PREDICTIVE MODELING PROJECT

VISHAL.S

G3 PGP DSBA JULY

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGENO |
|  | INTRODUCTION |  |
| A | PROBLEM 1 |  |
| 1 | EXPLORATORY DATA ANALYSIS (EDA) |  |
| 1.1 | HEAD |  |
| 1.2 | SHAPE |  |
| 1.3 | INFO |  |
| 1.4 | DESCRIBE |  |
| 1.5 | DATA TYPES PRESENT IN THE DATASET |  |
| 1.6 | NULL VALUE CHECK |  |
| 2 | UNIVARIATE ANALYSIS |  |
| 2.1 | HISTOGRAM |  |
| 2.2 | SKEWNESS |  |
| 3 | BIVARIATE ANALYSIS |  |
| 3.1 | PAIRPLOT |  |
| 3.2 | HEATMAP |  |
| 3.3 | EDA for Categorical Variable |  |
| B | PROBLEM 2 |  |
| 2.1 | EDA |  |
| 2.2 | UNIVARIATE ANALYSIS |  |
| 2.3 | BIVARIATE ANALYSIS |  |
| 2.4 | CONCLUSION |  |

INTRODUCTION

**Problem 1**: Linear Regression

You are hired by a company Gemstone co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

DATA DICTIONARY

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, 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 | The 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. |

* 1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

1 EXPLORATORY DATA ANALYSIS (EDA)

Let’s start the data exploration step with the head function to look at first 10 initial rows.

1.1 HEAD

Table

Description automatically generated

1.2 SHAPE



The above Dataset has 26967 rows and 11 columns

1.3 INFO

Graphical user interface, application, table

Description automatically generated

From the above info we can observe that data has three data types: 6 float variables, 3 object variables, and 1 integer variable and can observe missing some data points in the depth column.

1.4 DESCRIBE

Graphical user interface, application

Description automatically generated

We can observe that there is a huge difference between the ranges of means of each variable, one is in decimals. Few are in 1s, few are in 10s and price is in 1000s of range. This difference in means can influence the outcome of the analysis, with more weightages being given to the variable with the biggest mean. The above data points tell us the scaling is required

1.5 DATA TYPES PRESENT IN THE DATSET

Text

Description automatically generated

From the above figure we see the different data types present in the dataset

1.6 NULL VALUES

A picture containing text

Description automatically generated

We can observe that object depth has some missing values

2.UNIVARIATE ANALYSIS

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

2.1 Histograms

Chart

Description automatically generated

Fig 1.1: Histograms for variables carat, depth, table

Chart, histogram

Description automatically generatedFig 1.2: Histograms for variables x,y,z

Chart, histogram

Description automatically generated

Fig 1.3: Histograms for variable price

2.2 SKEWNESS

Text

Description automatically generated

* There is significant amount of outlier present in some variable.
* We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

1. BIVARIATE ANALYSIS

Bivariate analysis refers to the analysis of two variables. The main purpose of the bivariate analysis is to summarize and find the patterns and correlation in the data.

* 1. PAIR PLOTDiagram

     Description automatically generatedFig 1.4 Pair plot

From the above pair plot, we can observe how the data is distributed and relation and patterns between each variable.

3.2 HEAT MAP

Chart, histogram

Description automatically generated

Fig 1.5: Heatmap

Heat map tells the correlation and collinearity between variables from the above heat map we can observe that there is a heavy correlation and presence of multicollinearity. Multicollinearity is not acceptable in regression

* 1. EDA for Categorical variable

Chart, bar chart

Description automatically generated

Fig 1.6: Plot For variable cut

Chart, box and whisker chart

Description automatically generated

Fig 1.7: Box Plot for variable cut

Chart, bar chart

Description automatically generated

Fig 1.9: Plot for variable color

Chart, box and whisker chart

Description automatically generated Fig 1.10: Box plot for variable color

Chart, bar chart

Description automatically generated Fig 1.11: Plot for variable clarity

Chart, box and whisker chart

Description automatically generated Fig 1.12 Box Plot for Variable clarity

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning, or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

IMPUTING THE MISSING VALUES

A picture containing text

Description automatically generated

Replacing the missing data (Nan) in the column depth with the mean value of the same column

Table

Description automatically generated with medium confidence

We can observe from the above output that there are no missing values in the dataset

CHECKING FOR ZEROS

Table

Description automatically generated with medium confidence

We can observe from the above output that there are no null values in the dataset

DUPLICATES

Graphical user interface, text

Description automatically generated with medium confidence

There are 33 duplicate rows present in the dataset,After dropping the duplicatesTable

Description automatically generated

OUTLIERS

After removing the Outliers

Chart, histogram

Description automatically generated

Fig 1.13 Box Plot For carat variable after outlier removal

Chart, box and whisker chart

Description automatically generated

Fig 1.14 Box Plot For depth variable after outlier removal

Chart

Description automatically generated

Fig1.15: Box plot for table variable after outlier removal

Chart, box and whisker chart

Description automatically generated

Fig1.16: Box plot for x variable after outlier removal

Chart, box and whisker chart

Description automatically generated

Fig1.17: Box plot for y variable after outlier removal

Chart, box and whisker chart

Description automatically generated

Fig1.18: Box plot for z variable after outlier removal

Chart, histogram

Description automatically generated

Fig1.19: Box plot for price variable after outlier removal

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

ENCODE THE DATA (HAVING STRING VALUES)

Text

Description automatically generated with low confidence

### CONVERTING OBJECTS TO CATEGORICAL CODES

### Text Description automatically generated

### From the above output we find that except x,y,z variables all other object variables have been converted to categorical codes

### Text Description automatically generated

### From the above output we have found that all the object variables in the data set have been converted to categorical codes

### Table Description automatically generated

### LINEAR REGRESSION MODEL

### We are going to split the data into 70:30 ratios were 30% is the test data and 70% is the train data Now the data is spited and ready for the modelling Here we are doing Linear Regression Model using sklearn package

### COEFFICIENT OF INDEPENDENT VARIABLES

### Text, letter Description automatically generated

### INTERCEPT

### Now we have coefficients, to form an equation we need an intercept which is below



Now we have coefficient and intercept to form an equation but how to check the performance of the equation. We can check the performance of the equation by R Square or RMSE

R SQUARE VALUE FOR TRAIN DATA



The considered input variables which is going to explain the Variation in price is 91% which is acceptable.

R SQUARE VALUE FOR TEST DATA



ROOT MEAN SQUARE ERROR (RMSE) FOR TRAIN DATA



ROOT MEAN SQUARE ERROR (RMSE) FOR TEST DATA



LINEAR REGRESSION USING STATS MODEL

In this the train and test data are given for a single variable where the given data is stored

A picture containing text, receipt

Description automatically generated

The intercept and coefficient values for the sklearn model and stats model are same, only the approach and package are different. lML provides a simple, unified interface for building machine learning models, using those models to generate insights and to make accurate predictions. It is more advantage than the sklearn package because it gives more information and insights.

Table

Description automatically generated

From the above table we can observe that many insights are given at one place, the prob (F-Statistic) is 0.00 which tells us the equation is reliable and we can also check the p values for every independent variable if the p value is less than 0.5 then it is said to be reliable and we can observe one independent variable depth which has its p vale more than 0.5 except this remaining all parameters are reliable so we can drop the variable which is having p value more than 0.5. and we can know the autocorrelation and multicollinearity from Durbin-Watson and Cond.No.

SCATTER PLOT FOR ACTUAL VALUE AND PREDICTED VALUE

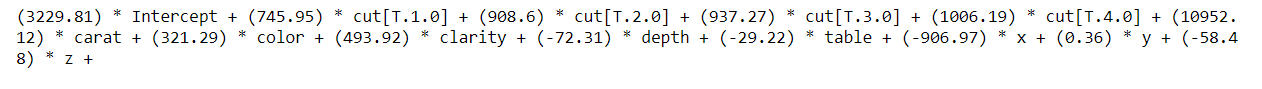
Chart, scatter chart

Description automatically generated

Fig 1.20 Scatter Plot

From the above scatter plot, we can observe that the correlation between the actual value and the predict value which are highly positively correlated

FINAL EQUATION FOR THE MODEL



THE FIVE ATTRIBUTES THAT ARE MOST IMPORTANT

1. CARAT

2. CUT\_IDEAL

3. COLOR\_E

4. CLARITY\_IF

5. DEPTH

INFERENCE

From the above equation carat + depth + cut\_ldeal + clarity\_IF will be the good combination to get the best price for the Zirconia diamond. We can observe that the color has all negative values which effects the increase in the prices so it’s better to avoid the color for a better price un the market There are mainly 4factors which gives the best price for the Zirconia diamond

PROBLEM: 2

LOGISTIC REGRESSION & LINEAR DISCRIMINANT ANALYSIS

PROBLEM STATEMENT: You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

2.1 EXPLORATORY DATA ANALYSIS (EDA)

Let’s start the data exploration step with the head function to look at first 5 initial rows

HEAD

Table

Description automatically generated with low confidence

SHAPE

A picture containing text

Description automatically generated

INFO

Graphical user interface, table

Description automatically generated

From the above info we can observe that data has three data types: 2 object variables, and 5 integer variables.

DESCRIBE

Table

Description automatically generated

The above data points tells us the scaling is not required

NULL VALUE

Text

Description automatically generated

We can observe that there are no missing values in the data set

OBJECT VARIABLES AND THEIR UNIQUE COUNTS

Graphical user interface, text

Description automatically generated

From the above output we can observe that data has 2 object variables: Holiday\_Package , Foreign and their unique values with counts

2.2 UNIVARIATE ANALYSIS

Univariate analysis refers to the analysis of a single variable. The main purpose of the univariate analysis is to summarize and find the patterns in the data. The key point is the only one variable is involved in the analysis.

Chart, box and whisker chart

Description automatically generated

Fig 2.1 Box plot and distplot for salary and age variable

Chart, histogram

Description automatically generated

Fig 2.2 Box plot and dist plot for educ and no\_young\_children variables

Chart

Description automatically generated

Fig 2.3 Box plot and Dist plot for no\_older\_children variable

2.3 BIVARIATE ANALYSIS

Bivariate analysis refers to the analysis of two variables. The main purpose of the bivariate analysis is to summarize and find the patterns and correlation in the data. The key point is the only two variables are involved in the analysis.

HEAT MAP

Graphical user interface, application

Description automatically generated

Fig 2.4 Heat Map

PAIR PLOT

Diagram

Description automatically generated

Fig 2.5: Pairplot

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis)

OUTLIERS

Chart, box and whisker chart

Description automatically generated

Fig 2.6 Box Plot with outliers

Chart, box and whisker chart

Description automatically generated

Fig 2.7 Box plot without Outliers

LABEL ENCODING

Table

Description automatically generated with medium confidence

TRAIN TEST SPLIT

A picture containing text

Description automatically generated

PREDICTING ON TRAIN AND TEST DATA

Graphical user interface, text, application, chat or text message

Description automatically generated

We got the predicted probabilities values instead of 0 and1

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized**.**

MODEL EVALUATION

Text

Description automatically generated with medium confidence

The model accuracy is 64%

AUC ROC CURVE FOR TRAIN DATA

Chart, line chart

Description automatically generated

Fig 2.8 AUC and ROC curve for Train set

AUC ROC CURVE FOR TEST DATA

Chart, line chart

Description automatically generated

Fig 2.9 AUC and ROC for Test set

The accuracy of train and test are same which indicates the model is not over fit

CONFUSION MATRIX FOR TRAIN DATA

A picture containing chart

Description automatically generated

Chart, treemap chart

Description automatically generated

By using the confusion matrix, we get f1 score which is not available in the AOC and ROC curve that’s way it is important to check the accuracy with confusion matrix

CLASSIFICATION REPORT TRAINING DATA

Table

Description automatically generated

From the above table we can observe the f1 score which tells use 0 is performing better than 1

CONFUSION MATRIX FOR TEST DATA

Graphical user interface, text, application

Description automatically generated

Chart, treemap chart

Description automatically generated

CLASSIFICATION REPORT OF TEST DATA

Table

Description automatically generated

The confusion matrix is same for the train and test data

GRIDSEARCH CV

By using grid search we can improve the accuracy and overcome the above-mentioned drawback by using the grid search model there is no change in the f1 values

LINEAR DISCRIMINANT ANALYSIS

CONFUSION MATRIX FOR TRAIN AND TEST DATA

Chart, waterfall chart

Description automatically generated

From the above out we can observe the confusion matrix for both the train and test data

Classification report for both Test and Train data

Table

Description automatically generated

From the classification report we can observe that both the accuracy and f1 score together and we can say that it is not over fit because both the values are similar, and we can observe that the model is fitting better for zeros and 1 is not much fitting the model.

AUC AND ROC CURVE FOR TEST AND TRAIN DATA

Chart, line chart

Description automatically generated

Fig 2.10 AUC and ROC curve for test and train data

From the above output we can observe the accuracy for the train and test is same We can use cut-off technique to maximize the accuracy

CONFUSION MATRIX

Chart, treemap chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

CLASSIFICATION REPORT

Table

Description automatically generated with medium confidence

We can observe that by using cut-off technique there is reducing in the accuracy, so it not recommended implementing for this model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  | LR Train | LR  Test | LDA Train | LDA Test |  |
|  | Accuracy | 0.64 | 0.63 | 0.64 | 0.63 |  |
|  | AUC | 0.67 | 0.66 | 0.67 | 0.66 |  |
|  | Recall | 0.45 | 0.44 | 0.43 | 0.44 |  |
|  | Precision | 0.68 | 0.62 | 0.68 | 0.62 |  |
|  | F1 Score | 0.54 | 0.52 | 0.53 | 0.52 |  |
|  |  |  |  |  |  |  |

Table 3 Comparison

Final Comparison:- On comparison with all the three models and all parameters like accuracy, precision, recall on positive side we find Logistic Regression as the better and the most optimized model as on all the parameters it stands the highest and the most accurate amongst the two models.

INFERENCE

By observing the data employees are not ready to opt for the holiday packages

2.4 Inference: Basis on these predictions, what are the insights and recommendations**.**

As per the Analysis, the employees with 1-2 young children tend to opt for the holiday package, hence focusing on number of young children of the employees can help the company to suggest holiday package. People with salary less than 1,50,000 are likely to take holiday package. Age and Education so not seem to be a good predictor when it comes to predicting Holiday package. Based on the above model, business can collect first-hand information and can set a cut-off prediction value above which the employee is likely to take the holiday package based on which the model scores will change drastically, and we can find the optimum model based on the business requirement and business best fit.

CONCLUSION

Based on the above model, business can collect first-hand information and can set a cut-off prediction value above which the employee is likely to take the holiday package based on which the model scores will change drastically, and we can find the optimum model based on the business requirement and business best fit.