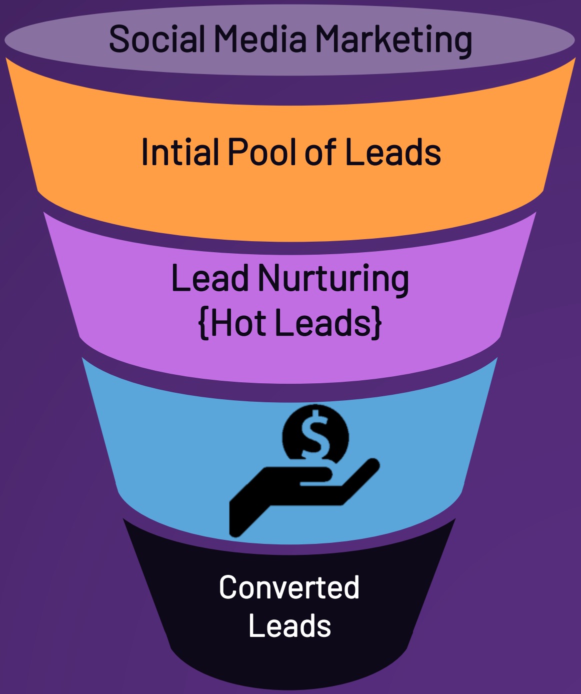
LEAD SCORING
CASE STUDY

By
Aswin Kumar S
Trisha Chandra

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as **'Hot Leads'**
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.



Business Objective:

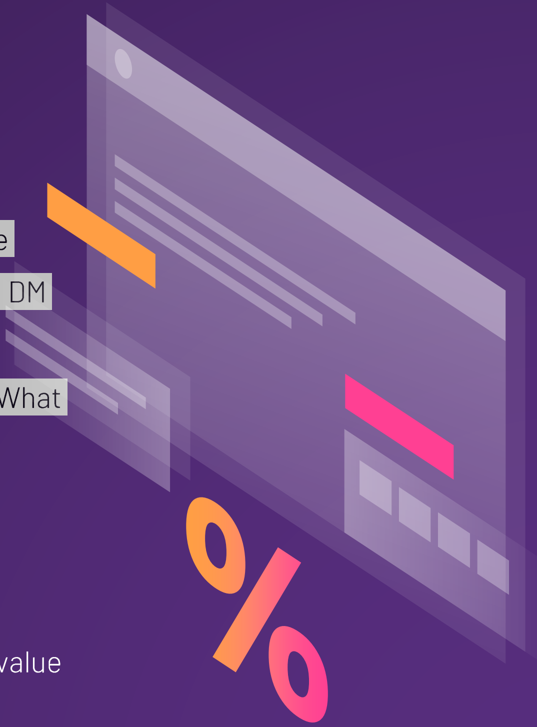
- X Education wants to know promising leads
- For that they want to build a model which identifies the hot leads
- Deployment of the model for future use

APPROACH OF THE ANALYSIS

- Data Cleaning and data manipulation
  - Checking and handling of duplicated values
  - Checking and handling of NA and missing values
  - Dropping columns, if it contains large amount of missing values and not useful for analysis
  - Imputation of missing values, if necessary
- Exploratory Data Analysis
  - Univariate data analysis: Value counts, distribution of variable etc.
  - Bivariate data analysis: Distribution of data with target variable
  - Multivariate data analysis: correlations and relation between variable
- Data Preparation for Model Building
- Classification technique: Logistic Regression used for model making and prediction
- Model Evaluation
- Reporting the final model
- Conclusions and Recommendations

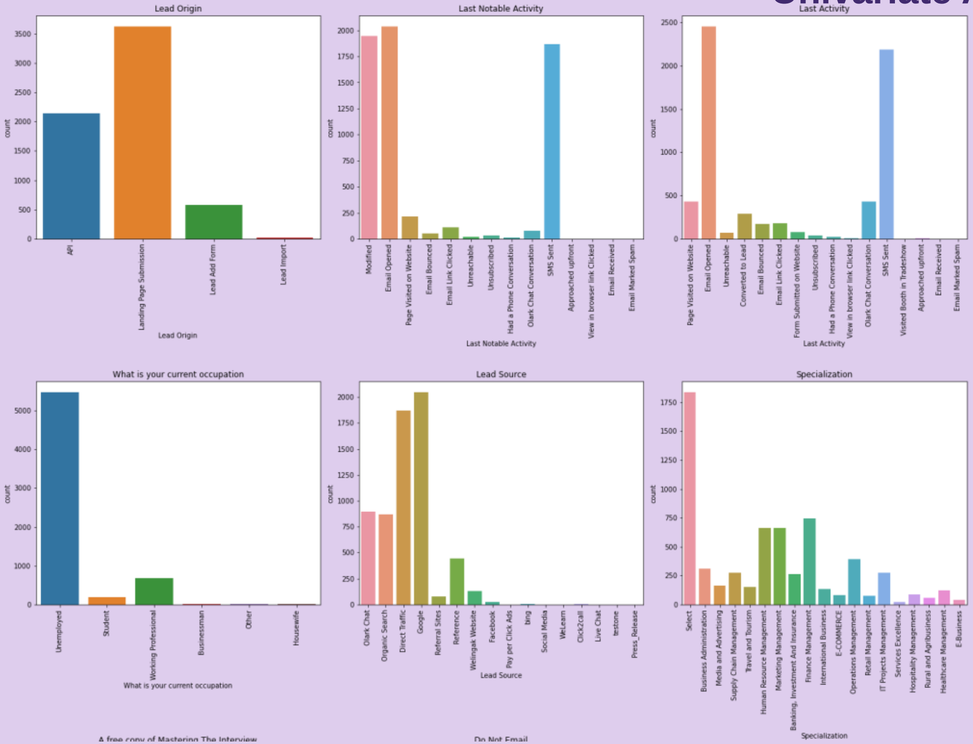
DATA CLEANING AND MANIPULATION

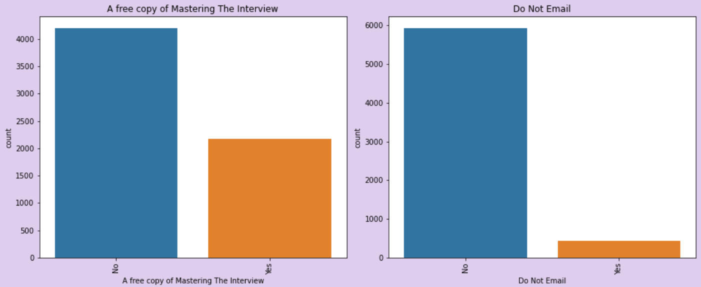
- Total Number of Rows = 9240, Total Number of Columns = 37
- Dropped the columns “Prospect ID”, “Lead Number” since they are unique value columns
- Geographical analysis is not important int this case, so “Country”, “City” columns dropped
- There were quite a few columns with single value, those columns are “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque”
- Some of the columns were containing very less variety, those columns are “Do Not Call”, “What matters most to you in choosing a course”, Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement”, “Through recommendations” those columns also removed
- Dropping 6 more columns which had more than 30% of null values
- "How did you hear about X Education", "Lead Profile” had more “Select” value which is null value removed those columns as well
- Missing value treatment carried by removing missing value rows



EXPLORATORY DATA ANALYSIS

Univariate Analysis



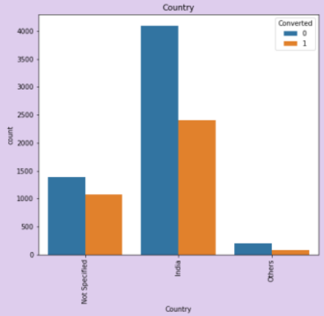


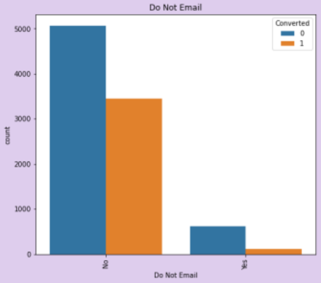
Inference:

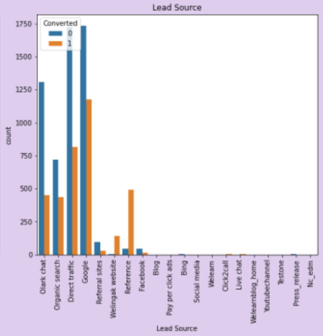
- Most of the customers are `Unemployed`
- The Source of the customer are mostly from `'Google’, 'Direct traffic', 'Olark chat’`
- The Majority of the `Last Activities` of the customers are `e-mail opened, SMS Sent and Modified`
- 31% of the customers wants a `free copy of Mastering the Interview`
- All the Numerical columns `'Total Time Spent on Website', 'TotalVisits', 'Page Views Per Visit’` skewed to the lower side

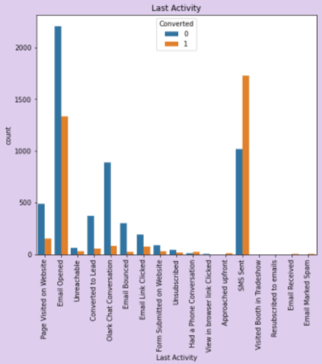
EXPLORATORY DATA ANALYSIS

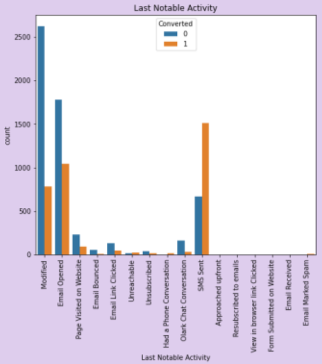
Bivariate Analysis

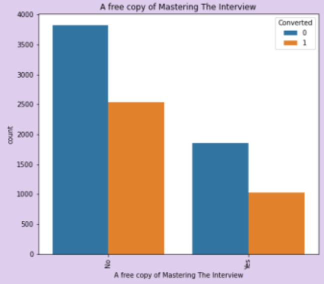


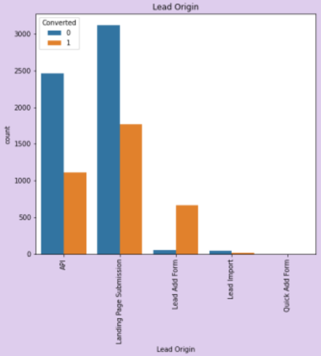


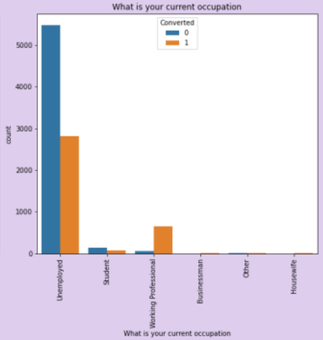










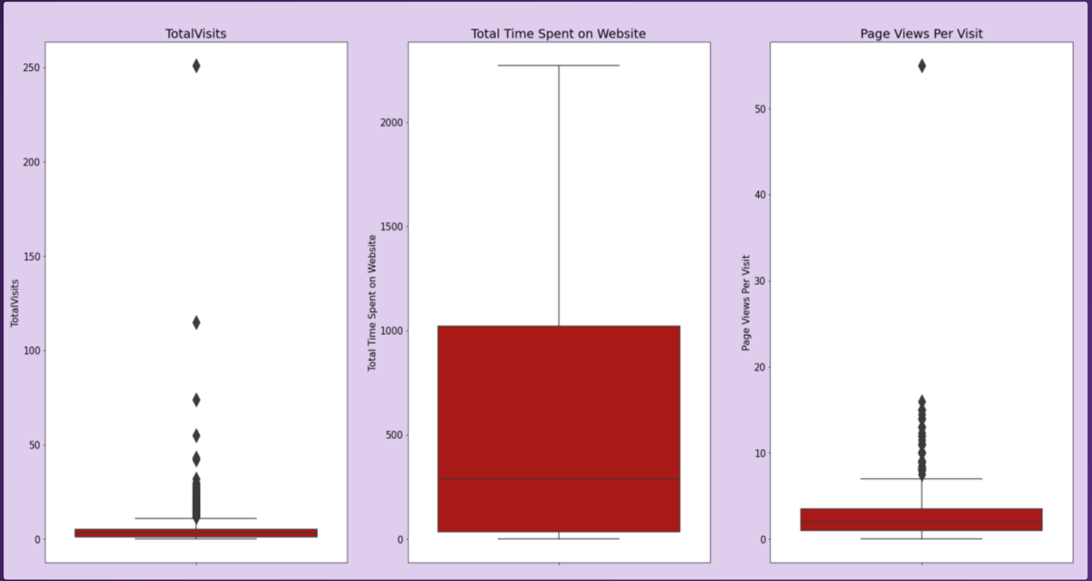


Inference:

- More `working professional` Leads are `converted`
- With `Reference` the leads are `converted` more compared to other `Lead Sources`
- Communicated through SMS leads are `converted` more compared to other `Last Activity and Last Notable Activity`
- Around 35`%` of Leads are being `converted` which is what stated in the problem statement

DATA PREPARATION

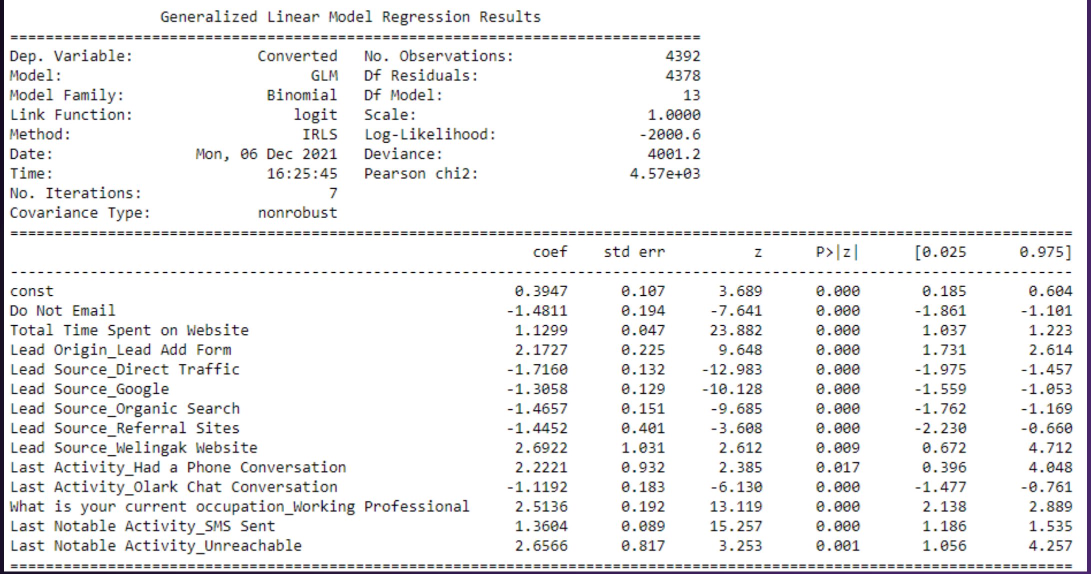
- After the EDA, started with the cleaned dataset by converting all the binary variables to '0' and '1' and multiple categories into dummy variables
- Next, checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side
- Outliers in Logistic Regression model is very sensitive hence the outliers treated, max value into 99th percentile minimum value to 1st percentile



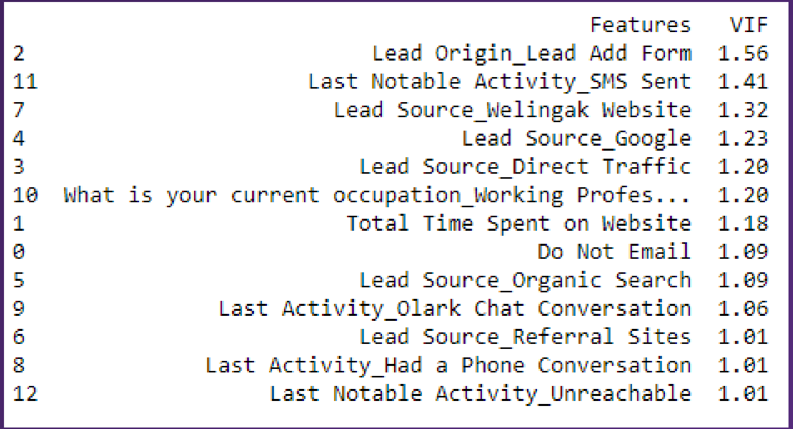
MODEL BUILDING

- Splitting the Dataset into training and test dataset with the ratio of 70:30
- Used RFE (Recursive Feature Elimination) for automatic feature elimination
- Selected 15 feature as output from RFE
- Model has been built by removing the variable whose p-value is grater than 0.05 which is insignificant, and VIF (variance inflation factor) grater than 5 to avoid multi collinearity
- With arbitrary cut-off of 0.5 got the result of
  - Accuracy – 79%
  - Sensitivity – around 74%
  - Specificity – around 83%

FINAL MODEL VISUALISATION WITH VIF



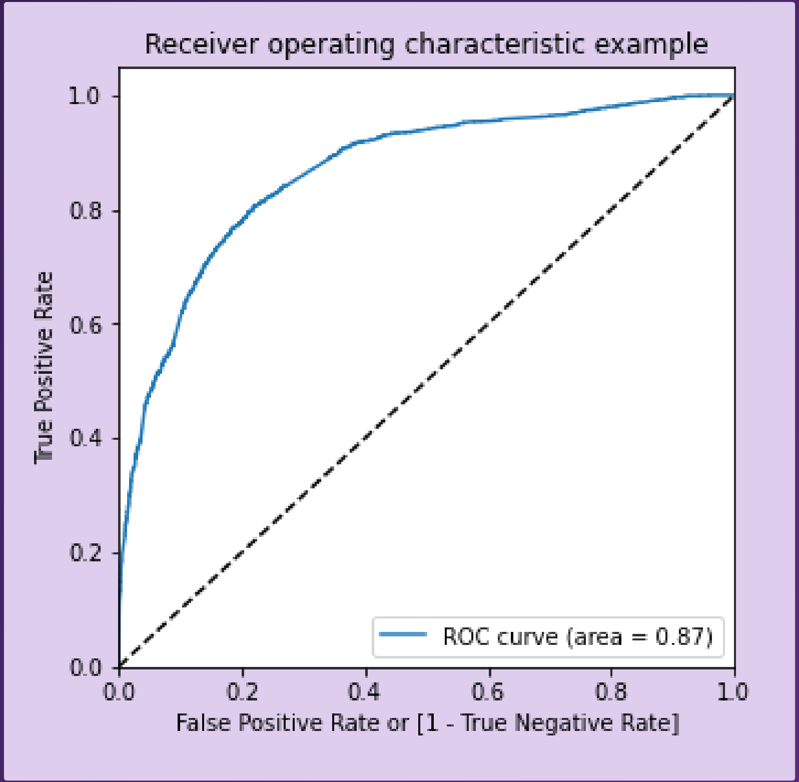
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4392
Model:                            GLM   Df Residuals:                     4378
Model Family:                Binomial   Df Model:                           13
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2000.6
Date:                Mon, 06 Dec 2021   Deviance:                       4001.2
Time:                        16:25:45   Pearson chi2:                 4.57e+03
No. Iterations:                     7
Covariance Type:            nonrobust
========================================================================================================================
                                                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                  0.3947      0.107      3.689      0.000       0.185       0.604
Do Not Email                                          -1.4811      0.194     -7.641      0.000      -1.861      -1.101
Total Time Spent on Website                            1.1299      0.047     23.882      0.000       1.037       1.223
Lead Origin_Lead Add Form                              2.1727      0.225      9.648      0.000       1.731       2.614
Lead Source_Direct Traffic                            -1.7160      0.132    -12.983      0.000      -1.975      -1.457
Lead Source_Google                                    -1.3058      0.129    -10.128      0.000      -1.559      -1.053
Lead Source_Organic Search                            -1.4657      0.151     -9.685      0.000      -1.762      -1.169
Lead Source_Referral Sites                            -1.4452      0.401     -3.608      0.000      -2.230      -0.660
Lead Source_Welingak Website                           2.6922      1.031      2.612      0.009       0.672       4.712
Last Activity_Had a Phone Conversation                 2.2221      0.932      2.385      0.017       0.396       4.048
Last Activity_Olark Chat Conversation                 -1.1192      0.183     -6.130      0.000      -1.477      -0.761
What is your current occupation_Working Professional   2.5136      0.192     13.119      0.000       2.138       2.889
Last Notable Activity_SMS Sent                         1.3604      0.089     15.257      0.000       1.186       1.535
Last Notable Activity_Unreachable                      2.6566      0.817      3.253      0.001       1.056       4.257
========================================================================================================================
```



```
                                             Features   VIF
2                           Lead Origin_Lead Add Form  1.56
11                     Last Notable Activity_SMS Sent  1.41
7                        Lead Source_Welingak Website  1.32
4                                  Lead Source_Google  1.23
3                          Lead Source_Direct Traffic  1.20
10  What is your current occupation_Working Profes...  1.20
1                         Total Time Spent on Website  1.18
0                                        Do Not Email  1.09
5                          Lead Source_Organic Search  1.09
9               Last Activity_Olark Chat Conversation  1.06
6                          Lead Source_Referral Sites  1.01
8              Last Activity_Had a Phone Conversation  1.01
12                  Last Notable Activity_Unreachable  1.01
```

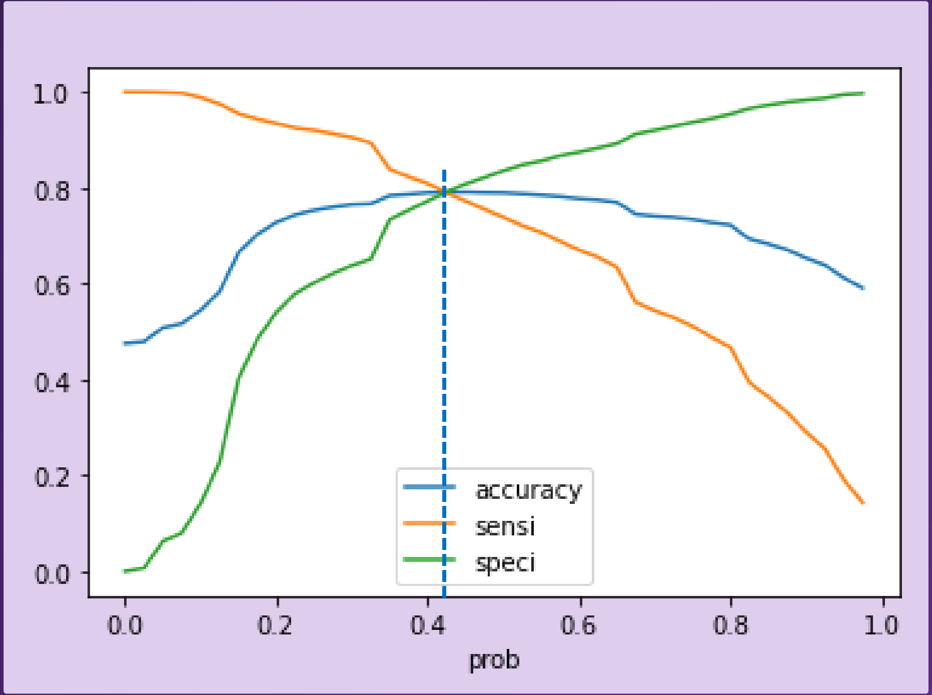
MODEL EVALUATION

- After building the final model making prediction on it (on train set), created ROC curve to find the model stability with AUC score (Area Under the Curve) as we can see from the graph plotted on the right side, the AUC score is 0.87 which is a good score
- The graph is leaned towards the left side of the border which means, the accuracy is good



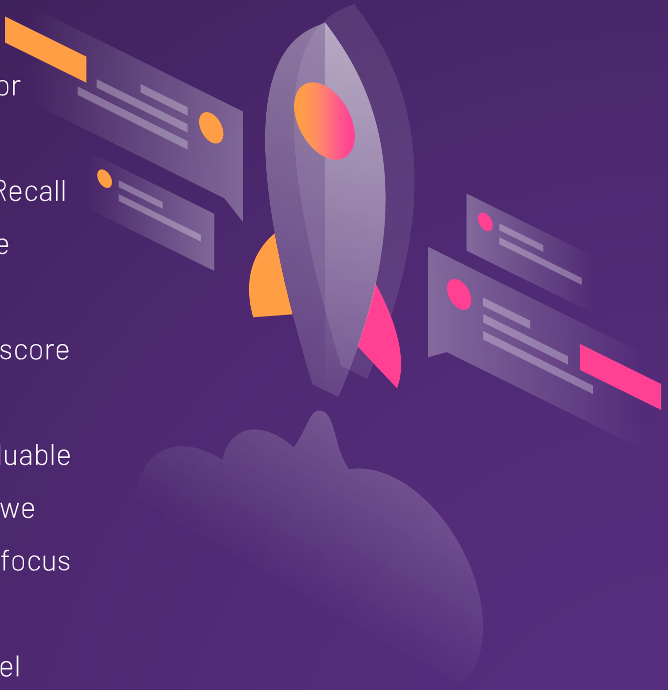
FINDING THE OPTIMAL CUT-OFF

- Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to choose for probability cut-off
- We found that on 0.425 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected
- To verify we plotted this is in a graph line plot which is on the right side all the lines are crossing 0.425 mark, hence we chose as the optimal cut-off



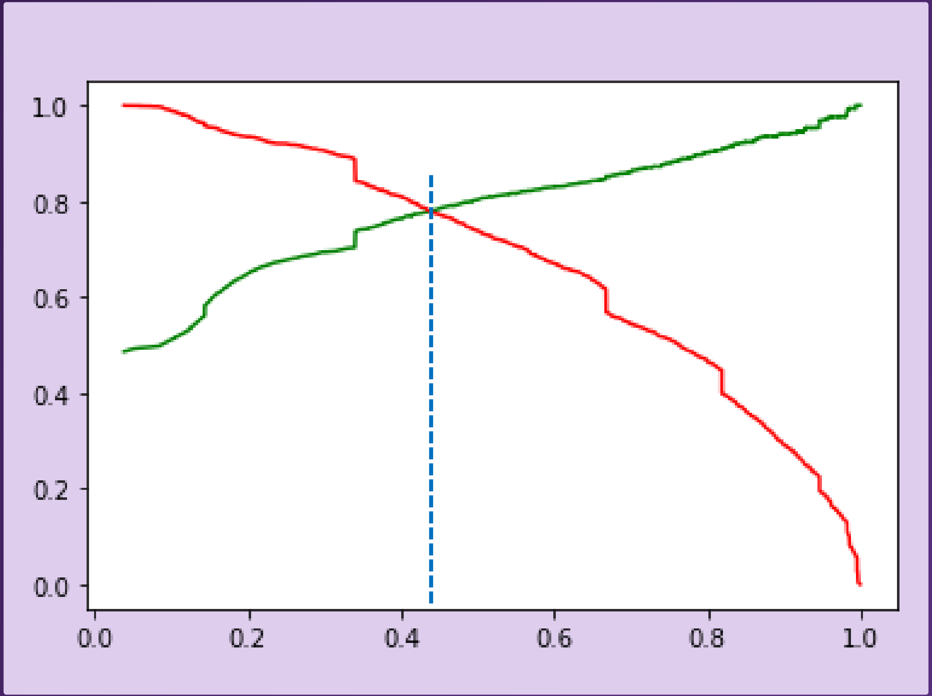
PRECISION AND RECALL

- With the probability cut-off of 0.35 to create a new column in our final dataset for predicting the outcomes
- After this we did another type of evaluation which is by checking Precision and Recall
- As we all know, Precision and Recall plays very important role in our model, more business oriented and it also tells how the model behaves
- Hence, we evaluated the precision and Recall for the final model and found the score as 0.77 for Precision and 0.79 for Recall
- Now, recall our business objective – the recall percentage will consider more valuable because it is okay if Precision is low which means less **'Hot Lead'** customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision
- i.e. we get more relevant results – as many as hot lead customers from the model



PRECISION AND RECALL TRADEOFF

- We plotted a graph which will show us the tradeoff between Precision and Recall
- Observed that there is a tradeoff between Precision and Recall and the meeting point is 0.44



PREDICTION ON TEST DATASET

- Before predicting on test set, we need to standardize the test set and need to have exact columns (columns trained by model) present in our final train dataset
- After doing that, predicted the test set and the new predictions value were saved in new data set
- After this we did model evaluation i.e. finding the accuracy, precision and recall
- The accuracy score was 0.77, precision 0.78 and recall 0.76 approximately
- This shows that the model is stable with good accuracy and recall/sensitivity.

FINAL RESULT OF THE MODEL

Train set	Test set
Accuracy – 0.79	Accuracy – 0.77
Sensitivity – 0.79	Sensitivity – 0.79
Specificity – 0.79	Specificity – 0.79
Precision – 0.77	Precision – 0.78
Recall – 0.79	Recall – 0.76
F1_Score – 0.78	F1_Score – 0.77

The above results drawn with optimal cut-off of 0.425

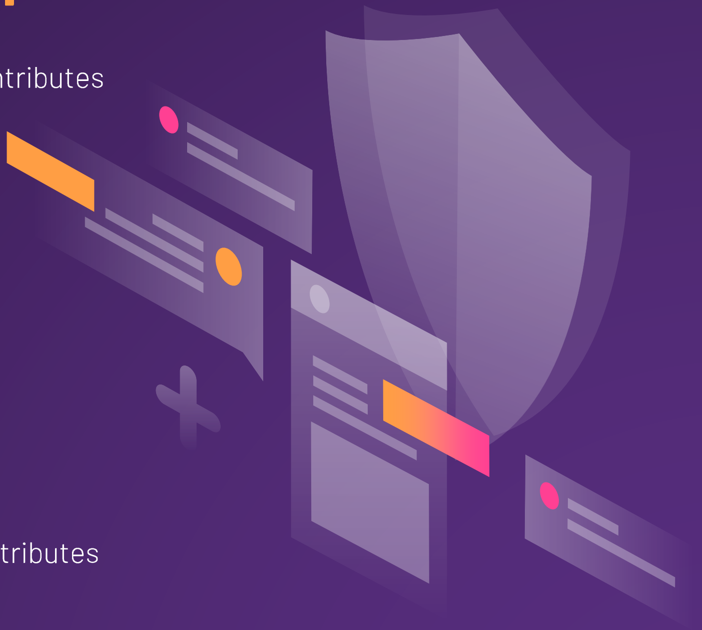
CONCLUSION

- The Specificity, Sensitivity, Accuracy, Precision and Recall score we got from test set in acceptable range.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.

RECOMMANDATIONS FOR COMPANY

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- Lead Source - Welingak website
- Last Notable Activity – Unreachable and SMS Sent
- What is your current occupation - Working Professional
- Last Activity – Had a Phone Conversation
- Lead Origin - Lead Add Form
- Total Time Spent on Website



Important features responsible for poor conversion rate or the ones' which contributes lesser probability of a lead getting converted are:

- Last Activity - Olark Chat Conversation
- Lead Source – Google, Referral Sites, Organic Search and Direct Traffic
- Do Not Email

THANK YOU!