LEAD SCORING CASE STUDY

- Badal Autade
- Akshay Kumar
- Nikhil Choithani

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Now, although 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.

AGENDA

- • To optimize and categorize the leads by assigning ranks to them basis multiple features from historical data so the company can focus more on highly convertible leads.

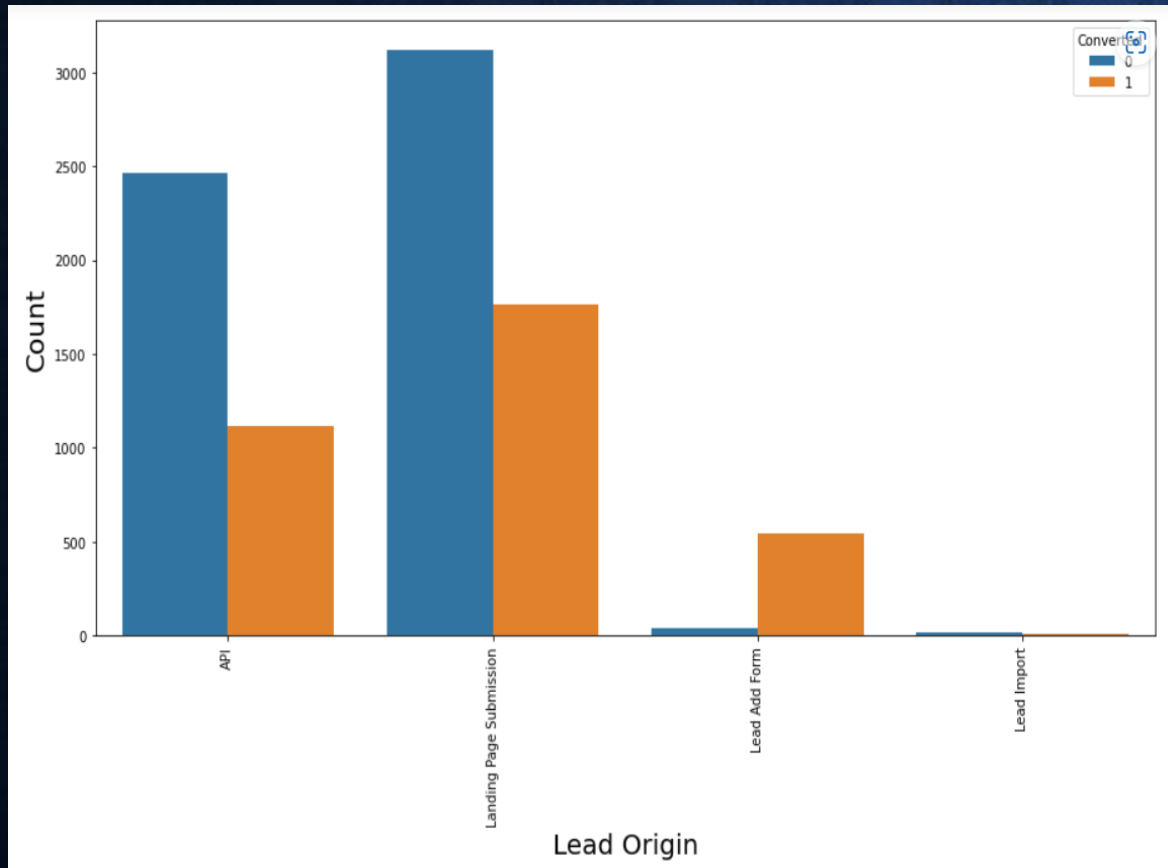
APPROACH

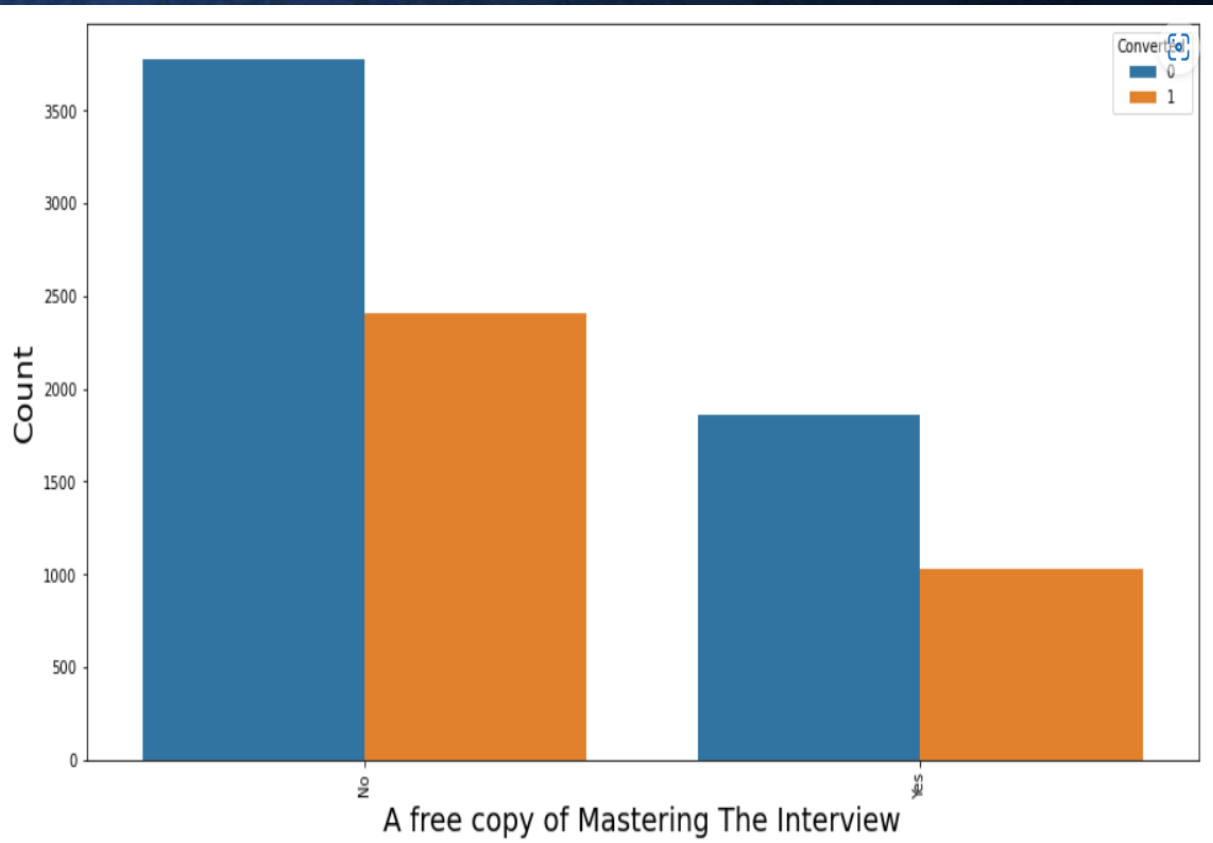
- Sourcing the dataset & Understanding the dataset
- Cleaning and Imputing the dataset
- Performing EDA
- Data Preparation for model building
- Splitting data into Train dataset and Test dataset
- Feature Scaling
- Model building and evaluation
- Prediction
- Conclusion and Recommendations

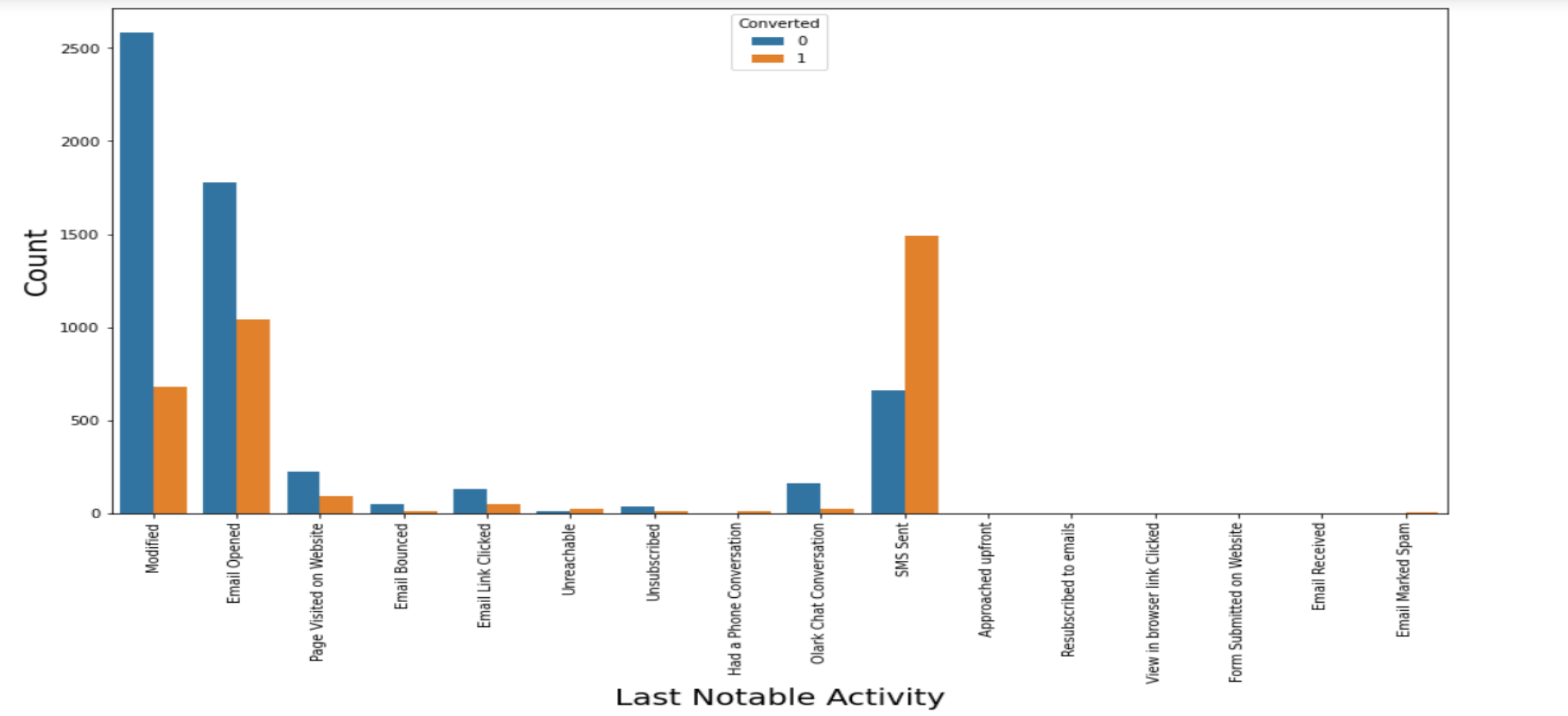
CLEANING AND IMPUTING THE DATASET

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like "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" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.
- Dropping the columns having more than 30% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

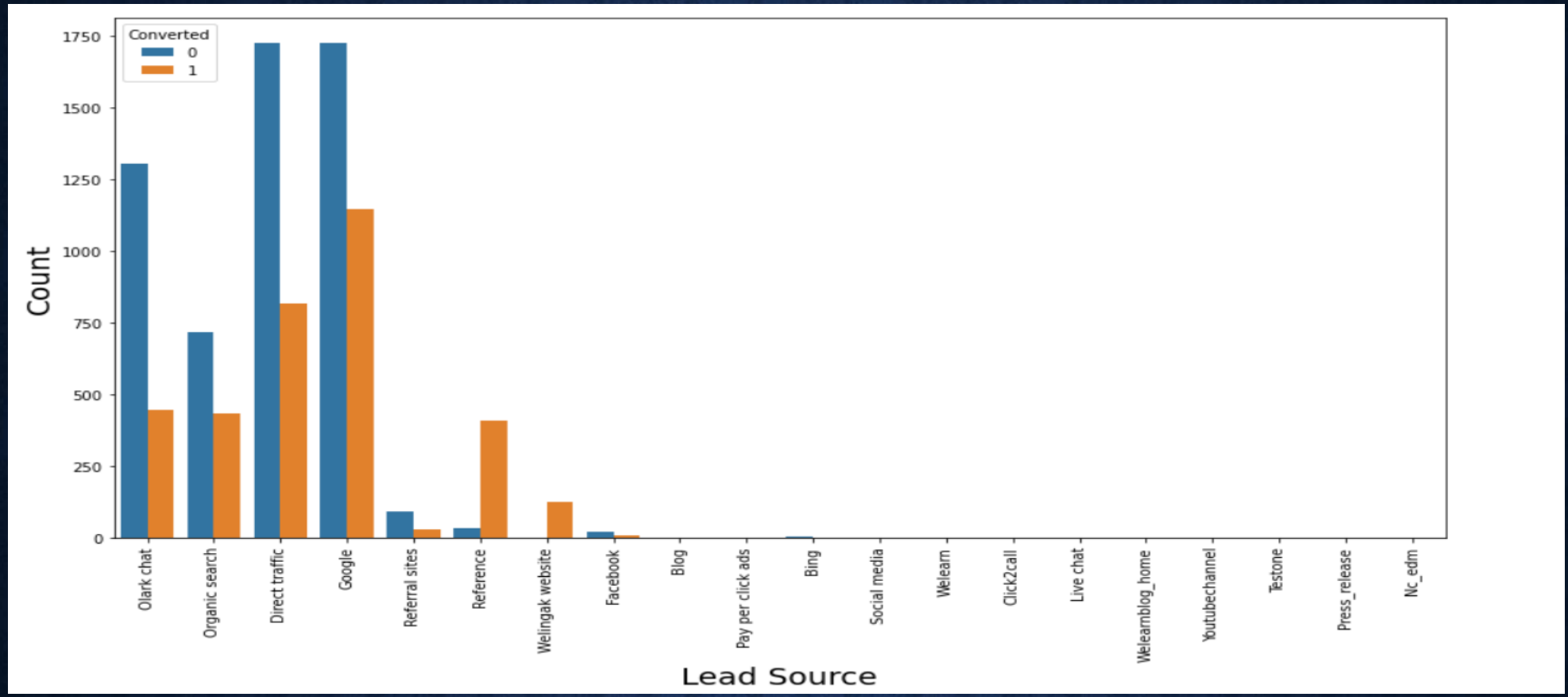
PERFORMING EDA







- SMS sent ctegory from above column has highest rate of conversion followed by Email opened category.



- Google as a Lead Source Category has highest numbers of converted people among all the rest of the categories.
- Whereas Referenced & Wellingak Website customers has highest rate of conversion in Lead Source column.

DATA PREPARATION

- Conversion of binary categorical variables into 0 & 1 categories
- Created dummy variables for remaining categorical variables
- Splitting data into 70:30 ratio to obtain Train and Test sub datasets
- Scaling the pending 3 numerical features for better modeling using standardization method.

MODEL BUILDING

- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5

FINAL MODEL

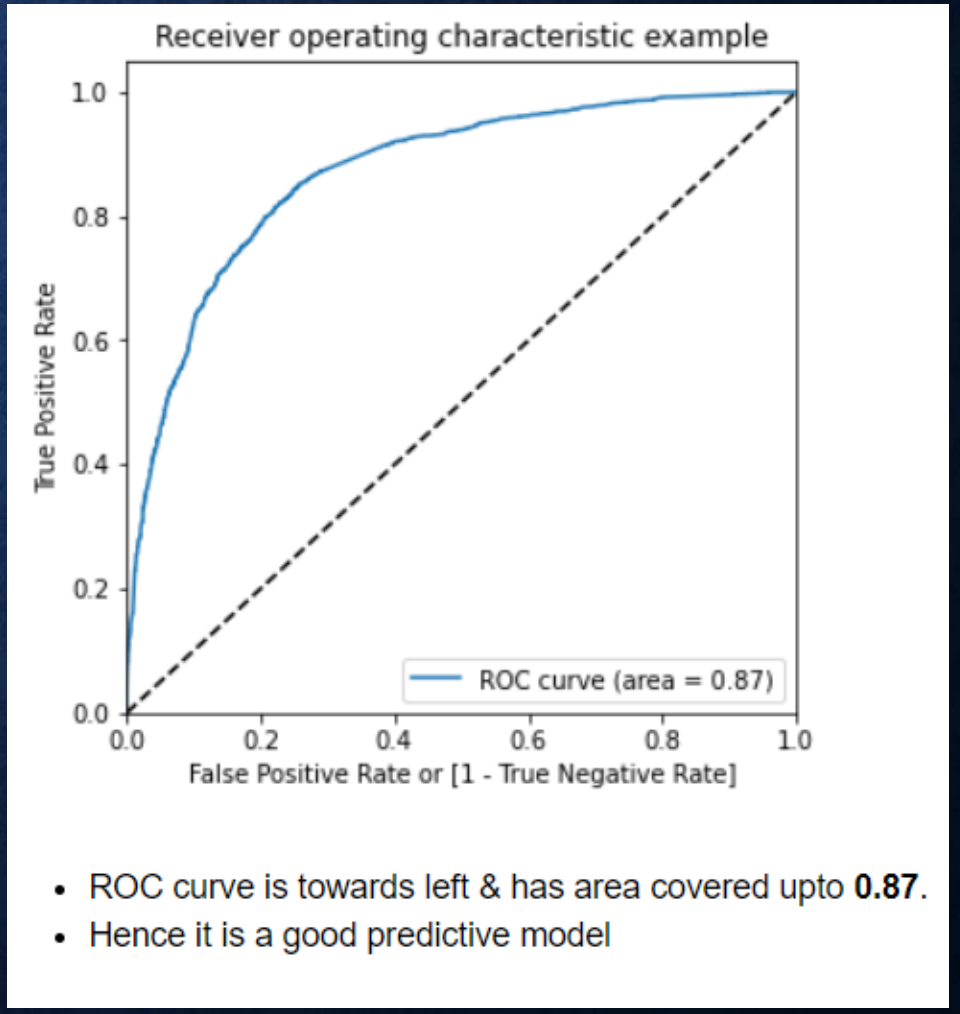
Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2784.3
Date:	Mon, 14 Nov 2022	**Deviance:**	5568.6
Time:	12:59:37	**Pearson chi2:**	6.35e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3663
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9943	0.448	2.217	0.027	0.115	1.873
Total Time Spent on Website	1.1168	0.039	28.518	0.000	1.040	1.194
Lead Origin_API	-1.0577	0.457	-2.317	0.021	-1.953	-0.163
Lead Origin_Landing Page Submission	-2.3587	0.453	-5.208	0.000	-3.246	-1.471
Lead Origin_Lead Add Form	2.3672	0.492	4.812	0.000	1.403	3.331
Lead Source_Olark chat	1.1790	0.121	9.765	0.000	0.942	1.416
Lead Source_Welingak website	2.4695	0.760	3.251	0.001	0.981	3.958
Last Activity_Email Bounced	-2.1261	0.367	-5.793	0.000	-2.845	-1.407
Last Activity_Olark Chat Conversation	-1.4227	0.161	-8.848	0.000	-1.738	-1.108
Specialization_Unknown	-1.4385	0.121	-11.886	0.000	-1.676	-1.201
Last Notable Activity_Had a Phone Conversation	3.4443	1.089	3.162	0.002	1.310	5.579
Last Notable Activity_SMS Sent	1.5745	0.077	20.498	0.000	1.424	1.725
Last Notable Activity_Unreachable	1.7747	0.454	3.909	0.000	0.885	2.664

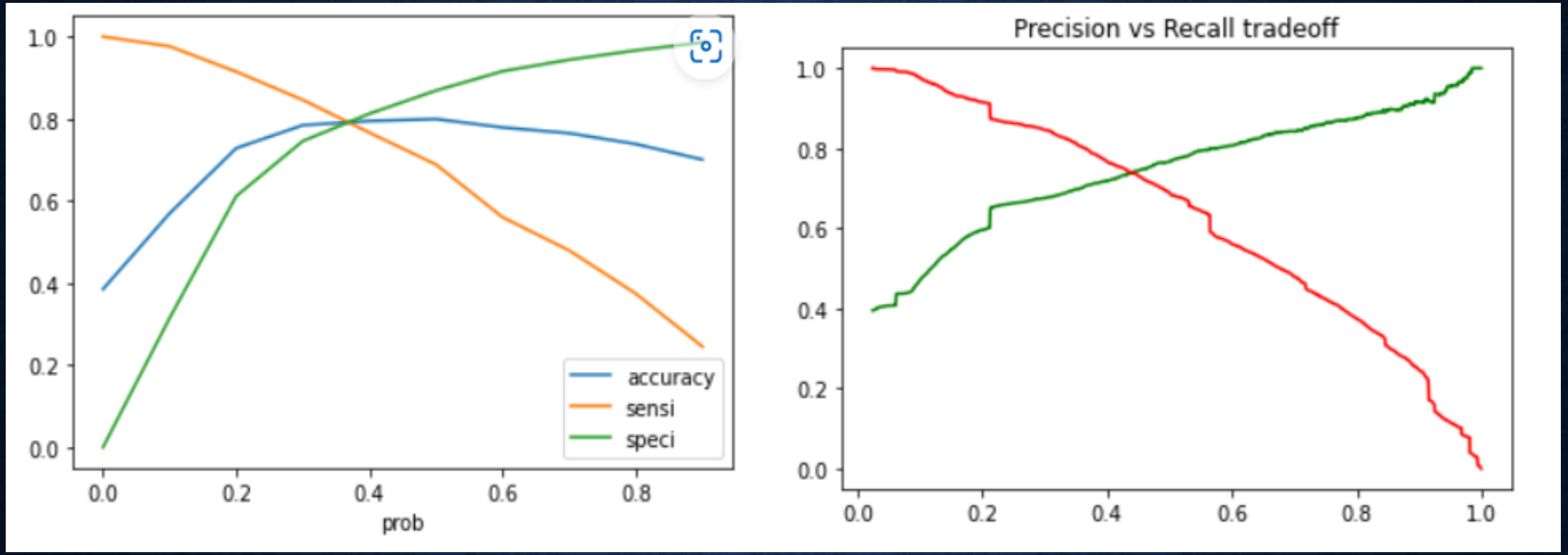
	Features	VIF
1	Lead Origin_API	4.94
8	Specialization_Unknown	4.44
4	Lead Source_Olark chat	2.31
3	Lead Origin_Lead Add Form	1.49
7	Last Activity_Olark Chat Conversation	1.45
10	Last Notable Activity_SMS Sent	1.45
5	Lead Source_Welingak website	1.38
2	Lead Origin_Landing Page Submission	1.33
0	Total Time Spent on Website	1.32
6	Last Activity_Email Bounced	1.07
9	Last Notable Activity_Had a Phone Conversation	1.01
11	Last Notable Activity_Unreachable	1.01

MODEL EVALUATION

- Training set :
  - Accuracy = 79.20%
  - Sensitivity= 81.11%
  - Specificity= 78.00%
  - Precision = 69.78%
  - Recall = 81.11%
  - Area under ROC curve is 0.87



- ROC curve is towards left & has area covered upto **0.87**.
- Hence it is a good predictive model



- The cutoff point determined from above graph is 0.35

PREDICTION ON TEST SET

- The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.
- Testing set:
  - Accuracy = 79.14%
  - Sensitivity= 79.77%
  - Specificity= 78.77%
  - Precision = 68.19%
  - Recall = 79.77%

This shows that our test prediction is having accuracy , precision and recall score in an acceptable range

CONCLUSION & RECOMMENDATIONS

- 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.
- Important features (in descending order) responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  - - Last Notable Activity_Had a Phone Conversation
  - - Lead Source_Welingak website
  - - Lead Origin_Lead Add Form
  - - Last Notable Activity_Unreachable
  - - Last Notable Activity_SMS Sent
- Company should focus on above recommended target groups to yeils maximum conversion and thereby profits.