot we can see that there are couple of outliers in the varaible. Lets treat the outliers in the further process.

***Analyzing Feature: Injury\_Propensity***

Chart, bar chart

Description automatically generated

Figure 33

* Basis above figure we can see that there are many students with low (0) Injury\_Propensity.
* Basis above figure we can see that there are low students with High (3) Injury\_Propensity.

*5).* ***Analyzing Feature: School\_Type***

***Chart, bar chart

Description automatically generated***

Figure 34

* Basis above figure we can see that 'C' Type school are more in number Followed by School Type 'B'.
* Basis above figure we can see that D' Type school are Low in number.

*6).* ***Analyzing Feature: Region***

Chart, bar chart

Description automatically generated

Figure 35

* Basis above figure we can see that there are many students from Easter Region.

**Bi-Variate Analysis.**

***Analyzing Feature:*** ***scholarship Vs ACADEMIC\_SCORE***

***Chart, box and whisker chart

Description automatically generated***

Figure 36

* Basis above figure we can see that students with high academic score have high chances to get full scholarship.
* Basis above figure we can also infer that as the academic score falls below 5 there are most of chances that he will not be eligible for full scholarship

***Analyzing Feature: scholarship Vs INJURY\_PROPENSITY***

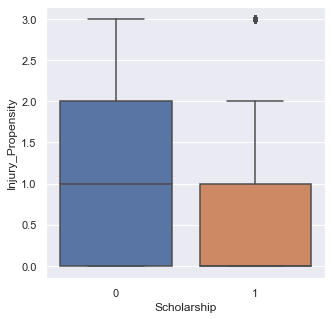


Figure 37

* Basis above figure we can see that students with INJURY\_PROPENSITY have low chances to get full scholarship.

***Analyzing Feature: scholarship Vs SCHOOL\_SCORE***

***Chart, box and whisker chart

Description automatically generated***

Figure 38

* Basis above figure we can see that students with school score above 1.5 have High chances to get full scholarship.

**Data Pre-Processing**

* Basis analysis of above we can see that there are outliers in the variables. Let’s go ahead and treat the outliers.
* sklearn in Python does not take the input of object data types when building linear regression model. So, we need to convert these variables into some numerical form. We shall perform one hot encoding for them (i.e. colour variable). Post Data Pre-processing the head of data looks like



* We are not scaling the dataset. And proceeded with dataset as it is.

**Logistic Regression**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

1. **Descriptive Logistic Regression**
2. **Predictive Logistic Regression**

**Descriptive Logistic Regression**

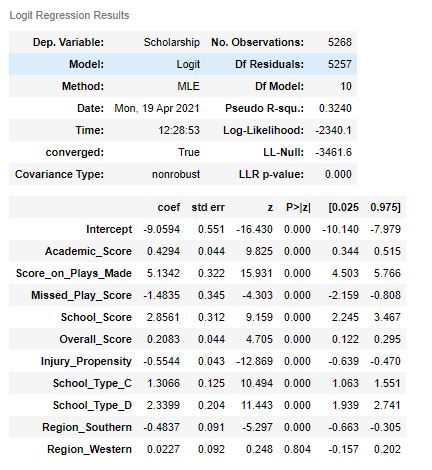
* Descriptive Logistic Regression – Main Objective of Descriptive Logistic Regression is to understand the relation between Features / Variables.
* In Descriptive Logistic Regression we need to look after assumptions i.e. VIF Values (for Multicollinearity). And we select variables basis significance of p value. Whereas for descriptive type assumptions and metric both stand important.
* For Descriptive type we don’t divide the dataset into train and test. Also, We Use statmodel for as python library for Descriptive Logistic Regression.

**Firstly, we will use descriptive logistic regression to understand which all variables are significant variables that impact the scholarship variable.**

**Descriptive logistic regression Models**

**Model 1 (Using all the variables)**

* Firstly, we will import statsmodels.formula.api as SM model
* In first model will run the model using all the variables (i.e. 'Scholarship~Academic\_Score+Score\_on\_Plays\_Made+Missed\_Play\_Score+School\_Score+Overall\_Score+Injury\_Propensity+School\_Type\_C+School\_Type\_D+Region\_Southern+Region\_Western').



adj\_pseudo\_r2 = (model.llf-model.df\_model)/model.llnull

* adj\_pseudo\_r2 = 0.32108239111187487
* Basis above model we can see that adjusted pseudo r square is 32.1%, this value cannot be read individually since it does not signify the model performance, but it can be used to compare different models.
* Let’s check the VIF values for the all the variables to identify multi-collinearity among the independent variables.

Variables VIF Value

Academic\_Score VIF = 1.68

Score\_on\_Plays\_Made VIF = 1.61

Missed\_Play\_Score VIF = 1.52

Injury\_Propensity VIF = 1.73

School\_Score VIF = 1.29

Overall\_Score VIF = 1.97

School\_Type\_C VIF = 2.92

School\_Type\_D VIF = 4.12

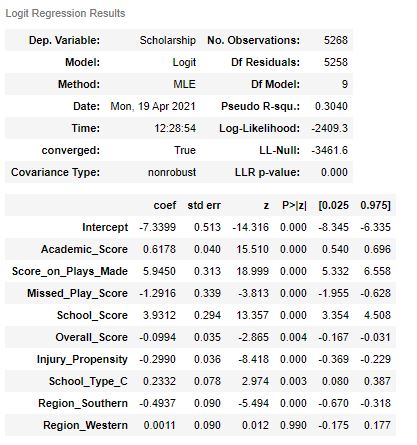
Region\_Southern VIF = 1.23

Region\_Western VIF = 1.25

**Considering the threshold value of 2 for VIF, we go ahead and drop the variable School\_Type\_D which has the maximum VIF value of 4.12**

**Model 2 (Dropping School\_Type\_D)**

* In second model will run the model using the variables (i.e. 'Scholarship~Academic\_Score+Score\_on\_Plays\_Made+Missed\_Play\_Score+School\_Score+Overall\_Score+Injury\_Propensity+School\_Type\_C+Region\_Southern+Region\_Western').

****

* adj\_pseudo\_r2 = 0.3014006091437501
* Basis above model we can see that adjusted pseudo r square is 30.1%, the model performance has gone down compared to model1, but p Value looks insignificant for variable’ Region\_Western’.
* Let us now check multi-collinearity using VIF value.

Variable VIF Value

Academic\_Score VIF = 1.39

Score\_on\_Plays\_Made VIF = 1.51

Missed\_Play\_Score VIF = 1.52

School\_Score VIF = 1.19

Overall\_Score VIF = 1.28

Injury\_Propensity VIF = 1.24

School\_Type\_C VIF = 1.13

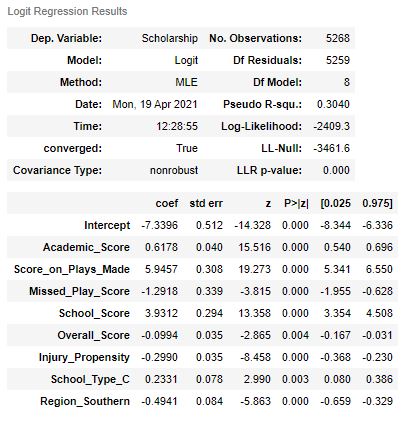
Region\_Southern VIF = 1.23

Region\_Western VIF = 1.25

* All the variables are below threshold of 2. Hence, we will go ahead and drop the insignificant variables by checking P value. (i.e,Region\_Western)

**Model 3 (Dropping School\_Type\_D & Region\_Western )**

* In third model will run the model using the variables (i.e. 'Scholarship~Academic\_Score+Score\_on\_Plays\_Made+Missed\_Play\_Score+School\_Score+Overall\_Score+Injury\_Propensity+School\_Type\_C+ Region\_Southern’).



* adj\_pseudo\_r2 - 0.30168947223257847
* Let us now check multi-collinearity using VIF value.

Variable VIF Value

Academic\_Score VIF = 1.39

Score\_on\_Plays\_Made VIF = 1.46

Missed\_Play\_Score VIF = 1.52

School\_Score VIF = 1.18

Overall\_Score VIF = 1.28

Injury\_Propensity VIF = 1.23

School\_Type\_C VIF = 1.12

Region\_Southern VIF = 1.09

**All the variables VIF Value is below threshold of 2. And by looking at p Value, all the variables are significant.**

**Model Evaluation**



**Inference -** Based on Adjusted Pseudo R square we can see that Model 3 seems to be good model with important variables and free from multi collinearity.

A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome. Its not used as Evaluation Metric

Basis Above descriptive Logistic regression we can say that below variables are the important variables in predicting the partial or full scholarship of students.

* Academic\_Score
* Score\_on\_Plays\_Made
* Missed\_Play\_Score
* School\_Score
* Overall\_Score
* Injury\_Propensity
* School\_Type\_C
* Region\_Southern'
* We will use Model 1, Model 2 and Model 3 to predict and check the model evaluation.

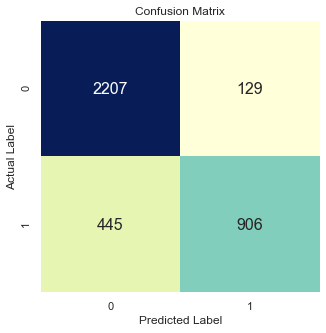
**Predictive** **Logistic Regression**

* Predictive Logistic Regression – Main Objective of Predictive Logistic Regression is predictive values for Features / Variables.
* In predictive Logistic Regression we no need to look after assumptions. Whereas for predictive type assumptions are not important only metric is important.

For predictive type we divide the dataset into train and test. Also, We Use sklearn & statmodel for as python library for predictive Logistic Regression.

**Logistic Regression - Model 1**

* Let’s import classification\_report from sklearn.metrics.
* Let us first evaluate on the **training data**.
* We will start by checking the confusion matrix and then the classification report as well.
* Firstly, Train Accuracy for Model is 0.844317873609981.
* Confusion\_matrix for train data.



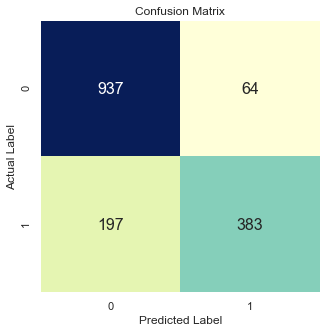
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.83 | 0.94 | 0.88 | 2336 |
| 1 | 0.88 | 0.67 | 0.76 | 1351 |
|  |  |  |  |  |
| accuracy |  |  | 0.84 | 3687 |
| macro avg | 0.85 | 0.81 | 0.84 | 3687 |
| weighted avg | 0.85 | 0.84 | 0.84 | 3687 |

* We can see 84% overall accuracy on the Training data.

**Let us now evaluate on the testing data**

* Test Accuracy for Model is 0.8349146110056926
* Confusion\_matrix for train data.



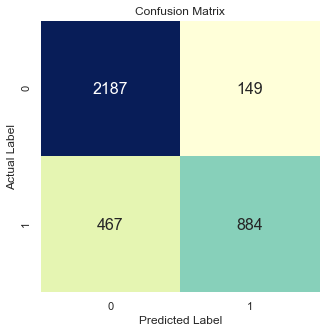
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.83 | 0.94 | 0.88 | 1001 |
| 1 | 0.86 | 0.66 | 0.75 | 580 |
|  |  |  |  |  |
| accuracy |  |  | 0.83 | 1581 |
| macro avg | 0.84 | 0.80 | 0.81 | 1581 |
| weighted avg | 0.84 | 0.83 | 0.83 | 1581 |

* We can see 83% overall accuracy on the Training data.

**Logistic Regression - Model 2**

* Let us first evaluate on the **training data**.
* Train Accuracy for Model is 0.8329264985082723.
* Confusion\_matrix for train data

****

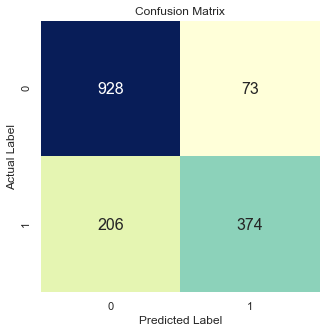
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.94 | 0.88 | 2336 |
| 1 | 0.86 | 0.65 | 0.74 | 1351 |
|  |  |  |  |  |
| accuracy |  |  | 0.83 | 3687 |
| macro avg | 0.84 | 0.80 | 0.81 | 3687 |
| weighted avg | 0.84 | 0.83 | 0.83 | 3687 |

* We can see 83% overall accuracy on the Training data.

**Let us now evaluate on the testing data**

* Test Accuracy for Model is 0.8235294117647058
* Confusion\_matrix for test data.



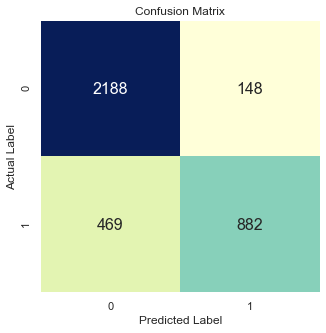
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.93 | 0.87 | 1001 |
| 1 | 0.84 | 0.64 | 0.73 | 580 |
|  |  |  |  |  |
| accuracy |  |  | 0.82 | 1581 |
| macro avg | 0.83 | 0.79 | 0.80 | 1581 |
| weighted avg | 0.83 | 0.82 | 0.82 | 1581 |

* We can see 82% overall accuracy on the Testing data.

**Logistic Regression - Model 3**

* Let us first evaluate on the **training data**.
* Train Accuracy for Model is 0.8326552752915649
* Confusion\_matrix for train data

****

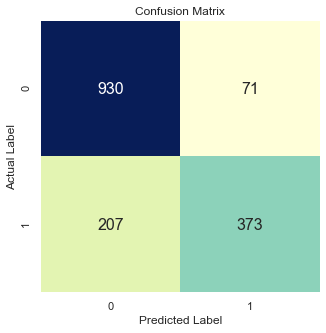
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.94 | 0.88 | 2336 |
| 1 | 0.86 | 0.65 | 0.74 | 1351 |
|  |  |  |  |  |
| accuracy |  |  | 0.83 | 3687 |
| macro avg | 0.84 | 0.79 | 0.81 | 3687 |
| weighted avg | 0.84 | 0.83 | 0.83 | 3687 |

* We can see a 83% overall accuracy on the Training data.

**Let us now evaluate on the testing data**

* Test Accuracy for Model is 0.8241619228336496
* Confusion\_matrix for test data.

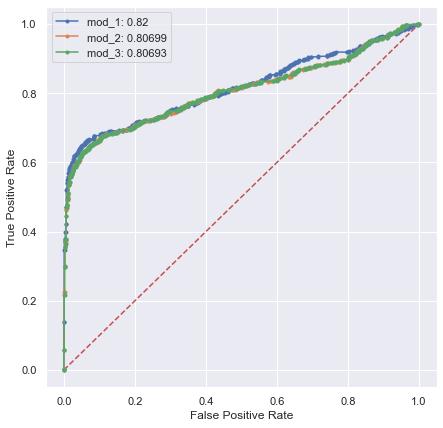
****

* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.93 | 0.87 | 1001 |
| 1 | 0.84 | 0.64 | 0.73 | 580 |
|  |  |  |  |  |
| accuracy |  |  | 0.82 | 1581 |
| macro avg | 0.83 | 0.79 | 0.80 | 1581 |
| weighted avg | 0.83 | 0.82 | 0.82 | 1581 |

* We can see a 82% overall accuracy on the Testing data.

**Check the summary statistics of the AUC-ROC curve for all the three Logistic Regression Models built. This is for the test data**

****

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. Random forest has the highest AUC for test of all the models.

**Logistic Model 1 has the Highest AUC amongst the all the Logistic Models.**

**Linear Discriminant Analysis (LDA)**

LDA uses linear combinations of independent variables to predict the class in the response variable of a given observation. LDA assumes that the independent variables(p) are normally distributed and there is equal variance / covariance for the classes. LDA is popular because it can be used for both classification and dimensionality reduction.

When these assumptions are satisfied, LDA creates a linear decision boundary. Note that based on many research studies, it is observed that LDA performs well when these assumptions are violated.

LDA is based upon the concept of searching for a linear combination of predictor variables that best separates the classes of the target variable.

Key Assumptions for LDA are

* Independent variables should be normally distributed.
* Each Independent variable must have the same variance across classes.
* LDA Does well even if these assumptions are flouted

**Linear Discriminant Analysis Model using the same models 1, Model 2 & Model 3.**

* Import LinearDiscriminantAnalysis from sklearn.discriminant\_analysis. Import confusion\_matrix from sklearn.metrics & import scale from sklearn.preprocessing
* Before building the model, we should split the data into Train and Test. We will thus build a model on the training data and use this model to predict on the test data.
* We will be doing a 70:30 split. 70% of the whole data will be used to train the data and then 30% of the data will be used for testing the model thus built.
* Importing train\_test\_split from sklearn.model\_selection to splitting data into training and test set for independent attributes.
* We Use Stratify in train\_test\_split. **stratification** means that the **train\_test\_split** method returns training and test subsets that have the same proportions of class labels as the input dataset.
* In LDA we use Bayes Theorem for calculating the probabilities. Using threshold, on probabilities calculated by Baysian Rules

Table

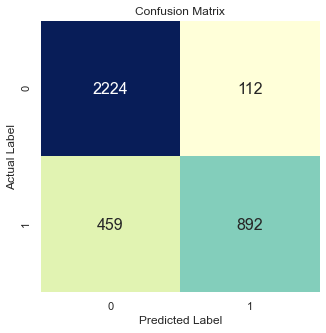
Description automatically generated

LDA makes predictions by estimating the probability that a new set of inputs belongs to each class. The class that gets the highest probability is the output class and a prediction is made.

The model uses Bayes Theorem to estimate the probabilities. Briefly Bayes’ Theorem can be used to estimate the probability of the output class (k) given the input (x) using the probability of each class and the probability of the data belonging to each class:

**Linear Discriminant analysis - Model 1**

* Let us first evaluate on the **training data**.
* Train Accuracy for Model is 0.8451315432601031.
* Confusion\_matrix for train data.



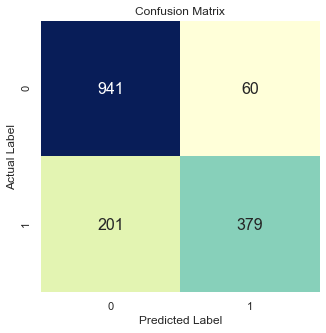
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.83 | 0.95 | 0.89 | 2336 |
| 1 | 0.89 | 0.66 | 0.76 | 1351 |
|  |  |  |  |  |
| accuracy |  |  | 0.85 | 3687 |
| macro avg | 0.86 | 0.81 | 0.82 | 3687 |
| weighted avg | 0.85 | 0.85 | 0.84 | 3687 |

* We can see 85% overall accuracy on the Training data.

**Let us now evaluate on the testing data**

* Test Accuracy for Model is 0.8349146110056926
* Confusion\_matrix for test data.



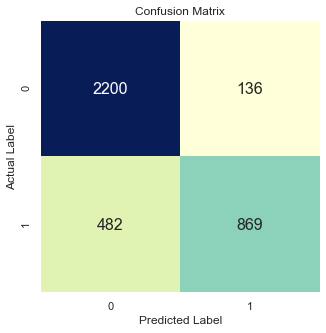
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.94 | 0.88 | 1001 |
| 1 | 0.86 | 0.65 | 0.74 | 580 |
|  |  |  |  |  |
| accuracy |  |  | 0.83 | 1581 |
| macro avg | 0.84 | 0.80 | 0.81 | 1581 |
| weighted avg | 0.84 | 0.83 | 0.83 | 1581 |

* We can see 83% overall accuracy on the Testing data.

**Linear Discriminant analysis - Model 2**

* Let us first evaluate on the **training data**.
* Train Accuracy for Model is 0.8323840520748577.
* Confusion\_matrix for train data.



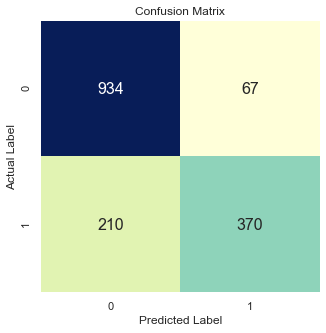
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.94 | 0.88 | 2336 |
| 1 | 0.86 | 0.64 | 0.74 | 1351 |
|  |  |  |  |  |
| accuracy |  |  | 0.83 | 3687 |
| macro avg | 0.84 | 0.79 | 0.81 | 3687 |
| weighted avg | 0.84 | 0.83 | 0.83 | 3687 |

* We can see a 83% overall accuracy on the Training data.

**Let us now evaluate on the testing data**

* Test Accuracy for Model is 0.8247944339025933
* Confusion\_matrix for test data.



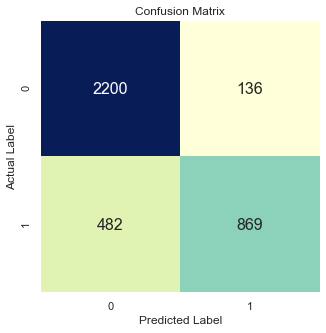
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.93 | 0.87 | 1001 |
| 1 | 0.85 | 0.64 | 0.73 | 580 |
|  |  |  |  |  |
| accuracy |  |  | 0.82 | 1581 |
| macro avg | 0.83 | 0.79 | 0.80 | 1581 |
| weighted avg | 0.83 | 0.82 | 0.82 | 1581 |

* We can see a 82% overall accuracy on the Testing data.

**Linear Discriminant analysis - Model 3**

* Let us first evaluate on the **training data**.
* Train Accuracy for Model is 0.8323840520748577.
* Confusion\_matrix for train data.



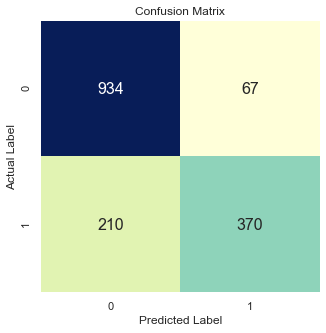
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.94 | 0.88 | 2336 |
| 1 | 0.86 | 0.64 | 0.74 | 1351 |
|  |  |  |  |  |
| accuracy |  |  | 0.83 | 3687 |
| macro avg | 0.84 | 0.79 | 0.81 | 3687 |
| weighted avg | 0.84 | 0.83 | 0.83 | 3687 |

* We can see a 83% overall accuracy on the Training data.

**Let us now evaluate on the testing data**

* Test Accuracy for Model is 0.8247944339025933
* Confusion\_matrix for test data.



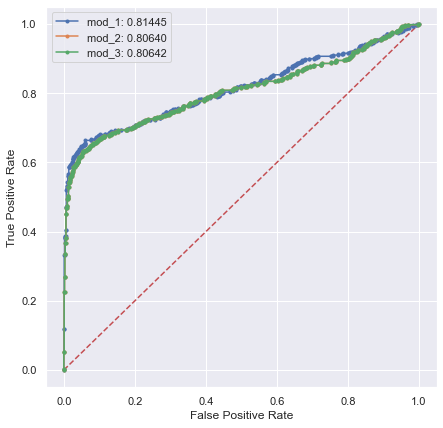
* Classification Report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.82 | 0.93 | 0.87 | 1001 |
| 1 | 0.85 | 0.64 | 0.73 | 580 |
|  |  |  |  |  |
| accuracy |  |  | 0.82 | 1581 |
| macro avg | 0.83 | 0.79 | 0.80 | 1581 |
| weighted avg | 0.83 | 0.82 | 0.82 | 1581 |

* We can see a 82% overall accuracy on the Testing data.

**Check the summary statistics of the AUC-ROC curve for all the three LDA Models built. This is for the test data.**

.



AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. Random forest has the highest AUC for test of all the models.

**LDA Model 1 has the Highest AUC amongst the all the three LDA Models**

**Model Evaluation**



**Inference** –

* For this problem we are using accuracy as metric for evaluation of the models.
* In comparison between Logistic regression model’s Logistic regression Model 1 has good Accuracy for both train and test compared to other models.
* In comparison between LDA Models basis above results we can see that LDA Model has good accuracy for both train and test models.
* When in comparison between Logistic model & Linear Discriminant Model We can see that LDA Model has a very good accuracy amongst the all the logistic & LDA Models. As the dataset is small ones. LDA Seems to out-preforming than logistic Regression.
* Logistic regression does not have as many assumptions and restrictions as discriminant analysis. But Logistic Regression lacks stability when the classes are well separated, Whereas the LDA Is good when the classes are well separated. this is when LDA comes into pictures

However, when discriminant analysis' assumptions are met, it is more powerful than logistic regression. Unlike logistic regression, discriminant analysis can be used with small sample sizes.