Summary

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%.

The company requires a model wherein a lead score is assigned 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Steps taken to build a model:

1.) **Data Reading and Cleanup:** Columns having missing values more than 30% of the data sets and categorical data sets having low variance have been removed from the data set.

2.) **Creation of Dummy Variables:** Dummy variables are created for columns having categorical variables.

3.) **Creation of train-test dataset:** Models are bifurcated using sklearn and continuous variables are rescaled

4.) **Model creation:**

- Model is created using RFE with 15 variables as output.
- **Variables having p-value > 0.05 and VIF greater than 5 are eliminated**
- **The significant variables and their p values**

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.1870	0.552	2.148	0.032	0.104	2.270
Total Time Spent on Website	1.0930	0.046	23.647	0.000	1.002	1.184
Lead Origin_Lead Add Form	4.0395	0.256	15.792	0.000	3.538	4.541
Lead Source_Olark Chat	1.2949	0.114	11.315	0.000	1.071	1.519
Lead Source_Welingak Website	2.0525	1.037	1.979	0.048	0.019	4.086
Do Not Email_Yes	-1.4186	0.193	-7.334	0.000	-1.798	-1.039
Last Activity_Had a Phone Conversation	2.8916	0.799	3.620	0.000	1.326	4.457
Last Activity_SMS Sent	0.9914	0.084	11.760	0.000	0.826	1.157
What is your current occupation_Student	-1.8569	0.593	-3.131	0.002	-3.019	-0.695
What is your current occupation_Unemployed	-1.9833	0.554	-3.581	0.000	-3.069	-0.898
What is your current occupation_Working Professional	0.6242	0.585	1.067	0.286	-0.523	1.771
Last Notable Activity_Modified	-0.8508	0.090	-9.502	0.000	-1.026	-0.675
Last Notable Activity_Unreachable	2.4648	0.807	3.054	0.002	0.883	4.046

	Features	VIF
5	Last Activity_Had a Phone Conversation	2.45
10	Last Notable Activity_Had a Phone Conversation	2.44
8	What is your current occupation_Unemployed	2.43
6	Last Activity_SMS Sent	1.68
1	Lead Origin_Lead Add Form	1.65
11	Last Notable Activity_Modified	1.57
2	Lead Source_Olark Chat	1.37
3	Lead Source_Welingak Website	1.33
9	What is your current occupation_Working Profes...	1.31
0	Total Time Spent on Website	1.28
4	Do Not Email_Yes	1.09
7	What is your current occupation_Student	1.04
12	Last Notable Activity_Unreachable	1.01

5.) Model Predication:

- The model predicts the probability of conversion of each lead.
- We set a cutoff as 0.5 Leads having probability > 0.5 is set 1 i.e. lead is expected to convert.
- Using above cut-off point accuracy of model is 79%
- Sensitivity of our logistic regression model 0.73
- Specificity: 0.83
- False positive rate - predicting converted when customer is not converted: 0.16
- Positive predictive value: 0.80
- Negitive predictive value: 0.77
- The cutoff point is taken randomly to optimize the cutoff point we calculate accuracy sensitivity and specificity for various probability cutoffs

	prob	accuracy	sensi	speci
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.572517	0.986040	0.188149
0.2	0.2	0.689980	0.949744	0.448529
0.3	0.3	0.765523	0.898557	0.641869
0.4	0.4	0.787267	0.805491	0.770329
0.5	0.5	0.791078	0.739879	0.838668
0.6	0.6	0.773593	0.668683	0.871107

0.7	0.7	0.740417	0.550954	0.916522
0.8	0.8	0.707241	0.442066	0.953720
0.9	0.9	0.659493	0.312238	0.982266

- As we can see that at 0.4 & 0.5 probability accuracy ; specificity and sensitivity have almost same values after plotting the above we observe intersection at 0.44 and
- Using above cut-off point accuracy of model is 79.2%
- Sensitivity of our logistic regression model 0.77
- Specificity: 0.80
- False positive rate - predicting converted when customer is not converted: 0.193
- Positive predictive value: 0.78
- Negitive predictive value: 0.79