Lead Scoring Case Study

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.

Business Requirement

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 Requirement

X Education has appointed you to help them select the most promising leads, i.e. the 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Case Study Outcome

Top Feature / Predictors

1. Lead Source: Welingak Website: 5.206469
2. Last Notable Activity: Had a Phone Conversation: 3.500256
3. What is your current occupation: Working Professional: 3.456307
4. Lead Source: Reference: 3.448623
5. Last Notable Activity: Unreachable: 1.661949
6. Last Activity: SMS Sent: 1.268065
7. Last Notable Activity: Olark Chat Conversation: 1.133033
8. Total Time Spent on Website: 1.093008
9. What is your current occupation: Unemployed: 0.936953
10. Lead Source: Olark Chat: 0.868793
11. Last Notable Activity: Modified: 0.709519
12. Lead Origin: Landing Page Submission: 0.447315
13. Specialization: Finance Management: 0.337836

Promising lead Information with Converted Probability

	Lead Number	Converted	Converted_Prob	Convert_Predicted	Converted_Probabilty%
0	660737	0	0.174337	0	17.43
1	660728	0	0.409382	1	40.94
2	660727	1	0.491288	1	49.13
3	660719	0	0.094377	0	9.44
4	660681	1	0.496148	1	49.61
5	660680	0	0.076410	0	7.64
6	660673	1	0.600747	1	60.07
7	660664	0	0.076410	0	7.64
8	660624	0	0.049419	0	4.94
9	660616	0	0.099967	0	10.00
10	660608	1	0.631839	1	63.18
11	660570	1	0.628110	1	62.81
12	660562	1	0.657438	1	65.74
13	660558	0	0.139218	0	13.92
14	660553	0	0.231507	0	23.15

Variable Scores:

Optimum Point: 0.35 (Cutoff Probability)

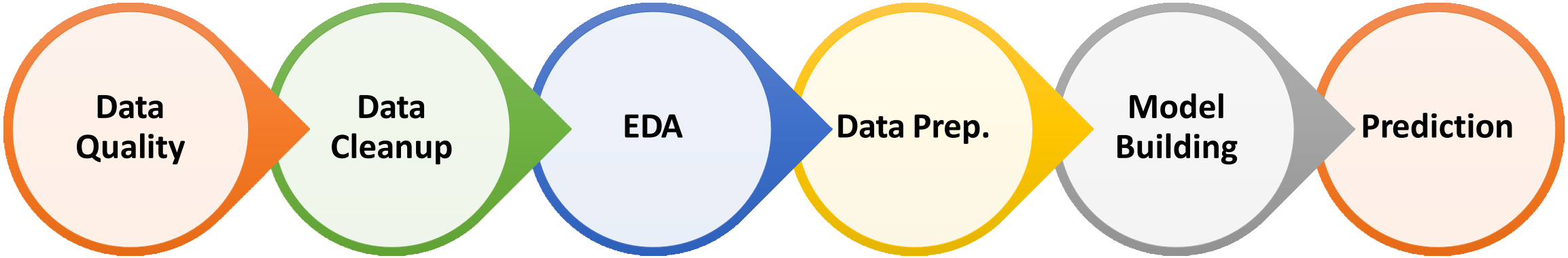
Train Data Set

- Overall accuracy: 0.806
- Sensitivity, recall, hit rate, or true positive rate (TPR): 0.806
- Precision or positive predictive value (PPV): 0.717
- F1 Score:0.759

Test Data Set

- Overall accuracy: 0.805
- Sensitivity, Recall, hit rate, or true positive rate (TPR): 0.797
- Precision or positive predictive value (PPV): 0.733
- F1 Score: 0.764

Approach & Action: Logistic Regression



- Data Load
- Check Data file, Shape and info
- Null value for columns

- Missing data treatment
- Skewed & Unique column treatment
- Data Imputation
- Outlier Treatment

- Data visualization
- Data Correlation
- Observation

- Dummy Creation
- Data split:
- Scaling of Data (Standard Scaler)

- Model building on all features
- Recursive Feature Elimination
- Manual model building
- Model finalized
- Confusion matrix & Metric calculation
- ROC Curve
- Optimal Cut-off point
- Precision and Recall graph
- F1 Score

- Prediction on Test data

Data Quality & Clean-up

37 Columns and 9240 Rows columns

Multiple missing values

Prospect ID	0
Lead Number	0
Lead Origin	0
Lead Source	36
Do Not Email	0
Do Not Call	0
Converted	0
TotalVisits	137
Total Time Spent on Website	0
Page Views Per Visit	137
Last Activity	103
Country	2461
Specialization	3380
How did you hear about X Education	7250
What is your current occupation	2690
What matters most to you in choosing a course	2709
Search	0
Magazine	0
Newspaper Article	0
X Education Forums	0
Newspaper	0
Digital Advertisement	0
Through Recommendations	0
Receive More Updates About Our Courses	0
Tags	3353
Lead Quality	4767
Update me on Supply Chain Content	0
Get updates on DM Content	0
Lead Profile	6855
City	3669
Asymmetrique Activity Index	4218
Asymmetrique Profile Index	4218
Asymmetrique Activity Score	4218
Asymmetrique Profile Score	4218
I agree to pay the amount through cheque	0
A free copy of Mastering The Interview	0
Last Notable Activity	0

Data Clean-up and Treatment:

Dropping of columns: High percentage of missing values & sales team

- **Dropping of columns: Single field columns:**
- 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

vity Index

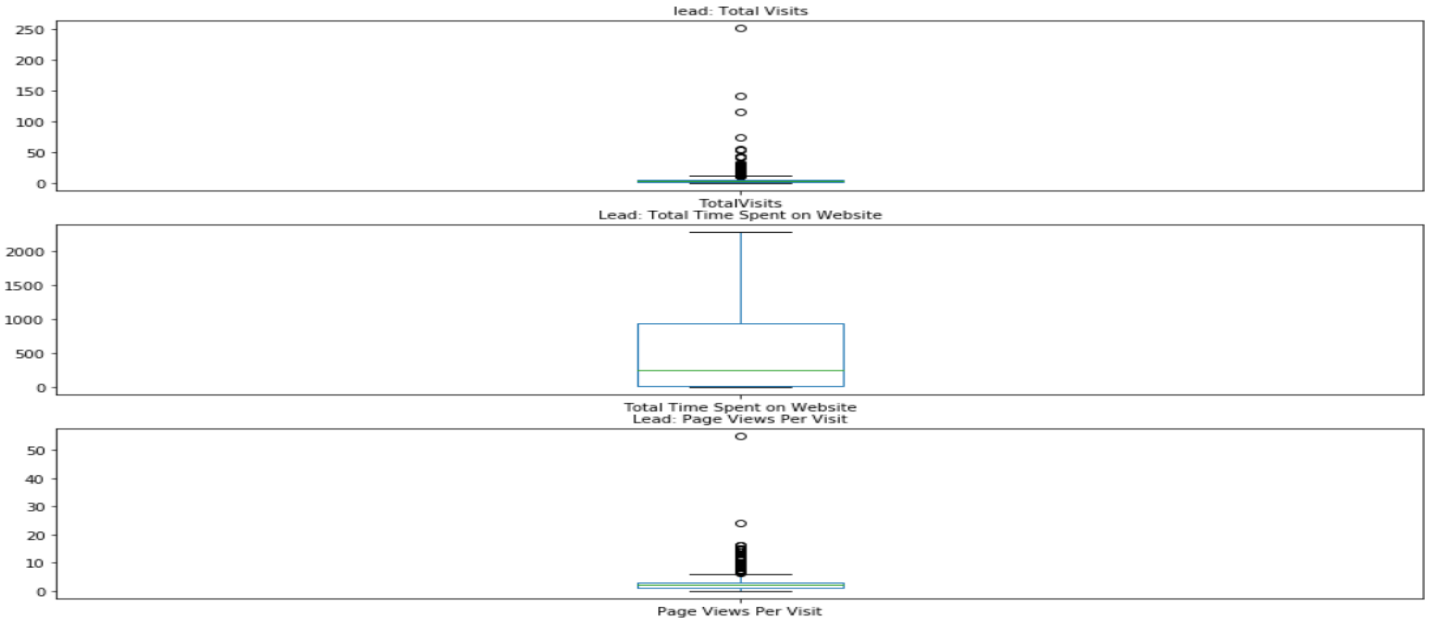
ile Index

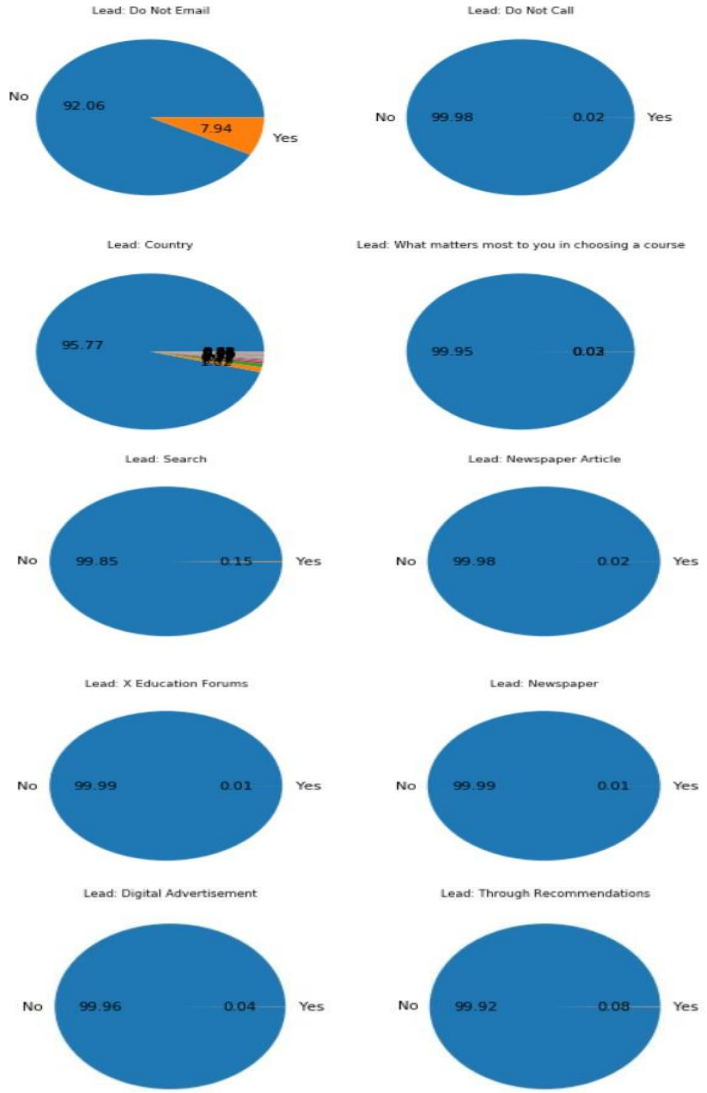
- Asymmetrique Activity Score

Data Quality & Clean-up

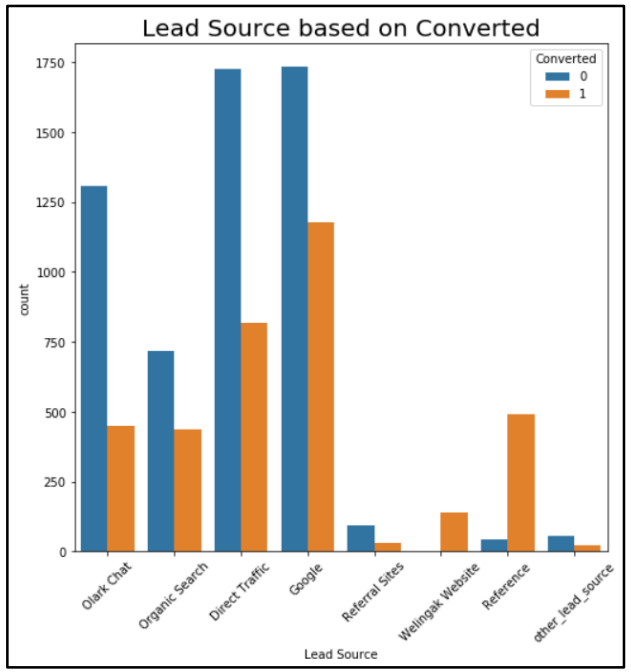
- **Dropping of Skewed columns:**
- Do Not Email
- Do Not Call
- Country
- Search
- What matters most to you in choosing a course
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Through Recommendations

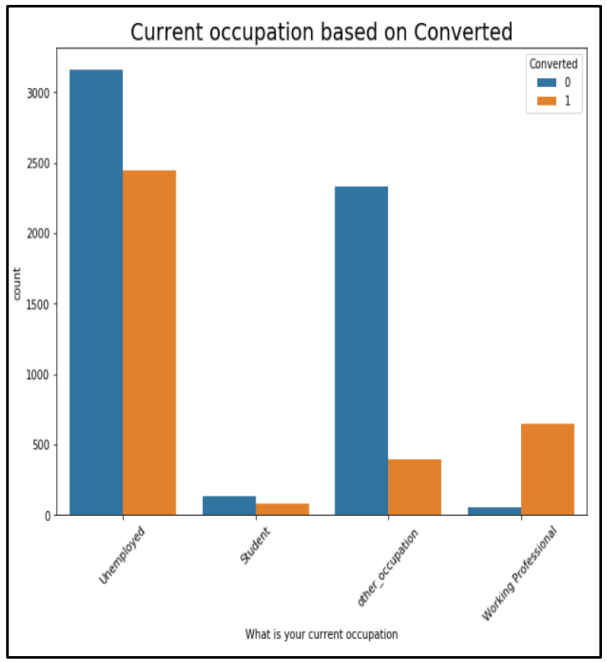
- **Outlier Treatment: Soft capping of 1% & 99%**

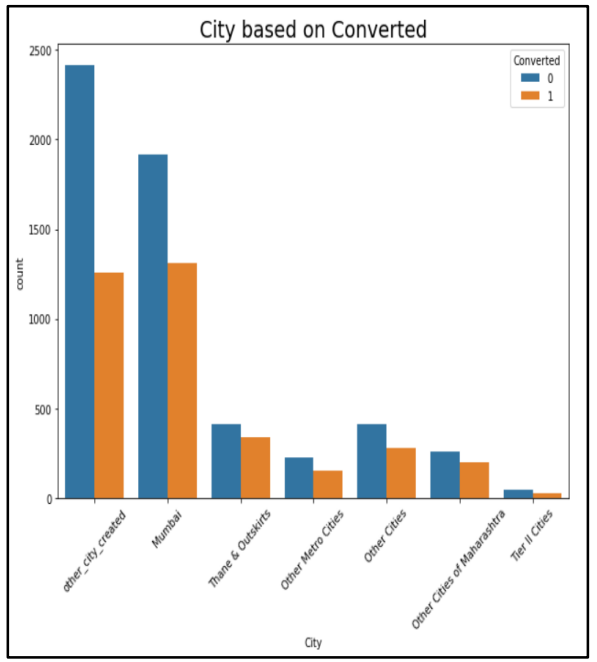


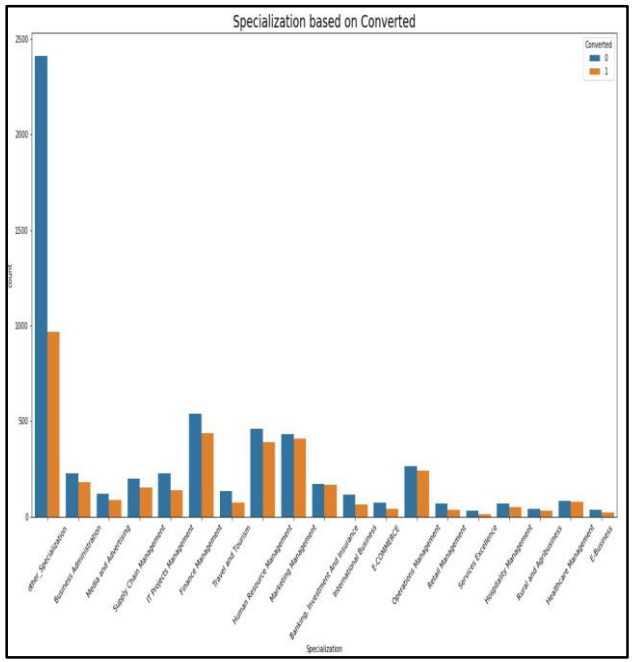


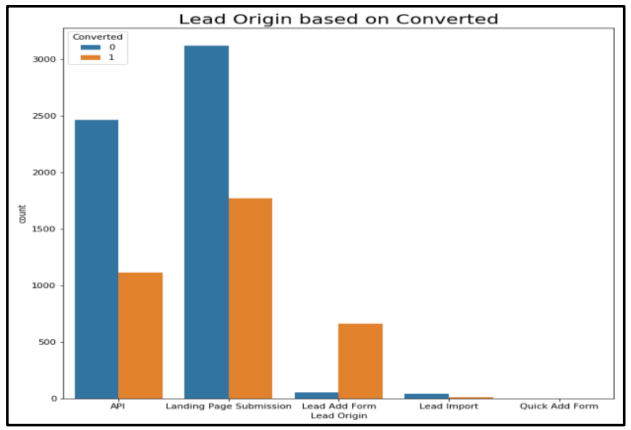
Exploratory Data Analysis: EDA

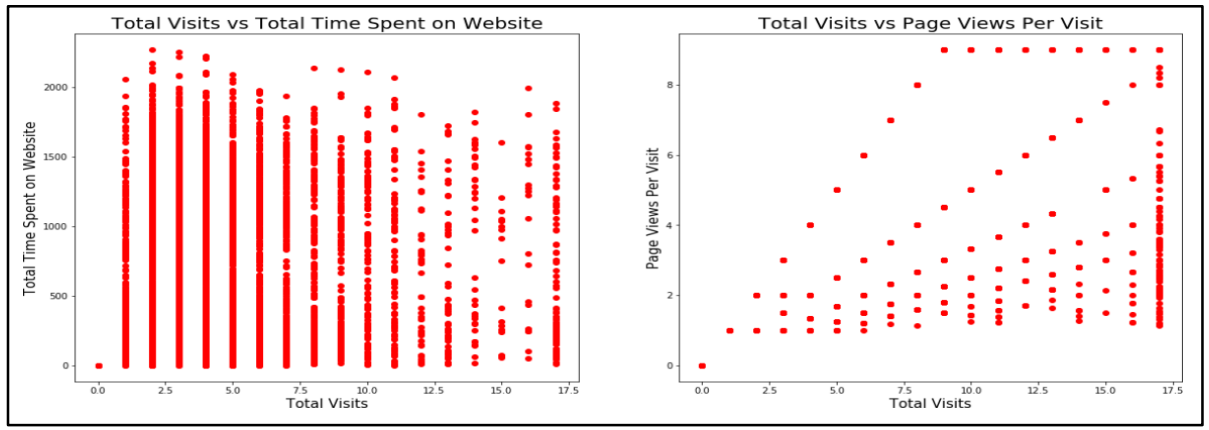


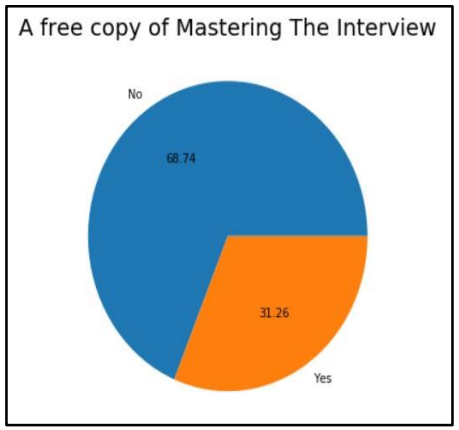




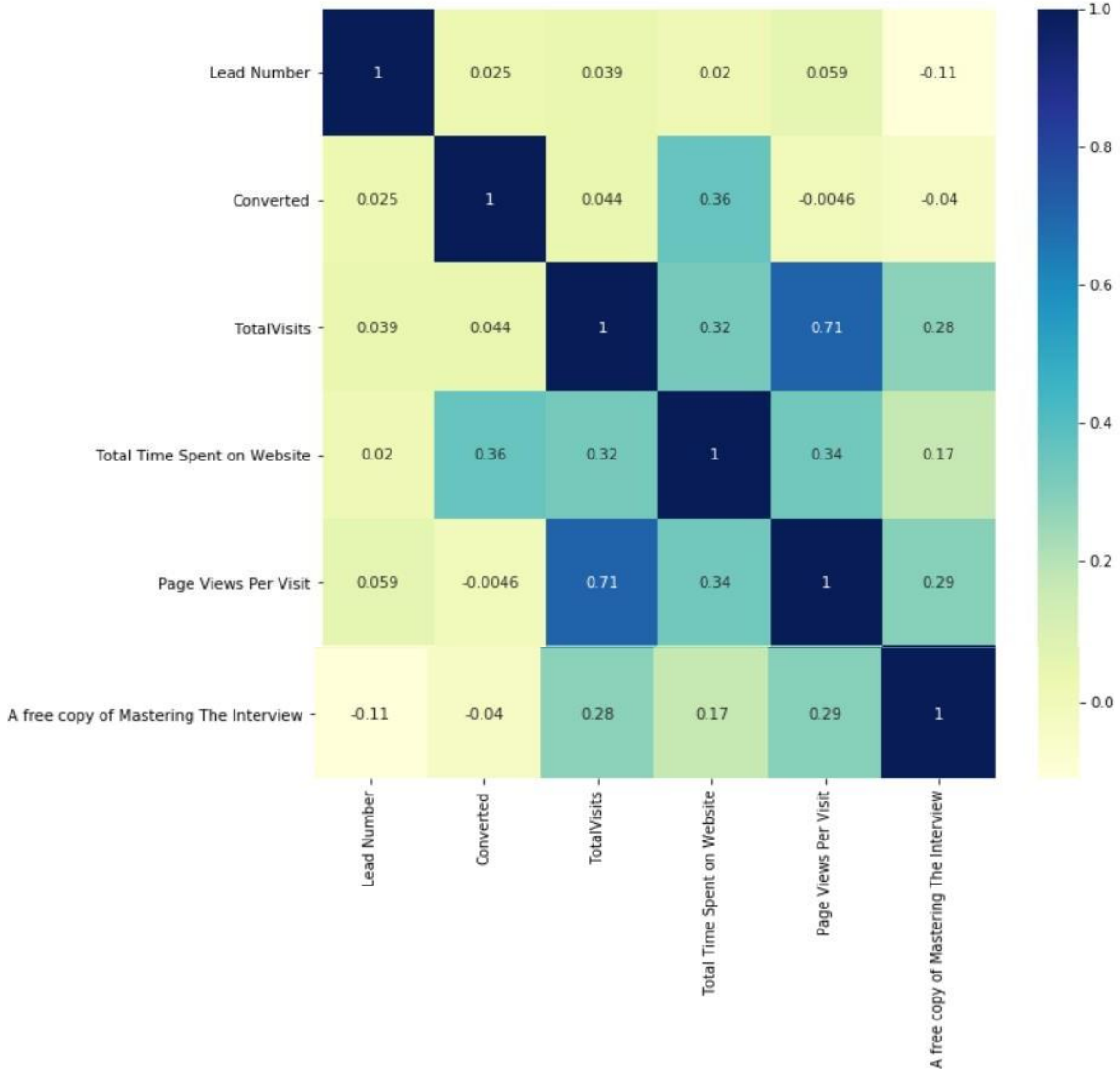








Exploratory Data Analysis: EDA



Observation: EDA

1.Lead Origin: Landing pagesSubmission has highest count with maximum number of submission followed by API.
2.Lead Source: Google has the highest count with maximum conversion. They should put more advertising on Google. In terms of Converted percentage, Other_Lead_Source has highest percentage but problem is source is missing to identify. Direct traffic is also good contributor for lead conversion.
3.Current Occupation: seems like, unemployed is highest group with most conversion. In terms of Working Profession, conversion percentage is very high. For longer run, Working Professional will be good focus area for company.
4.City: Seems like, company should focus on Mumbai but there is lot of opportunity to expand to other cities.
5.Specialization: This seems to very scattered with most of the specialization under other section or very small section.
6.Positive correlation between Converted and Time spend on website.

Data Preparation

1. Dummy Creation
2. Data split:
   - Train Data Set: 70%
   - Test Data Set: 30%
3. Scaling of Data (Standard Scaler)

Logistic Regression Model Building

1. Model building on all features
2. Recursive Feature Elimination (RFE): 20 features
3. Manual model building
4. Model finalized with 13 features
   - P value show Significance of features
   - VIF for all features below 3.0
   - Coefficient shows importance of features
5. Confusion matrix & Metric calculation
6. ROC Curve
7. Optimal Cut-off point
8. Precision and Recall graph
9. F1 Score

Final Model

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:**	-2705.9
Date:	Mon, 26 Oct 2020	**Deviance:**	5411.8
Time:	18:19:20	**Pearson chi2:**	6.92e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

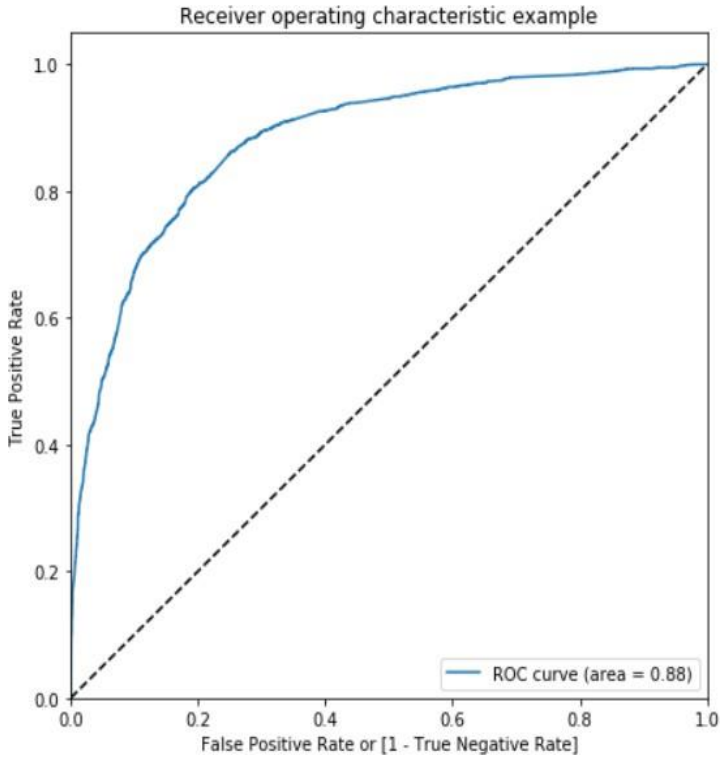
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6837	0.104	-16.160	0.000	-1.888	-1.480
Total Time Spent on Website	1.0930	0.039	27.919	0.000	1.016	1.170
Landing Page Submission	-0.4473	0.088	-5.069	0.000	-0.620	-0.274
Olark Chat	0.8688	0.115	7.532	0.000	0.643	1.095
Reference	3.4486	0.212	16.247	0.000	3.033	3.865
Welingak Website	5.2065	0.726	7.173	0.000	3.784	6.629
Finance Management	0.3378	0.112	3.005	0.003	0.117	0.558
Unemployed	0.9370	0.082	11.494	0.000	0.777	1.097
Working Professional	3.4563	0.195	17.701	0.000	3.074	3.839
Had a Phone Conversation	3.5003	1.112	3.149	0.002	1.322	5.679
Modified	-0.7095	0.083	-8.544	0.000	-0.872	-0.547
Olark Chat Conversation	-1.1330	0.333	-3.398	0.001	-1.786	-0.480
SMS Sent	1.2681	0.084	15.033	0.000	1.103	1.433
Unreachable	1.6619	0.530	3.136	0.002	0.623	2.701

VIF

	Features	VIF
6	Unemployed	2.40
1	Landing Page Submission	2.33
9	Modified	1.62
11	SMS Sent	1.60
2	Olark Chat	1.57
3	Reference	1.34
7	Working Professional	1.34
0	Total Time Spent on Website	1.24
5	Finance Management	1.18
4	Welingak Website	1.07
10	Olark Chat Conversation	1.07
12	Unreachable	1.01
8	Had a Phone Conversation	1.00

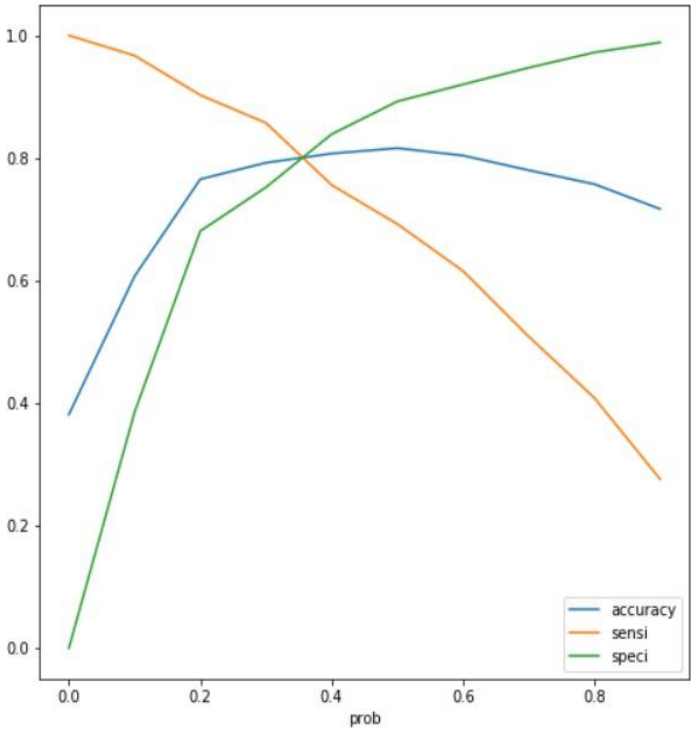
Model Building

ROC Curve



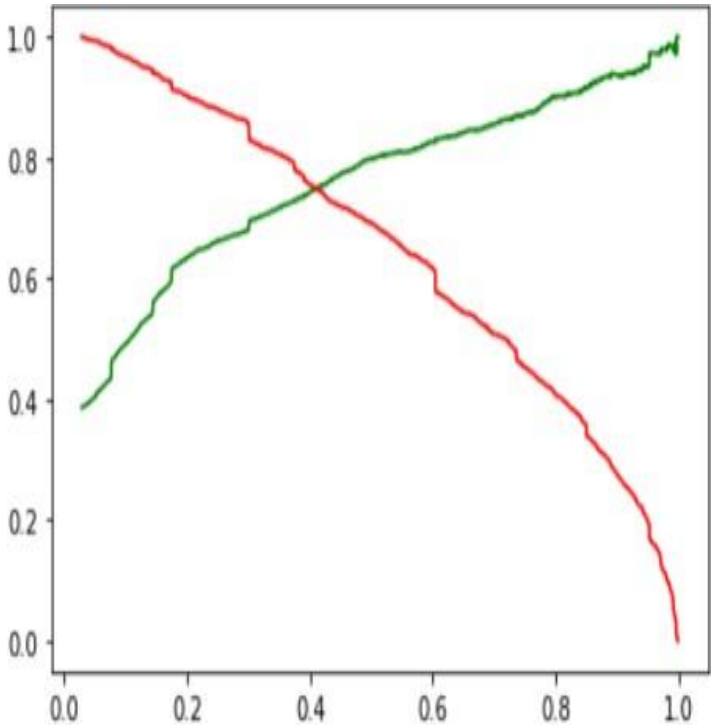
ROC Score: 0.88

Optimal Cut-off Graph



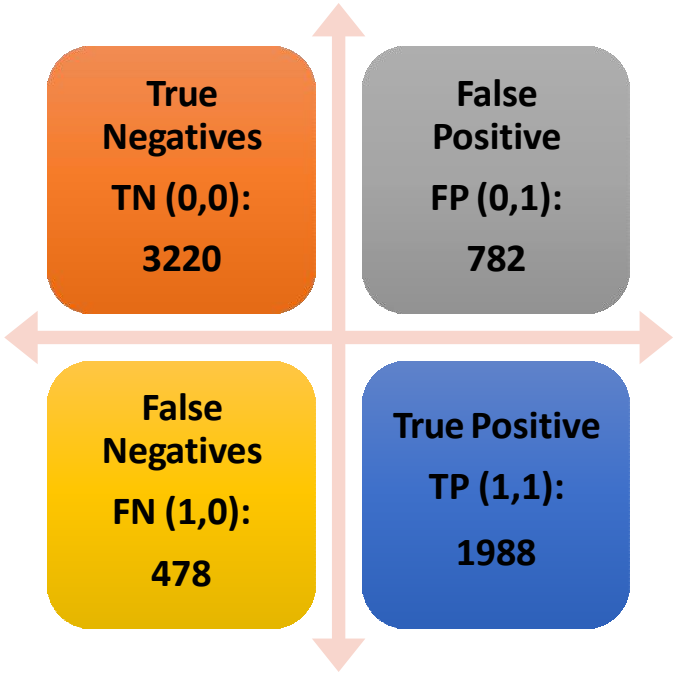
Optimal Cut-off: 0.35

Precision and Recall Graph



Precision and Recall: 0.40

Confusion Matrix



Prediction on Test Data

	Converted	Converted_Prob	Sr. Number	Convert_Predict	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Convert_predicted2
0	0	0.300340	1871	0	1	1	1	1	0	0	0	0	0	0	0
1	0	0.233649	6795	0	1	1	1	0	0	0	0	0	0	0	0
2	0	0.346675	3516	0	1	1	1	1	0	0	0	0	0	0	0
3	0	0.827240	8105	1	1	1	1	1	1	1	1	1	1	0	1
4	0	0.174337	3934	0	1	1	0	0	0	0	0	0	0	0	0
5	1	0.990335	4844	1	1	1	1	1	1	1	1	1	1	1	1
6	0	0.120067	3297	0	1	1	0	0	0	0	0	0	0	0	0
7	1	0.985599	8071	1	1	1	1	1	1	1	1	1	1	1	1
8	0	0.150459	987	0	1	1	0	0	0	0	0	0	0	0	0
9	1	0.935805	7423	1	1	1	1	1	1	1	1	1	1	1	1

Final Results: Train Data Set

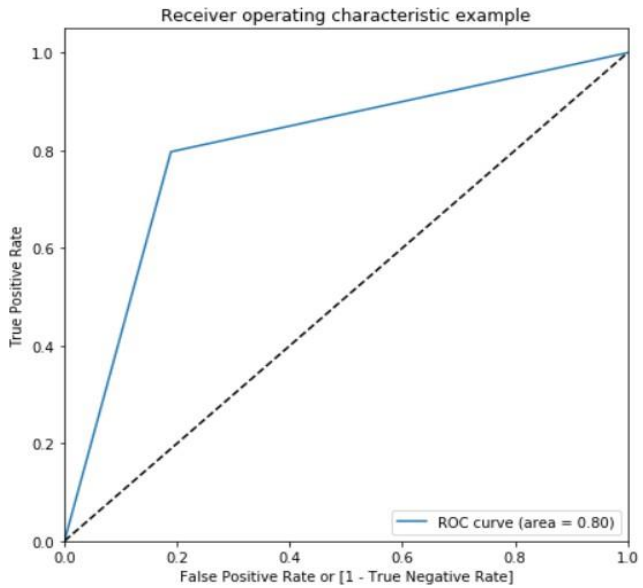
- Overall Accuracy: 0.806
- Sensitivity, Recall, hit rate, or true positive rate (TPR): 0.806
- Precision or positive predictive value (PPV): 0.717
- F1 Score:0.759

Prediction on Test Data Set

Confusion Matrix

True Negatives TN (0,0): 1360	**False Positive FP (0,1): 317**
False Negatives FN (1,0): 222	**True Positive TP (1,1): 873**

ROC Curve



Prediction on Test Data

	Converted	Sr. Number	Converted_Prob	Convert_Predicted
0	1	4269	0.654429	1
1	1	2376	0.952674	1
2	1	7766	0.929879	1
3	0	9199	0.174337	0
4	1	4359	0.849942	1

Final Results: Test Data Set

- Overall Accuracy: 0.805
- Sensitivity, Recall, hit rate, or true positive rate (TPR): 0.797
- Precision or positive predictive value (PPV): 0.733
- F1 Score:0.764

Thank You