LEAD SCORING LOGISCTIC REGRESSION ASSIGNMENT

SUMMARY REPORT

To work on this business problem, the approach that was followed was:

- **Data Preparation, Preprocessing & Missing value treatment – Since this dataset contains certain columns with missing values, outliers, simple "Select" field, binaries in the form of yes/no etc. So in order to fit them into our logistic regression model, these were treated in the manner required for our model.**

- **Exploratory Data Analysis and visualization – This is one of the essential part before starting the model training. The EDA over different variables (categorical and numerical) gave a clear picture of their pattern in the dataset also their relation with the other variables.**

- **Dummy Variable Creation – The dataset contains columns with different value ranging from single digits to doubles. So for non-binary categorical columns we perform this operation and omitting one variable(for k variables, we require k-1 dummy variables).**

- **Test-Train Split – Then the sample dataset was splitted based on train- test splitting (after the necessary libraries were imported). The ration kept was conventional 70-30.**

- **Model Building – Then model was built using statsmodel.**

- **Feature Selection Using RFE – Since the number of variables is very large so recursive feature elimination was used to get the best 15 variables.**

- **Checking VIFs – It was done to remove the variables with high correlation. The threshold was considered to be 5.**

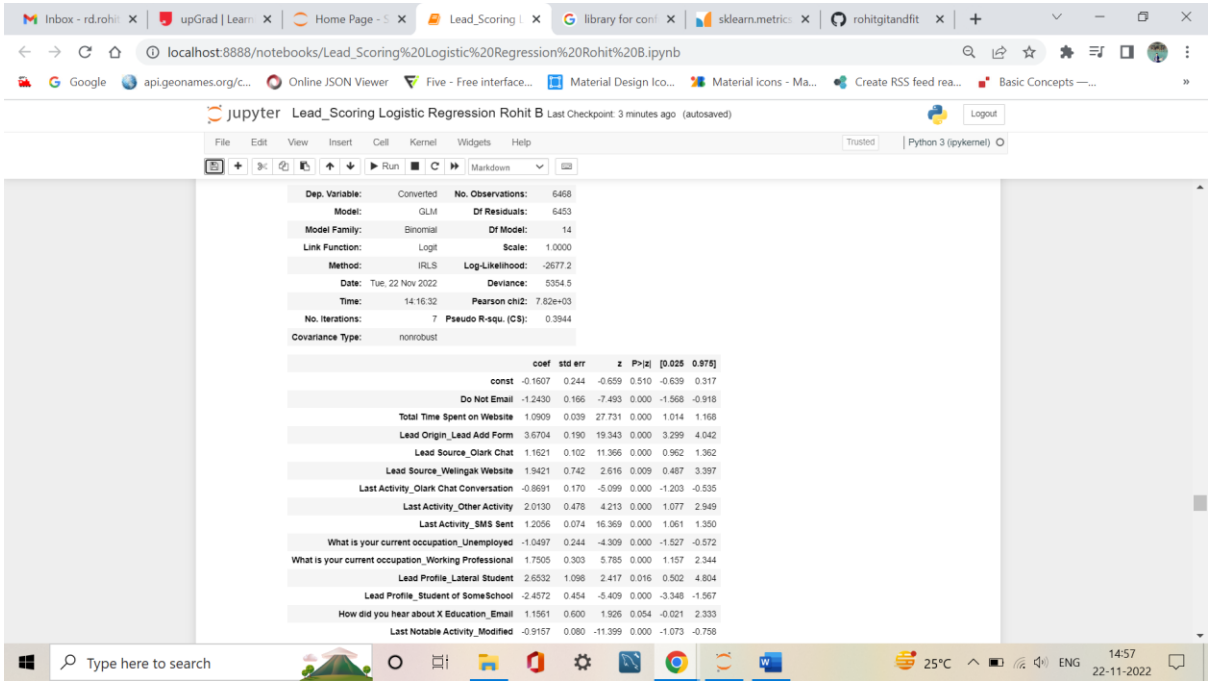
-

- **Plotting the Confusion Matrix, Accuracy, Sensitivity, Specificity**

- **Plotting ROC Curve - Shows tradeoff between sensitivity and specificity (increase in one will cause decrease in other). The closer the curve follows the y-axis and then the top border of the ROC space, means more area under the curve and the more accurate the test.**

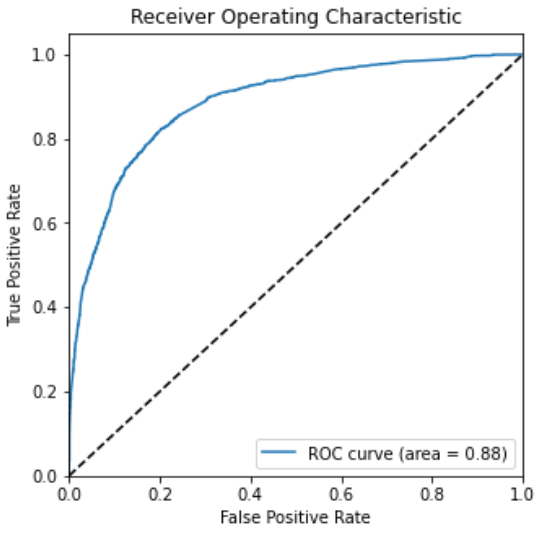
- **Finding optimal value of the cut off - In Sensitivity-Specificity-Accuracy plot 0.35 probability looked optimal.**
- **Precision-Recall Trade off - In Precision-Recall Curve 0.4 looked optimal**
- **Model Building - Then model was built using statsmodel.**

The Final Model

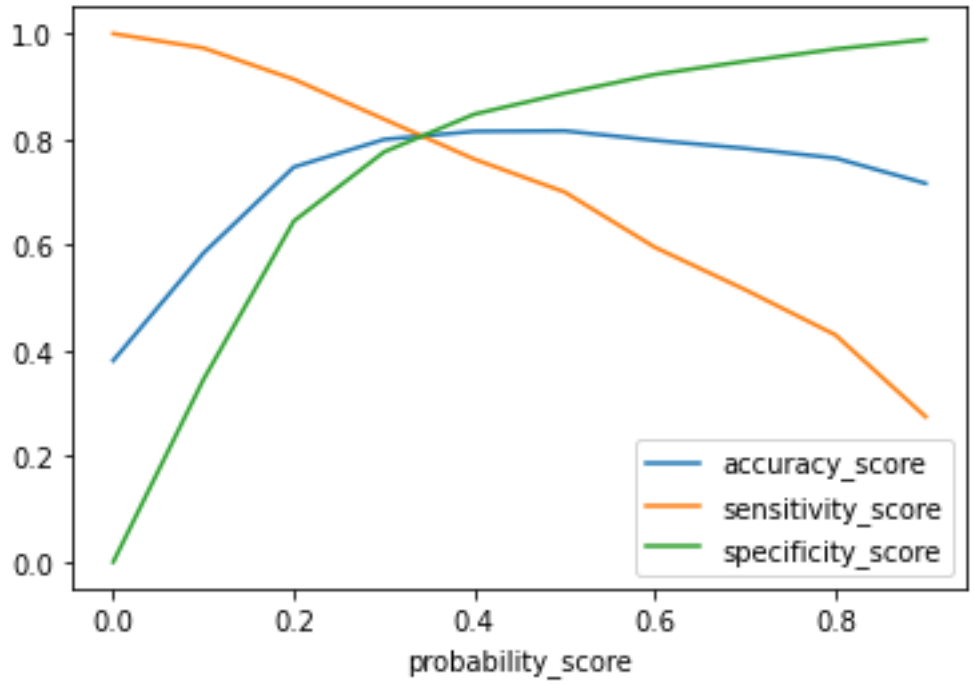


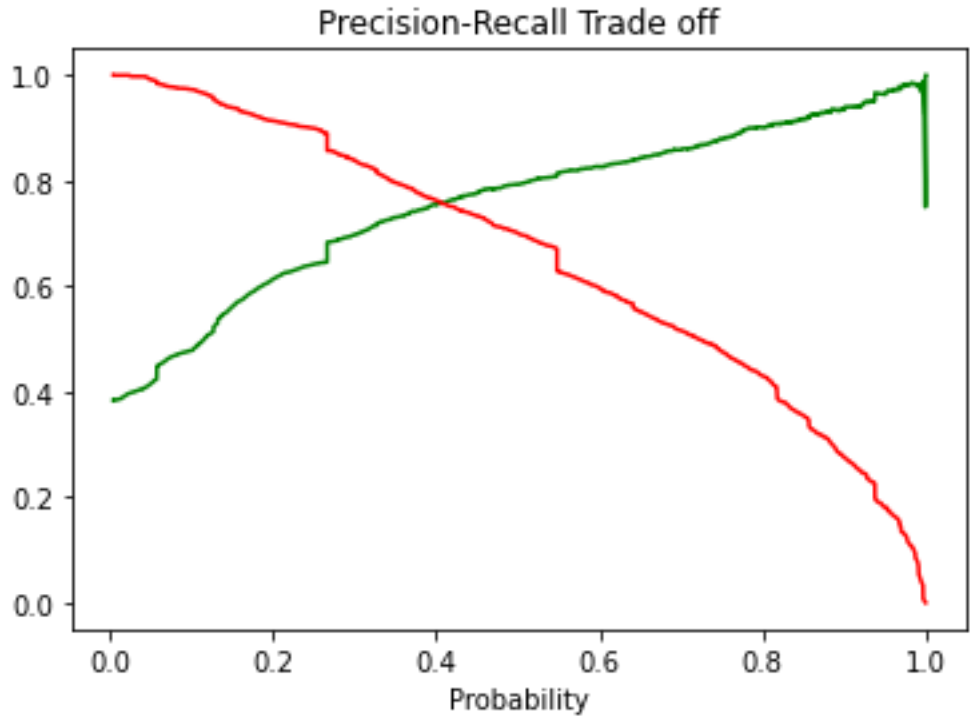
Model Evaluation

ROC Curve:



Finding optimal value of the cut off

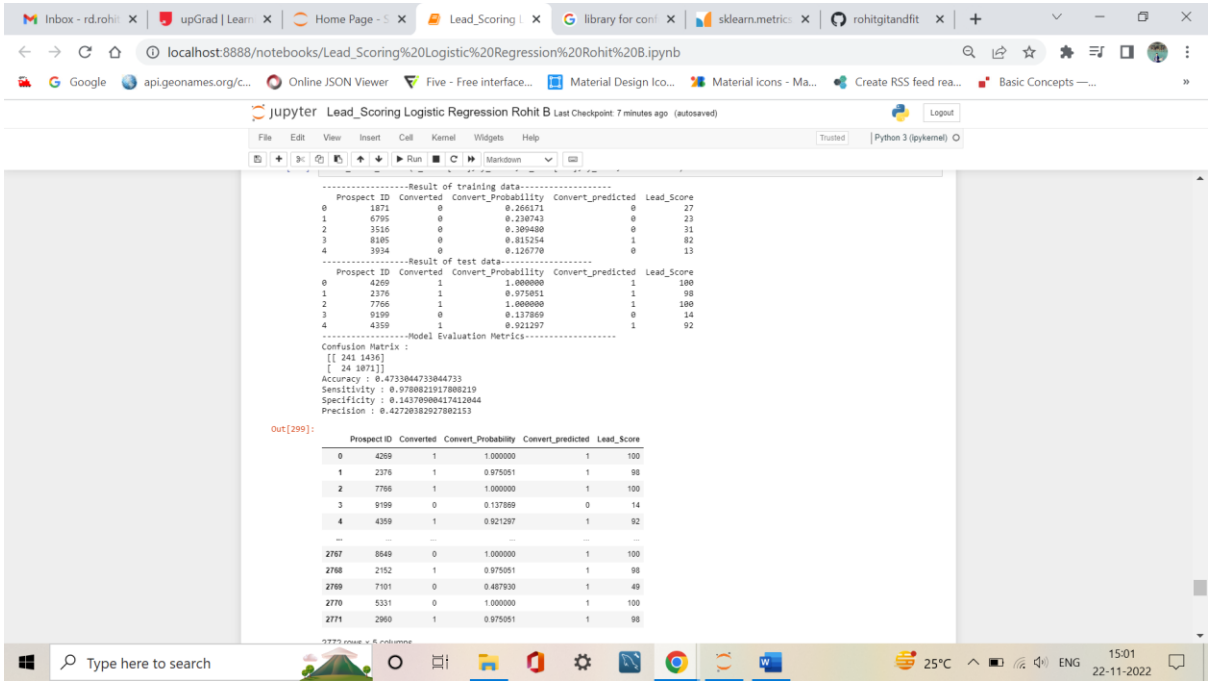




- In Sensitivity-Specificity-Accuracy plot 0.35 probability looks optimal. In Precision-Recall Curve 0.4 looks optimal.

- We are taking 0.35 is the optimum point as a cutoff probability and assigning Lead Score in training data

Final Model



Conclusion

The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable. Here, the logistic regression model is used to predict the probabilty of conversion of a customer. Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert) Our final Logistic Regression Model is built with 14 features.

Features used in final model and the aspects business should focus on are:

['Do Not Email', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Olark Chat Conversation', 'Last Activity_Other Activity', 'Last Activity_SMS Sent', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Lead Profile_Lateral Student', 'Lead Profile_Student of SomeSchool', 'How did you hear about X Education_Email', 'Last Notable Activity_Modified']