Logistic Regression Case Study

Lead Scoring Case Study

by
Bipin Joseph Odattil

Step 1

Reading and Understanding Data

Available Data Sheets

1. Leads.csv

Contains information about past Leads of X Education

9240 rows, 37 columns

2. Leads Data Dictionary.xlsx

Contains description of the 37 columns

Note: We will not be loading this sheet as a python dataframe as this is a reference data sheet

Leads.csv

Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content
7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No
2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No
8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No
0cc2df48-7cf4-4e39-9de9-	660719	Landing Page	Direct Traffic	No	No	0	1.0	305	1.0	...	No

Problem Statement: Use the Leads data to build a regression model that can get a lead conversion probability(Lead Score) and predict potential leads.

Data Understanding- Columns

Leads.csv

Data Types of Columns

- There are 37 columns of object, float and int datatypes
- Key points:
  - Each lead is identified by Prospect ID and Lead Number
  - Converted (1/0)- This is the target variable
    - 1 means the lead has converted
    - 0 means the lead did not convert

Column	Non-Null Cour
Prospect ID	9240 non-nul
Lead Number	9240 non-nul
Lead Origin	9240 non-nul
Lead Source	9204 non-nul
Do Not Email	9240 non-nul
Do Not Call	9240 non-nul
Converted	9240 non-nul
TotalVisits	9103 non-nul
Total Time Spent on Website	9240 non-nul
Page Views Per Visit	9103 non-nul
Last Activity	9137 non-nul
Country	6779 non-nul
Specialization	7802 non-nul
How did you hear about X Education	7033 non-nul
What is your current occupation	6550 non-nul
What matters most to you in choosing a course	6531 non-nul
Search	9240 non-nul
Magazine	9240 non-nul
Newspaper Article	9240 non-nul

‘Select’ value in categorical variables

- Many of the categoric columns have the value 'Select' in it like below
- This is probably when the user did not select anything from the drop down.
- This is same as an empty or null value.
- Replaced these ‘Select’ with NaN

	Lead Profile	City
0	Select	Select
1	Select	Select
2	Potential Lead	Mumbai
3	Select	Mumbai
4	Select	Mumbai

Handling null values

These nulls in each column was handled separately

- **Country**- Most of the value in the column is 'India', then null. Other values are very less. This column was dropped.
- **Specialization**- Most of the values are null. This column was dropped.
- **How did you hear about X Education**- Most of the values are null. This column was dropped.
- **What is your current occupation**- The nulls were filled with the value 'Other' which was available
- **What matters most to you in choosing a course**- Most of the value are same('Better Career Prospects'). Other values are negligible in number. The column was dropped.
- **Tags**- Most values are null. The column was dropped.
- **Lead Quality**- This seemed like an important column. Nulls were filled with 'Not Sure' which seemed like a neutral value.
- **Lead Profile**- This seemed like an important column. Nulls were filled with 'Other Leads'
- **City**- Most of the values are null and other values are similar. Dropped this column.
- **Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score**- About half of the values are null in these columns. Columns were dropped.

Handling null values

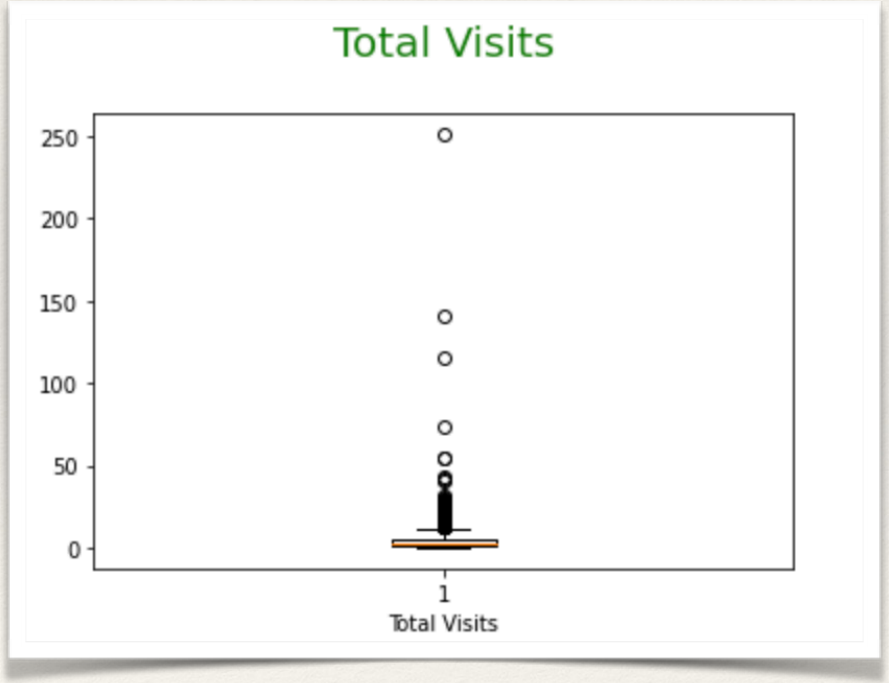
Some columns had very few count of nulls.

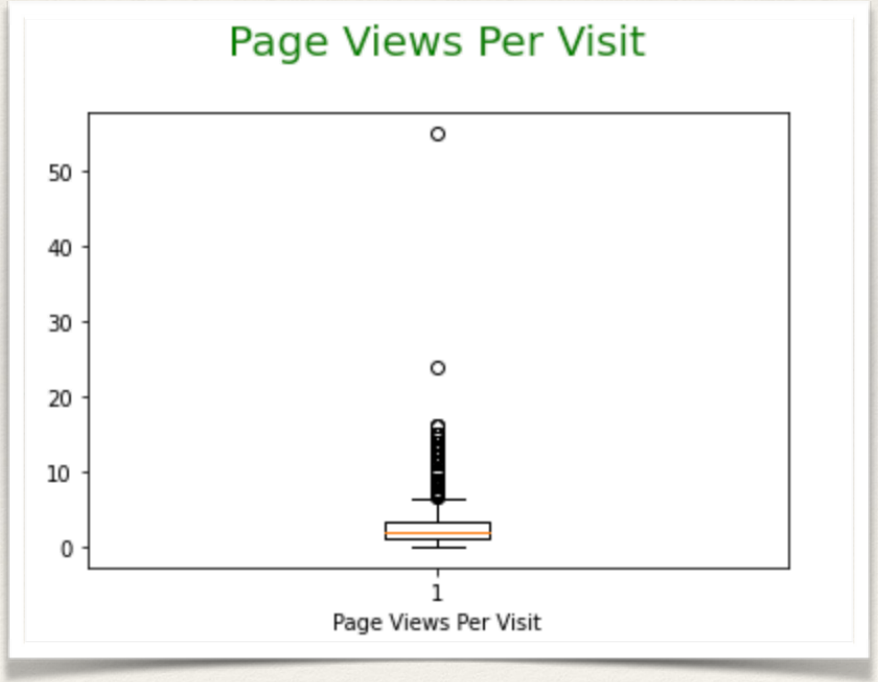
- Below columns had very low null value percentage.
- Lead Source, TotalVisits, Page Views Per Visit, Last Activity.
- These rows with null values were dropped since the count was less.

- There are no more null values

Handling Outliers

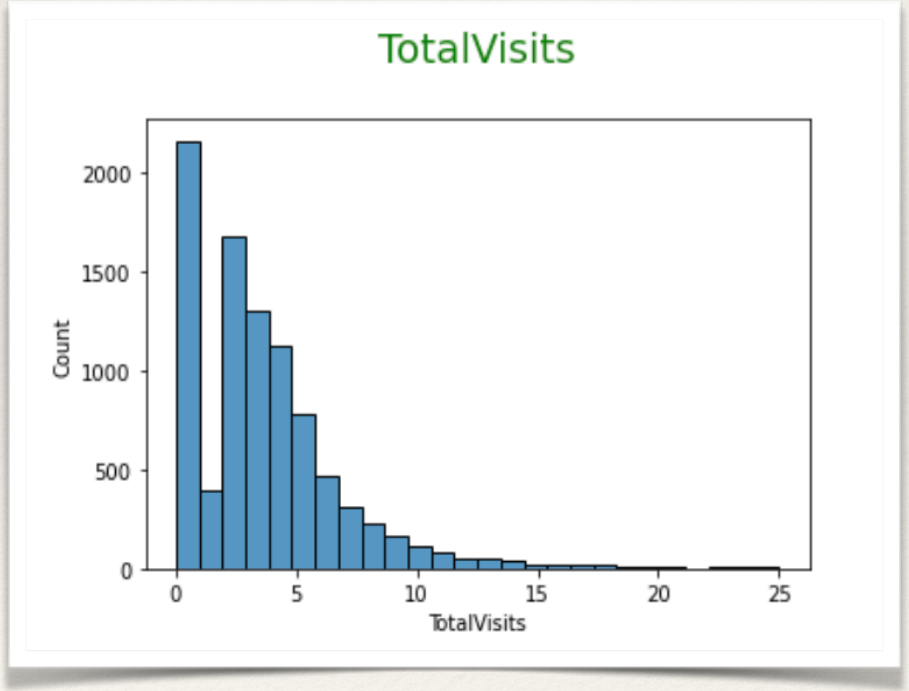
- There are outliers for the columns- TotalVisits and Page Views Per Visit
- The outliers less than and above 1.5 * IQR were removed from the data set

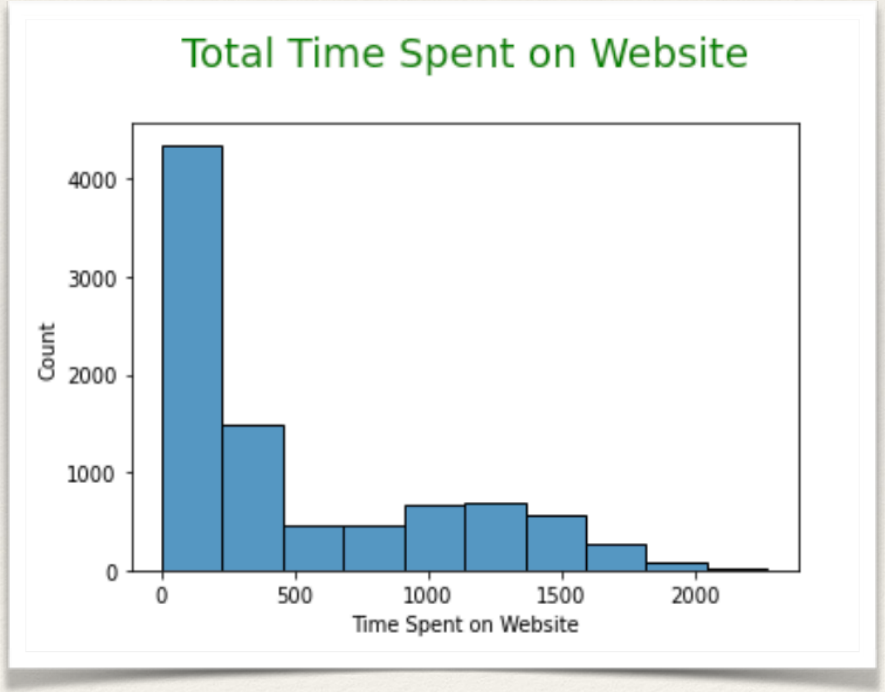




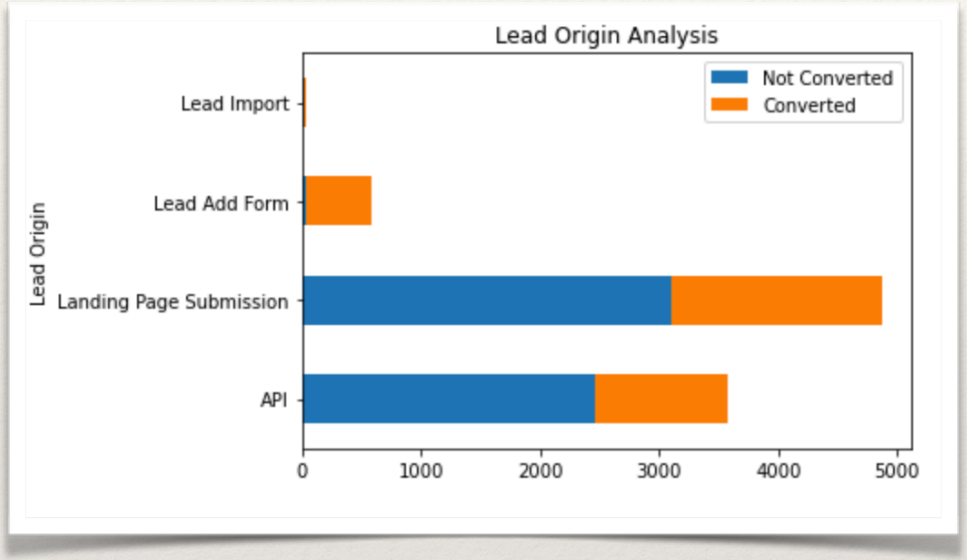
Step 2: EDA- Visualisation

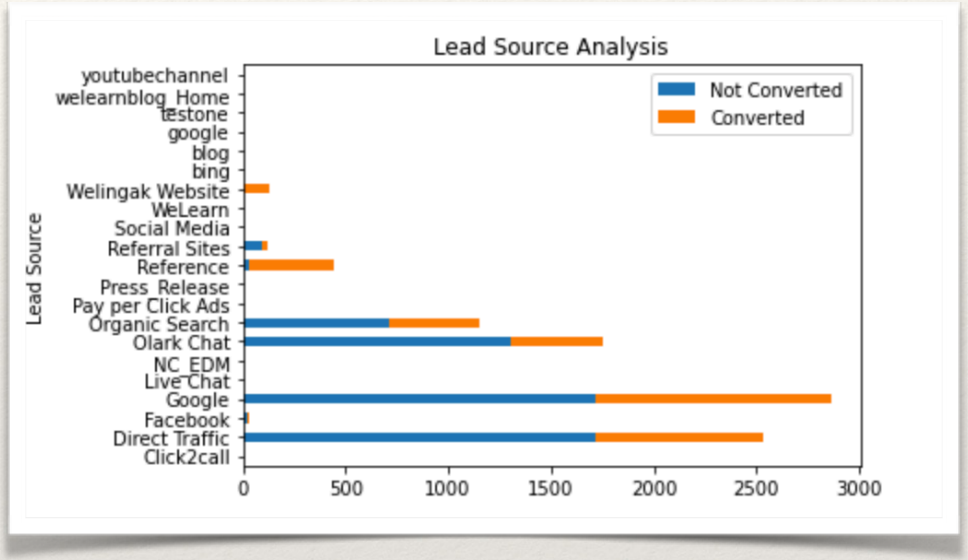
- Nearly 40% of the leads have converted
- The frequency of users making lesser number of visits and lesser time spent are more and frequency decreases for higher values



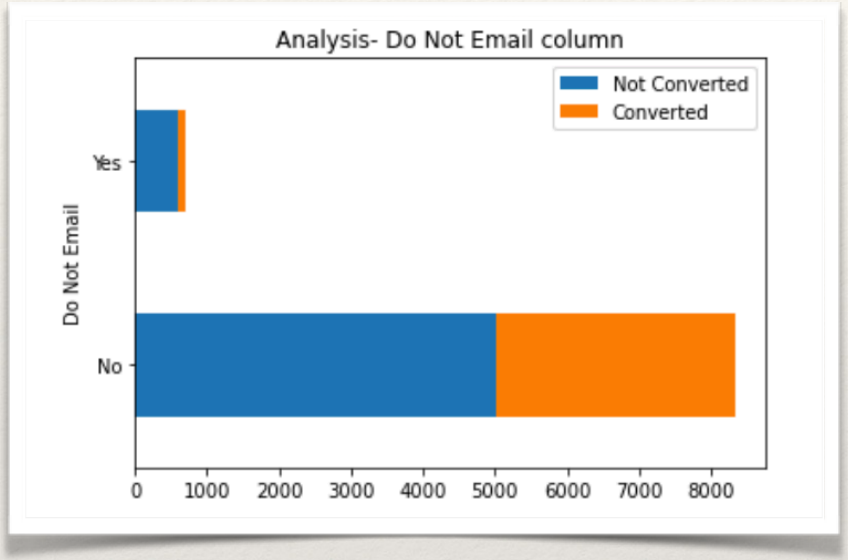


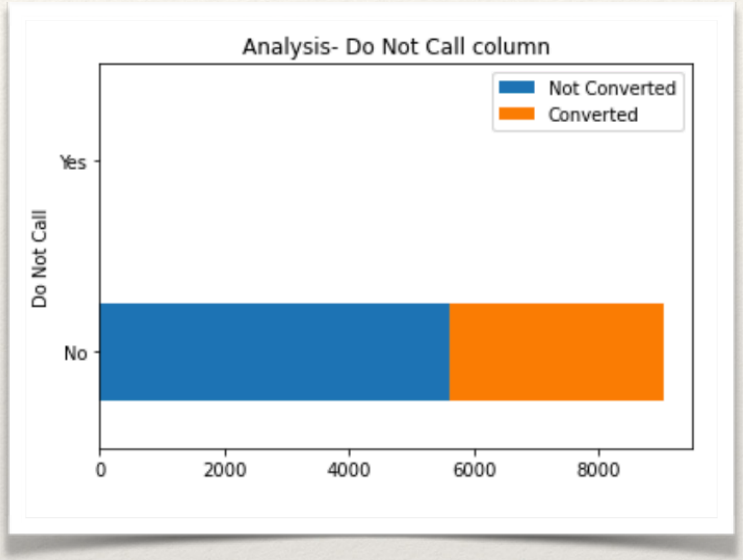
- Lead Origin - Lead Add Form has more converted than not converted
- Leads through references and Welingak website have mostly converted



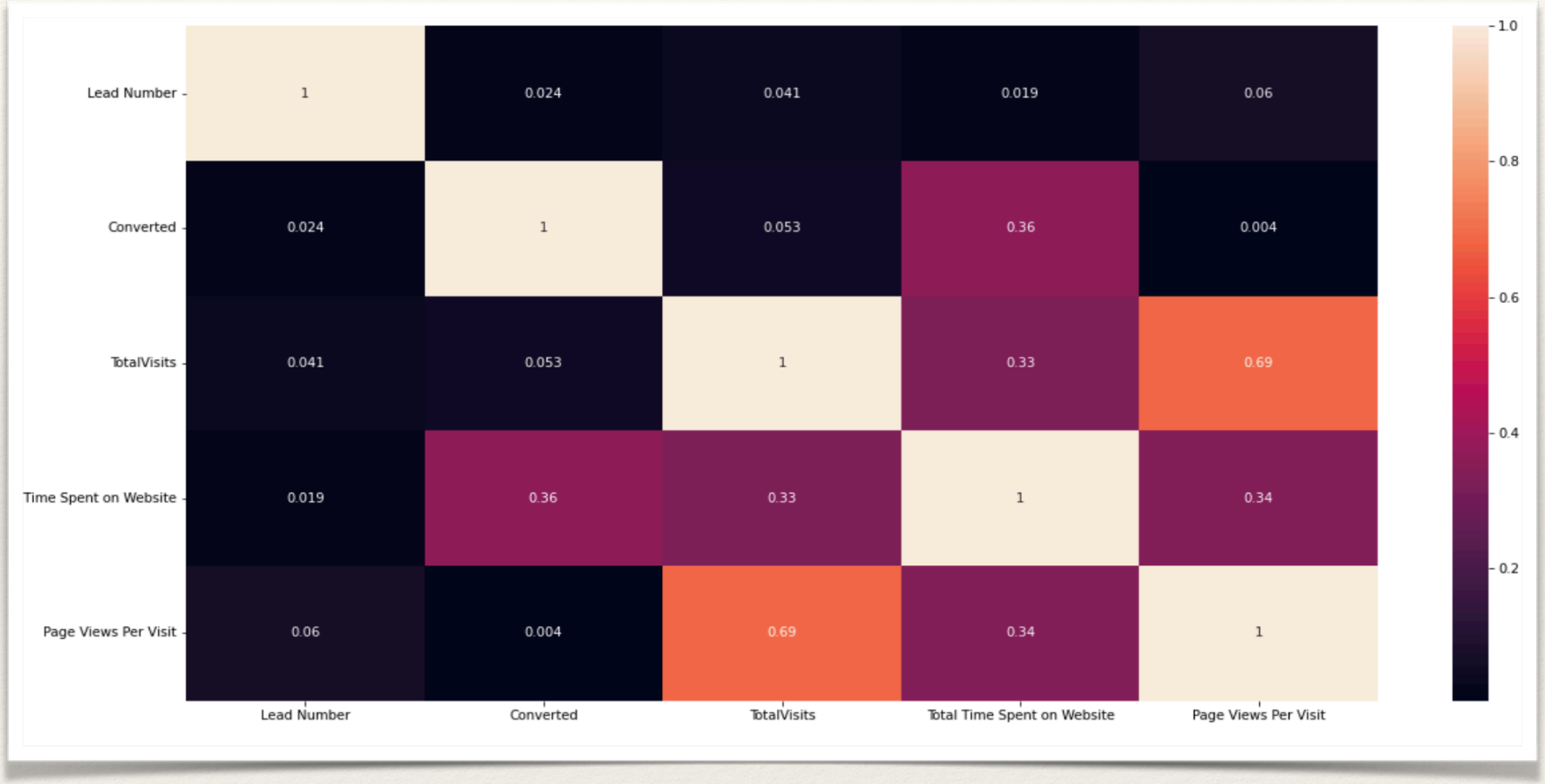


- Contrary to what it initially seems- there are good amount of conversion in the category ‘Do not Email
- Similarly, ‘Do not call’ category also has good percentage of conversion





- There is good correlation between Total Visits and Page Views per visit



Step 3: Data Preparation

Binary Variables conversion

- Prospect ID and Lead Number are both unique values to identify leads. Dropped Prospect ID since we need only one variable for identifying records in some case.
- Some columns have values- 'Yes' or 'No'
- Below columns only had one value- 'No'. Dropped these since there is no variance
  - **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**
- Below remaining Yes/No columns were converted to 1/0
  - **Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, A free copy of Mastering The Interview**

Before conversion

	Do Not Email	Do Not Call	Search	Magazine	Newspaper Article	X Education Forums	Newspaper	Digital Advertisement	Through Recommendations	More Updates About Our Courses	Update me on Supply Chain Content	Get updates on DM Content	to pay the amount through cheque	A free copy of Mastering The Interview
0	No	No	No	No	No	No	No	No	No	No	No	No	No	No
1	No	No	No	No	No	No	No	No	No	No	No	No	No	No
2	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes
3	No	No	No	No	No	No	No	No	No	No	No	No	No	No
4	No	No	No	No	No	No	No	No	No	No	No	No	No	No

After conversion

	Do Not Email	Do Not Call	Search	Newspaper Article	X Education Forums	Newspaper	Digital Advertisement	Through Recommendations	A free copy of Mastering Interv
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	

Dummy Variable Creation

- Below category variables were converted to dummy variables so that they can be used in model creation
  - **Lead Origin, Lead Source, Last Activity, What is your current occupation, Lead Quality, Lead Profile, Last Notable Activity**
- The first of the dummy variables were dropped since we need one less to explain all categories
- The resulting data set now has only numeric fields which can be used to build the model

	Lead Number	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Search	Newspaper Article	X Education Forums	...	Last Notable Activity_Form Submitted on Website	Last Notable Activity_Had a Phone Conversation	Last Notable Activity_Modified	Last Nota Activity_Ol C Conversat
0	660737	0	0	0	0.0	0	0.0	0	0	0	...	0	0	1	
1	660728	0	0	0	5.0	674	2.5	0	0	0	...	0	0	0	
2	660727	0	0	1	2.0	1532	2.0	0	0	0	...	0	0	0	
3	660719	0	0	0	1.0	305	1.0	0	0	0	...	0	0	1	
4	660681	0	0	1	2.0	1428	1.0	0	0	0	...	0	0	1	

Final Data Set

9048 records

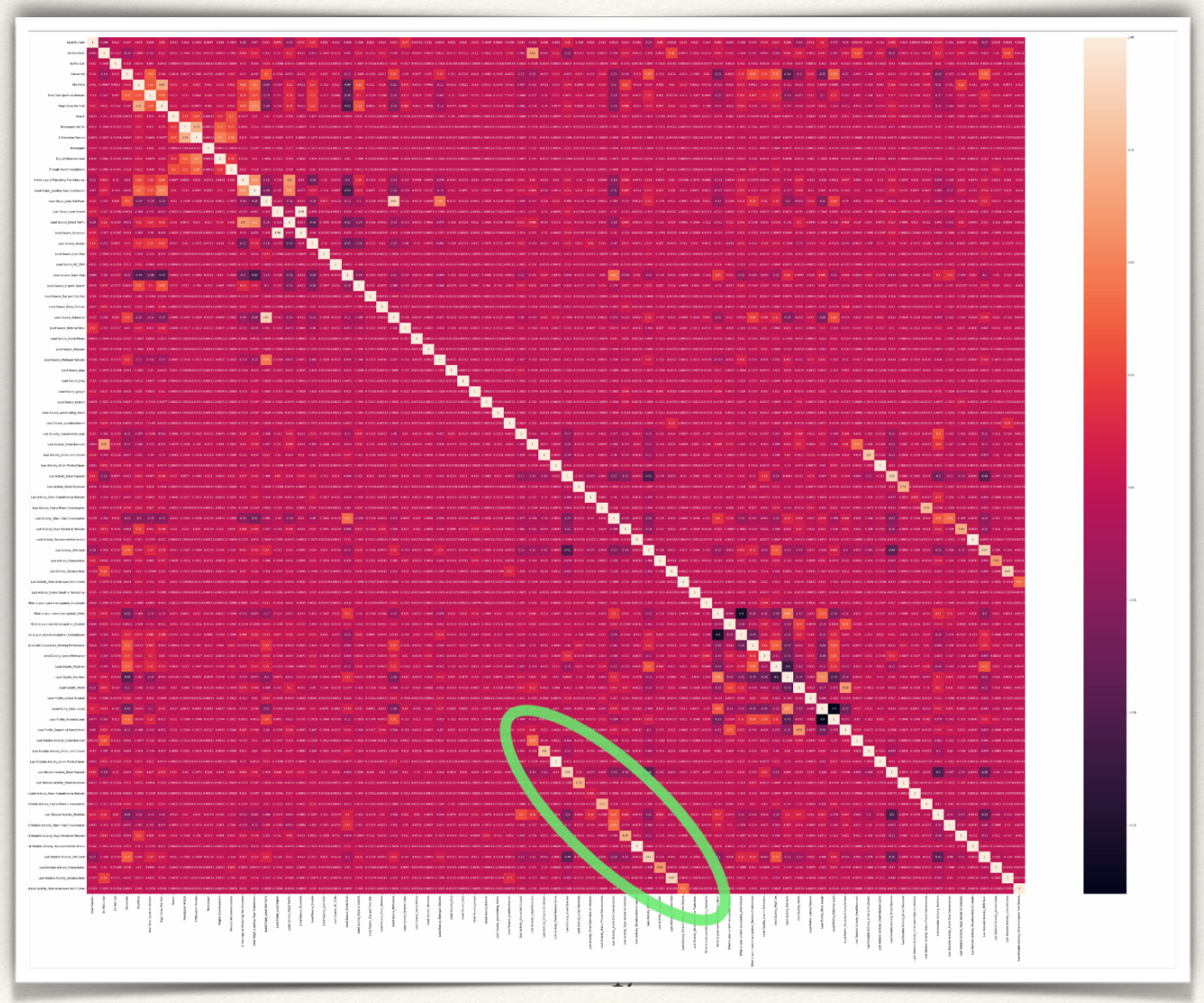
72 columns

```
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Number                             9048 non-null   int64
 1   Do Not Email                            9048 non-null   int64
 2   Do Not Call                             9048 non-null   int64
 3   Converted                               9048 non-null   int64
 4   TotalVisits                             9048 non-null   float64
 5   Total Time Spent on Website             9048 non-null   int64
 6   Page Views Per Visit                    9048 non-null   float64
 7   Search                                  9048 non-null   int64
 8   Newspaper Article                       9048 non-null   int64
 9   X Education Forums                      9048 non-null   int64
 10  Newspaper                               9048 non-null   int64
 11  Digital Advertisement                   9048 non-null   int64
 12  Through Recommendations                 9048 non-null   int64
 13  A free copy of Mastering The Interview  9048 non-null   int64
 14  Lead Origin_Landing Page Submission     9048 non-null   uint8
 15  Lead Origin_Lead Add Form               9048 non-null   uint8
 16  Lead Origin_Lead Import                 9048 non-null   uint8
```

Step 4: Looking at Correlations

Heat map of correlations of variables

- Few of the dummy variables from the set- Last Activity and Last Notable Activity are highly correlated. These were removed



Step 5: Test-Train Split

- The data set features are put into X without the converted(target) variable and Lead Number(unique id)
- The converted(target) variable is put into Y
- The data set was split into train(70%) and test(30%) set
- Train set contains 6333 records
- Test set contains 2715 records

Step 6: Feature Scaling

- The continuous numeric fields were scaled
  - **TotalVisits, Total Time Spent on Website, Page Views Per Visit**
- Standard scaling was used

Before scaling

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	6333.000000	6333.000000	6333.000000
mean	3.312174	477.066161	2.336973
std	3.263096	543.083086	2.039782
min	0.000000	0.000000	0.000000
25%	1.000000	9.000000	1.000000
50%	3.000000	244.000000	2.000000
75%	5.000000	904.000000	3.000000
max	25.000000	2253.000000	15.000000

After scaling

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	6.333000e+03	6.333000e+03	6.333000e+03
mean	-5.140019e-17	1.512203e-16	7.496153e-17
std	1.000079e+00	1.000079e+00	1.000079e+00
min	-1.015121e+00	-8.785098e-01	-1.145788e+00
25%	-7.086390e-01	-8.619365e-01	-6.555006e-01
50%	-9.567570e-02	-4.291877e-01	-1.652135e-01
75%	5.172876e-01	7.861919e-01	3.250736e-01
max	6.646921e+00	3.270354e+00	6.208519e+00

Step 7: Model Building

- Building first model using statsmodels with all 67 features
- But lot of P values were high. So we need a better way to build model

eneralized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6265
Model Family:	Binomial	**Df Model:**	67
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2075.1
Date:	Tue, 12 Jul 2022	**Deviance:**	4150.1
Time:	03:07:36	**Pearson chi2:**	6.21e+03
No. Iterations:	23		
ovariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	
const	23.7459	1.24e+05	0.000	1.000	-2.43e+05	2.4
Do Not Email	-1.0842	0.235	-4.606	0.000	-1.546	
Do Not Call	23.3160	5.58e+04	0.000	1.000	-1.09e+05	1.0
TotalVisits	0.2566	0.056	4.569	0.000	0.147	
Total Time Spent on Website	1.0532	0.046	23.092	0.000	0.964	
Page Views Per Visit	-0.2192	0.063	-3.467	0.001	-0.343	
Search	-0.6594	1.110	-0.594	0.553	-2.835	
Newspaper Article	24.7726	1.26e+05	0.000	1.000	-2.46e+05	2.4
X Education Forums	-34.8581	1.09e+05	-0.000	1.000	-2.14e+05	2.1
Newspaper	-25.9427	1.56e+05	-0.000	1.000	-3.06e+05	3.0
Digital Advertisement	-34.8581	1.09e+05	-0.000	1.000	-2.14e+05	2.1
Through Recommendations	22.1653	4.45e+04	0.000	1.000	-8.71e+04	8.7

Step 8: Feature Selection Using RFE

- Using RFE from SKLearn to eliminate features.
- Have selected 20 features from the 67 features
- Built second model using these features

Model 2
20 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6312
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2141.1
Date:	Mon, 11 Jul 2022	**Deviance:**	4282.2
Time:	21:01:03	**Pearson chi2:**	6.55e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4752	0.143	10.307	0.000	1.195	1.756
Do Not Email	-1.1073	0.194	-5.721	0.000	-1.487	-0.728
Do Not Call	23.8279	5.16e+04	0.000	1.000	-1.01e+05	1.01e+05
Total Time Spent on Website	1.0713	0.045	23.846	0.000	0.983	1.159
Lead Origin_Lead Add Form	2.6323	0.249	10.586	0.000	2.145	3.120
Lead Origin_Lead Import	1.5807	0.581	2.721	0.007	0.442	2.719
Lead Source_Olark Chat	1.5027	0.114	13.163	0.000	1.279	1.726
Lead Source_Welingak Website	4.2960	1.042	4.124	0.000	2.254	6.338
Last Activity_Had a Phone Conversation	1.5046	0.884	1.702	0.089	-0.229	3.238
Last Activity_SMS Sent	1.3446	0.084	15.937	0.000	1.179	1.510

Model 3

- Feature- **Do Not Call** had high P value
- Removed it from the list of features and built new model

Model 3
19 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6313
Model Family:	Binomial	**Df Model:**	19
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2142.7
Date:	Tue, 12 Jul 2022	**Deviance:**	4285.4
Time:	03:25:56	**Pearson chi2:**	6.55e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4730	0.143	10.293	0.000	1.193	1.753
Do Not Email	-1.1090	0.194	-5.729	0.000	-1.488	-0.730
Total Time Spent on Website	1.0729	0.045	23.885	0.000	0.985	1.161
Lead Origin_Lead Add Form	2.6479	0.248	10.664	0.000	2.161	3.135
Lead Origin_Lead Import	1.5807	0.581	2.722	0.006	0.442	2.719
Lead Source_Olark Chat	1.5024	0.114	13.163	0.000	1.279	1.726
Lead Source_Welingak Website	4.2804	1.042	4.110	0.000	2.239	6.322
Last Activity_Had a Phone Conversation	1.5026	0.884	1.699	0.089	-0.230	3.236
Last Activity_SMS Sent	1.3423	0.084	15.914	0.000	1.177	1.508
What is your current occupation_Housewife	22.7028	2.51e+04	0.001	0.999	-4.91e+04	4.91e+04
What is your current occupation_Working Professional	1.6178	0.218	7.429	0.000	1.191	2.045

Model 4

- Feature- **What is your current occupation_Housewife** had high P value
- Removed it from the list of features and built new model

Model 4
18 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6314
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2145.5
Date:	Tue, 12 Jul 2022	**Deviance:**	4291.0
Time:	03:28:12	**Pearson chi2:**	6.56e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4871	0.143	10.398	0.000	1.207	1.767
Do Not Email	-1.1125	0.194	-5.744	0.000	-1.492	-0.733
Total Time Spent on Website	1.0720	0.045	23.879	0.000	0.984	1.160
Lead Origin_Lead Add Form	2.6609	0.248	10.740	0.000	2.175	3.146
Lead Origin_Lead Import	1.5762	0.581	2.713	0.007	0.437	2.715
Lead Source_Olark Chat	1.4986	0.114	13.135	0.000	1.275	1.722
Lead Source_Welingak Website	4.2655	1.041	4.096	0.000	2.224	6.307
Last Activity_Had a Phone Conversation	1.4926	0.885	1.687	0.092	-0.242	3.227
Last Activity_SMS Sent	1.3381	0.084	15.870	0.000	1.173	1.503
What is your current occupation_Working Professional	1.6097	0.218	7.387	0.000	1.183	2.037
Lead Quality_Might be	-1.1043	0.167	-6.628	0.000	-1.431	-0.778
Lead Quality_Not Sure	-2.7951	0.156	-17.968	0.000	-3.100	-2.490
Lead Quality_Worst	-4.6212	0.431	-10.714	0.000	-5.467	-3.776
Lead Profile_Lateral Student	20.0251	1.17e+04	0.002	0.999	-2.29e+04	2.29e+04
Lead Profile_Other Leads	-0.6505	0.130	-5.002	0.000	-0.905	-0.396

Model 5

- Feature- **Lead Profile_Lateral Student** had high P value
- Removed it from the list of features and built new model

Model 5
17 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6315
Model Family:	Binomial	**Df Model:**	17
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2146.7
Date:	Tue, 12 Jul 2022	**Deviance:**	4293.4
Time:	03:31:29	**Pearson chi2:**	6.58e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.5060	0.143	10.563	0.000	1.227	1.785
Do Not Email	-1.1124	0.194	-5.746	0.000	-1.492	-0.733
Total Time Spent on Website	1.0725	0.045	23.890	0.000	0.985	1.160
Lead Origin_Lead Add Form	2.6629	0.248	10.754	0.000	2.178	3.148
Lead Origin_Lead Import	1.5746	0.582	2.707	0.007	0.434	2.715
Lead Source_Olark Chat	1.4985	0.114	13.133	0.000	1.275	1.722
Lead Source_Welingak Website	4.2660	1.041	4.097	0.000	2.225	6.307
Last Activity_Had a Phone Conversation	1.4870	0.885	1.680	0.093	-0.248	3.222
Last Activity_SMS Sent	1.3377	0.084	15.866	0.000	1.172	1.503
What is your current occupation_Working Professional	1.6076	0.218	7.377	0.000	1.180	2.035
Lead Quality_Might be	-1.1164	0.167	-6.703	0.000	-1.443	-0.790

Model 6

- Feature- **Last Activity_Had a Phone Conversation** had high P value
- Removed it from the list of features and built new model

Model 6
16 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6316
Model Family:	Binomial	**Df Model:**	16
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2148.4
Date:	Tue, 12 Jul 2022	**Deviance:**	4296.8
Time:	03:34:44	**Pearson chi2:**	6.59e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.5223	0.142	10.699	0.000	1.243	1.801
Do Not Email	-1.1145	0.194	-5.754	0.000	-1.494	-0.735
Total Time Spent on Website	1.0712	0.045	23.877	0.000	0.983	1.159
Lead Origin_Lead Add Form	2.6529	0.248	10.708	0.000	2.167	3.138
Lead Origin_Lead Import	1.5673	0.582	2.691	0.007	0.426	2.709
Lead Source_Olark Chat	1.4931	0.114	13.094	0.000	1.270	1.717
Lead Source_Welingak Website	4.2724	1.041	4.103	0.000	2.231	6.313
Last Activity_SMS Sent	1.3312	0.084	15.806	0.000	1.166	1.496
What is your current occupation_Working Professional	1.6125	0.218	7.404	0.000	1.186	2.039
Lead Quality_Might be	-1.1145	0.167	-6.693	0.000	-1.441	-0.788
Lead Quality_Not Sure	-2.8150	0.156	-18.098	0.000	-3.120	-2.510
Lead Quality_Worst	-4.6432	0.431	-10.770	0.000	-5.488	-3.798
Lead Profile_Other Leads	-0.6614	0.130	-5.091	0.000	-0.916	-0.407
Lead Profile_Student of SomeSchool	-1.5116	0.725	-2.086	0.037	-2.932	-0.092
Last Notable Activity_Modified	-1.0152	0.089	-11.399	0.000	-1.190	-0.841
Last Notable Activity_Olark Chat Conversation	-1.4278	0.360	-3.965	0.000	-2.134	-0.722
Last Notable Activity_Unreachable	1.5618	0.607	2.574	0.010	0.373	2.751

Checking VIF for Model 6

- All P values for Model 6 are below 0.05
- Checked VIF for the model to see if any features can be explained by others

VIF for Model 6

	Features	VIF
11	Lead Profile_Other Leads	7.99
9	Lead Quality_Not Sure	7.56
8	Lead Quality_Might be	2.42
10	Lead Quality_Worst	2.24
12	Lead Profile_Student of SomeSchool	1.80
6	Last Activity_SMS Sent	1.67
13	Last Notable Activity_Modified	1.66
2	Lead Origin_Lead Add Form	1.55
4	Lead Source_Olark Chat	1.55
5	Lead Source_Welingak Website	1.33
1	Total Time Spent on Website	1.30
7	What is your current occupation_Working Profes...	1.30
0	Do Not Email	1.12
14	Last Notable Activity_Olark Chat Conversation	1.09
3	Lead Origin_Lead Import	1.01
15	Last Notable Activity_Unreachable	1.01

Model 7

- Feature- **Lead Profile_Other Leads** had high VIF
- Removed it from the list of features and built new model

Model 7
15 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6317
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2161.2
Date:	Tue, 12 Jul 2022	**Deviance:**	4322.4
Time:	03:39:30	**Pearson chi2:**	6.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2899	0.132	9.768	0.000	1.031	1.549
Do Not Email	-1.1391	0.193	-5.897	0.000	-1.518	-0.760
Total Time Spent on Website	1.0684	0.045	23.923	0.000	0.981	1.156
Lead Origin_Lead Add Form	2.7424	0.248	11.081	0.000	2.257	3.228
Lead Origin_Lead Import	1.5995	0.566	2.827	0.005	0.491	2.708
Lead Source_Olark Chat	1.4748	0.114	12.985	0.000	1.252	1.697
Lead Source_Welingak Website	4.1964	1.041	4.031	0.000	2.156	6.237
Last Activity_SMS Sent	1.3355	0.084	15.926	0.000	1.171	1.500
What is your current occupation_Working Professional	1.7007	0.215	7.912	0.000	1.279	2.122
Lead Quality_Might be	-1.4179	0.155	-9.135	0.000	-1.722	-1.114
Lead Quality_Not Sure	-3.2108	0.137	-23.454	0.000	-3.479	-2.943
Lead Quality_Worst	-4.9928	0.425	-11.750	0.000	-5.826	-4.160
Lead Profile_Student of SomeSchool	-0.9619	0.724	-1.329	0.184	-2.380	0.456
Last Notable Activity_Modified	-1.0047	0.088	-11.357	0.000	-1.178	-0.831
Last Notable Activity_Olark Chat Conversation	-1.3937	0.358	-3.893	0.000	-2.095	-0.692
Last Notable Activity_Unreachable	1.5797	0.599	2.636	0.008	0.405	2.754

Model 8

- Feature- **Lead Profile_Student of SomeSchool** had high P value
- Removed it from the list of features and built new model

Model 8
14 features

Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6318
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2162.2
Date:	Tue, 12 Jul 2022	**Deviance:**	4324.4
Time:	03:44:12	**Pearson chi2:**	6.48e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2851	0.132	9.741	0.000	1.026	1.544
Do Not Email	-1.1376	0.193	-5.889	0.000	-1.516	-0.759
Total Time Spent on Website	1.0661	0.045	23.918	0.000	0.979	1.153
Lead Origin_Lead Add Form	2.7426	0.248	11.070	0.000	2.257	3.228
Lead Origin_Lead Import	1.5972	0.566	2.824	0.005	0.489	2.706
Lead Source_Olark Chat	1.4693	0.113	12.951	0.000	1.247	1.692
Lead Source_Welingak Website	4.1926	1.041	4.027	0.000	2.152	6.233
Last Activity_SMS Sent	1.3387	0.084	15.964	0.000	1.174	1.503
What is your current occupation_Working Professional	1.6971	0.215	7.889	0.000	1.275	2.119
Lead Quality_Might be	-1.4150	0.155	-9.122	0.000	-1.719	-1.111
Lead Quality_Not Sure	-3.2074	0.137	-23.452	0.000	-3.475	-2.939
Lead Quality_Worst	-5.3074	0.390	-13.601	0.000	-6.072	-4.543
Last Notable Activity_Modified	-0.9995	0.088	-11.310	0.000	-1.173	-0.826
Last Notable Activity_Olark Chat Conversation	-1.3926	0.357	-3.896	0.000	-2.093	-0.692
Last Notable Activity_Unreachable	1.5819	0.599	2.640	0.008	0.408	2.756

Checking VIF for Model 8

- All P values for Model 8 are below 0.05
- Checked VIF for the model again
- All VIF are below 5.

VIF for Model 8

	Features	VIF
9	Lead Quality_Not Sure	2.08
6	Last Activity_SMS Sent	1.67
11	Last Notable Activity_Modified	1.65
8	Lead Quality_Might be	1.63
2	Lead Origin_Lead Add Form	1.55
4	Lead Source_Olark Chat	1.54
5	Lead Source_Welingak Website	1.33
1	Total Time Spent on Website	1.29
7	What is your current occupation_Working Profes...	1.29
10	Lead Quality_Worst	1.14
0	Do Not Email	1.12
12	Last Notable Activity_Olark Chat Conversation	1.09
3	Lead Origin_Lead Import	1.01
13	Last Notable Activity_Unreachable	1.01

Creating Lead Score

- Model 8 is our final model
- Using model 8 got the probabilities for conversion
- Multiply probability by 100 and round to 2 decimal points to get the Lead score
- Lead score is a value between 0 and 100. Higher the score, higher the chance of lead conversion

	Converted	Conversion_Prob	Lead Number	Lead_Score
0	0	0.199471	8425	19.95
1	0	0.199471	8341	19.95
2	0	0.016318	7376	1.63
3	0	0.011105	7414	1.11
4	1	0.985218	3521	98.52

Predictions using Lead Score

- Using the model 8 Lead score made predictions on Train set by taking cut off score as 50 arbitrarily
- Model and predictions looks good

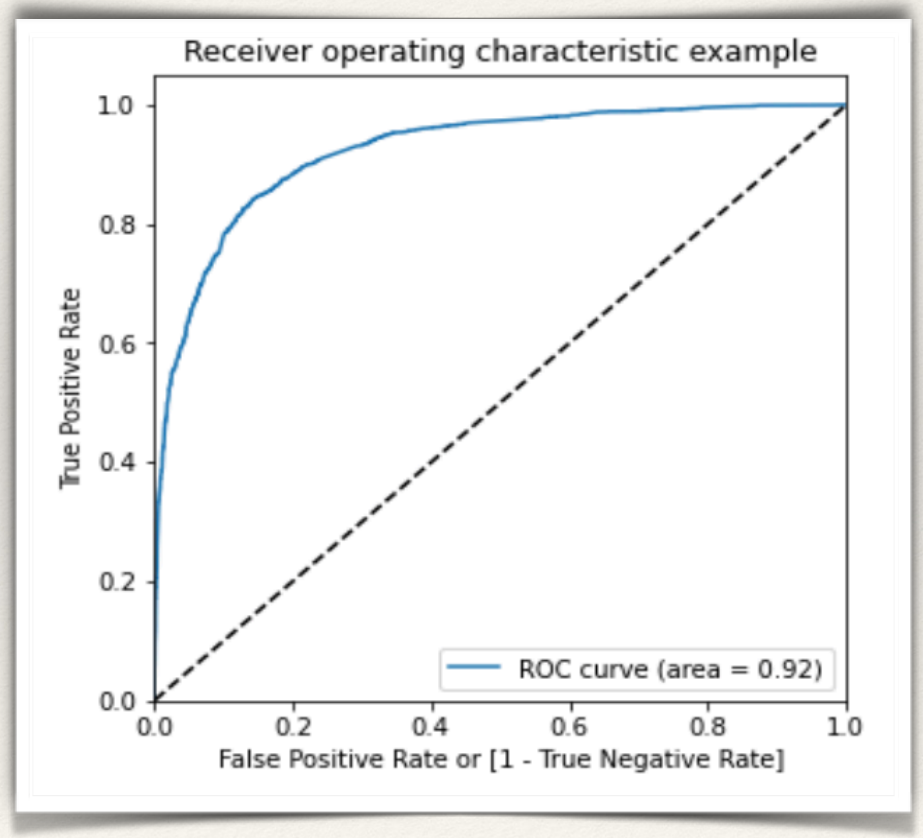
	Converted	Conversion_Prob	Lead Number	Lead_Score	Predicted
0	0	0.199471	8425	19.95	0
1	0	0.199471	8341	19.95	0
2	0	0.016318	7376	1.63	0
3	0	0.011105	7414	1.11	0
4	1	0.985218	3521	98.52	1

Step 9: Model Evaluation

- Model was evaluated on several factors. Below are the various findings
- Overall Accuracy- 84.98%
- Sensitivity- 74.94%
- Specificity- 91.09%
- False Positive Rate- 8.91%
- Positive Predictive Value- 83.64%

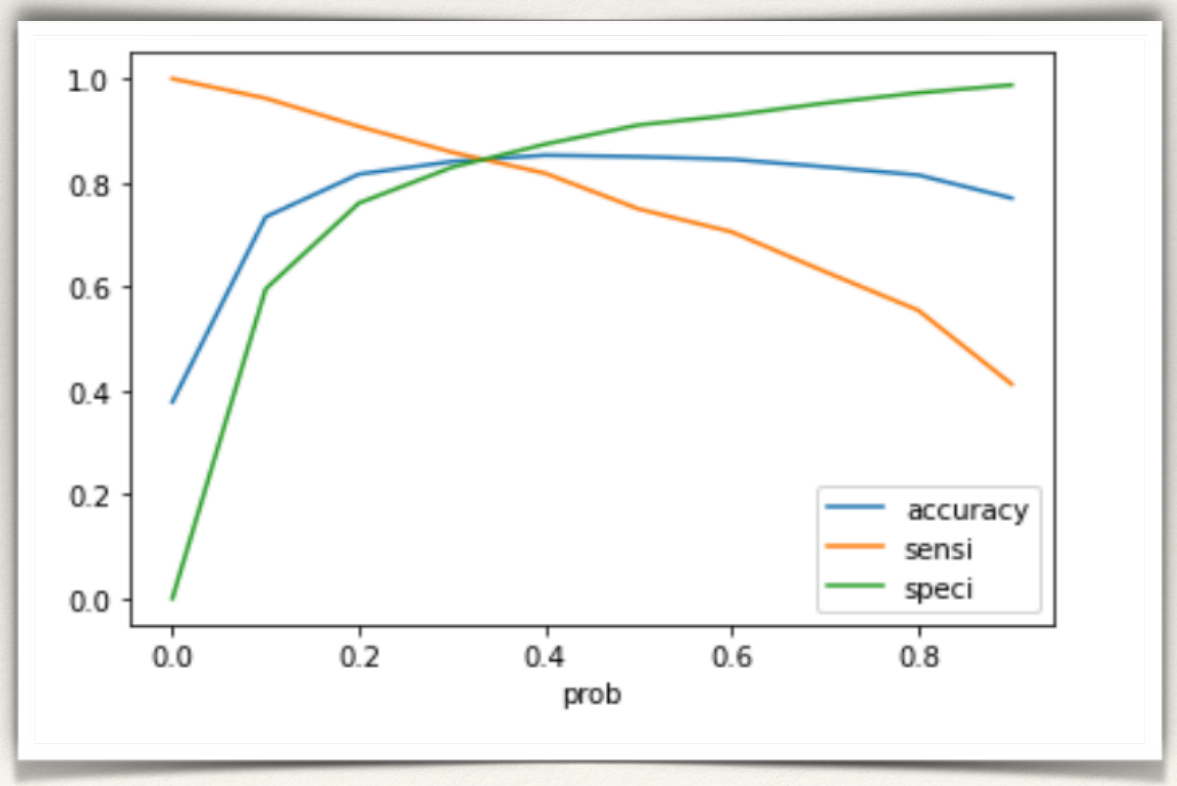
Step 10: Plotting the ROC Curve

- ROC Curve was plotted for the model
- The curve is good and has an area of 0.92



Step 11: Finding Optimal Cutoff Point

- Optimal cut off was found out by plotting graph of different evaluation metrics for different cut off values
- Accuracy, Sensitivity and Specificity was plotted for different cut off probabilities from 0 to 1
- The ideal cut off is where all 3 meet, which is near to 0.3
- Ideal Lead score cut off is 30.



Step 12: Making predictions on the test set

- Using the same model, predictions were made on the test set that we had made earlier
- The ideal cut off that we found which was 30 % was used for predictions
- Below are the evaluation metrics for the test set.
- Accuracy- 83.28%
- Sensitivity- 85.11%
- Specificity- 82.15%

- Overall Metrics look good for both train set and test set.
- We have a good final logistic regression model

Conclusion

- We have a good final model that can predict the potential leads well.
- Based on the business requirement an ideal cut off Lead score can be chosen and the predicted leads can be worked upon by staffs
- Below are the main features from the model that can positively affect the lead conversion
  - **Lead Source_Welingak Website, Lead Origin_Lead Add Form, Lead Origin_Lead Import and working professionals**
  - Leads with these features should be targeted more.
- Below are the main features that can negatively affect the lead conversion
  - **Lead Quality_Worst, Lead Quality_Not Sure, Lead Quality_Might be, Last Notable Activity_Olark Chat Conversation**
  - Leads with these features can be less targetted
  - Brief summary report with detailed findings and action points are also attached.

Thank You