Project - Predictive Modelling

|  |
| --- |
| **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. **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. - 9** |
| 1. Provided data set has 26,927 records with 10 Features. 2. Features, Cut, Color and Clarity are Object data types. 3. Features Carat, Depth, Table, X, Y, Z are Numeric Data Types. 4. Length(X), Width(Y) and Height(Z) should not be 0 for a diamond to exist, so clearly there are some anomalies in all the 3 attributes with more anomalies found in Height(Z). 5. X-Zero 7. Y-Zero 9. Z-Zero 11. **Uni Variate Analysis** 13. Carat has a right-skewed distribution with a long right tail with outliers. With Median hugging to the lower quartile. Depth has a perfect normal distribution; it has outliers to the lower and upper quartiles. 15. Table data is almost normal distribution with few outliers, Average diameter seems to be grouped into an even bucket with the median hugging to the left quartile. However, from this stand point the distribution seems normal. 17. Length(X) is right skewed, it specifies most of the population has maximum length of the diamond. 18. Height(Z) and Width(Y) are normally distributed with the values hugging between 0-5 with an occasional outlier, this specifies that the values are on different scales. 19. **Bi-Variate Analysis** 21. Features, Carat and the Length(X) have positive Correlation in a clean transitioning into a diagonal line. 22. Carat, Width(Y) and Height(Z) and weakly correlated to each other. 23. Depth and Table are not Correlated with Carat. 24. Depth has no good correlation with any Feature. 25. Table has no good correlation with any Feature. 26. X has a good positive correlation with Carat and a weak correlation with y and z and dense positive correlation with Price. 27. Y and Z does not have good correlation with any feature. 28. Price has positive correlation with Carat and X and flat weak relations with depth and table and weak corelation with y and z. 29. **Correlation Plot** 31. Actual variables that affect the model will be shown after preparing the model |
|  |

|  |
| --- |
| **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? - 5** |
| 1. We do have 697 null values for depth. 2. We have 3 records with X, Y being 0 and Z is zero for 9 records. 3. Depth is a cumulative result to determine overall size of diamond. Having this value null is incorrect, so I have imputed with the Mean, and it has a normal distribution. 4. The number of records in which X, Y and Z are 0 is small, we can drop them as size of data set is huge. However, I did impute with the Mean as there’s no such case where the Length, Width and Height of the Diamond can be a 0. 5. Price feature has values with a very high number, Scaling Price to a logarithmic scale results in a small presentable number and we can use this to build the Linear Regression model. |
|  |

|  |
| --- |
| **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. - 7** |
| 1. Encoded the Features ‘cut’, ‘color’, ‘clarity’ using the One-Hot Encoding. 2. I had tried the Ordinal Encoder, but it’s clear, One-Hot-Encoding gives more control to identify which Cut, Color or Clarity affects the Price. 3. Data split is at 70:30 with X and Y, Train and Test sets. 5. Top 5 Features that affect Price are Carat, Clarity\_IF, Clarity\_VVS1, Clarity\_VVS2, Clarity\_VS1. Customer can focus on the Carat of the Diamond along with a best Clarity. Also, we can see the worst Color is not being sold and Price is very low for it. 6. **Performance Metrics** 7. **R Square-Training: 0.92053:** R2 ranges between 0-1 and the obtained value is near to 1 and this specifies the **model is good**. 8. **R Square-Testing: 0.92344:** R2 ranges between 0-1 and the obtained value is near to 1 and this specifies the **model is good**. 9. **R Square** values of Train and Test are close/almost equal. There is no deviation for the test set, I believe this model is indeed good for predictions. 10. **RMSE:** Best fit line will give a minimum value of the best fit line. 11. **RMSE-Training Data:** **0.2809**: This value is close to zero and will have a best fit line. 12. **RMSE-Test Data:** **0.2787**: This value is close to zero and is less and proposes to have a best fit line, also this value is close to the Training RMSE. 13. Best fit line for the Training and Test set are really close to each other and I will recommend this model for Predictions. |
|  |

|  |
| --- |
| **1.4 Inference: Basis on these predictions,**  **what are the business insights and recommendations? - 6** |
| 1. Business Insights:    1. Top 5 factors that are directly proportional to Price are **Carat, Clarity\_IF, Clarity\_VVS1, Clarity\_VVS2, Clarity\_VS1**.    2. Top priced diamonds are the ones with the Top 5 coefficients as explained in the above bar plot.    3. This data set does not have information on the Number of sales vs Price.    4. **R Square** value for the **Train & Test** models are ~0.93, it is near to 1 and this score is good to consider the Model for Prediction. It explains 93% of variance in the Price is explained by the Model.    5. **RMSE** on the **Train & Test models** are ~0.28, it is near to 0 and a less Mean Squared Error specifies less deviation from the y hat and y bar thereby a better model and we can use this for Prediction on the real data. 2. Recommendations:    1. Company can get a good profitable share by investing in Cubic Zirconia diamonds with better Carat value and with following Clarity (IF, VVS1, VVS2, VS1, VS2, SI1 and SI2).    2. Company should focus on the Color of the diamonds, Color J does impact in negative Pricing and we should over all remove such Colors from the diamonds and focus on Colors E, F and G for a profitable share.    3. Company can obtain a profitable share from all Cuts (Ideal, Premium, Good and Very Good). |
|  |

|  |
| --- |
| **Problem 2: Linear 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. - 7** |
| 1. Provided data set has 872 records with no-null items in it and it has 7 features. 2. There are no duplicates in the data set. 3. There are no nulls in the data set. 4. Target Class is the Holiday Package and Opted Yes/No is 54% & 46%, almost half and half, this data set has good separability of the Target variable, will be a good candidate for better Prediction model. 5. Features no\_young\_children and no\_older\_children have 25%-50% zeros, however, these are not anomalies, it is natural to have or not have children. 6. Education, Age, Salary are well defined. 7. **Uni-Variate Analysis:**     2. Salary Feature is a slight right-skewed distribution with the majority of the data falls near Mean.    3. Age is a normal distribution where the tails are not touching zero and both limits start at 50 counts.    4. Education is a normal distribution with evened out tails.    5. No\_young\_children feature has the data classified in 3 buckets, more like a classification.    6. No\_older\_children feature has the data classified into 7 buckets and it’s a classification and Kde shows it as a right skewed distribution. 8. Box plot:     2. Salary has outliers, but this cannot be considered as incorrect values, truly we can have high salaried customers opting for holiday packages. 9. **Bi-Variate analysis:**     2. None of the Features have positive or Negative correlations, It’s all scatter.    3. **Independent Variable Comparison:**    4. Holiday\_Package vs Salary:        2. Median-Salary of the Opted/Not-Opted customers is almost near with customers who opted has less salary compared to the one’s who have not opted.    5. Holiday\_Package vs Age:        2. Median-Salary of the Opted/Not-Opted is equal. Customers who opted for the holiday package fall under the low range of age in compared with not-opted.    6. Holiday\_Package vs Education:        2. Less in years of formal education opt for the Holiday Package and the education group density is almost same for the opted and not-opted groups.    7. Holiday\_Package vs number\_of\_young\_children:        2. Most holiday packages were opted with no young children.    8. Holiday\_Package vs number\_of\_older\_children:        2. Holiday package was opted/not-opted have equal range of older children.    9. Holiday\_Package vs Foriegner:        2. Equal proportions of Holiday proportions are accepted by foreigner or not.       3. Most foreigners have opted for holiday package.       4. Non foreigners have not accepted the holiday package. 10. Correlation-Plot:      2. Features are not correlating as in the above pair plot. |
|  |

|  |
| --- |
| **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). – 7** |
| 1. Logistic Regression:    1. Holiday\_package variable is dropped from X and only Holiday\_package was selected in Y.    2. Train-Test split is proportioned at 70-30% with random\_state as 1 and stratify applied for Holiday\_Package.    3. y\_train and y\_test value counts are at 54% and 46%.    5. Iterations are capped at 10,000 and used 2 processors with ‘none’ as penalty and random\_state set to 1, ‘newton-cg’ is chosen as the solver, tolerance is set at 1e-06.    7. With default probability of 0.5 the above is head of 10 records.     10. Grid Search CV gave the above model and scoring will be optimizing the F1 score. 2. LDA:    1. Holiday\_package variable is dropped from X and only Holiday\_package was selected in Y.    2. Train-Test split is proportioned at 70-30% with random\_state as 1 and stratify applied for Holiday\_Package.    3. y\_train and y\_test value counts are at 54% and 46%. |
|  |

|  |
| --- |
| **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. - 7** |
| **Logistic Regression** |
| |  | | --- | | **AUC and ROC – Train & Test - Logistic Regression** | | AUC and ROC curve for Train and Test data set is at 74% and 72% respectively. This is a good score, capturing most of the data falling under the curve and overlapping Test and Train Curves. | |  | | **Important Coefficients – Logistic Regression** | |  | | Important coefficients or Betas are no\_young\_children and foreign, other features seem to have no much importance for the Holiday\_Package. | |  | | **Confusion Matrix for Training Data – Logistic Regression** | | Type 1 Error(Alpha) is lesser than the False Negative Type 2 Error (Beta), chance of incorrectly identifying a customer as Not Opted is more. | |  | | **Confusion Matrix for Testing Data – Logistic Regression** | | Type 1 Error(Alpha) is lesser than the False Negative Type 2 Error (Beta), chance of incorrectly identifying a customer as Not Opted is more. | |  | | **Classification Report for Training Data – Logistic Regression** | |
| As explained by the Confusion matrix above, the Recall Score (Type2 Error) is at 58% and the Overall Test Accuracy is at 67%. This is not a great score and using this model will lead company to wrongly predict customers as Not-Opted, even if they would Opt for the Holiday Package. |
|  |
| **Classification Report for Testing Data – Logistic Regression** |
| As explained by the Confusion matrix above, the Recall Score (Type2 Error) is at 58% and the Overall Test Accuracy is at 65%. The results are closely identical with the Training data, if more data is provided, we can get better results and Predictions. However, this is not a great score and using this model will lead company to wrongly predict customers as Not-Opted, even if they would Opt for the Holiday Package. |
|  |
| **GridSearchCV for Logistic Regression** |
| I have applied GridSearchCV to optimize the model, variety of grid parameters were provided. |
|  |
| **GridSearchCV best model** |
| GridSearchCV provided the optimal model to have max\_iterations of 10000 and the l1 to be the penalty with solver as liblinear. |
|  |
| **GridSearchCV confusion matrix for Training Data** |
| Type 1 Error(Alpha) is lesser than the False Negative Type 2 Error (Beta), chance of incorrectly identifying a customer as Not Opted is more even with GridSearchCV. |
|  |
| **GridSearchCV confusion matrix for Testing Data** |
| Type 1 Error(Alpha) is lesser than the False Negative Type 2 Error (Beta), chance of incorrectly identifying a customer as Not Opted is more even with GridSearchCV. |
|  |
| **GridSearchCV Classification Report for Training Data** |
| As explained by the Confusion matrix above, the Recall Score (Type2 Error) is at 57% and the Overall Test Accuracy is at 67%. This is not a great score and using this model will lead company to wrongly predict customers as Not-Opted, even if they would Opt for the Holiday Package. |
| GridSearchCV Classification Report for Testing Data |
| As explained by the Confusion matrix above, the Recall Score (Type2 Error) is at 53% and the Overall Test Accuracy is at 66%. The results are closely identical with the Training data, if more data is provided, we can get better results and Predictions. However, this is not a great score and using this model will lead company to wrongly predict customers as Not-Opted, even if they would Opt for the Holiday Package. |
|  |
| **Important Coefficients obtained from the best model of GridSearchCv** |
| Importance of the coefficients is also similar to the Logistic Regression, Even with GridSearchCV I could not come up with a better model even with better hyper parameters. If I have more data to mine, then the model will yield better results. |
|  |
| **Linear Discriminant Analysis** |
|  |
| **Important Coefficients** |
| As with the Logisitc Regression, LDA also points us to the foreign and no\_young\_children to be the top features amongst all. |
|  |
| **Confusion Matrix on Training Data** |
| Type 1 Error(Alpha) is lesser than the False Negative Type 2 Error (Beta), chance of incorrectly identifying a customer as Not Opted is more. |
| **Confusion Matrix on Testing Data** |
| Type 1 Error(Alpha) is lesser than the False Negative Type 2 Error (Beta), chance of incorrectly identifying a customer as Not Opted is more. |
|  |
| **Classification Report Training Data** |
| As explained by the Confusion matrix above, the Recall Score (Type2 Error) is at 58% and the Overall Test Accuracy is at 66%. This is not a great score and using this model will lead company to wrongly predict customers as Not-Opted, even if they would Opt for the Holiday Package. |
|  |
| **Classification Report Testing Data** |
| As explained by the Confusion matrix above, the Recall Score (Type2 Error) is at 49% and the Overall Test Accuracy is at 64%. The results are closely identical with the Training data, moreover the Recall Score dropped, if more data is provided, we can get better results and Predictions. However, this is not a great score and using this model will lead company to wrongly predict customers as Not-Opted, even if they would Opt for the Holiday Package. |
|  |
| **ROC\_AUC curve for Training and Testing data** |
| AUC and ROC curve for Train and Test data set is at 73% and 71% respectively. This is a good score, capturing most of the data falling under the curve and overlapping Test and Train Curves. |
|  |
| **Threshold adjustment and Model Performance Measures** |
| For various Threshold values the Accuracy, F1 and Recall scores were calculated, found that 0.3 threshold results in best Model Performance Measures. |
|  |
| **Classification Report after the Threshold adjustment of 0.3 instead of 0.5** |
| Above classification reports are for 0.5 and 0.3 Threshold respectively. With 0.3 threshold we can push the recall score to 90% with F1 Score to 66% and with a dip in precision and accuracy. If Company is only focussing on identifying Wrongly-Not-Opted classes (Type 2 Error) We can use the model with a Threshold of 0.3. |

|  |
| --- |
| **2.4 Inference:**  **Basis on these predictions, what are the insights and recommendations.**  **Please explain and summarise the various steps performed in this project.**  **There should be proper business interpretation and actionable insights present. - 6** |
| 1. Insights:    1. Chance of predicting a customer Opting for Holiday Package was at 49% while Testing the model, which will only predict 50 of 100 customers correctly.    2. F1 Score is better than the recall score for Holiday Package Opted Class, which specifies the Model is accurate to Predict the Class.    3. Overall Accuracy is at 64% which is not great, we can improve it by altering the Threshold.    4. AUC\_ROC curve for Train and Test data set is 73% and 71%, 70% of the data falls under the curve which is a good thing for the Model.    5. Changing the Threshold to 0.3 from 0.5 will improve the accuracy to 0.67, F1 to 0.68 and Recall Score to 0.758. By doing the threshold shift we can improve the numbers, but we are allowing data to be considered as Opted by the system, when it was not! This will cause inconsistencies in model prediction on production data. 2. Steps applied for Logistic Regression and LDA    1. Performed EDA: Described data, null values identified, duplicate values identified, and anomalies were checked.    2. Proportion of the Target Classes were checked.    3. Converted the categorical columns to numeric equivalent.    4. Performed UniVariate and BiVariate analysis to understand relationships and the distribution.    5. Understood the correlation between the data using Correlation Plots    6. Performed Logistic Regression and LDA and concluded with a good model. To help we have used Confusion Matrix and Classification Report for the Train and Test data sets.    7. Applied alteration of threshold and improved the Model Performance Measures to explain the better predictability of the Model. 3. Recommendations and Actions:    1. Company should focus on selling the holiday package to foreign customers as the Coefficient is very high. They can introduce special packages to customer interested countries and curate a custom itinerary if required.    2. Company can focus on customers with no or less young children and curate a destination packages for them, like a beach resort or hill station resort in the itinerary.    3. The above two features have significant impact on Customer Holiday Package Opted or Not Opted decision, other features have very less significance.    4. Company can proceed with existing offers for the Education, Age and Older Children and Salary Features, but focus should be on Foreign and no\_young\_children. |
|  |