Lead Scoring Case Study

NIMIT BHAWANI

Business Problem Statement

Business 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.

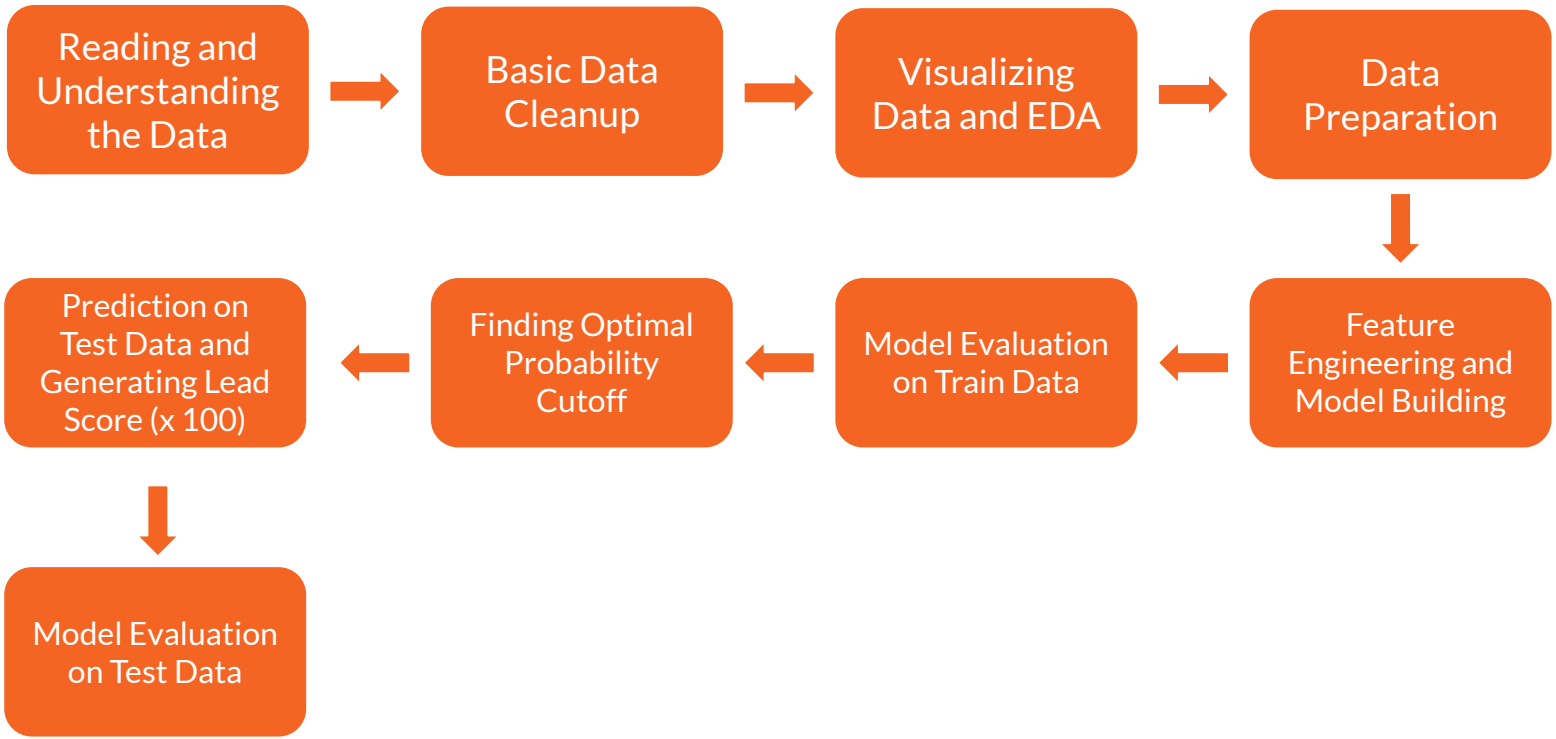
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified 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%.

Goal:

1. To identify the features that contributes to predict Lead Conversion.

2. Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

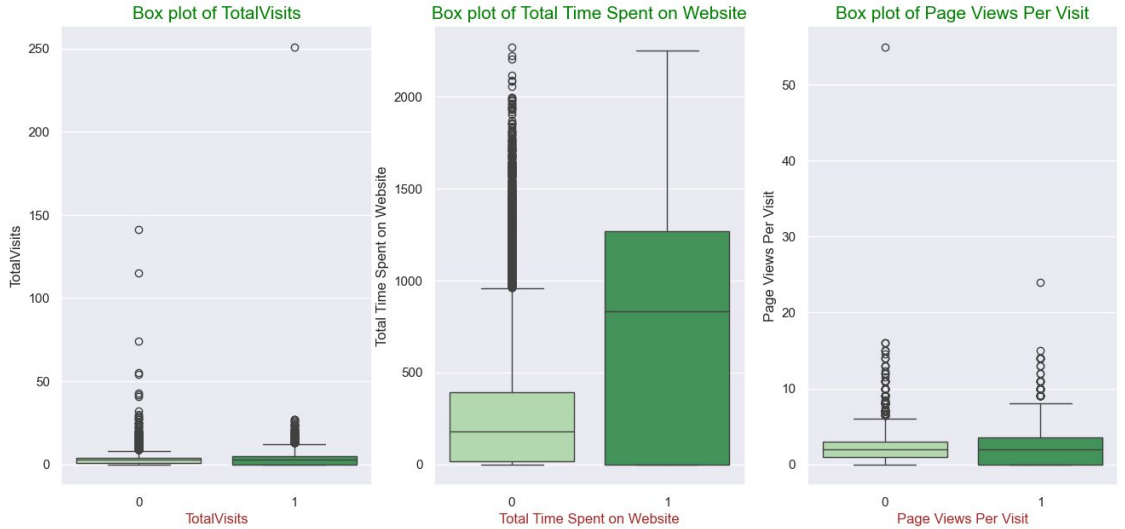
Overall Approach

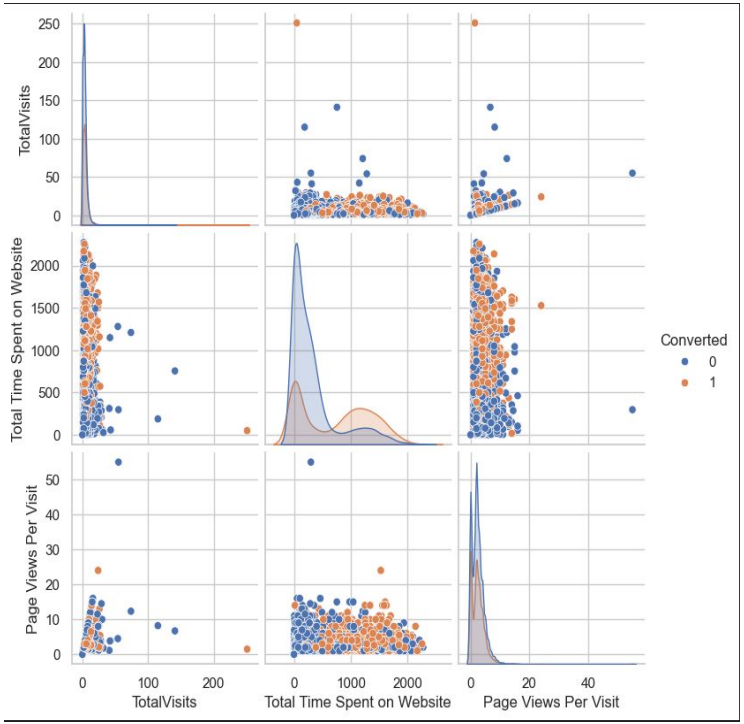


Understanding the Data & Basic Data Cleanup

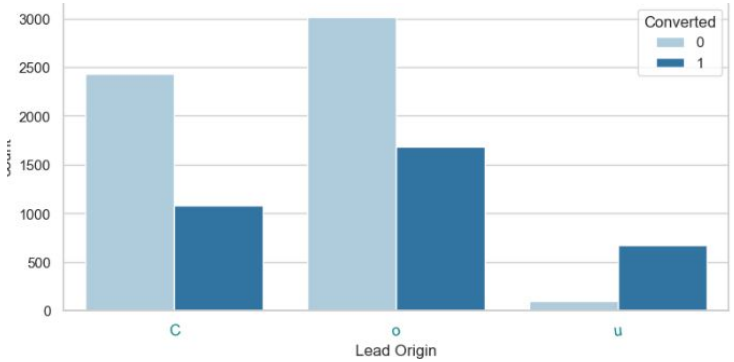
- There are 37 columns (30 categorical and 7 Numeric) and 9240 observations in the dataset.
- Select is present as a class in different columns like:
  - ➢ Specialization
  - ➢ How did you hear about X Education
  - ➢ Lead Profile
  - ➢ City
- As Select is not a valid class, we can conclude that the Select might be the default value set in the form dropdown and if the user has not selected any option from the dropdown, then the value remained as Select. We replaced Select with NaN.
- 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 - These columns have no missing data and have only one unique value. So, these columns have no variance and not helpful for our EDA or model building, hence we dropped these columns.
- How did you hear about X Education, Lead Profile, Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Index , Asymmetrique Activity Score, Asymmetrique Profile Score – These columns have more than 40% missing value. So, we have dropped these columns from our EDA and model building.
- There is no datapoint/ observation (rows) in our dataset having more than 70% missing values.
- We have created new buckets/bins for the categorical variables having very high numbers of classes with few datapoints: Lead Origin, Lead Source, Last Activity, Last Notable Activity, Country, Specialization, What is your current occupation.
- Performed missing value treatment using Business Understanding. For Specialization and Occupation NaN values are replaced with a new category Not Disclosed.
- We renamed What is your current occupation column to Occupation and What matters most to you in choosing a course to Reason_choosing for our convenience during EDA and Model building .

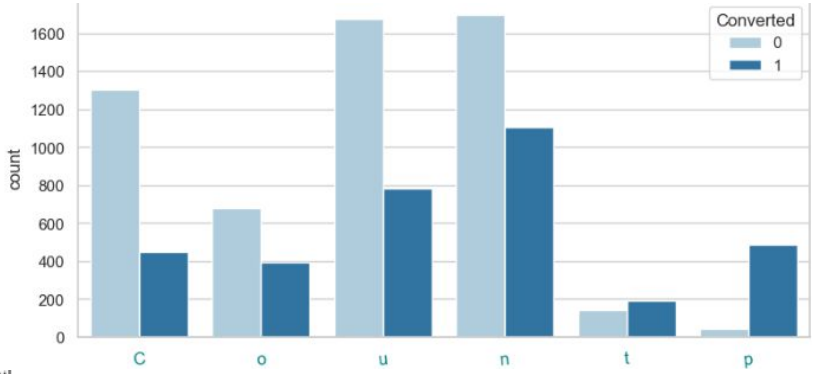
Visualizing Data and EDA : Numerical variables



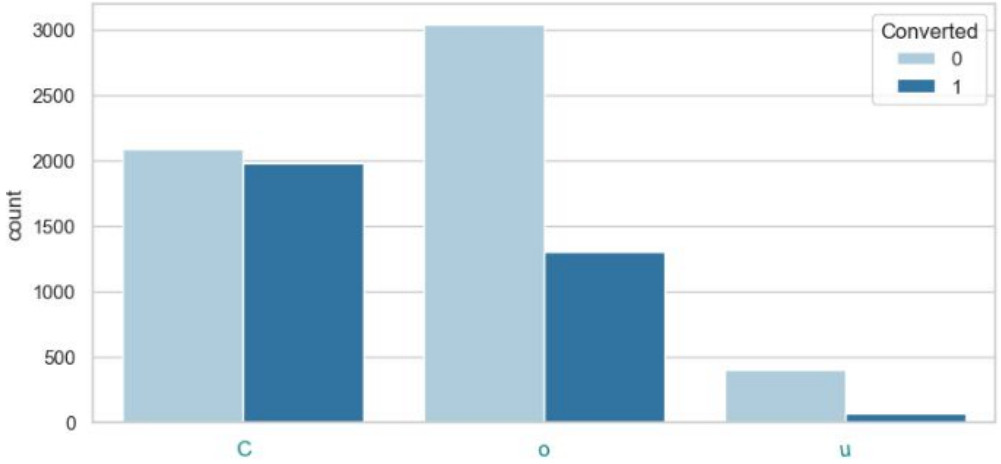


Visualizing Data and EDA : Categorical Variables I

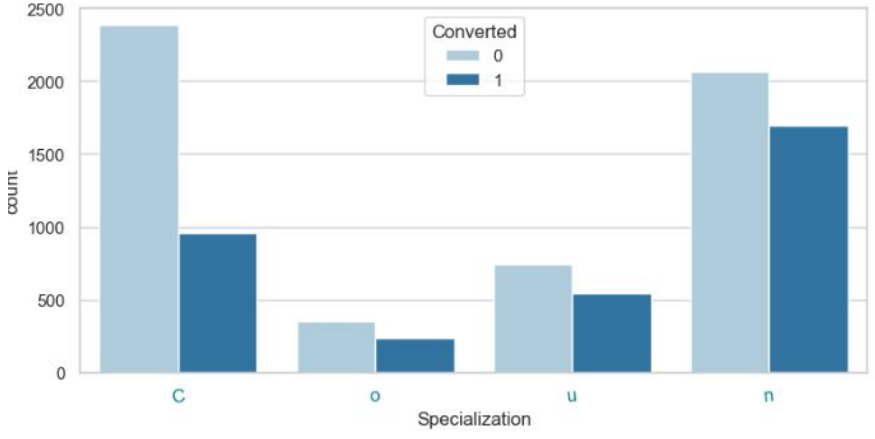


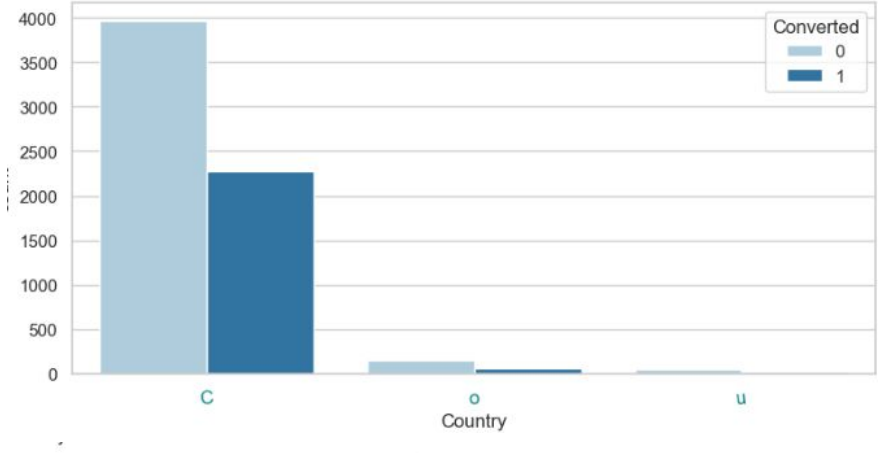


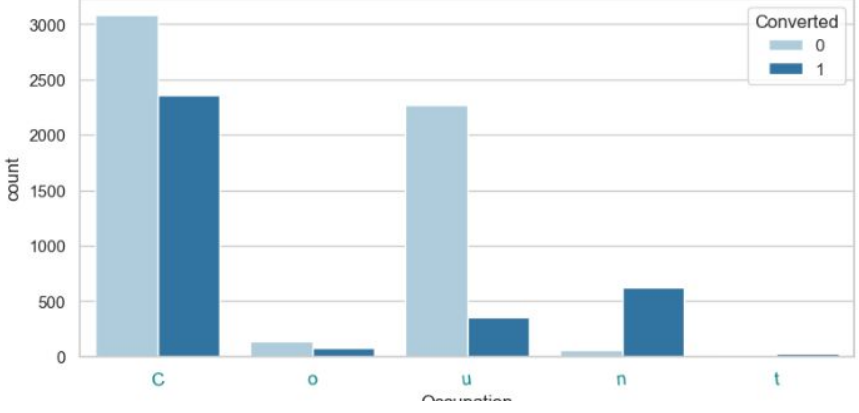
Do Not Email



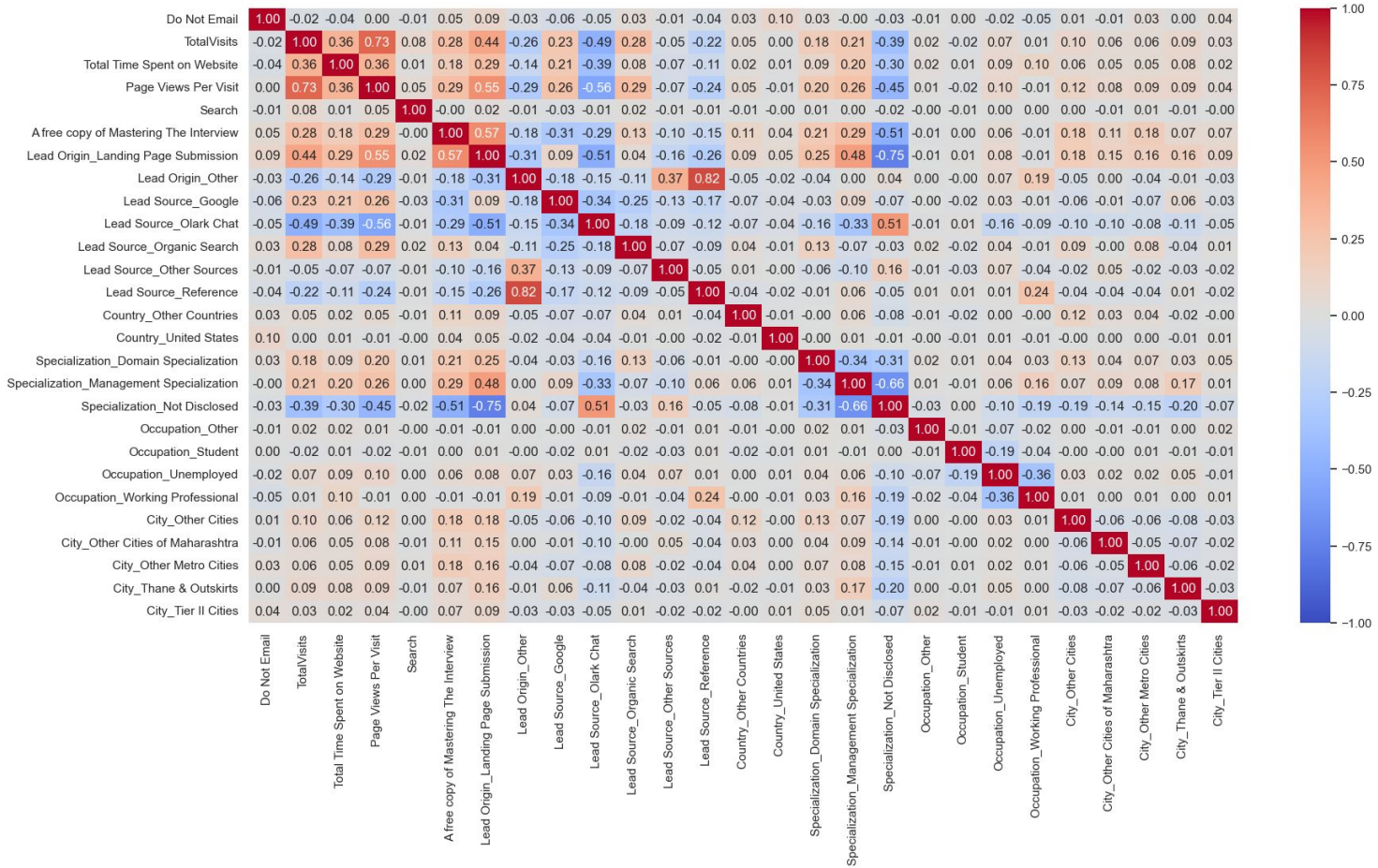
Visualizing Data and EDA : Categorical Variables II







Data Preparation: Pairwise Correlation



Model Building : Approach

1. Recursive Feature Elimination (RFE) has been used to get top 16 features

➢ Do Not Email : An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
➢ TotalVisits: The total number of visits made by the customer on the website.
➢ Total Time Spent on Website: The total time spent by the customer on the website.
➢ Page Views Per Visit: Average number of pages on the website viewed during the visits.
➢ Lead Origin_Landing Page Submission: Dummy variable for Landing Page category of the origin identifier with which the customer was identified to be a lead.
➢ Lead Origin_Other: Dummy variable for the Other category of the origin identifier with which the customer was identified to be a lead.
➢ Lead Source_Olark Chat: Dummy variable for the Olark Chat category of the source of the lead.
➢ Lead Source_Other Sources: Dummy variable for the Other category (other than Google, Direct Traffic, Olark Chat, Organic Search, Reference) of the source of the lead.
➢ Country_Other Countries: Dummy variable for the Other category (other than India and United States) of the country of the customer.
➢ Specialization_Domain Specialization: Dummy variable for Domain Specialization bin of Specialization variable.
➢ Specialization_Management Specialization: Dummy variable for Management Specialization bin of Specialization variable.
➢ Occupation_Other: Dummy variable for 'Other' category of customer's occupation.
➢ Occupation_Student: Dummy variable for 'Student' category of customer's occupation.
➢ Occupation_Unemployed: Dummy variable for 'Unemployed' category of customer's occupation.
➢ Occupation_Working Professional: Dummy variable for 'Working Professional' category of customer's occupation.
➢ City_Tier II Cities: Dummy variable for 'Tier II Cities' category of customer's city.

2. We built first Logistic Regression model using GLM (Generalized Linear Model) in statsmodels with these 16 features.

3. Then manually fine tuned the model to get statistically significant features (by checking the p-values) and removed multicollinearity (By checking Variance Inflation Factors) simultaneously. Accepted p-value is lower than .05 and accepted VIF is lower than 5.

4. Total 7 models were built and after each model building p-values of all beta coefficients and VIFs have been checked and identified feature has been removed in next model building. We have also checked Overall model accuracy and Confusion Matrix after each new model, to understand how the new model is performing in compared to the previous one.

Model Building : Model 1

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.5408	0.150	-10.277	0.000	-1.835	-1.247
TotalVisits	-0.2658	0.251	-1.059	0.290	-0.758	0.226
Total Time Spent on Website	2.9075	0.126	23.068	0.000	2.660	3.155
Page Views Per Visit	-2.8306	0.247	-11.439	0.000	-3.316	-2.346
Lead Origin_Landing Page Submission	-1.1655	0.098	-11.889	0.000	-1.358	-0.973
Lead Origin_Other	1.5778	0.181	8.696	0.000	1.222	1.933
Lead Source_Olark Chat	-1.3287	0.076	-17.453	0.000	-1.478	-1.179
Lead Source_Other Sources	-0.8808	0.201	-4.374	0.000	-1.275	-0.486
Country_Other Countries	-0.6534	0.220	-2.974	0.003	-1.084	-0.223
Specialization_Domain Specialization	0.3693	0.118	3.137	0.002	0.139	0.600
Specialization_Management Specialization	0.2861	0.095	3.025	0.002	0.101	0.471
...						
Occupation_Unemployed	0.3231	0.065	4.942	0.000	0.195	0.451
Occupation_Working Professional	2.8928	0.188	15.384	0.000	2.524	3.261
City_Tier II Cities	0.2280	0.342	0.666	0.506	-0.443	0.899

```
Confusion Matrix:
True Negative: 3248      False Positive: 643
False Negative: 878      True Positive: 1519

Overall model accuracy: 0.7581106870229007
```

	Features	VIF
3	Page Views Per Visit	6.06
4	Lead Origin_Landing Page Submission	5.04
1	TotalVisits	4.78
10	Specialization_Management Specialization	3.68
13	Occupation_Unemployed	2.82
2	Total Time Spent on Website	2.20
9	Specialization_Domain Specialization	1.92
5	Lead Origin_Other	1.66
14	Occupation_Working Professional	1.43
7	Lead Source_Other Sources	1.25
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
12	Occupation_Student	1.07
8	Country_Other Countries	1.04
15	City_Tier II Cities	1.02
11	Occupation_Other	1.01

Model Building : Model 2

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.5381	0.150	-10.262	0.000	-1.832	-1.244
TotalVisits	-0.2660	0.251	-1.059	0.290	-0.758	0.226
Total Time Spent on Website	2.9076	0.126	23.070	0.000	2.661	3.155
Page Views Per Visit	-2.8311	0.247	-11.441	0.000	-3.316	-2.346
Lead Origin_Landing Page Submission	-1.1606	0.098	-11.875	0.000	-1.352	-0.969
Lead Origin_Other	1.5783	0.181	8.699	0.000	1.223	1.934
Lead Source_Olark Chat	-1.3284	0.076	-17.451	0.000	-1.478	-1.179
Lead Source_Other Sources	-0.8809	0.201	-4.375	0.000	-1.276	-0.486
Country_Other Countries	-0.6548	0.220	-2.980	0.003	-1.085	-0.224
Specialization_Domain Specialization	0.3697	0.118	3.142	0.002	0.139	0.600
Specialization_Management Specialization	0.2841	0.095	3.006	0.003	0.099	0.469
...						
Occupation_Student	0.1512	0.202	0.750	0.453	-0.244	0.546
Occupation_Unemployed	0.3228	0.065	4.938	0.000	0.195	0.451
Occupation_Working Professional	2.8944	0.188	15.393	0.000	2.526	3.263

```
Confusion Matrix:
True Negative: 3250      False Positive: 641
False Negative: 876      True Positive: 1521

Overall model accuracy: 0.7587468193384224
```

	Features	VIF
3	Page Views Per Visit	6.06
4	Lead Origin_Landing Page Submission	5.01
1	TotalVisits	4.78
10	Specialization_Management Specialization	3.68
13	Occupation_Unemployed	2.82
2	Total Time Spent on Website	2.20
9	Specialization_Domain Specialization	1.92
5	Lead Origin_Other	1.66
14	Occupation_Working Professional	1.43
7	Lead Source_Other Sources	1.25
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
12	Occupation_Student	1.07
8	Country_Other Countries	1.04
11	Occupation_Other	1.01

Model Building : Model 3

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.5434	0.150	-10.303	0.000	-1.837	-1.250
TotalVisits	-0.2745	0.251	-1.095	0.274	-0.766	0.217
Total Time Spent on Website	2.9070	0.126	23.082	0.000	2.660	3.154
Page Views Per Visit	-2.8374	0.247	-11.477	0.000	-3.322	-2.353
Lead Origin_Landing Page Submission	-1.1746	0.098	-12.044	0.000	-1.366	-0.983
Lead Origin_Other	1.5915	0.182	8.764	0.000	1.236	1.947
Lead Source_Olark Chat	-1.3277	0.076	-17.449	0.000	-1.477	-1.179
Lead Source_Other Sources	-0.8987	0.201	-4.474	0.000	-1.292	-0.505
Specialization_Domain Specialization	0.3743	0.118	3.183	0.001	0.144	0.605
Specialization_Management Specialization	0.2814	0.094	2.982	0.003	0.096	0.466
Occupation_Other	0.6217	0.523	1.189	0.234	-0.403	1.647
Occupation_Student	0.1621	0.202	0.804	0.422	-0.233	0.557
Occupation_Unemployed	0.3194	0.065	4.892	0.000	0.191	0.447
Occupation_Working Professional	2.8849	0.187	15.404	0.000	2.518	3.252

	Features	VIF
3	Page Views Per Visit	6.06
4	Lead Origin_Landing Page Submission	5.00
1	TotalVisits	4.78
9	Specialization_Management Specialization	3.68
12	Occupation_Unemployed	2.82
2	Total Time Spent on Website	2.20
8	Specialization_Domain Specialization	1.92
5	Lead Origin_Other	1.66
13	Occupation_Working Professional	1.43
7	Lead Source_Other Sources	1.25
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
11	Occupation_Student	1.07
10	Occupation_Other	1.01

```
Confusion Matrix:
True Negative: 3249      False Positive: 642
False Negative: 879      True Positive: 1518

Overall model accuracy: 0.7581106870229007
```

Model Building : Model 4

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.5592	0.149	-10.444	0.000	-1.852	-1.267
TotalVisits	-2.1609	0.203	-10.641	0.000	-2.559	-1.763
Total Time Spent on Website	2.6996	0.122	22.046	0.000	2.460	2.940
Lead Origin_Landing Page Submission	-1.4111	0.095	-14.908	0.000	-1.597	-1.226
Lead Origin_Other	1.7969	0.182	9.880	0.000	1.440	2.153
Lead Source_Olark Chat	-1.2400	0.075	-16.600	0.000	-1.386	-1.094
Lead Source_Other Sources	-1.0475	0.199	-5.257	0.000	-1.438	-0.657
Specialization_Domain Specialization	0.3057	0.116	2.636	0.008	0.078	0.533
Specialization_Management Specialization	0.2303	0.093	2.468	0.014	0.047	0.413
Occupation_Other	0.4798	0.516	0.930	0.353	-0.532	1.491
Occupation_Student	0.0649	0.201	0.324	0.746	-0.328	0.458
Occupation_Unemployed	0.1633	0.063	2.593	0.010	0.040	0.287
Occupation_Working Professional	2.7132	0.185	14.696	0.000	2.351	3.075

	Features	VIF
3	Lead Origin_Landing Page Submission	4.69
8	Specialization_Management Specialization	3.67
1	TotalVisits	2.87
11	Occupation_Unemployed	2.74
2	Total Time Spent on Website	2.18
7	Specialization_Domain Specialization	1.92
4	Lead Origin_Other	1.64
12	Occupation_Working Professional	1.43
6	Lead Source_Other Sources	1.24
5	Lead Source_Olark Chat	1.20
0	Do Not Email	1.09
10	Occupation_Student	1.07
9	Occupation_Other	1.01

```
Confusion Matrix:
True Negative: 3149      False Positive: 742
False Negative: 877      True Positive: 1520

Overall model accuracy: 0.7425254452926209
```

Model Building : Model 5

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.7697	0.147	-12.066	0.000	-2.057	-1.482
TotalVisits	-2.8068	0.201	-13.940	0.000	-3.201	-2.412
Total Time Spent on Website	2.5114	0.119	21.064	0.000	2.278	2.745
Lead Origin_Other	2.3223	0.175	13.258	0.000	1.979	2.666
Lead Source_Olark Chat	-1.0493	0.072	-14.569	0.000	-1.190	-0.908
Lead Source_Other Sources	-1.0829	0.205	-5.290	0.000	-1.484	-0.682
Specialization_Domain Specialization	-0.5971	0.096	-6.188	0.000	-0.786	-0.408
Specialization_Management Specialization	-0.6668	0.069	-9.624	0.000	-0.803	-0.531
Occupation_Other	0.5932	0.511	1.161	0.246	-0.409	1.595
Occupation_Student	-0.1197	0.194	-0.616	0.538	-0.501	0.261
Occupation_Unemployed	0.0294	0.061	0.484	0.628	-0.090	0.148
Occupation_Working Professional	2.7405	0.185	14.785	0.000	2.377	3.104

	Features	VIF
1	TotalVisits	2.70
10	Occupation_Unemployed	2.68
7	Specialization_Management Specialization	2.27
2	Total Time Spent on Website	2.16
3	Lead Origin_Other	1.50
6	Specialization_Domain Specialization	1.46
11	Occupation_Working Professional	1.42
5	Lead Source_Other Sources	1.24
4	Lead Source_Olark Chat	1.17
0	Do Not Email	1.06
9	Occupation_Student	1.06
8	Occupation_Other	1.01

```
Confusion Matrix:
True Negative: 3222      False Positive: 669
False Negative: 900      True Positive: 1497

Overall model accuracy: 0.7504770992366412
```

Model Building : Model 6

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.8074	0.145	-12.448	0.000	-2.092	-1.523
TotalVisits	-3.1592	0.195	-16.179	0.000	-3.542	-2.776
Total Time Spent on Website	2.4330	0.118	20.596	0.000	2.201	2.665
Lead Origin_Other	2.2725	0.175	13.005	0.000	1.930	2.615
Lead Source_Olark Chat	-1.0276	0.072	-14.319	0.000	-1.168	-0.887
Lead Source_Other Sources	-0.9733	0.203	-4.790	0.000	-1.372	-0.575
Specialization_Management Specialization	-0.4924	0.063	-7.811	0.000	-0.616	-0.369
Occupation_Other	0.4937	0.517	0.954	0.340	-0.520	1.508
Occupation_Student	-0.2028	0.195	-1.042	0.297	-0.584	0.179
Occupation_Unemployed	-0.0485	0.059	-0.819	0.413	-0.165	0.068
Occupation_Working Professional	2.6307	0.185	14.243	0.000	2.269	2.993

```
Confusion Matrix:
True Negative: 3232      False Positive: 659
False Negative: 904      True Positive: 1493

Overall model accuracy: 0.7514312977099237
```

	Features	VIF
9	Occupation_Unemployed	2.55
1	TotalVisits	2.48
2	Total Time Spent on Website	2.15
6	Specialization_Management Specialization	1.88
3	Lead Origin_Other	1.49
10	Occupation_Working Professional	1.39
5	Lead Source_Other Sources	1.23
4	Lead Source_Olark Chat	1.16
0	Do Not Email	1.06
8	Occupation_Student	1.06
7	Occupation_Other	1.01

Model Building : Model 7

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.8163	0.143	-12.696	0.000	-2.097	-1.536
TotalVisits	-3.5042	0.192	-18.237	0.000	-3.881	-3.128
Total Time Spent on Website	2.2953	0.116	19.791	0.000	2.068	2.523
Lead Origin_Other	2.1961	0.174	12.609	0.000	1.855	2.538
Lead Source_Olark Chat	-1.0003	0.072	-13.974	0.000	-1.141	-0.860
Lead Source_Other Sources	-0.8346	0.201	-4.147	0.000	-1.229	-0.440
Occupation_Other	0.3911	0.511	0.766	0.444	-0.609	1.392
Occupation_Student	-0.3052	0.195	-1.567	0.117	-0.687	0.077
Occupation_Unemployed	-0.1654	0.057	-2.903	0.004	-0.277	-0.054
Occupation_Working Professional	2.4277	0.183	13.254	0.000	2.069	2.787

	Features	VIF
8	Occupation_Unemployed	2.39
1	TotalVisits	2.34
2	Total Time Spent on Website	2.11
3	Lead Origin_Other	1.49
9	Occupation_Working Professional	1.32
5	Lead Source_Other Sources	1.22
4	Lead Source_Olark Chat	1.16
0	Do Not Email	1.05
7	Occupation_Student	1.05
6	Occupation_Other	1.01

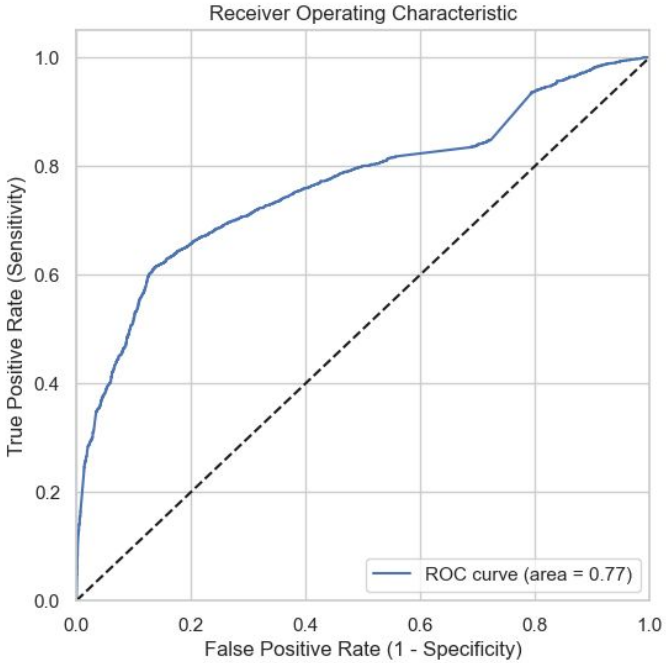
```
Confusion Matrix:
True Negative: 3272      False Positive: 619
False Negative: 887      True Positive: 1510

Overall model accuracy: 0.7604961832061069
```

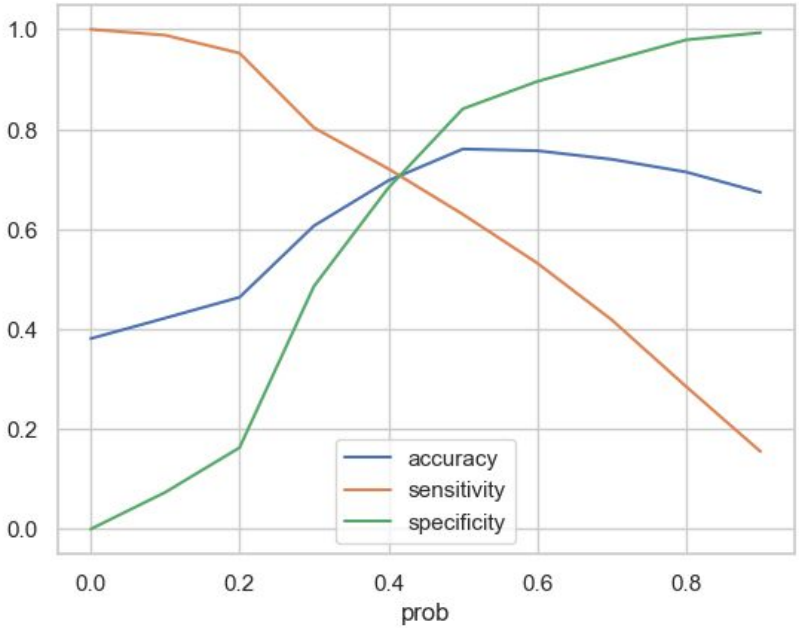
Prediction & Model Evaluation : (on Training data - cutoff .5)

```
Overall model accuracy: 0.6262722646310432
Sensitivity / Recall:  0.7909887359198998
Specificity:  0.5248008224106914
False Positive Rate:  0.4751991775893087
Positive Predictive Value:  0.5062750333778371
Positive Predictive Value:  0.8029885961462839
```

```
Confusion Matrix:
True Negative: 2042      False Positive: 1849
False Negative: 501      True Positive: 1896
```



Finding Optimal Probability cutoff & Evaluating on Train Data




```
Confusion Matrix:
True Negative: 2042      False Positive: 1849
False Negative: 501      True Positive: 1896

Overall model accuracy: 0.6262722646310432
Sensitivity / Recall:  0.7909887359198998
Specificity:  0.5248008224106914
False Positive Rate:  0.4751991775893087
Positive Predictive Value:  0.5062750333778371
Positive Predictive Value:  0.8029885961462839
```

Prediction & Generating Lead Score (Business Requirement)

	Lead Number	Converted	pred_Converted	prob	Lead Score
2656	634047	1	1	0.997795	99.779495
3478	627106	1	1	0.997700	99.769958
6383	600952	1	1	0.997621	99.762093
5921	604411	1	1	0.996650	99.664997
7579	591536	1	1	0.996531	99.653124
6751	598055	1	1	0.996336	99.633613
7211	594089	1	1	0.995963	99.596274
8081	588013	1	1	0.995799	99.579920
9015	581257	1	1	0.995563	99.556311
120	659283	1	1	0.995290	99.529046

	Lead Number	Converted	pred_Converted	prob	Lead Score
8074	588037	1	1	0.997591	99.759112
3428	627462	1	1	0.996899	99.689865
7187	594369	1	1	0.996225	99.622488
8063	588075	1	1	0.995825	99.582523
4613	615524	1	1	0.994379	99.437907
2984	631268	1	1	0.994071	99.407100
8057	588097	0	1	0.992265	99.226470
79	659710	1	1	0.992207	99.220701
4782	614077	1	1	0.990280	99.028011
5784	605335	1	1	0.990280	99.028011

Model Evaluation : (on Test data) & Interpretation

```
Model Evaluation Metrics on Test dataset
#################################################
Confusion Matrix:
True Negative: 896      False Positive: 764
False Negative: 237      True Positive: 798

Overall model accuracy: 0.6285714285714286
Sensitivity / Recall:  0.7710144927536232
Specificity:  0.5397590361445783
False Positive Rate:  0.4602409638554217
Positive Predictive Value:  0.5108834827144686
Positive Predictive Value:  0.7908208296557812
```

	coef	std err	z	P>|z|	[0.025	0.975]
Do Not Email	-1.8163	0.143	-12.696	0.000	-2.097	-1.536
TotalVisits	-3.5042	0.192	-18.237	0.000	-3.881	-3.128
Total Time Spent on Website	2.2953	0.116	19.791	0.000	2.068	2.523
Lead Origin_Other	2.1961	0.174	12.609	0.000	1.855	2.538
Lead Source_Olark Chat	-1.0003	0.072	-13.974	0.000	-1.141	-0.860
Lead Source_Other Sources	-0.8346	0.201	-4.147	0.000	-1.229	-0.440
Occupation_Other	0.3911	0.511	0.766	0.444	-0.609	1.392
Occupation_Student	-0.3052	0.195	-1.567	0.117	-0.687	0.077
Occupation_Unemployed	-0.1654	0.057	-2.903	0.004	-0.277	-0.054
Occupation_Working Professional	2.4277	0.183	13.254	0.000	2.069	2.787

Conclusion and Recommendations

1. As per business requirement Lead Score (between 0 to 100) of the leads have been calculated by using this Logistic Regression model. A higher score means hot lead (most likely to convert), lower score implies cold lead (mostly not get converted).

2. This Lead Score would help to identify the hot leads faster and efficiently, that would result in decrease in lead conversion time and increase in lead conversion rate. Leads should be sorted in descending order according to their Lead Scores.

3. Phone calls or contact should be made to the leads having higher Lead Score first. Some special attentions should be provided to these hot leads (may be assigning a dedicated support SPOC for a small batch of hot leads that have higher Lead Scores), as there is a very high chance of Lead conversion.

4. Leads having medium Lead Score are also potentially good candidates for Lead conversion. They also should be contacted, and right questions should be asked, so that business can understand their requirements and problem areas and can take necessary actions. Few to mention: some changes in existing courses, introducing new courses, some change in class schedules, introducing easy financial options for fees etc. may help to successfully convert these leads.

5. Cold Leads should be contacted after business gets very good Conversion rate with the leads having High and Medium Lead scores. As chance of conversion is very less here, they could be part of company's aggressive marketing policy.