Leads scoring case study

Problem Statement/Objective

- X Education company 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.
- Identifying the most promising leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

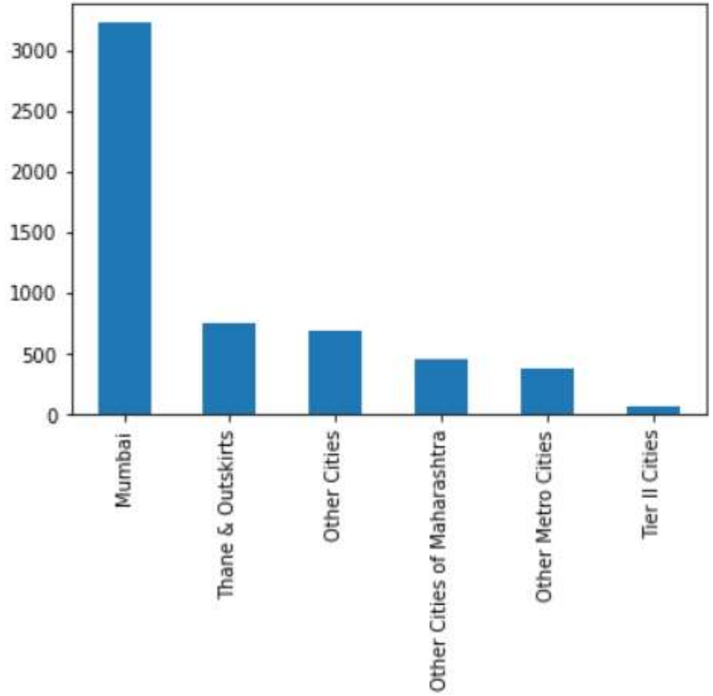
Analysis Approach

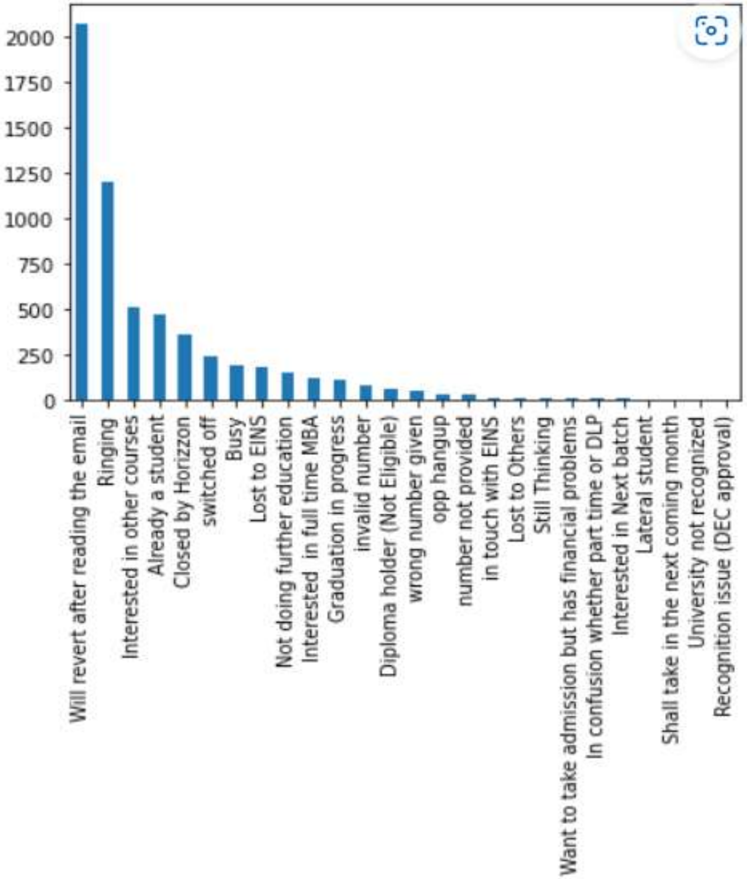
- Data Reading
- Data Understanding
- Data Cleaning
- Data preparation
- EDA
- Model Building
- Model evaluation
- Predictions on Test Data
- Recommendations

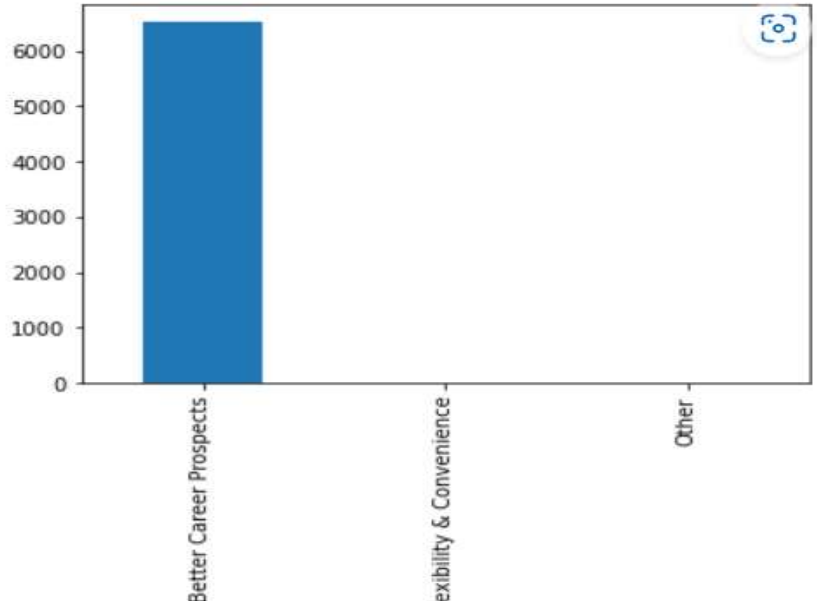
Analysis

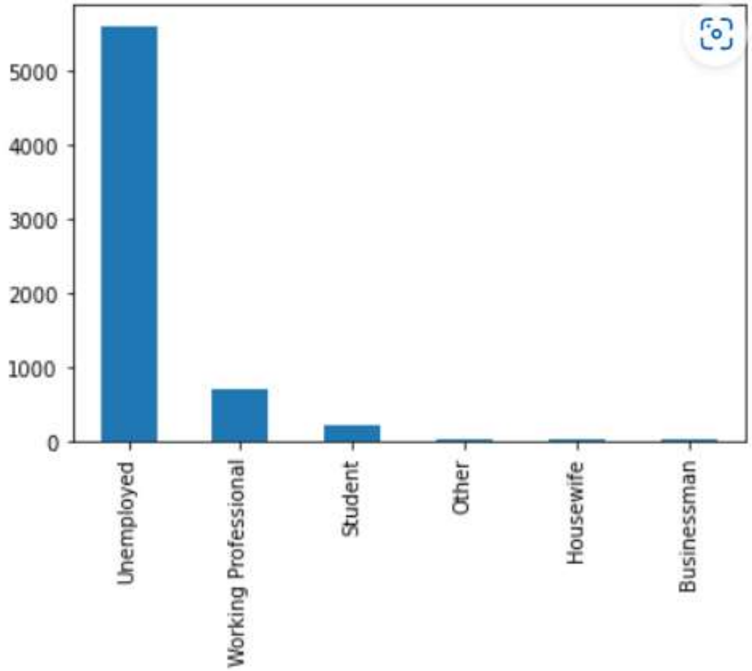
- Initially Cleaned data and handled null and Missing values
- Dropped columns with high null values and reduced irregularities.
- Converted Binary variables with Yes and No to 0 & 1.
- Created Dummy variables and dropped duplicated columns
- Split data into test and train data set and performed feature scaling.
- Found correlation by EDA
- Performed Model building, Selected Feature by RFE
- Built model and made Predictions

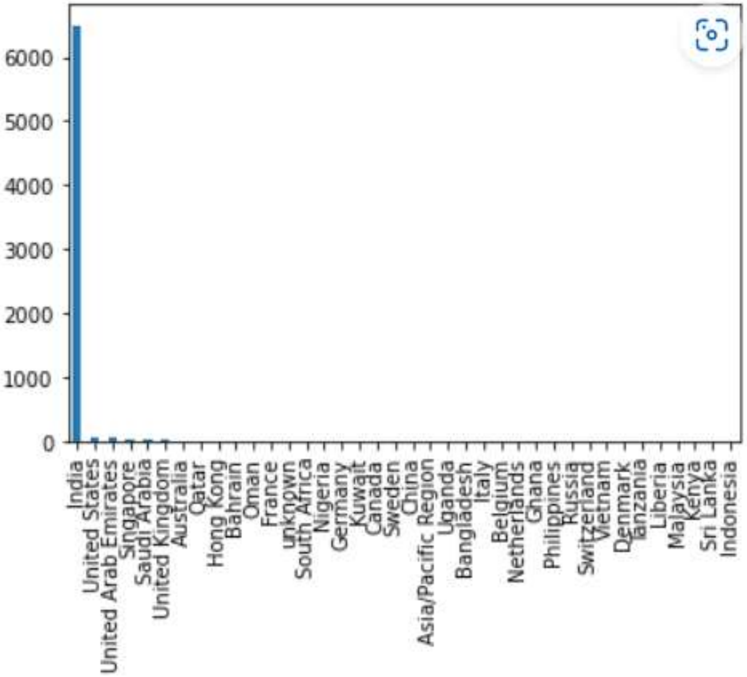
Handled Null values

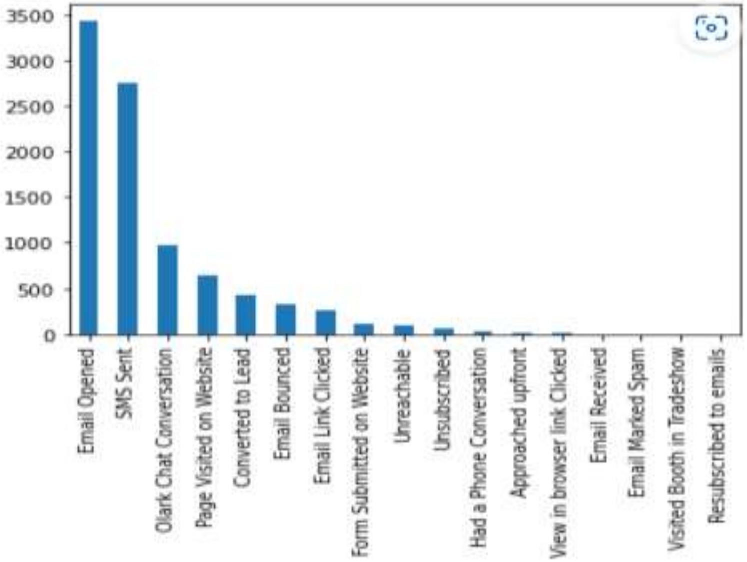





6000
5000
4000
3000
2000
1000
0
Better Career Prospects
exibility & Convenience
Other


5000
4000
3000
2000
1000
0
Unemployed
Working Professional
Student
Other
Housewife
Businessman

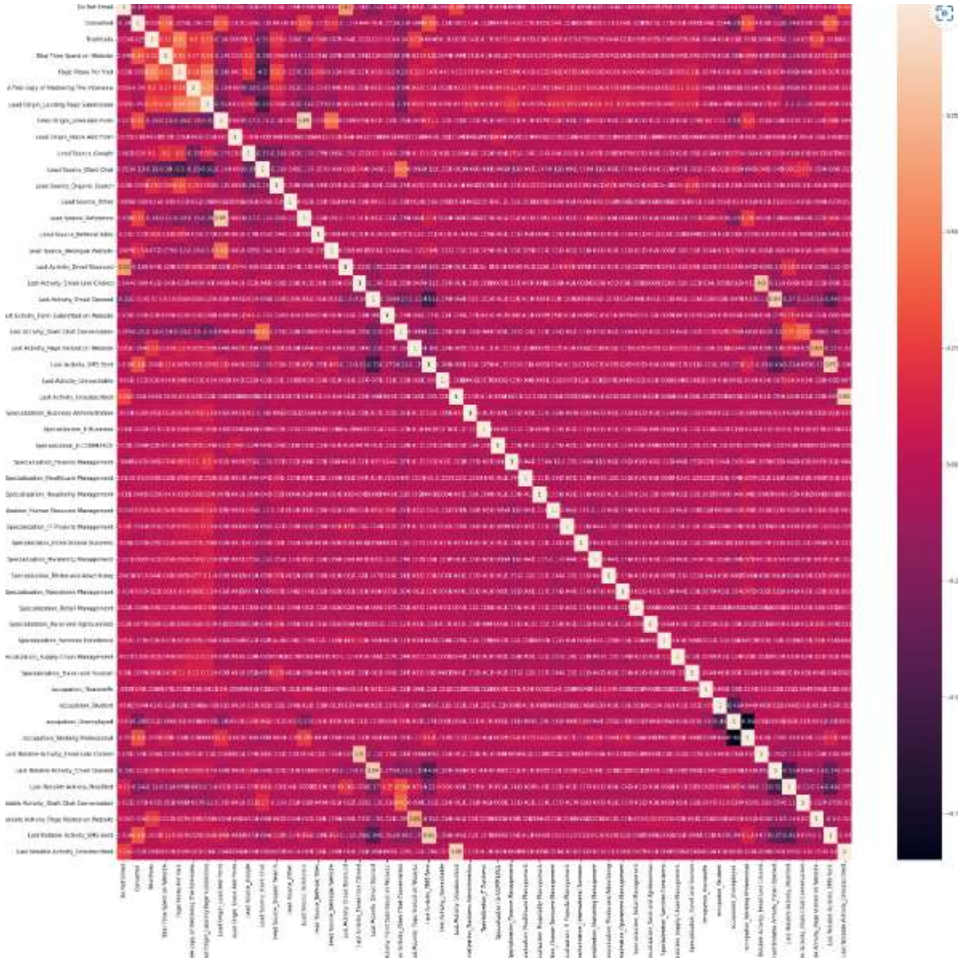

6000
5000
4000
3000
2000
1000
0
India
United States
United Arab Emirates
Singapore
Saudi Arabia
United Kingdom
Australia
Qatar
Hong Kong
Bahrain
Oman
France
unknown
South Africa
Nigeria
Germany
Kuwait
Canada
Sweden
China
Asia/Pacific Region
Uganda
Bangladesh
Italy
Belgium
Netherlands
Ghana
Philippines
Russia
Switzerland
Vietnam
Denmark
Tanzania
Liberia
Malaysia
Kenya
Sri Lanka
Indonesia


3500
3000
2500
2000
1500
1000
500
0
Email Opened
SMS Sent
Olark Chat Conversation
Page Visited on Website
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unreachable
Unsubscribed
Had a Phone Conversation
Approached upfront
View in browser link Clicked
Email Received
Email Marked Spam
Visited Booth in Tradeshow
Resubscribed to emails

- As the Above images handled the null values and unique values segregated into different useful variables and handled accordingly to get proper insights
- Null values with more than 25 % are removed and adjusted
- also replaced less frequency values in columns like Lead Source, Last Activity etc replaced with Others.

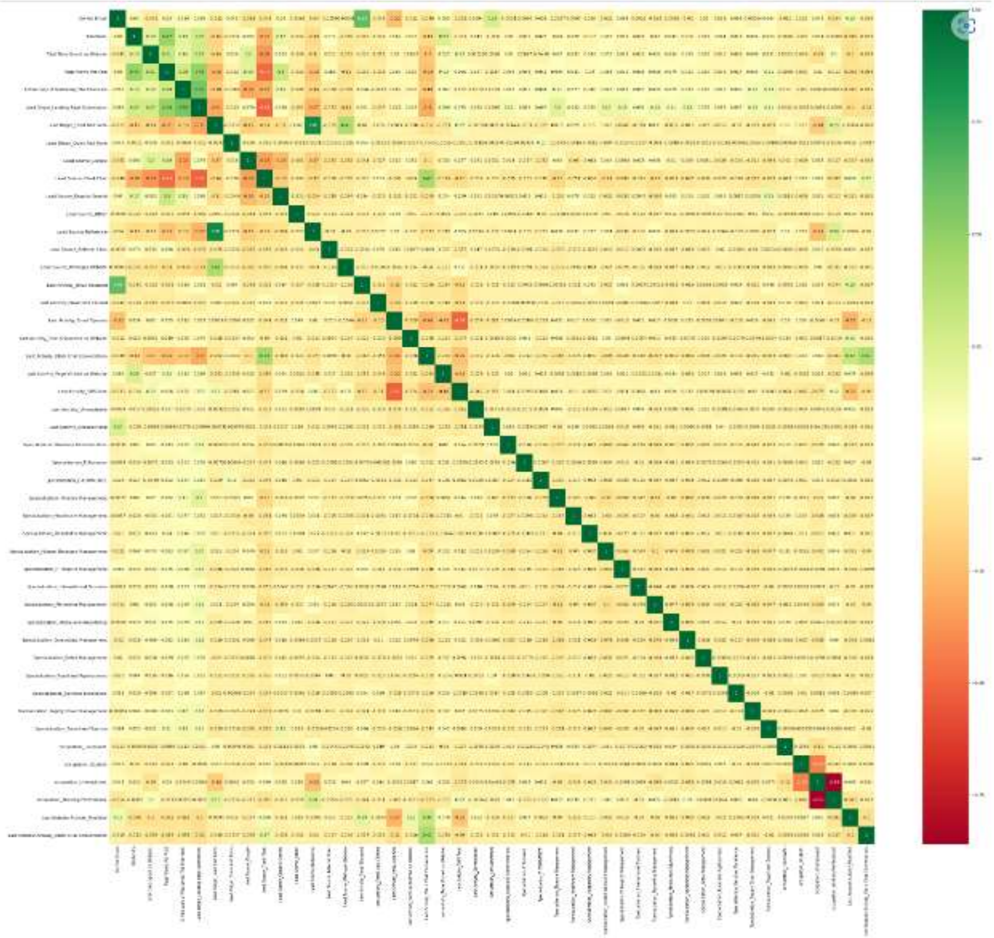
Correlation by EDA

Found the correlation on data between variables , and removed the highly co-related dummy variables



Final Correlation

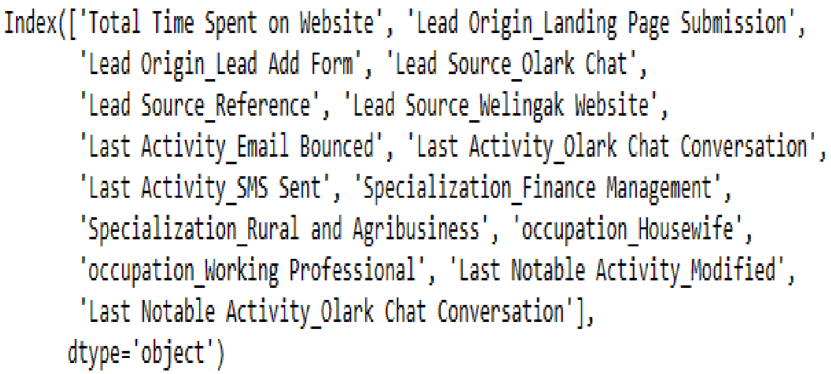
- Handled some of the categorical variables of Yes and No values and converted to 1 and 0 for ease of process.
- Created dummy variables for the remaining categorical variables and dropped the duplicate columns.
- Checked for missing values and outliers and imputed them



Feature selection by RFE

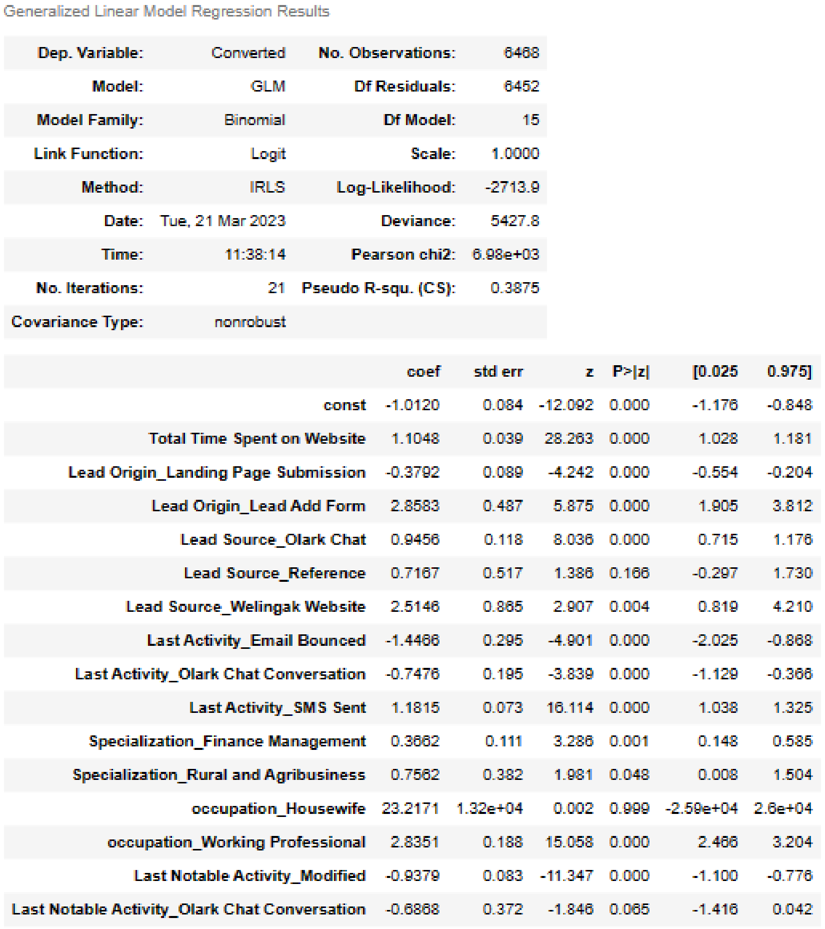
- Here we divided the data into test and train data, with 70 % to train model and 30 % to test model.
- For Feature selection chose RFE model .
- As per process by RFE we chose and deducted insignificant variables
- Then we built a stats model.
- Column selection and stats model are pictured below.

Feature Scaling



```
Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Reference', 'Lead Source_Welingak Website',
       'Last Activity_Email Bounced', 'Last Activity_Olark Chat Conversation',
       'Last Activity_SMS Sent', 'Specialization_Finance Management',
       'Specialization_Rural and Agribusiness', 'occupation_Housewife',
       'occupation_Working Professional', 'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation'],
      dtype='object')
```

- Regression model with features selection are executed
- Data has irregularities with high P values and Coefficients.



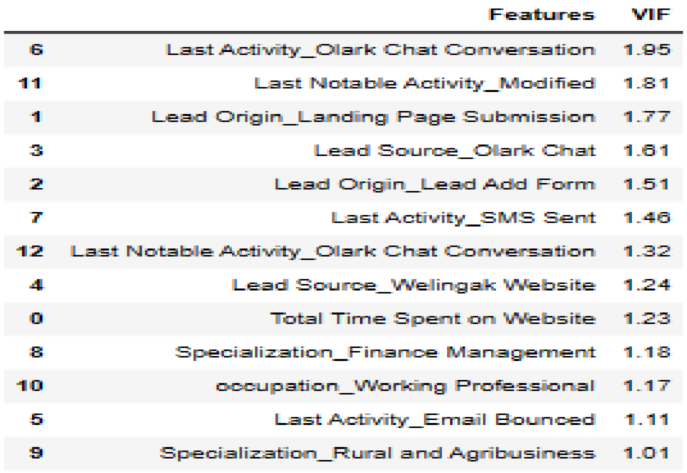
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2713.9
Date:	Tue, 21 Mar 2023	Deviance:	5427.8
Time:	11:38:14	Pearson chi2:	6.98e+03
No. Iterations:	21	Pseudo R-squ. (CS):	0.3875
Covariance Type:	nonrobust		

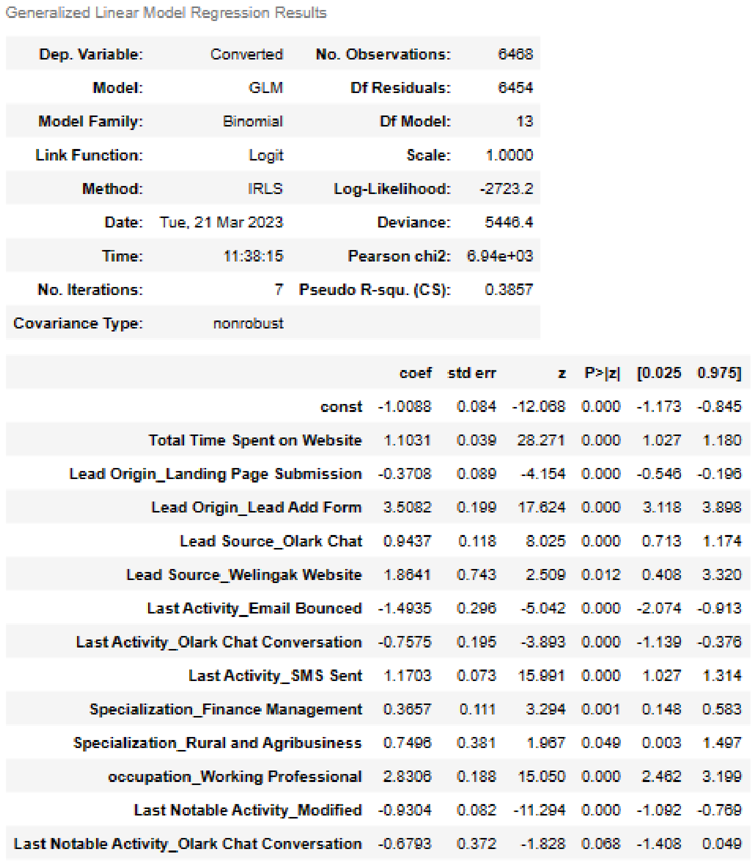
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0120	0.084	-12.092	0.000	-1.176	-0.848
Total Time Spent on Website	1.1048	0.039	28.263	0.000	1.028	1.181
Lead Origin_Landing Page Submission	-0.3792	0.089	-4.242	0.000	-0.554	-0.204
Lead Origin_Lead Add Form	2.8583	0.487	5.875	0.000	1.905	3.812
Lead Source_Olark Chat	0.9456	0.118	8.036	0.000	0.715	1.176
Lead Source_Reference	0.7167	0.517	1.386	0.166	-0.297	1.730
Lead Source_Welingak Website	2.5146	0.865	2.907	0.004	0.819	4.210
Last Activity_Email Bounced	-1.4466	0.295	-4.901	0.000	-2.025	-0.868
Last Activity_Olark Chat Conversation	-0.7476	0.195	-3.839	0.000	-1.129	-0.366
Last Activity_SMS Sent	1.1815	0.073	16.114	0.000	1.038	1.325
Specialization_Finance Management	0.3662	0.111	3.286	0.001	0.148	0.585
Specialization_Rural and Agribusiness	0.7562	0.382	1.981	0.048	0.008	1.504
occupation_Housewife	23.2171	1.32e+04	0.002	0.999	-2.59e+04	2.6e+04
occupation_Working Professional	2.8351	0.188	15.058	0.000	2.466	3.204
Last Notable Activity_Modified	-0.9379	0.083	-11.347	0.000	-1.100	-0.776
Last Notable Activity_Olark Chat Conversation	-0.6868	0.372	-1.846	0.065	-1.416	0.042

Model Building

- After removing high P values and we arrive at variables with good values of VIF say threshold of less than 5 and we go ahead using these for making predictions.
- Final features after using VIF and essential columns remained , final model built



	Features	VIF
6	Last Activity_Olark Chat Conversation	1.95
11	Last Notable Activity_Modified	1.81
1	Lead Origin_Landing Page Submission	1.77
3	Lead Source_Olark Chat	1.61
2	Lead Origin_Lead Add Form	1.51
7	Last Activity_SMS Sent	1.46
12	Last Notable Activity_Olark Chat Conversation	1.32
4	Lead Source_Welingak Website	1.24
0	Total Time Spent on Website	1.23
8	Specialization_Finance Management	1.18
10	occupation_Working Professional	1.17
5	Last Activity_Email Bounced	1.11
9	Specialization_Rural and Agribusiness	1.01



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6454
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2723.2
Date:	Tue, 21 Mar 2023	Deviance:	5446.4
Time:	11:38:15	Pearson chi2:	6.94e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3857
Covariance Type:	nonrobust		

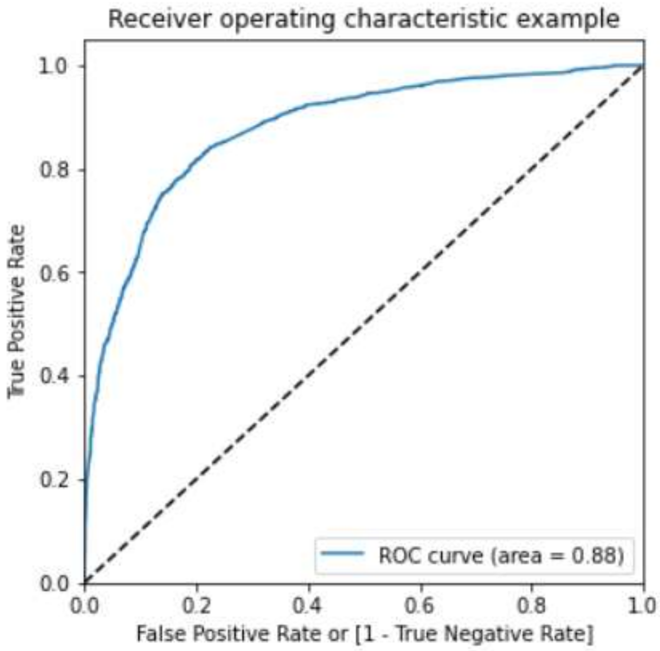
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0088	0.084	-12.068	0.000	-1.173	-0.845
Total Time Spent on Website	1.1031	0.039	28.271	0.000	1.027	1.180
Lead Origin_Landing Page Submission	-0.3708	0.089	-4.154	0.000	-0.546	-0.196
Lead Origin_Lead Add Form	3.5082	0.199	17.624	0.000	3.118	3.898
Lead Source_Olark Chat	0.9437	0.118	8.025	0.000	0.713	1.174
Lead Source_Welingak Website	1.8641	0.743	2.509	0.012	0.408	3.320
Last Activity_Email Bounced	-1.4935	0.296	-5.042	0.000	-2.074	-0.913
Last Activity_Olark Chat Conversation	-0.7575	0.195	-3.893	0.000	-1.139	-0.376
Last Activity_SMS Sent	1.1703	0.073	15.991	0.000	1.027	1.314
Specialization_Finance Management	0.3657	0.111	3.294	0.001	0.148	0.583
Specialization_Rural and Agribusiness	0.7496	0.381	1.967	0.049	0.003	1.497
occupation_Working Professional	2.8306	0.188	15.050	0.000	2.462	3.199
Last Notable Activity_Modified	-0.9304	0.082	-11.294	0.000	-1.092	-0.769
Last Notable Activity_Olark Chat Conversation	-0.6793	0.372	-1.828	0.068	-1.408	0.049

Model Building

- we go ahead using these for making predictions with stable VIF and features
- the confusion matrix overall accuracy is at 0.81 i.e 81%.
- sensitivity :0.69 i.e 69%
- specificity:0.88 i.e 88%

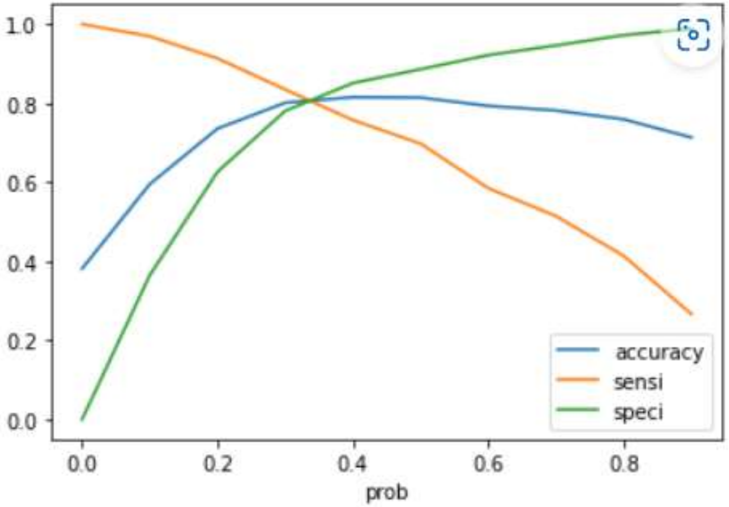
Model Evaluation

- After building final model making prediction on it we created ROC curve to find the model stability.
- Below are Train data Accuracy, sensitivity and specificity
- Accuracy : 80%
- Sensitivity : 83.45%
- Specificity : 78.03%



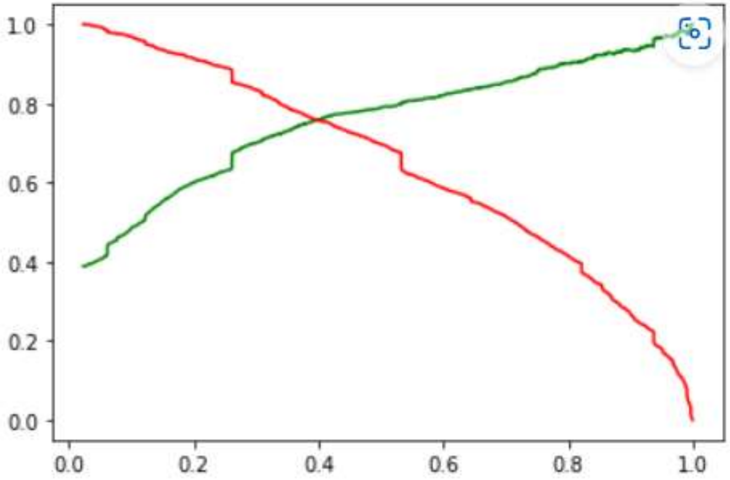
Finding optimal cutoff point

- Now we have created range of points to find accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cut-off.
- We found that on 0.3 accuracy, sensitivity and specificity nearly close.
- sensitivity :83% , specificity:78% , accuracy: 80%



Precision and Recall tradeoff

- We created a graph which will show trade-off between precision and recall and the meeting point is approximately at 0.4
- precision_score= 71%
- recall_score = 82%



Predictions on Test Data

- We also standardized the test set and started predicting test set and save those values in data frame.
- We did model evaluation and find accuracy, sensitivity and specificity for test data
- Accuracy : 80%
- Sensitivity : 82.73%
- Specificity : 78.29%

Recommendations

- Below are the important features responsible for good conversion rate
- Lead Add Form (from Lead Origin)
- Working Professional (from occupation)
- Welingak website(from Lead Source)
- Total time spent on Website
- SMS Sent(from Last Activity)

Thankyou