**Predictive modelling**

# Project Report

**Problem 1**

Linear Regression  
You are hired by a company Gem Stones 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.

**1.1 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.**

We also import all the necessary libraries to fit the linear regression model

We import the data using the pd.read function.

**Exploratory Data Analysis**

**A screenshot of a cell phone

Description automatically generated**

The data has 10 variables, carat, cut, clarity, depth, table x,y,z and price. The data contains multiple data types such as object, float and integer. We can find out the same using the df.info() function. The shape of the data is 26967 rows and 10 columns.

Descriptive Statistics for the data set-

A screenshot of a cell phone

Description automatically generated

We use the df.describe function to get the 5 point summary for all the variables.

We check for null values using the df.isnull().sum() function

A screenshot of a cell phone

Description automatically generated

We can see that there are null values only for one variable. There are 697 null values in the dataset.

We also check for duplicate rows in the data using the df.duplicated function. We see there are 34 duplicate rows. On further analysis we note that these values are not actually duplicate hence we don’t drop them

**Univariate analysis**

Using the df.hist function we see the distribution of each of the variables. This helps us to identify if any variables are skewed or not

A screenshot of text

Description automatically generated

From the histograms we can see that variables carat, price and y are highly right skewed

Depth, table and x are more normally distributed

The range of values for y and z is very less and therefore it not very distributed. From the histogram we can observe that both are highly skewed.

We also plot the boxplot to identify if the data has any outliers.

A picture containing clock

Description automatically generated

From the boxplot we can see that all the variables have outliers. However since the scale of the values varies significantly, we can plot separate boxplots to see the outliers clearly (Refer Appendix)

**Bivariate analysis**

We use the pairplot function and heatmap for bivariate analysis

A close up of a map

Description automatically generated

We only look at one side of the diagonal in order to spot any relationships between the variables. We can clearly see on the basis of the scatter plot that there exists linear relationships between the variables.

Further to quantify the correlation, we plot the heatmap using the sns.heatmap function

A picture containing drawing

Description automatically generated

From the heat map we can observe very highly correlation between the following variables

1. Carat and price, x,y and z
2. Price and x,y,z
3. x,y and z

With extremely high correlation between the variables, we have the issue of multicollinearity, therefore we need to be mindful of the same.

**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? Do you think scaling is necessary in this case?**

i) As noted before we have null values only in the depth column. Given the large size of the data, we could choose to drop these null values as well as they only constitute to 2.5% of the rows of the dataset

However, we would be imputing the null values. We can observe that there isn’t much difference between the mean and median values – Mean: 61.7 , median – 61.8. This implies that there may not be outliers present in the data. Therefore we can impute with the mean or the median value.

We use the fill.na() function to impute the null values.

ii) As seen in the histogram, we also observed that the column x,y and z contained 0 values. Given that these values represent the length, width and height of the cubic zirconia, 0 value for the same seems improbable. We identify the rows with 0 values and note there are only 9 such line items. We will drop these rows before proceeding the with data analysis.

iii) Yes I do think scaling is necessary as the units of each variable are different. Scaling will help us to standardize the data. We will scale the continuous variables.

**1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE**

i) Since we have categorical data in the data frame, we will first convert these values to numerical data and then proceed to scale the data.

We use the pd.categorical.codes functions to convert object type data to numerical. This picks up the unique the values in respective columns and assigns a code to them, which is essentially replacing them with a numerical value. In the given data set, we have cut, clarity and color columns as object data type. We generate codes and assign the data to a new data frame df\_new. We can see below the values have been replaced by codes and datatype has been changed to int64

A picture containing flower

Description automatically generated A close up of a piece of paper

Description automatically generated

We then scale the continuous variables in the data using the standard scaler.

We also treat the outliers in the data by replacing the values with Q1 and Q3 values by defining a new function to calculate the lower range and upper range values. After replacing the values we can observe that there are no more outliers in the data

Encoded variables are excluded from scaling and outlier treatment

Further we proceed to fit the model on the data set.

We first need to split the data into training and testing data in the proportion of 70% train and 30% test data. We use the sklearn.train\_test\_split function to split the data into train and test. The **training** set contains a known output and the model learns on this **data** in order to be generalized to other **data** later on. We have the **test** dataset (or subset) in order to **test** our model's prediction on this subset

Once we do all this, we define our X and y dataframes by dropping and populating the target variable in X and y respectively

We then split the data into test and train

For the linear regression model, we follow the following steps-

1.We split the data into test and train

2. We fit the Linear regression model on the train data.

3. We then calculate the intercept and coefficient values for the data

4. Check the performance by calculating the R square and RMSE values

We calculate the R square value as 91%. This means that 91% of the variation in the data is explained by the predictors in the model. R square is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable. Instead we use adjusted R^2 which removes the statistical chance that improves R^2 Scikit does not provide a facility for adjusted R^2 so we use statsmodel, a library that gives results. In this case, our adjusted R square value is also 91%

Further we also check the RMSE value. We have MSE value of 0.06. Further on taking the square root of this value we have the avg variance between predicted and actual values.

We get a value of 0.26. Which means that here is avg of 0.3 (roundoff) unit price difference from real price on an avg

Please refer the appendix for code for each of the steps for respective classification models

We also have an accuracy of 90% in the data. (This could be due to overfitting or presence of multicollinearity)

**1.4 Inference: Basis on these predictions, what are the business insights and recommendations**

The final Linear Regression equation is

price = b0 + b1*carat +* b2*cut* + b3*color + b4*clarity + b5*depth+ b6*table + b7*x+ b8*y+ b9*z*

price = **-0.1578 + 1.097*carat +* 0.0098*cut* -0.057color *+ 0.0633*clarity -0.0142*depth-0.04100*table -0.4533*x+ 0.4535*y-0.1858*z***

When carat increases by 1 unit, price increases by 1.097units, keeping all other predictors constant.  
similarly, when cut increases by 1 unit, price increases by 0.0098 units, keeping all other predictors constant.

There are also some negative co-efficient values, for instance, color, depth, x,z has its corresponding co-efficient as negative. In an ideal scenario this implies, for a change of 1 unit in depth, the price inversely changed by 0.0142 units, keeping all other predictors constant.

In this case we can identify that the top 5 variables for predicting the price of a cubic zirconia are –

1. Carat
2. Y (width)
3. X(length)
4. Z(height)
5. clarity

**Note: As we noted earlier, there was presence of multicollinearity in the data as there was high correlation between the variables. This could lead to the model not being accurate for prediction of the price.**

**As we are predicting the price of the cubic zirconia it is not possible that the price can be negative as we can see from the value of the intercept. Similarly, we know that length, and height of the cubic zirconia would logically not have an inverse relationship with price. Therefore this model must be revisited in order to deal with multicollinearity in the variables to create a model that predicts a more accurate relationship between the variables and price of the cubic zirconia**

**Problem 2:** Logistic Regression and LDA

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.**

We import the data using the pd.read function. We also import all the necessary libraries to fit the logistic regression and linear discriminant analysis model along with Classification reports, ROC-AUC curves

We use the df.head function to see if the data has been imported properly.

A screenshot of a cell phone

Description automatically generated

We drop any unwanted columns (such as Unnamed:0)

Further we look at the summary of the data, using the df.info and df.describe function

df. describe tells us the 5 point summary of the continuous variables

A screenshot of a cell phone

Description automatically generated

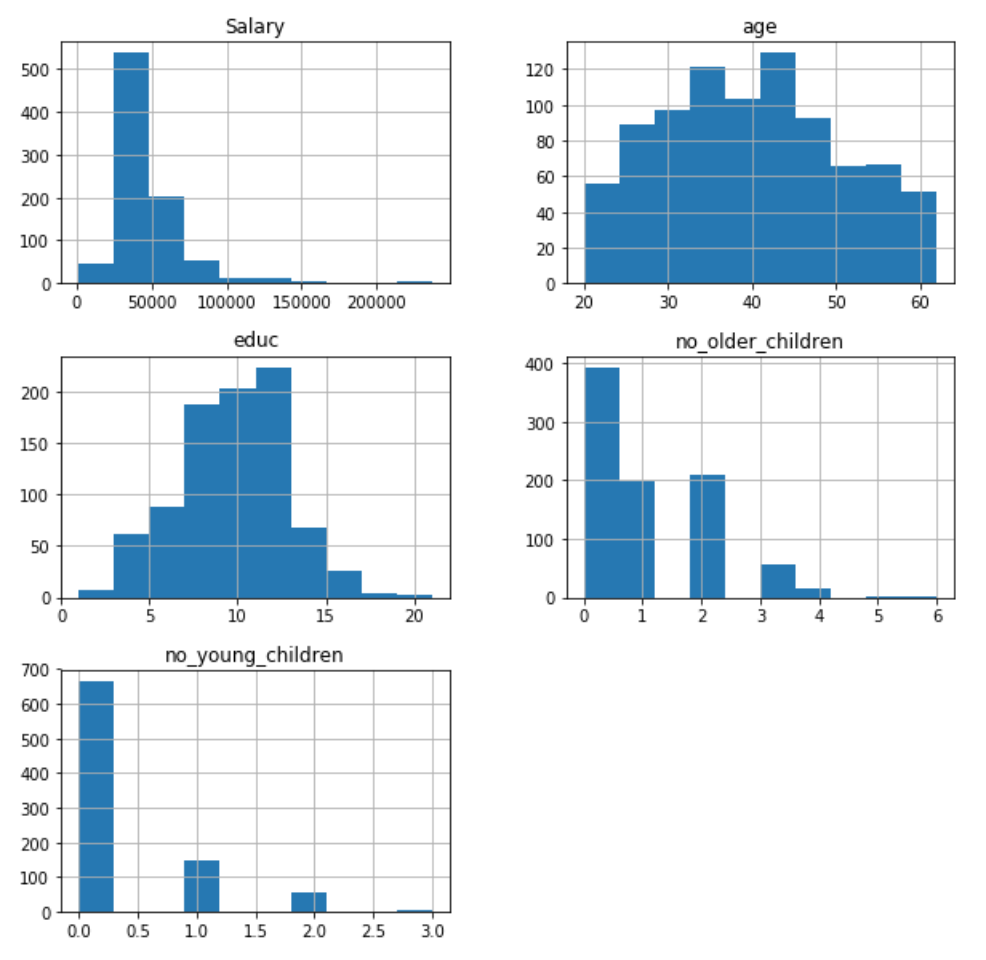
df.info function gives us the data types of each variable and number of non-null values

A screenshot of a cell phone

Description automatically generated

We also check for any null values and if there are any duplicate values in the dataset using the df.isnull().sum() and df.duplicated function

We can perform univariate analysis by plotting the histogram to understand the distribution of the data. The df. hist function helps us visualize the distribution of the data and understand if there are any highly skewed variables



From the histogram we can see that education and age are normally distributed. Salary is right skewed.

As no. of children can’t be continuous data we have breaks in the histogram however both are right skewed

We also plot the boxplot to identify if the data has any outliers. From the boxplot we can observe that all the numerical variables have outliers. In case of presence of outliers we need to either impute them or remove them from the data so that it does not affect our insights. We can see that the salary variable has a large number of outliers whereas the other variables have only a few outliers. Therefore we impute the outliers only for the salary variable

A picture containing clock

Description automatically generated

A picture containing text, sitting, ready, black

Description automatically generated

We use the sns.pairplot() function to plot the relationship between all the numerical variables in the dataset. The histograms on the diagonal shows distribution of each variable whereas the scatter plots on the upper and lower triangles show the relationship between two variables. We can see that all the plots show a linear relationship between the variables

A close up of text on a white background

Description automatically generated

We also plot a heatmap to view the correlation between the continuous variables. WE can observe that there is no high correlation between the variables

A picture containing drawing

Description automatically generated

We also look at the proportion of values in the target variable. We are looking at whether an employee will buy a holiday package or not. Hence the column “Holliday\_Package” is our target variable. We can see that the dataset is more or less balanced as proportion of both values is approximately the same

Yes (Employee will buy holiday package) – 46%

No (Employee will not buy holiday package) – 54%

**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).**

To perform the modelling, we need to convert the categorical variables into numerical ones, using codes. This assigns numerical values to each unique value in the respective variables, making it easier for us to analyse. We have two categorical variables in the data Holliday package and foreign. We use the codes function to assign numerical values.

Further, we define our x and y by dropping and populating the target variable In x and y respectively. We then split the data into test and train in the proportion of 70% train and 30% test data

**Logistic Regression**

We then fit the logistic regression model on the train data. Further we predict the target variables using the model for both train and test data.

We then calculate the predicted classed and their respective probabilities

A screenshot of a cell phone

Description automatically generated

**Linear Discriminant Analysis**

For linear discriminant analysis we fit the model on X and y without splitting intro train and test data

We then calculate the predicted values and attach it to the X dataframe

Please refer the appendix for code for each of the steps for respective classification models

**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**

We then perform the model evaluation using the ROC and AUC for the training and test data respectively – This gives us the area under the curve. We then calculate the confusion matrix for train and test data for Logistic regression and separately for Linear Discriminant analysis. Finally we execute the command for the classification report and accuracy for both train and test data. Based on the AUC, Accuracy, sensitivity, precision and f1-score values, of the train and test data, we draw our conclusions about the model

A screenshot of a cell phone

Description automatically generated

Accuracy – The proportion of the total predictions that were correct

AUC- Area under the curve

Precision – The proportion of positive cases that were correctly identified

Recall – The proportion of actual positive cases correctly identified

F1 – Harmonic mean of precision and recall values

Basis the table we can see the that LDA model has the higher accuracy at 67%.

Looking at all metrics, we can see that the logistic model has very poor precision and recall in comparison to the LDA model in addition to the fact that both test and train data have very low accuracy.

Between the two models, the LDA model would be better at predicting whether an employee will buy a holiday package or not

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

On the basis of the above table the LDA model with 67% accuracy is a better predictor to identify which employee would purchase the holiday package and which won’t.

We have 6 factors available in the dataset, Salary, age, education, no of young children, no. of old children and foreign. In order to identify which features are of most importance we would have to calculate the feature importance of these variables. Once we know which variables influence an employee’s decisions the most, strategies can be employed by the agency in order to further increase their sales and marketing by being able to target the correct market segment

Overall both models are not reasonably stable enough to be used for making any future predictions. Both have low accuracy scores. Further, looking at the shape of the data, we can observe that we only have 872 observations. This may not be enough data to make the analysis. It could also have to led to under fitting of the model

Further we could also consider using classification models to make the predictions.