PGP - Data Science and Business Analysis - Jan’ 21

Predictive Modeling Project Report

*by Sarang Manohar*

**

Table of Contents 1

Introduction 2

Problem 1: Linear Regression 3

Data Dictionary: 3

Question 1.1. 3

Question 1.2 12

Question 1.3 13

Question 1.4 15

Problem 2: Logistic Regression and LDA 17

Data Dictionary: 17

Question 2.1 17

Question 2.2 22

Question 2.3 24

Question 2.4 25

# Introduction

This project report is in response to the final assignment for completion of Predictive Modeling module. In this module, we learned regression techniques both linear and logistic, as well as linear discriminant analysis for multi-class classification problems. We also learned the short-comings of ordinary linear regression and how Ridge and Lasso methods can help us overcome the short comings.

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.

**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 best and J the worst. |
| Clarity | cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3 |
| Depth | The Height of a cubic zirconia, 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. |

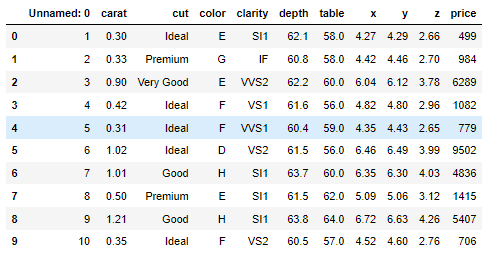
*Table 1: ProblemStatement1\_DataDictionary*

**Question 1.1.**

Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Uni-variate and Bi-variate Analysis.

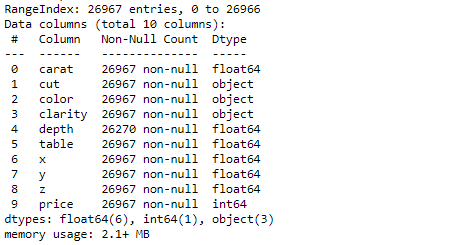
We start by importing the data and checking the first few records. A snapshot of the same is below

Some quick observations on the data after looking at the sample records.

Illustration 1: First 10 records from the data

1. There are 11 features in the data, of which the first feature 'Unnamed: 0' seemed to be the record id which should be dropped
2. cut, color and clarity are the categorical features
3. The carat, depth, table, x, y, & z are the continuous features
4. The price is the target feature which we need to predict

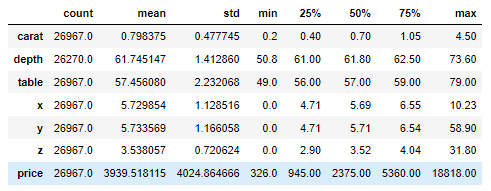
Moving ahead with EDA, we run a info() check on the data and compare the results with our assumptions by looking at the sample data. However, we should not proceed further unless the 'Unnamed: 0' feature is dropped.

Illustration 2: Basic information about the data frame

1. There are 26,967 records in the data set
2. Except depth which has a few null values, all the features have not null values across all records
3. As expected, cut, color & clarity have object datatype which is as per the data dictionary definition
4. All the remaining features are float64 or int64 which is as per the expectations as well

Next we run describe method on the continuous features to get am insight on how the data is distributed.

1. The carat, table, x, y, & z appear to be NOT normally distributed with some outliers appearing in the data as well. This can be further asserted by other means in later in this section

Illustration 3: Results of describe method on continuous features

1. Only depth appears to be normally distributed
2. The features x, y, & z have minimum value as zero which should be deemed as missing data, as they are physical dimensions of zirconia. We will be imputing this later considering correlation across other the features

**Uni-variate analysis**

Our focus in this sub-section will be to analyze each feature individually. For continuous features we will be using Shapiro-Wilk test, Distribution plot, Skewness test, Box plots. For categorical features we will use count plots.

**Continuous features**

| Feature | Shapiro-Wilk test | % age Outliers | Skewness | Distribution plot | Box plot |
| --- | --- | --- | --- | --- | --- |
| carat | Stats = 0.893  p value = 0.000 | 2.455% | 1.12 | Illustration 4: Distribution plot and box plot for carat | |
| depth | Stats = NaN  p value = 1.000 | 2.455% | -0.03 | Illustration 5: Distribution plot and box plot for depth | |
| table | Stats=0.955  p=value=0.000 | 3.63% | 0.77 | Illustration 6: Distribution plot and box plot for table | |
| x | Stats=0.955  p=value=0.000 | 3.69% | 0.39 | Illustration 7: Distribution plot and box plot for x | |
| y | Stats=0.894  p=value=0.000 | 3.745% | 3.85 | Illustration 8: Distribution plot and box plot for y | |
| z | Stats=0.912  p=value=0.000 | 3.831% | 2.57 | Illustration 9: Distribution plot and box plot for z | |

*Table 2. Results of uni-variate analysis of continuous features*

As we can see from the above analysis, except depth rest all the features do not follow Gaussian curve, in fact they are multi-modal and hence are NOT normally distributed. We can also observe that the data has lot of outliers. However, we will not be treating the outliers as in jewelry industry quiet often we see stone that stand out from the rest and have unusually high carat value and price relatively.

**Categorical features**

| *Feature* | *Bar plot against mean price* | *Count plot* |
| --- | --- | --- |
| **color** | Illustration 10: Bar plot with mean price for color | Illustration 11: Count plot for color |
| **clarity** | Illustration 12: Bar plot with mean price for clarity | Illustration 13: Count plot for clarity |
| **cut** | Illustration 14: Bar plot with mean price for cut | Illustration 15: Count plot for cut |

*Table 3. Results of uni-variate analysis of categorical features*

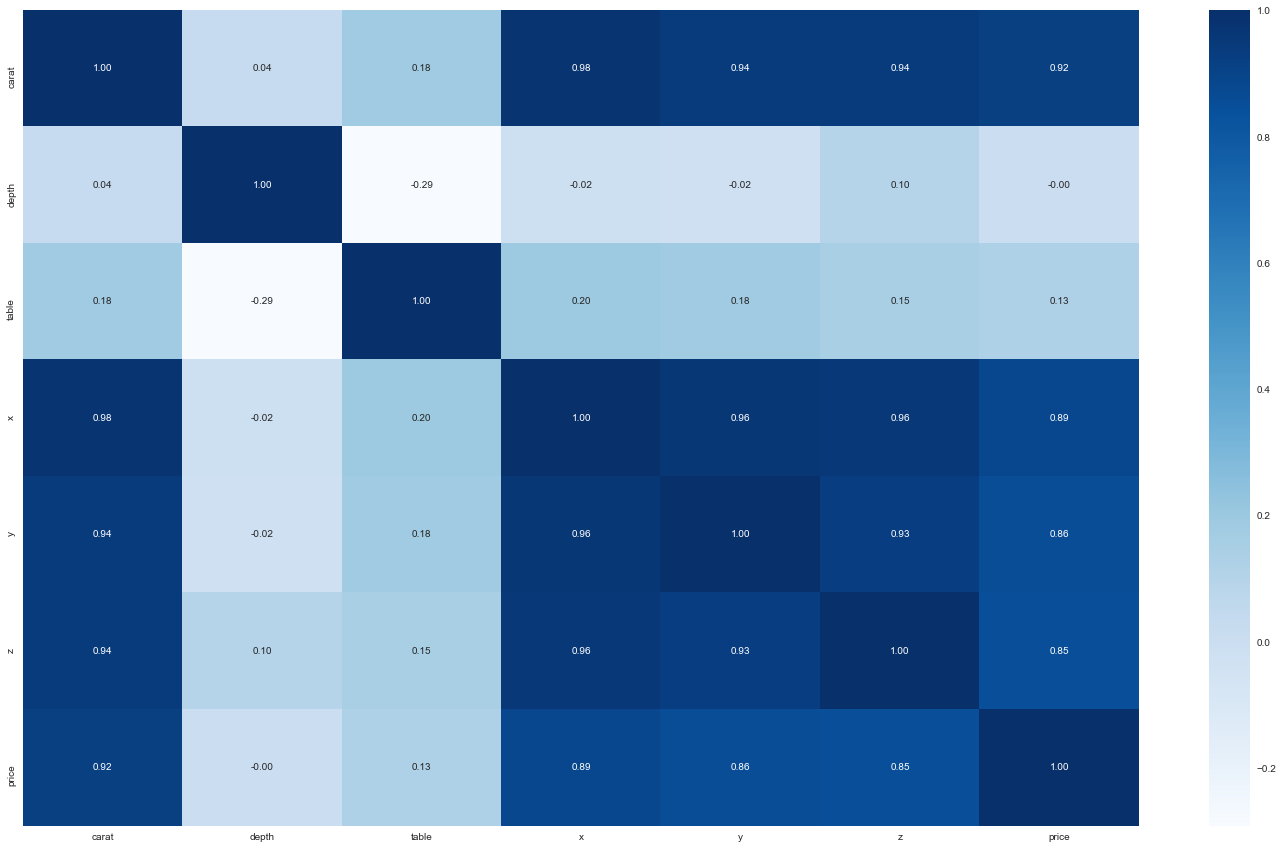
Looking at the bar plots for the mean price & record count plots on the categorical features, below are the observations when arranged in increasing order of ratings

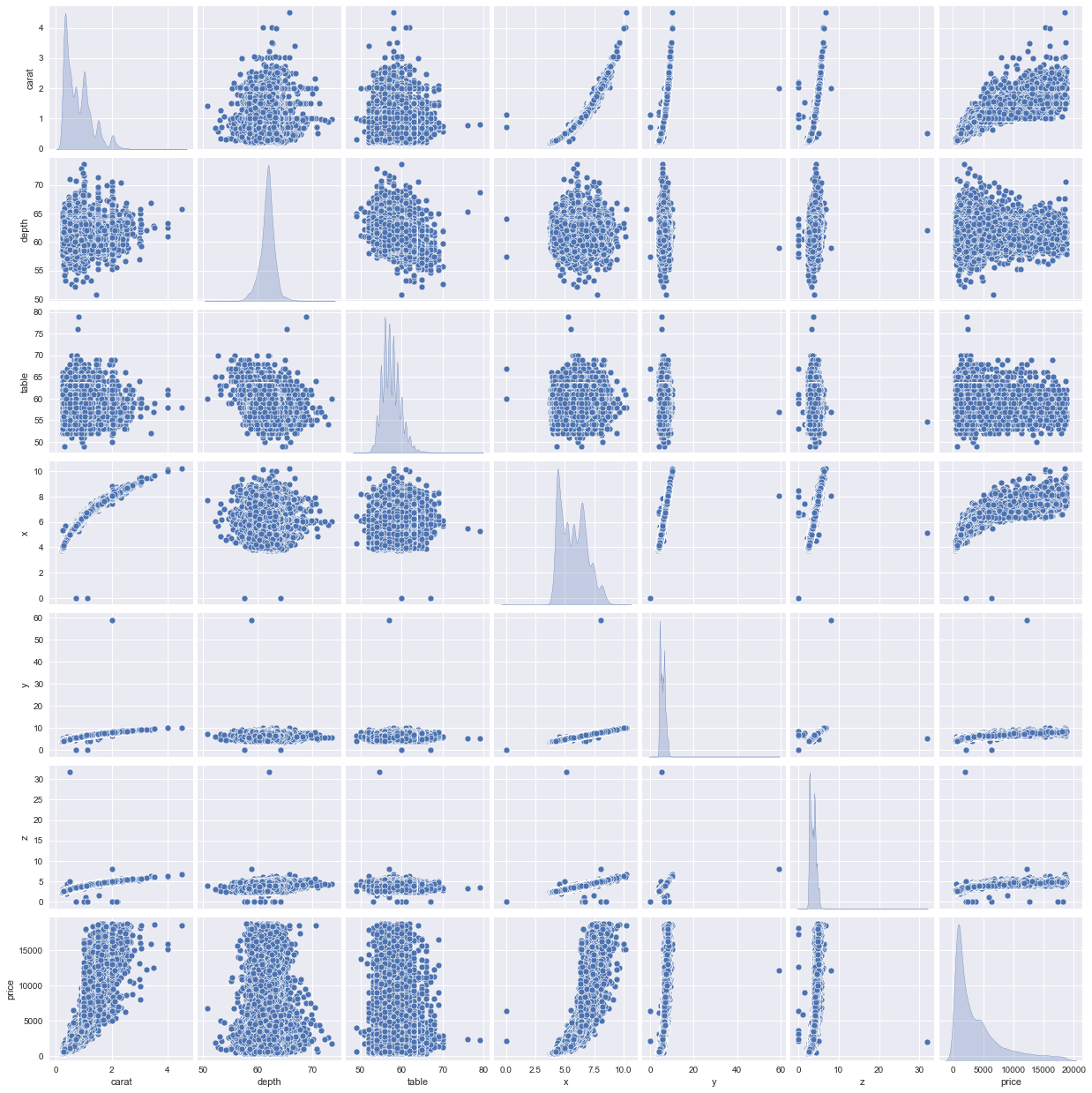
1. The gem stones with the top 3 best colors have higher mean price than the rest of the colors. The count distribution is slightly left skewed with color 'G' having the maximum number of records
2. The clarity feature appears to slightly follow a normal distribution with the mean price when arranged in the increasing order of desirability. The maximum number of observations are in the mid-range, where as there are no observations for the worst and best of the best qualities. The count plot however, is heavily right skewed
3. The good and ideal cut gems have the best mean price. The record counts steadily increases from fair to ideal with ideal having maximum number of records

**Multi-variate analysis**

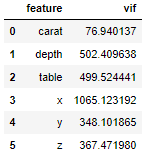
We will use pair plots, correlation heat map & variance inflation factor to determine the relationship between the continuous features and Chi-Squared contingency analysis for categorical features.

**Continuous features**

Illustration 17: Heat map showing correlation between continuous features using Pearson method

Illustration 16: Pair plot of continuous features

The results the pair plot, correlation heat map created using Pearson method and the variance inflation factor helps us draw the following conclusions

Illustration 18: Variance inflation factor calculation

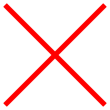
1. carat has very strong correlation with x, y, z and eventually on the price
2. The heat map shows that depth and table have no correlation with any of the other features and nor between themselves, but they would bring in a lot of variance while predicting the results and may not be good features to continue with
3. x, y, & z have very strong correlation among themselves and also with carat. We would later need to decide which of these features should be dropped, else we will risk building a model with high multi-collinearity

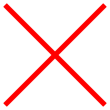
**Categorical features**

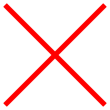
The results of Chi-Squared contingency test suggest all the categorical features are correlated. Further below the section, a set of cross tabs have been put in place that capture the differences between observed values and expected values for each pair wise of combination of clarity, cut and color. Below is a table with results from chi-squared test

| Combination | Statistics | P value |
| --- | --- | --- |
| Color vs cut | 181.61 | 3.61E-026 |
| Clarity vs color | 954.17 | 1.02E-172 |
| Cut vs clarity | 1967.76 | 0 |

*Table 4: Results of Chi-Squared test between categorical features*

Illustration 19: color vs cut, difference between observed values and expected values

Illustration 20: color vs clarity, Difference between observed values and expected values

Illustration 21: clarity vs cut, difference between observed values and expected values

Question 1.2

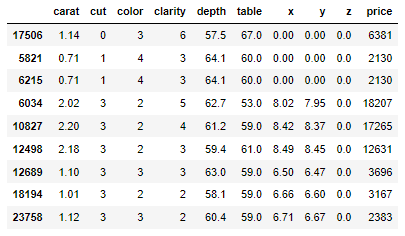
Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

There are 697 records where 'depth' feature had null/missing values. They were imputed using the median value of the feature across different values of 'cut'. The values used to impute missing values are in the table below.

|  |  |
| --- | --- |
| **Cut** | **Median value of depth** |
| Fair | 65 |
| Good | 63.3 |
| Very Good | 62.1 |
| Premium | 61.4 |
| Ideal | 61.8 |

*Table 5: Reference data used for imputing null values of 'depth'*

There are additional 8 records which had zero's for z. Within these 8 records three had zero values for x, & y as well. These features represent the physical dimensions of the gem stones and hence cannot be zero.

Illustration 22: Records with zero for either x, y, & z

The zero values were imputed using the median(as there are outliers in the data) of respective features.

'depth' & 'table' features have values in comparable range to each other. Whereas carat, x, y, & z have values in comparable range to each other but is different from 'depth' & 'table'. The categorical features are of ordinal nature. As our intent is to predict the price of the stones using these features and also identify the important features which drive the price, **for better interpretability we should not scale the data**.

**Question 1.3**

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

The first step in building the model is to convert the categorical features into numerical fields. Since all the categorical features are of ordinal nature, replacing them with 1 to *n*(*n* being the count of unique values of respective categorical feature) would be suffice the use case here. Approaches like OneHotEncoding, get\_dummies etc. would further increase the dimensions of the data which would negatively impact the complexity and the performance of the model.

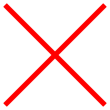
Before we proceed with the linear regression model building, it would be good to know the importance of the features by running decision tree and random forest models on the processed data. Below are the results of decision tree and random forest models built on the entire data set.

| Decision Tree | Random forest |
| --- | --- |
| Illustration 23: Percentage contribution for features in deciding the price using decision tree | Illustration 24: Percentage contribution for features in deciding the price using random forest |
| Illustration 25: Cumulative sum of contribution for features in deciding the price using decision tree | Illustration 26: Cumulative sum of contribution for features in deciding the price using random forest |

*Table 6: Feature importance using decision tree and random forest regressors*

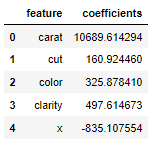
Both the models indicate that the features carat, y, clarity, & color are the most important features in predicting the price. The rest of the features contribute very little, almost close to zero percentage in predicting the price. The features x, & z were highly correlated to y and carat, so it is understandable that their contribution is minimal. Depth and table has almost no correlation with price and hence don't seem to contribute in predictions. Interestingly, cut also doesn't seem to be contributing to price prediction.

Next step, we split the data in the training and test with 70:30 ratio and build a linear regression model. After running multiple versions of the model with different combinations of features included, we can see that iteration#5 seems to be best combination when the RMSE, Adjusted R-squared and AIC scores are compared. Both for training and testing data.

Illustration 27: Comparison of model performances with different combinations of features

The finalized sets of the coefficients of the model is as below.

The intercept comes out to be -4893.502239.

Illustration 28: Coefficients of features generated from model

The mathematical formulae of the model would be:

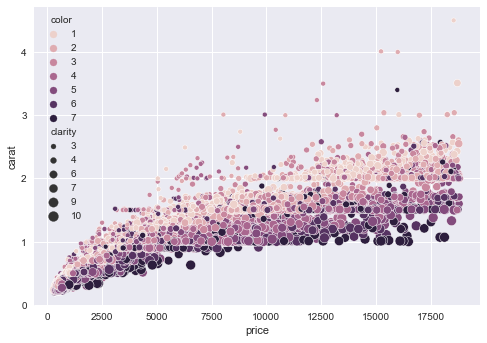
p*rice =* ***-4893.502239*** *+* ***10689.614294*** *x* ***carat*** *+* ***497.614673*** *x* ***clarity*** *+* ***497.614673*** *x* ***clarity*** *+* ***160.924460*** *x* ***cut*** *+* ***325.878410*** *x* ***color*** *-* ***835.107554*** *x* ***X***

Question 1.4

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

From the final version of linear regression model, we can make following insights and recommendations.

1. Insight 1 – The top 5 features contributing to prediction of the price of the stones are – carat, clarity, color, cut and length of the stone

Illustration 29: Scatterplot of carat vs price with color and clarity

1. Insight 2 – Carat is the biggest factor in determining the price of the stones.

Recommendation: High carat stones encourage higher prices. carat has multi-modal distribution of the information. Price buckets can be created and tagged to each mode of carats say a price bucket for stones less than 1 carat, greater than 1 carat but less than 1.5 carat, greater than 1.5 but less than 2 and so on

1. Insight 3 - Color and Clarity together play an important factor determining the price point of stones

Recommendation: A small carat stone can still be high priced if it is of best color and clarity. Price buckets identified on the basis of carat earlier can be further segregated considering all relevant combinations of clarity and color.

1. Insight 4 – Cut has the least importance among the 4 C's but we also saw that majority of the stones are of ideal(the best) cut.

Recommendation: Each step increase in the quality of the cut increases the price by a smaller bit as compared to other features. For obvious, reasons the price point of the stones should get influenced by quality but it shouldn't be a critical factor.

**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.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package - **yes/no**? |
| Salary | Employee salary |
| age | Age in years |
| educ | Years of formal education |
| No\_young\_children | The number of young children (younger than 7 years) |
| No\_older\_children | Number of older children |
| foreign | foreigner – y**es/no?** |

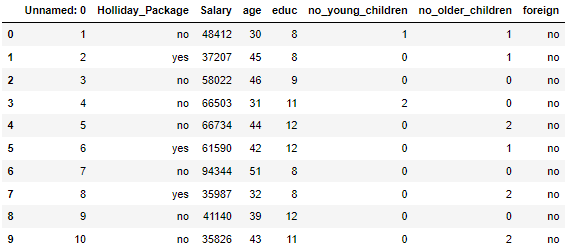
*Table 7: ProblemStatement2\_DataDictionary*

**Question 2.1**

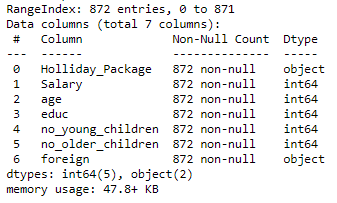
Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

We start by importing the data and checking the first few records. A snapshot of the same is below

Some quick observations on the data after looking at the sample records.

Illustration 30: First 10 records from the data

1. There are 8 features in the data, of which the first feature 'Unnamed: 0' seemed to be the record id which should be dropped
2. Holliday\_Package is the target feature
3. Salary, age, no\_young\_children and no\_older\_children are continuous features
4. educ is an ordinal feature
5. foreign is a categorical feature

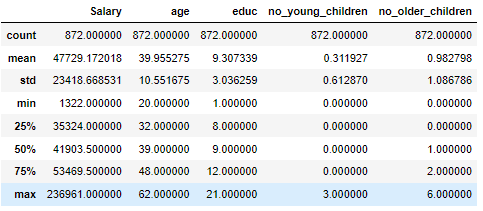
Illustration 31: Basic information about the data frame

Moving ahead with EDA, we run a info() check on the data and compare the results with our assumptions by looking at the sample data. However, we should not proceed further unless the 'Unnamed: 0' feature is dropped.

1. There are 872 records in the data set
2. There are no records with null values
3. foreign and Holliday\_Package have object datatype which is as per the data dictionary definition
4. Rest of the features are int64 which is as per the expectations as well

Next we run describe method on the continuous features to get am insight on how the data is distributed.

1. Except age none of the features appear to be normally distributed as there is sizable gap between mean and 50%-ile values of the features

Illustration 32: Results of describe method on continuous features

1. The zero values in no\_young\_children and no\_older\_children is expected and are not missing values

**Uni-variate analysis**

Our focus in this sub-section will be to analyze each feature individually. For continuous features we will be using Shapiro-Wilk test, Distribution plot, Skewness test, Box plots. For categorical features we will use count plots.

**Continuous features**

| Feature | Shapiro-Wilk test | % age Outliers | Skewness | Distribution plot | Box plot |
| --- | --- | --- | --- | --- | --- |
| Salary | Stats = 0.749  p value = 0.000 | 6.54% | 3.1 | Illustration 33: Distribution plot and box plot for Salary | |
| age | Stats = 0.975  p value = 0.000 | 6.54% | 0.15 | Illustration 34: Distribution plot and box plot for age | |
| no\_young\_children | Stats = 0.559  p value = 0.000 | 30.734% | 1.95 | Illustration 35: Distribution plot and box plot for no\_young\_children | |
| no\_young\_children | Stats = 0.812  p value = 0.000 | 30.96% | 0.95 | Illustration 36: Distribution plot and box plot for no\_older\_children | |

*Table 8: Results of uni-variate analysis of continuous features*

As we can see from the above analysis, none of the features follow Gaussian curve, in fact they are multi-modal and hence are NOT normally distributed. We can also observe that the data has a few percentage of outliers.

**Non-continuous features**

| Feature | Count plot |
| --- | --- |
| educ | Illustration 37: Countplot for educ |
| foreign | Illustration 38: Countplot for foreign |

*Table 9: Countplots for non continuous features*

Looking at the bar plots for record count plots on the non-continuous features, below are the observations when arranged in increasing order of ratings

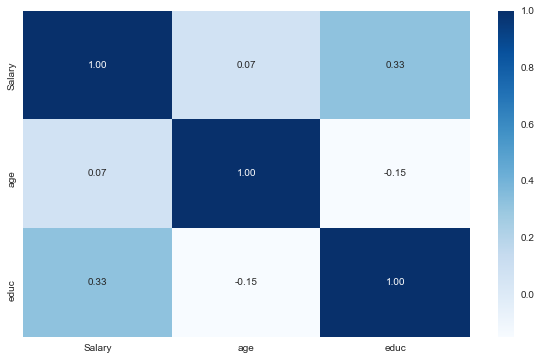
1. The majority of employees have completed either middle schools, or high schools.
2. There are very few employees who have completed any graduation. The post graduation is even lesser.
3. Majority of the employees are natives(non-foreigners)

**Multi-variate analysis**

We will use pair plots, correlation heat map to determine the relationship between the continuous features.

**Continuous features**

Illustration 40: Pair plot of continuous features

Illustration 39: Heat map showing correlation between continuous features using Pearson method

The results the pair plot, and correlation heat map created using Pearson method helps us draw following conclusions

1. Salary and age have no correlation between themselves
2. No\_young\_children and no\_older\_children though being continuous have, understandably, a small range and behave as ordinal categorical features on when put on pair plot

Question 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).

The first step in building the model is to convert the categorical features into numerical fields. Foreign is the only categorical independent feature and a simple replace of 'no' with 0 and 'yes' with 1 would do the trick. The approaches like OneHotEncoding, get\_dummies etc. would not be required.

Before we proceed with logistic regression and linear discriminant analysis model building, it would be good to know the importance of features by running a random forest model. Below are the results of decision tree and random forest models built on the entire data set.

| Percentage contribution | Cumulative Importance |
| --- | --- |
| Illustration 41: Percentage contribution of features using random forest | Illustration 42: Cumulative importance of features using random forest |

*Table 10: Feature importance using random forest classifiers*

The random forest model indicate that the features Salary, age & educ are the most important features in predicting if the employees will buy holiday package or not. The rest of the features contribute around remaining 20% of decision making process.

Next step, we step we split the data in training and test in 70:30 ratio and build logistic regression and linear discriminant analysis models.

To build the model GridsearchCV was used to pick the most optimal set of hyper parameters for both the models.

For Logistic Regression following set of parameters were used

* + *C=0.009111627561154896,*
  + *max\_iter=100,*
  + *multi\_class='multinomial',*
  + *penalty='l2',*
  + *random\_state=1,*
  + *solver='newton-cg',*
  + *tol=0.0001*

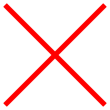
For linear discriminant analysis , we used the solver as *'svd'*

**Question 2.3**

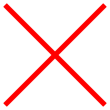
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

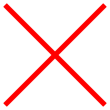
Confusion matrix create using logistic model

Confusion matrix create using linear discriminant analysis model

Illustration 43: Confusion matrix for logistic regression

Below is the grid with precision, recall and F1-score of both the models on training and test data respectively.

Illustration 44: Confusion matrix for linear discriminant analysis

Illustration 45: Performance metrics comparison between training and test for both the models

The scores of both the models across different criteria are in the same range. The overall accuracy and AUC of logistic regression is slightly better than LDA. The recall scores are also very similar for the class of interest. Overall logistic regression, appears to perform only slightly better than LDA.

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression | Linear Discriminant Analysis |
| Training data | Accuracy = 68%  Illustration 46: ROC of Logistic Regression model using training data  AUC = 74.3% | Accuracy = 67.2%  Illustration 47: ROC of LDA model using training data  AUC=74.2% |
| Test data | Accuracy = 63.7%  Illustration 48: ROC of Logistic Regression model using test data  AUC = 70.5% | Illustration 49: ROC of LDA model using test data  Accuracy = 64.1%  AUC = 70.3% |

*Table 11: Comparison between the accuracy of both the models*

**Question 2.4**

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

**Insight 1:** Employees with Salary below 50,000 are more likely to buy a holiday package

**Recommendation**: Targeting employees with salary less than 50,000 would increase the likelihood of selling the holiday package

**Insight 2:** Employees within the age bracket of 30 and 50 years are more likely to buy holiday package

**Recommendation**: The company should prioritize selling holiday package to employees between 30 and 50 years of age

Illustration 50: Scatterplot between Salary and Age

**Insight 3:** The prediction models don't have very high accuracy. Feature engineering didn't improve the performance of the models either. This can be primarily attributed to the data at hand. Better data can help train an efficient model and improve accuracy of the predictions

**Recommendation**: The company should seek to gather more information of the employees. For example, previous holiday details, when, where, if package was taken, marital status, spouse employment details, etc.

Illustration Index

Illustration 1: First 10 records from the data 4

Illustration 2: Basic information about the data frame 4

Illustration 3: Results of describe method on continuous features 5

Illustration 4: Distribution plot and box plot for carat 5

Illustration 5: Distribution plot and box plot for depth 6

Illustration 6: Distribution plot and box plot for table 6

Illustration 7: Distribution plot and box plot for x 6

Illustration 8: Distribution plot and box plot for y 6

Illustration 9: Distribution plot and box plot for z 7

Illustration 10: Bar plot with mean price for color 7

Illustration 11: Count plot for color 7

Illustration 12: Bar plot with mean price for clarity 8

Illustration 13: Count plot for clarity 8

Illustration 14: Bar plot with mean price for cut 8

Illustration 15: Count plot for cut 8

Illustration 16: Pair plot of continuous features 9

Illustration 17: Heat map showing correlation between continuous features using Pearson method 10

Illustration 18: Variance inflation factor calculation 10

Illustration 19: color vs cut, difference between observed values and expected values 11

Illustration 20: color vs clarity, Difference between observed values and expected values 11

Illustration 21: clarity vs cut, difference between observed values and expected values 12

Illustration 22: Records with zero for either x, y, & z 12

Illustration 23: Percentage contribution for features in deciding the price using decision tree 13

Illustration 24: Percentage contribution for features in deciding the price using random forest 13

Illustration 25: Cumulative sum of contribution for features in deciding the price using decision tree 14

Illustration 26: Cumulative sum of contribution for features in deciding the price using random forest 14

Illustration 27: Comparison of model performances with different combinations of features 14

Illustration 28: Coefficients of features generated from model 15

Illustration 29: Scatterplot of carat vs price with color and clarity 15

Illustration 30: First 10 records from the data 17

Illustration 31: Basic information about the data frame 18

Illustration 32: Results of describe method on continuous features 18

Illustration 33: Distribution plot and box plot for Salary 19

Illustration 34: Distribution plot and box plot for age 19

Illustration 35: Distribution plot and box plot for no\_young\_children 20

Illustration 36: Distribution plot and box plot for no\_older\_children 20

Illustration 37: Countplot for educ 20

Illustration 38: Countplot for foreign 21

Illustration 39: Heat map showing correlation between continuous features using Pearson method 21

Illustration 40: Pair plot of continuous features 22

Illustration 41: Percentage contribution of features using random forest 23

Illustration 42: Cumulative importance of features using random forest 23

Illustration 43: Confusion matrix for logistic regression 24

Illustration 44: Confusion matrix for linear discriminant analysis 24

Illustration 45: Performance metrics comparison between training and test for both the models 24

Illustration 46: ROC of Logistic Regression model using training data 25

Illustration 47: ROC of LDA model using training data 25

Illustration 48: ROC of Logistic Regression model using test data 25

Illustration 49: ROC of LDA model using test data 25

Illustration 50: Scatterplot between Salary and Age 26

Table Index

Table1\_ProblemStatement1\_DataDictionary 3

Table2\_Results\_of\_uni-variate\_analysis\_of\_continuous\_features 5

Table3\_Results\_of\_uni-variate\_analysis\_of\_categorical\_features 7

Table4\_Results\_of\_Chi-Squared\_test\_between\_categorical\_features 11

Table5\_Reference\_data\_used\_for\_imputing\_null\_values\_of\_'depth' 12

Table6\_Feature\_importance\_using\_decision\_tree\_and\_random\_forest\_regressors 13

Table7\_ProblemStatement2\_DataDictionary 17

Table8\_Results\_of\_uni-variate\_analysis\_of\_continuous\_features 19

Table9\_Countplots\_for\_non\_continuous\_features 20

Table10\_Feature\_importance\_using\_random\_forest\_classifiers 23

Table11\_Comparison\_between\_the\_accuracy\_of\_both\_the\_models 25