Predictive modeling Project

Name: Saswat Sidhant Prusty

# Project: Predictive Modelling

# Batch: PGP Batch B February

# Name: Saswat Sidhant Prusty

Contents

[4. Problem Statement: 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. 2](#_Toc78741178)

[5. 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. 2](#_Toc78741179)

[6. 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? 6](#_Toc78741180)

[7. 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. 7](#_Toc78741181)

[8. Inference: Basis on these predictions, what are the business insights and recommendations. 7](#_Toc78741182)

[9. 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. 8](#_Toc78741183)

[10. 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. 8](#_Toc78741184)

[11. 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). 10](#_Toc78741185)

[12. 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. 11](#_Toc78741186)

[13. Inference: Basis on these predictions, what are the insights and recommendations. 13](#_Toc78741187)

[Figure 1 Reading Data 4](#_Toc78749726)

[Figure 2 Data Types 4](#_Toc78749727)

[Figure 3 Checking for Null Values 5](#_Toc78749728)

[Figure 4 Describe Dataset 5](#_Toc78749729)

[Figure 5 Univariate Analysis 5](#_Toc78749730)

[Figure 6 Univariate Analysis of Price 6](#_Toc78749731)

[Figure 7 Checking for outliers 6](#_Toc78749732)

[Figure 8 After Treatment of outliers 7](#_Toc78749733)

[Figure 9 Box plot of Price 7](#_Toc78749734)

[Figure 10 Bi-Variate Analysis 8](#_Toc78749735)

[Figure 11 Correlation 8](#_Toc78749736)

[Figure 12 Null Values after imputing 9](#_Toc78749737)

[Figure 13 Checking for multi-collinearity 9](#_Toc78749738)

[Figure 14 Solving Collinearity 9](#_Toc78749739)

[Figure 15 Splitting Dataset 10](#_Toc78749740)

[Figure 16 Coefficient for the first model 10](#_Toc78749741)

[Figure 17 coefficient for second model 10](#_Toc78749742)

[Figure 18 Score for 1st model 10](#_Toc78749743)

[Figure 19 Score for second Model 10](#_Toc78749744)

[Figure 20 RMSE for first model 10](#_Toc78749745)

[Figure 21 RMSE for second model 10](#_Toc78749746)

[Figure 22 Data Reding 11](#_Toc78749747)

[Figure 23 Data info 11](#_Toc78749748)

[Figure 24 Null values checking 11](#_Toc78749749)

[Figure 25 Education Boxplot 12](#_Toc78749750)

[Figure 26 Age Boxplot 12](#_Toc78749751)

[Figure 27 Number of Children 12](#_Toc78749752)

[Figure 28 Salary Boxplot 12](#_Toc78749753)

[Figure 29 Univariate Analysis 13](#_Toc78749754)

[Figure 30 Label Encoded Holiday Package 13](#_Toc78749755)

[Figure 31 Inserting dummy variable in foreginers 14](#_Toc78749756)

[Figure 32 Splitting dataset 14](#_Toc78749757)

[Figure 33 Creating Logistic Regression Model 14](#_Toc78749758)

[Figure 34 Creating LDA Model 14](#_Toc78749759)

[Figure 35 Score for LDA test mode 14](#_Toc78749760)

[Figure 36 Score for LDA train model 14](#_Toc78749761)

[Figure 37 Score for Logistic Regression Test model 14](#_Toc78749762)

[Figure 38 Score for Logistic Regression train model 14](#_Toc78749763)

[Figure 39 LDA Test Model 15](#_Toc78749764)

[Figure 40 LDA Train Model 15](#_Toc78749765)

[Figure 41 Logistic Regression Test Model 15](#_Toc78749766)

[Figure 42 Logistic Regression Model Train 15](#_Toc78749767)

[Figure 43 LDA Test Model 16](#_Toc78749768)

[Figure 44 LDA Train Model 16](#_Toc78749769)

[Figure 45 LR Test Model 16](#_Toc78749770)

[Figure 46 Lr Train Model 17](#_Toc78749771)

# Problem Statement: 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.

## 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.

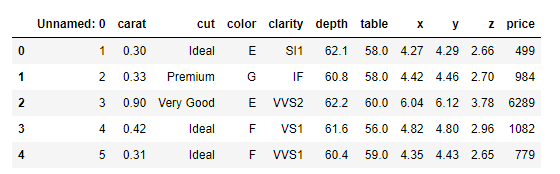


Figure Reading Data

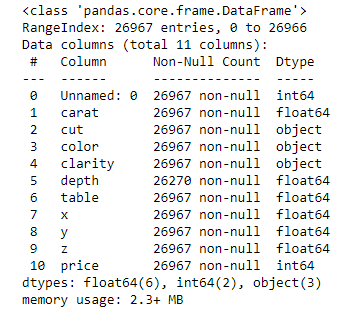


Figure Data Types

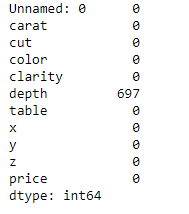


Figure Checking for Null Values

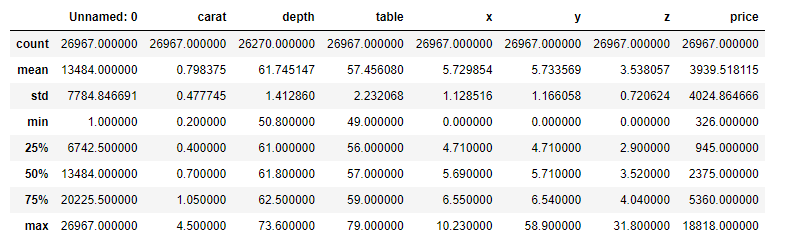


Figure Describe Dataset

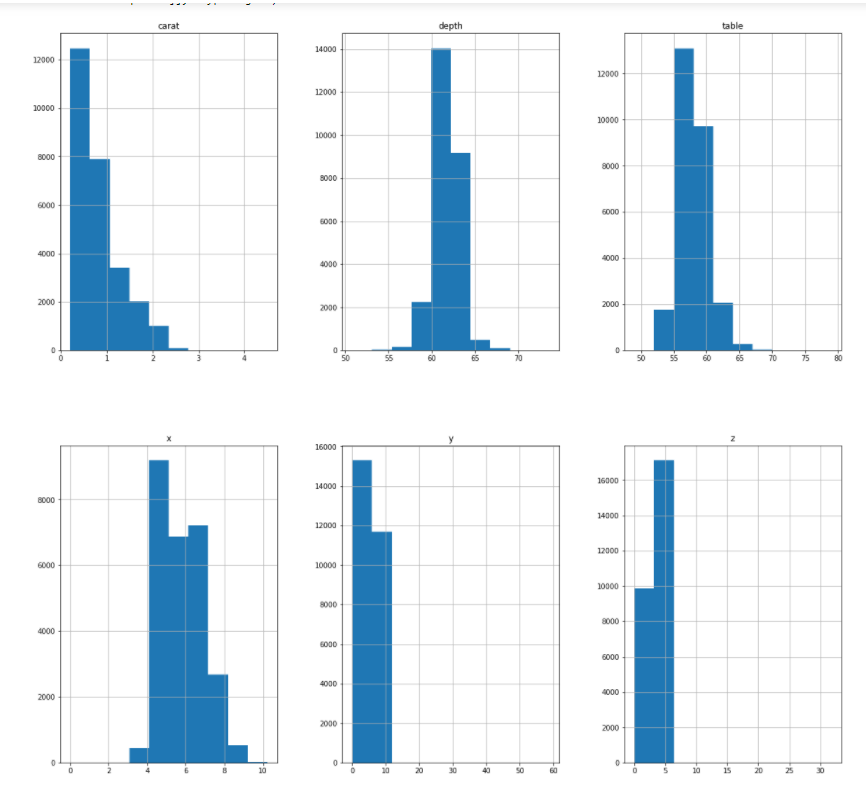


Figure Univariate Analysis

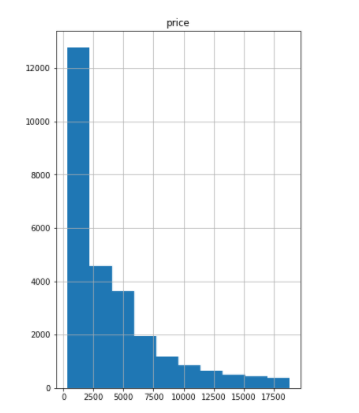


Figure Univariate Analysis of Price

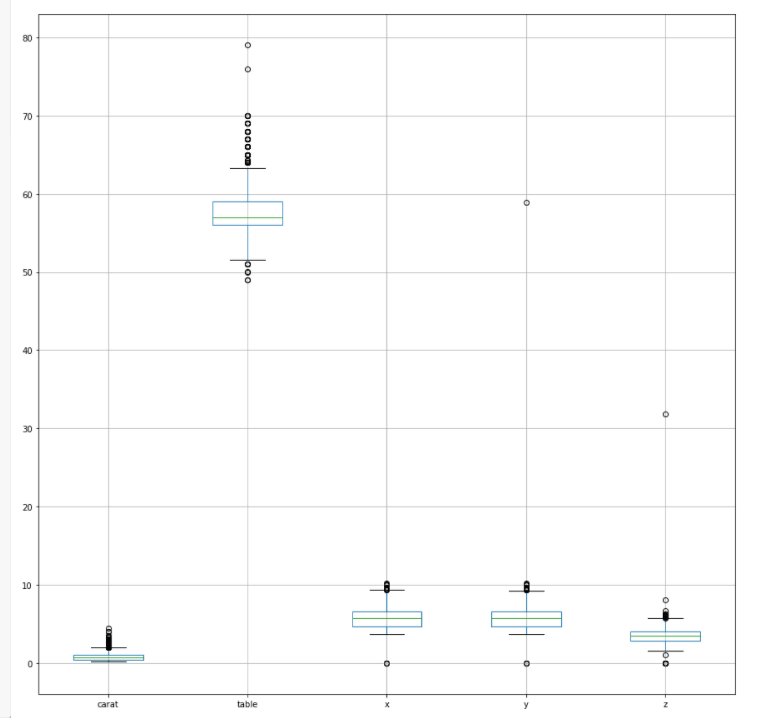


Figure Checking for outliers

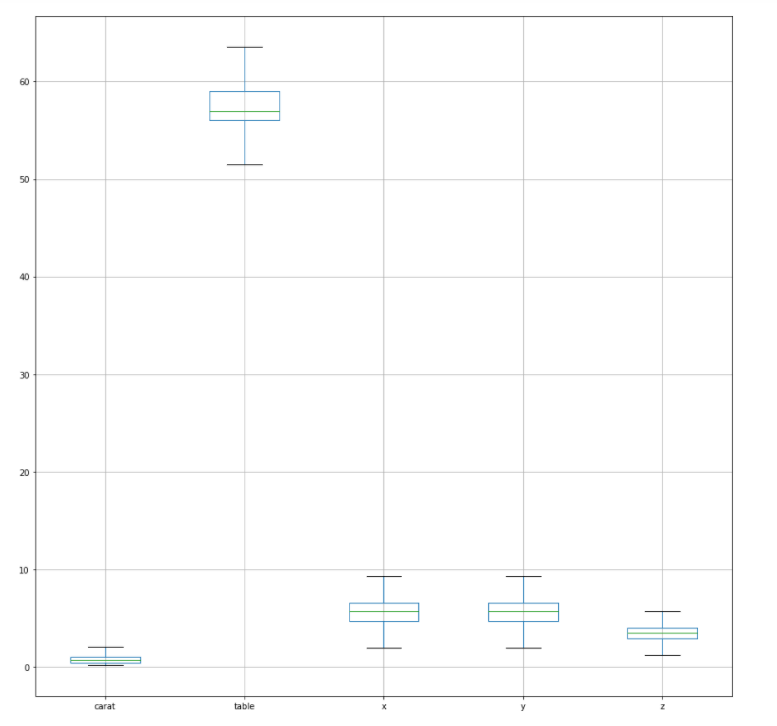


Figure After Treatment of outliers

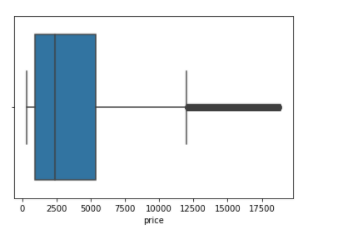


Figure Box plot of Price

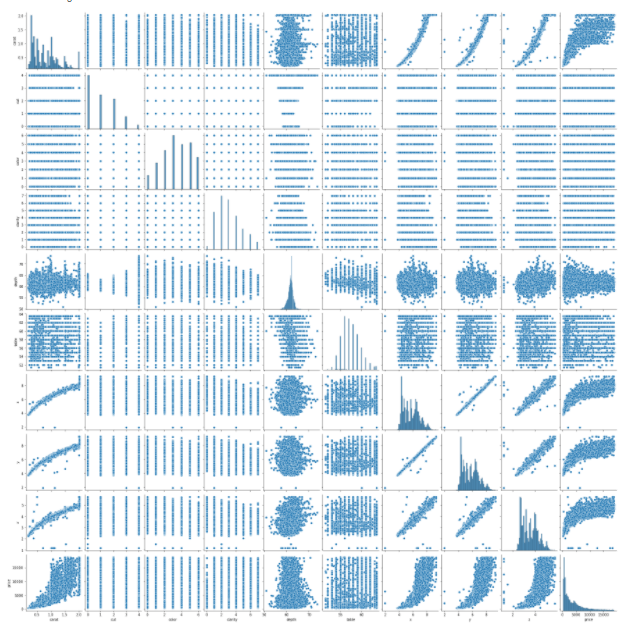


Figure Bi-Variate Analysis

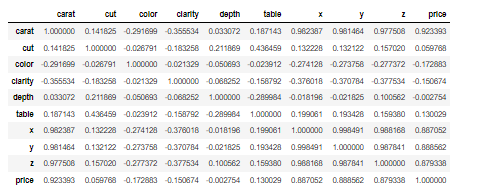


Figure Correlation

Dropped the “Unnamed:0” column. Treated the outliers but did not treat it for Price. Omitting such values might hide real information about the dataset.

## 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?

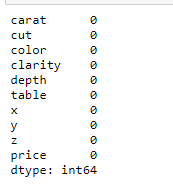


Figure Null Values after imputing

Imputed values for depth. There was error while entering the value of 62.1. Processed the dataset and changed the data type of depth to numeric.

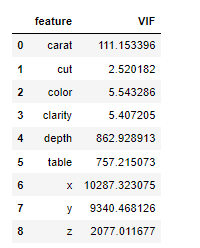


Figure Checking for multi-collinearity

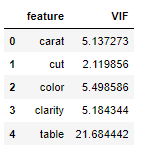


Figure Solving Collinearity

The dataset had a high level of collinearity as can be seen from figure 13. Hence used Variance Inflation Factor to deal with it. In figure 14, one can see, only the important features to be considered for model building. But however, it was seen that we can build two model one using table and the other using depth to better build a better model.

## 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.



Figure Splitting Dataset

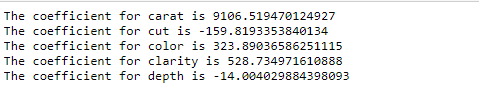


Figure Coefficient for the first model

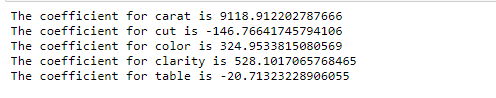


Figure coefficient for second model



Figure Score for 1st model



Figure Score for second Model

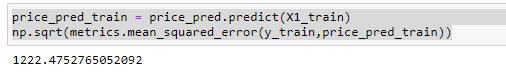


Figure RMSE for first model

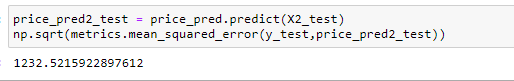


Figure RMSE for second model

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

Some of the insights are:

1. The most important factor determining the price is carat.
2. The order is carat > clarity > colour > cut > table/depth
3. To maximise profits, the business should optimise the buying and selling based on carat.

# 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.

## 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.

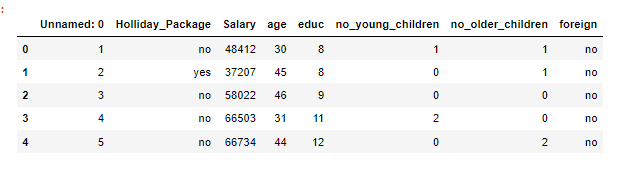


Figure Data Reding

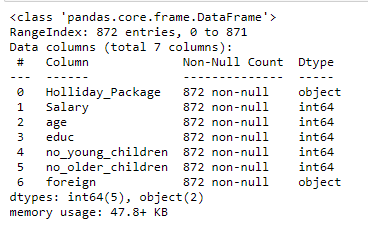


Figure Data info

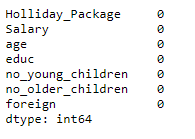


Figure Null values checking

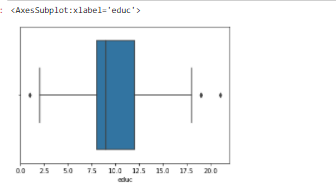


Figure Education Boxplot

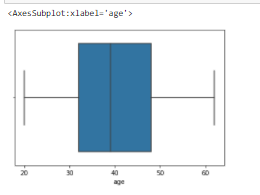


Figure Age Boxplot

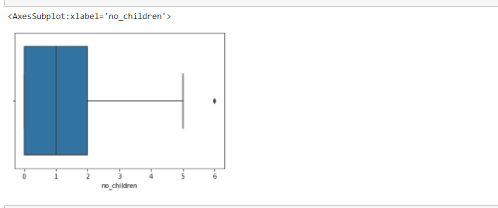


Figure Number of Children

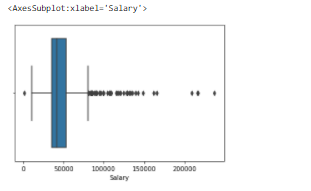


Figure Salary Boxplot

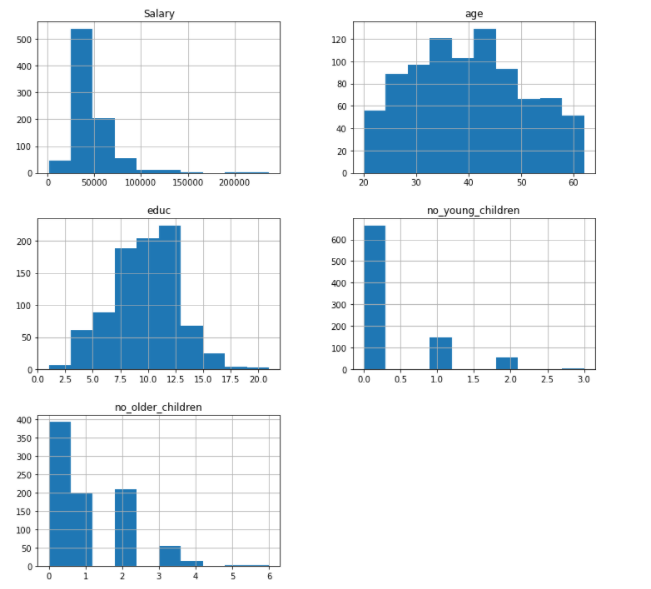


Figure Univariate Analysis

Dropped the “Unnamed: 0” columns, and merged two columns “no\_older\_children” and “no\_younger\_children” into no\_children. Also, seeing the data, it would be wise not to treat outliers, as there are not many outliers and in the salary, where there is a lot of outliers, is best left untreated as it risks omitting information.

## 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).

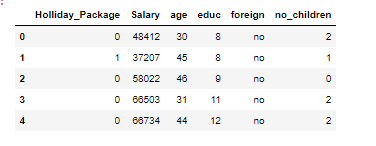


Figure Label Encoded Holiday Package

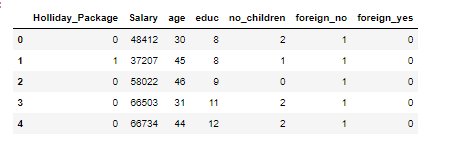


Figure Inserting dummy variable in foreginers



Figure Splitting dataset

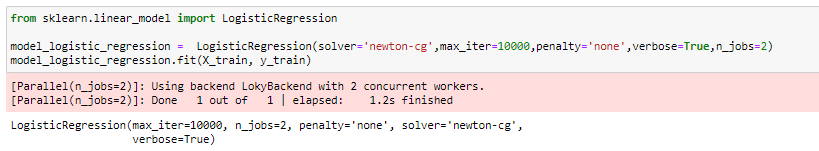


Figure Creating Logistic Regression Model



Figure Creating LDA Model

## 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.



Figure Score for LDA test mode



Figure Score for LDA train model



Figure Score for Logistic Regression Test model



Figure Score for Logistic Regression train model

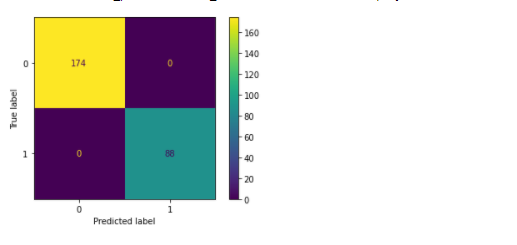


Figure LDA Test Model

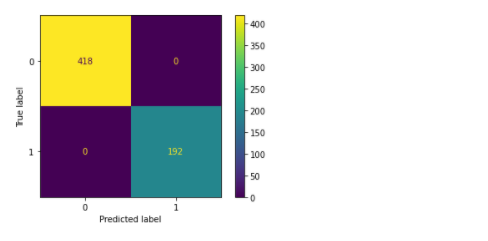


Figure LDA Train Model

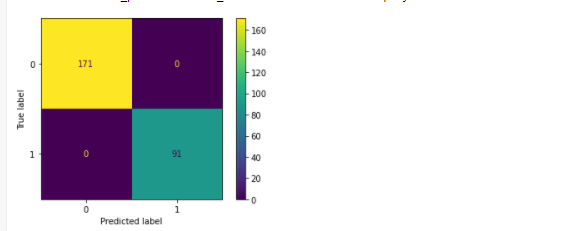


Figure Logistic Regression Test Model

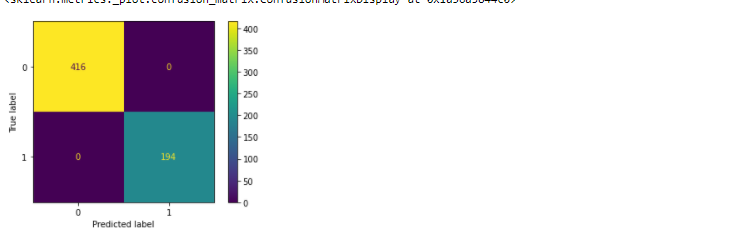


Figure Logistic Regression Model Train

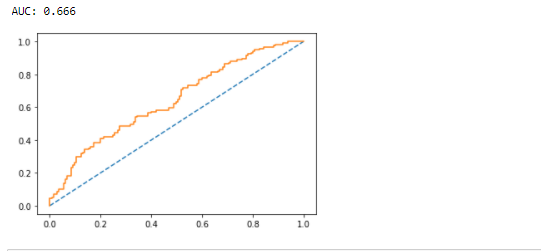


Figure LDA Test Model

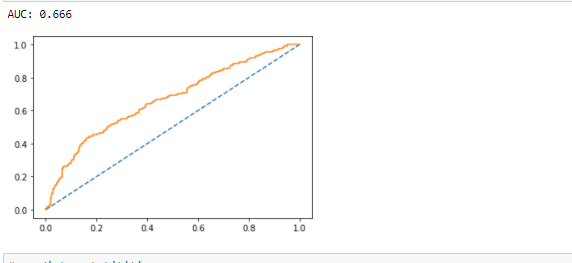


Figure LDA Train Model

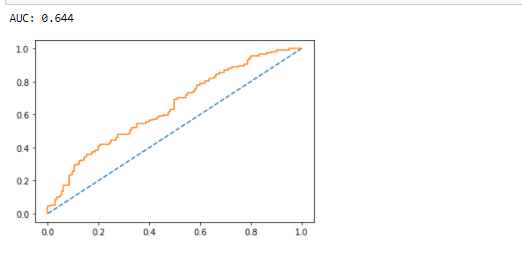


Figure LR Test Model

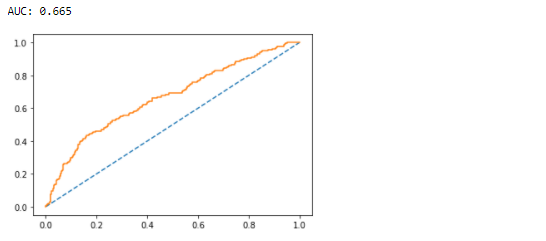


Figure Lr Train Model

As one can see both the models have been able to successfully classify the dataset. But the ROC-AUC curve of LDA is slightly better than logistic regression and for this reason we shall choose the LDA model.

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

Some business insights are:

1. Salary is the most important factor in deciding if employee will opt for the travel package.
2. The order of preferences for the business is as follows:
3. Salary
4. Foreigner
5. Number of Children
6. So, based on these three factors, the travel agencies can target such people. And ideal customer is one with high children, few children and is a foreigner.