upGrad

LEAD SCORE CASE STUDY

Group Members:

1. Advaith Radhakrishnan
2. P Deb Sarma

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. 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.

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.

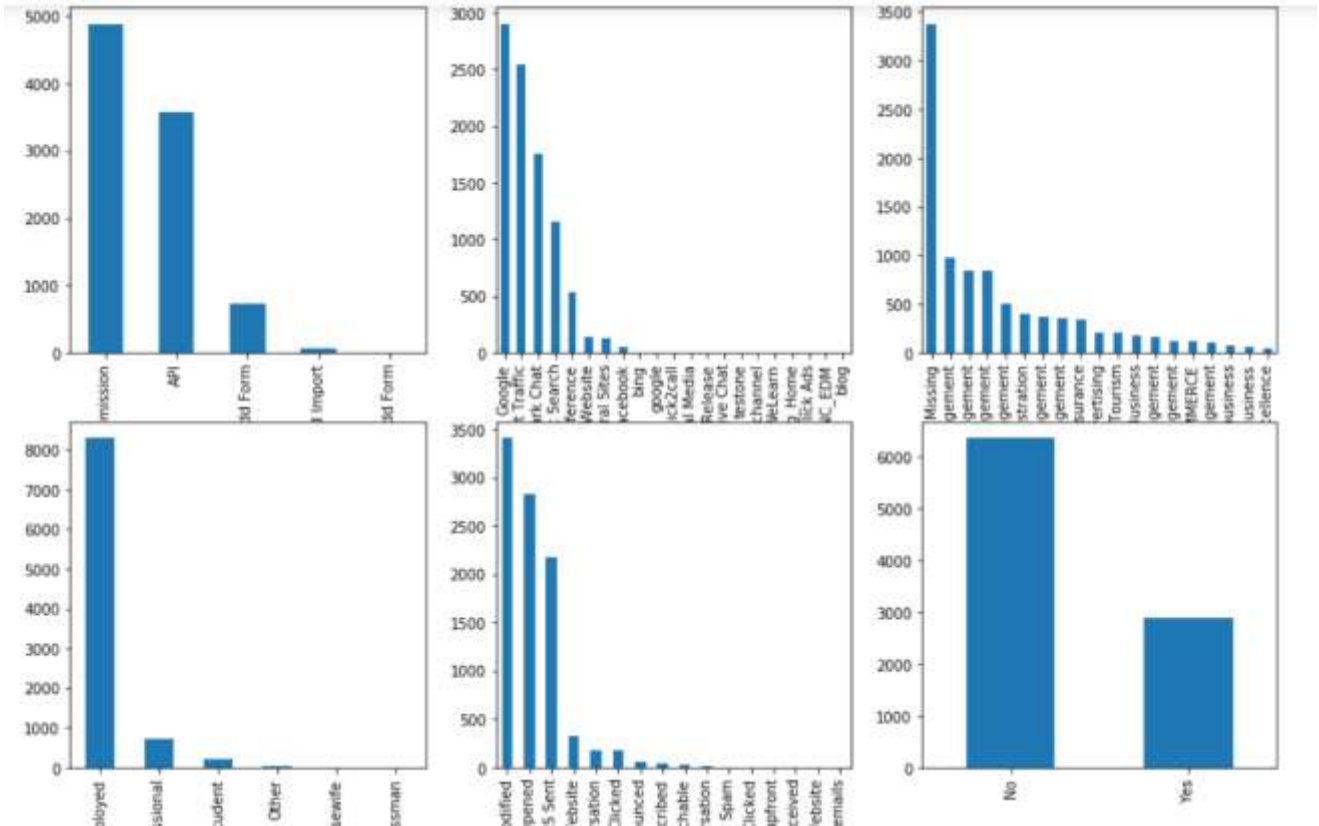
ANALYSIS APPROACH

- After importing the dataset, we first replaced the select values into NaN, as they are the as good as null values and won't give any insights into the data.
- We dropped columns having null percentage greater than 45%
- We dropped columns like country as these columns tend to have a very high percentage of only one particular value and the remaining values have very less number of rows.
- In some of the columns, there are a lot of categories with very less number of rows. Hence we club them all in one category called 'others'.
- Imputation of values into the columns that have very less missing percentage with either mean, median or mode.
- Data Visualisation (Performing EDA with various variables) to gain insight of the data and for any outliers (if any outliers present, we cap them).
- Creating dummy variables for all the categorical columns.
- Splitting the data into test and train sets.

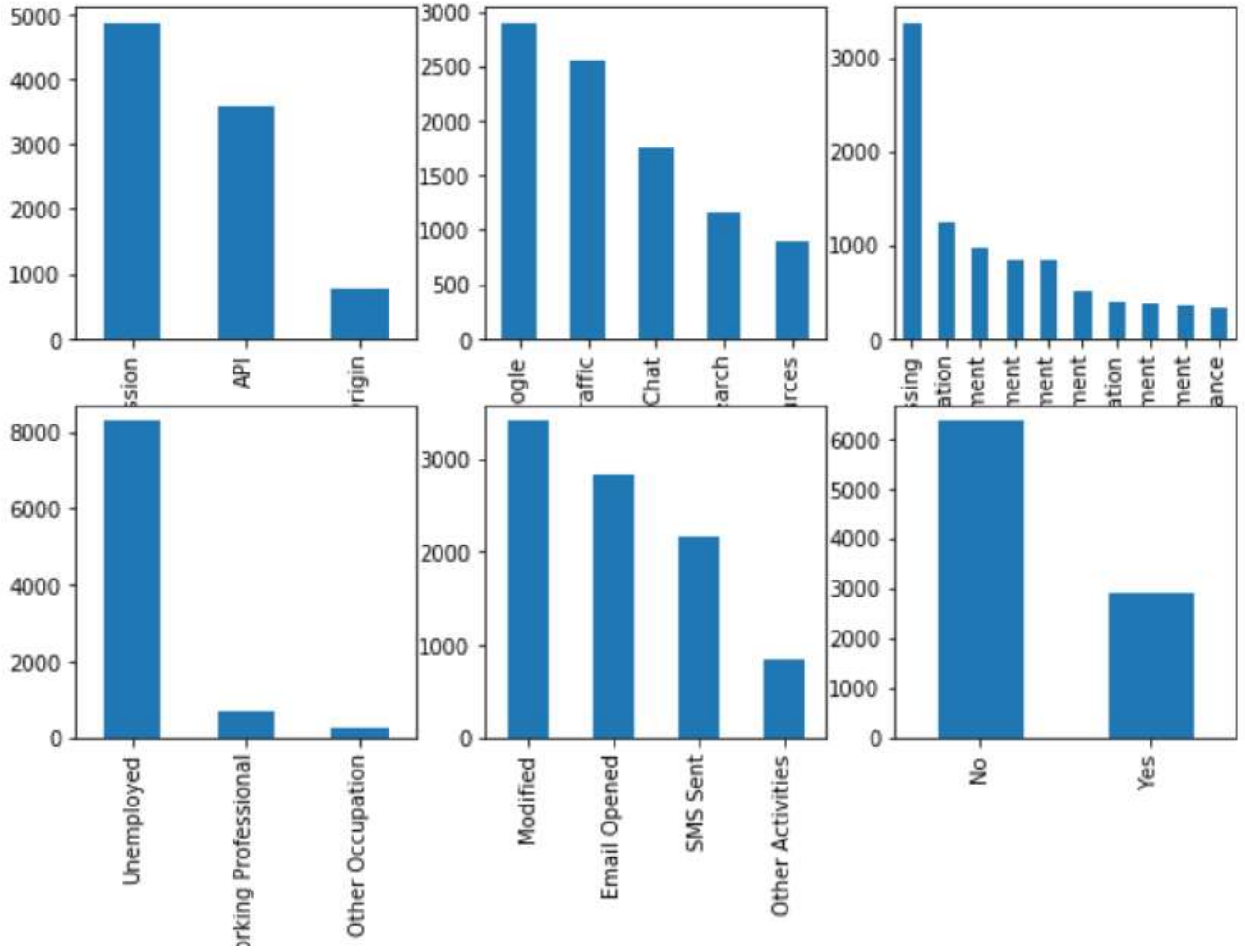
- Rescaling all variables except dummies, using MinMaxScalar so that the numerical values are similar in terms of magnitudes, units, and range.
- Building the logistic regression model.
- Select the top 20 features using RFE.
- Then, using manual approach, that is VIF scores and the p-values, we remove features one by one until VIF score is less than 0.5 and p-value is between 0 to 0.05. The range can vary according to the business aspects.
- We then find the optimal probability cut-off. For this, we calculate accuracy, sensitivity and specificity for various probability cut-offs. And we plot them in order to find the probability cut-off. The point at which all the 3 curves meet is the optimal cut-off, in our case, being 0.35.
- We find out the final prediction value, according to the optimal probability cut-off.
- We create new column 'Lead Score' with where lead Score= Converted_prob*100. Hence we find out what the lead score for each lead will be.

EXPLORATORY DATA ANALYSIS

After removing all the unnecessary columns, we can see that a lot of these columns have categories with less values in them, as shown in the figure below

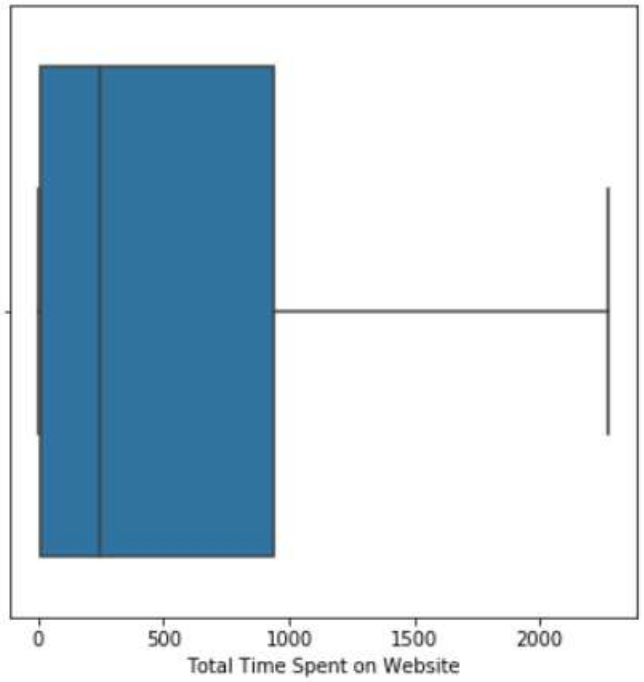


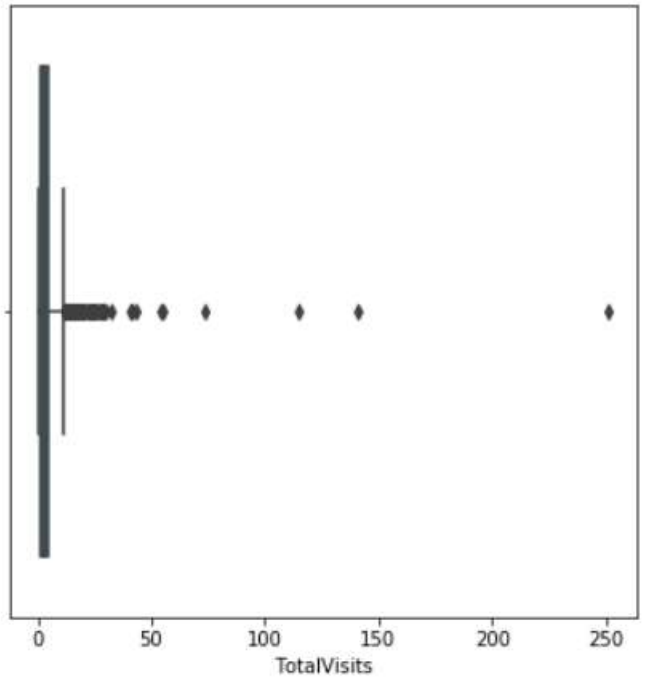
All this is clubbed together, to form a different category called others in each of the columns.

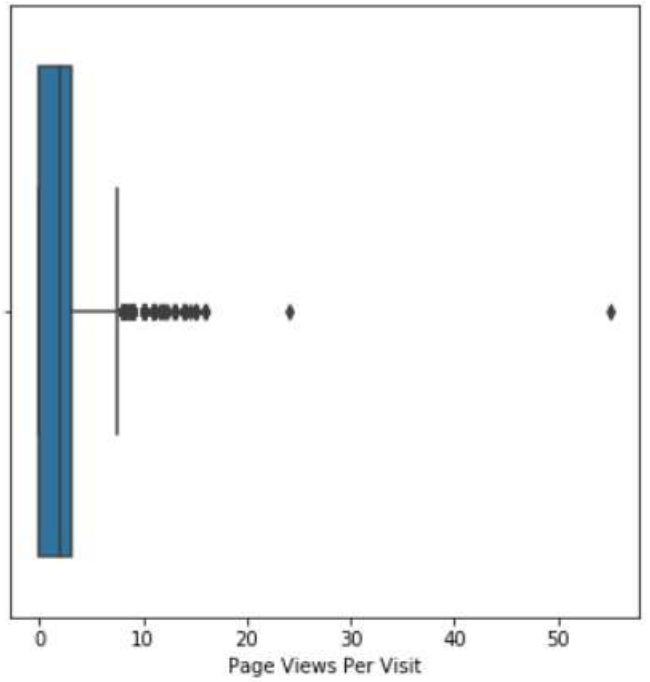


OUTLIER ANALYSIS

2 of the numerical columns contain outliers, as shown below, hence we cap them to their highest percentile







FINAL MODEL ANALYSIS

The picture in the slide shows the final model details.

According to the model, the top 3 features that contributes the most towards the probability of the lead getting converted to a customer are:

1. Total Time Spent on Website
2. Other Origin
3. Working Professional

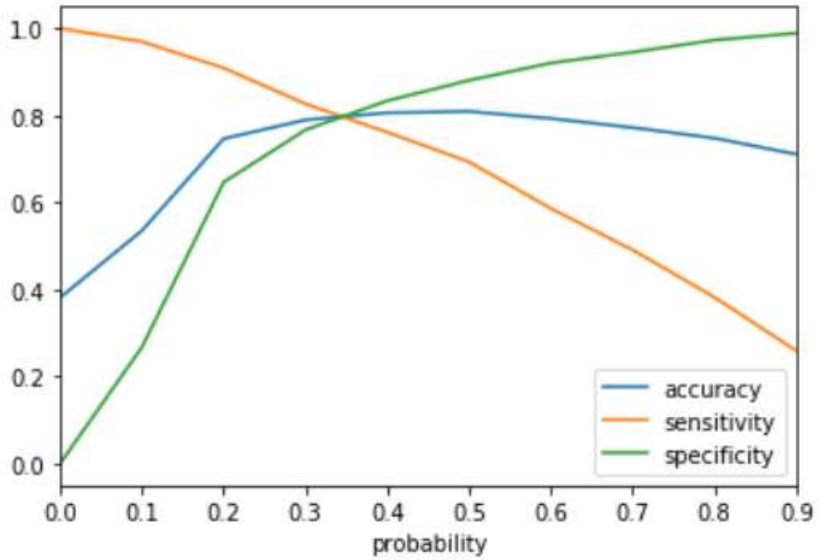
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6457
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2807.5
Date:	Mon, 07 Sep 2020	**Deviance:**	5614.9
Time:	17:27:04	**Pearson chi2:**	7.01e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5046	0.099	-25.256	0.000	-2.699	-2.310
TotalVisits	0.9771	0.205	4.776	0.000	0.576	1.378
Total Time Spent on Website	4.5268	0.162	27.943	0.000	4.209	4.844
Finance Management	0.2851	0.110	2.602	0.009	0.070	0.500
Marketing Management	0.2092	0.115	1.819	0.069	-0.016	0.435
Working Professional	2.7985	0.186	15.067	0.000	2.434	3.163
Other Origin	3.8160	0.160	23.872	0.000	3.503	4.129
Olark Chat	1.2519	0.111	11.267	0.000	1.034	1.470
Modified	-0.8060	0.084	-9.548	0.000	-0.972	-0.641
Other Activities	-0.5110	0.129	-3.964	0.000	-0.764	-0.258
SMS Sent	1.2847	0.086	14.921	0.000	1.116	1.453

RESULTS(OPTIMAL PROBABILITY CUT-OFF)

Plotting the accuracy, sensitivity and specificity for various probabilities, we get the cut-off to be 0.33 to 0.35.

This means that the probability of a lead converting to a customer is 0.35 or greater.



RESULTS (LEAD SCORE)

We create a new column 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.

That is, for example. Lead Number 8105 with a lead score of 82 has a higher chance of converting into a customer than Lead Number 1871, whos respective lead score is 26.

Converted	Converted_Prob	Lead Number	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead_Score
0	0.260483	1871	0	1	1	1	0	0	0	0	0	0	0	0	26.048267
0	0.213436	6795	0	1	1	1	0	0	0	0	0	0	0	0	21.343564
0	0.319942	3516	0	1	1	1	1	0	0	0	0	0	0	0	31.994165
0	0.820269	8105	1	1	1	1	1	1	1	1	1	1	0	1	82.026891
0	0.113179	3934	0	1	1	0	0	0	0	0	0	0	0	0	11.317863