Summary

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. 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 Objectives:

Our Objective is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Then suggest this to CEO so that these leads could be converted.

Approach:

- Data Understanding (Understanding Problem statement, Columns provided , data dictionary, Variable identification, uniqueness of variables.)
- Data Cleaning/Standardizing (Dropping unnecessary variables or variables with imbalance, Renaming columns, duplicates handling, Imputing missing values.
- Creating dummy variables and Outlier treatments
- Model creation -- Train Test split, feature scaling, Eliminating multi co-linearity on highly correlated data making feature selection using RFE.
- Model building – On train set building model and scaling it to the finest using P values and VIFs. Getting the predicated values, plotting ROC , metrics finding
- Making predictions on Test set – evaluating both train and test set, finding metrics.

Data Understanding and Cleaning:

We read the data from csv file and gone through the data dictionary for understanding the data.

We checked for conversion rate in the given data set which is around 40%.

Identified categorical and numerical variables and checked for their unique value counts. This helped us understand the variance present inside them.

Get updates on DM Content	1
I agree to pay the amount through cheque	1
Receive More Updates About Our Courses	1
Magazine	1
Update me on Supply Chain Content	1
Through Recommendations	2
Digital Advertisement	2
Newspaper	2
X Education Forums	2
A free copy of Mastering The Interview	2
Search	2
Newspaper Article	2
Converted	2
Do Not Call	2
Do Not Email	2
What matters most to you in choosing a course	4
Asymmetrique Activity Index	4
Asymmetrique Profile Index	4
Lead Origin	5
Lead Quality	6
What is your current occupation	7
Lead Profile	7
City	8
How did you hear about X Education	11
Asymmetrique Profile Score	11
Asymmetrique Activity Score	13
Last Notable Activity	16
Last Activity	18
Specialization	20

Lead Source	22
Tags	27
Country	39
TotalVisits	42
Page Views Per Visit	115
Total Time Spent on Website	1731
Lead Number	9240
Prospect ID	9240

Dropping Columns with Single/All Different Values:

'Prospect ID', and 'Lead Number' are columns which contains unique values for each record. They are like primary keys to identify each record. We can drop them as they won't serve any purpose.

Below columns have "unique value count" of 1. They don't have any variance associated with them. Hence we can drop them for our analysis.

1. Get updates on DM Content
2. I agree to pay the amount through cheque
3. Receive More Updates About Our Courses
4. Magazine
5. Update me on Supply Chain Content

Dropping Columns with Very High Imbalance:

Below columns have a very high number of a single value compared to the other hence not useful for the analysis.

- Through Recommendations
- Digital Advertisement
- Newspaper
- X Education Forums
- Search
- Newspaper Article
- What matters most to you in choosing a course

- Do Not Call

Dropping Insignificant Columns:

If we observe the values of city column, most of them are from India which makes the column **Country** not useful for the analysis.

Standardizing Values:

Converting all text to lower helps in identifying duplicate values due to case. There are 1282 duplicate rows in the data set. We can drop them.

Treating Missing Values:

Dropping columns whose missing value % is >= 30% as they won't help in the analysis and imputing them would only add more bias.

Lead_Quality	46.56
Profile_Score	44.16
Activity_Score	44.16
Profile_Index	44.16
Activity_Index	44.16
Tags	30.18
Lead_Profile	24.75
Occupation	24.52
Hear	18.45
Specialization	8.78
City	8.57
Page_Views	1.72
Total_Visits	1.72
Last_Activity	1.29
Lead_Source	0.41
Converted	0
No_Email	0
Last_Notable_Activity	0
Time_On_Website	0
Free_Copy	0
Lead_Origin	0

Many of the categorical variables have a level called **'Select'** which needs to be handled because it is as good as a null value.

We are going to apply a general strategy here to impute nulls, which is to impute them with the 'select' value which is equivalent to null. At a later stage, dummy variables are created and the select dummy variable will be dropped.

This will take care of nulls and as well as preserves the information.

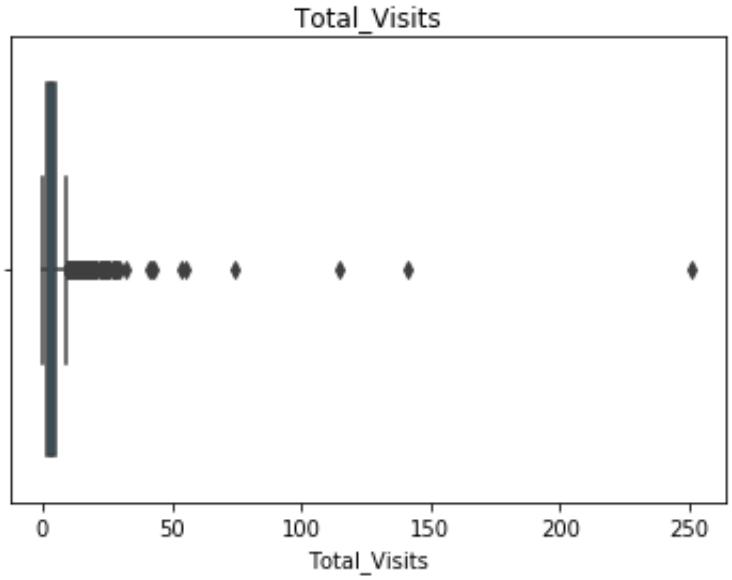
Same is the case with other columns (Occupation, Last_Activity, Lead_Source) where the nulls are imputed with 'unspecified' and that dummy variable is dropped later.

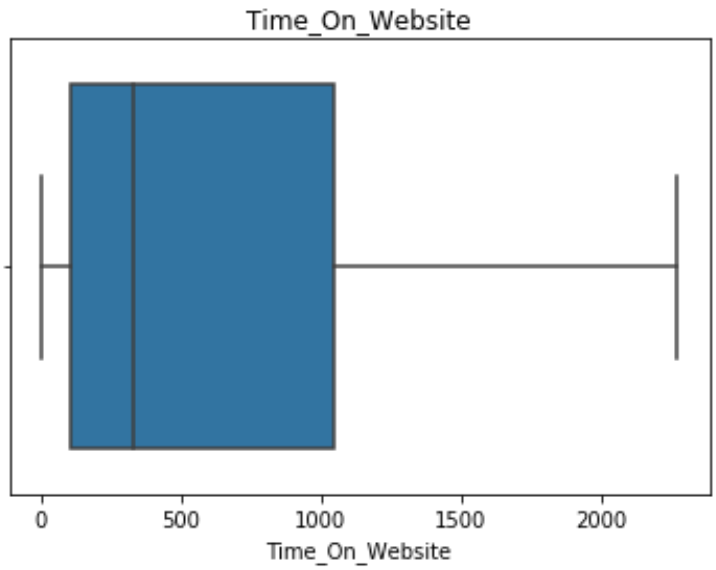
Creating Dummy variables:

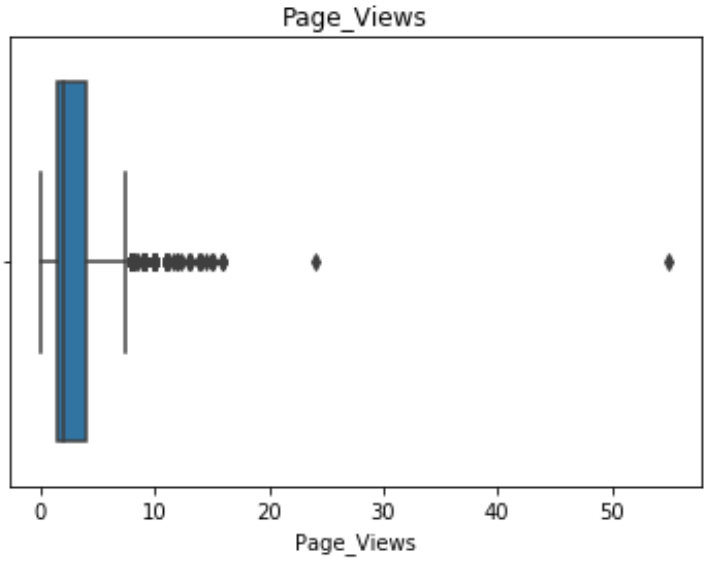
Going with plain old dummy variable creation method and creating dummy variables for categorical variables and dropping the first one. We finally are left with 106 columns to deal with for our analysis.

Check for Outliers in the numeric columns:

We performed univariate analysis and “boxplotted” numeric variables such as **‘**Total_Visits', 'Time_On_Website', and 'Page_Views'.





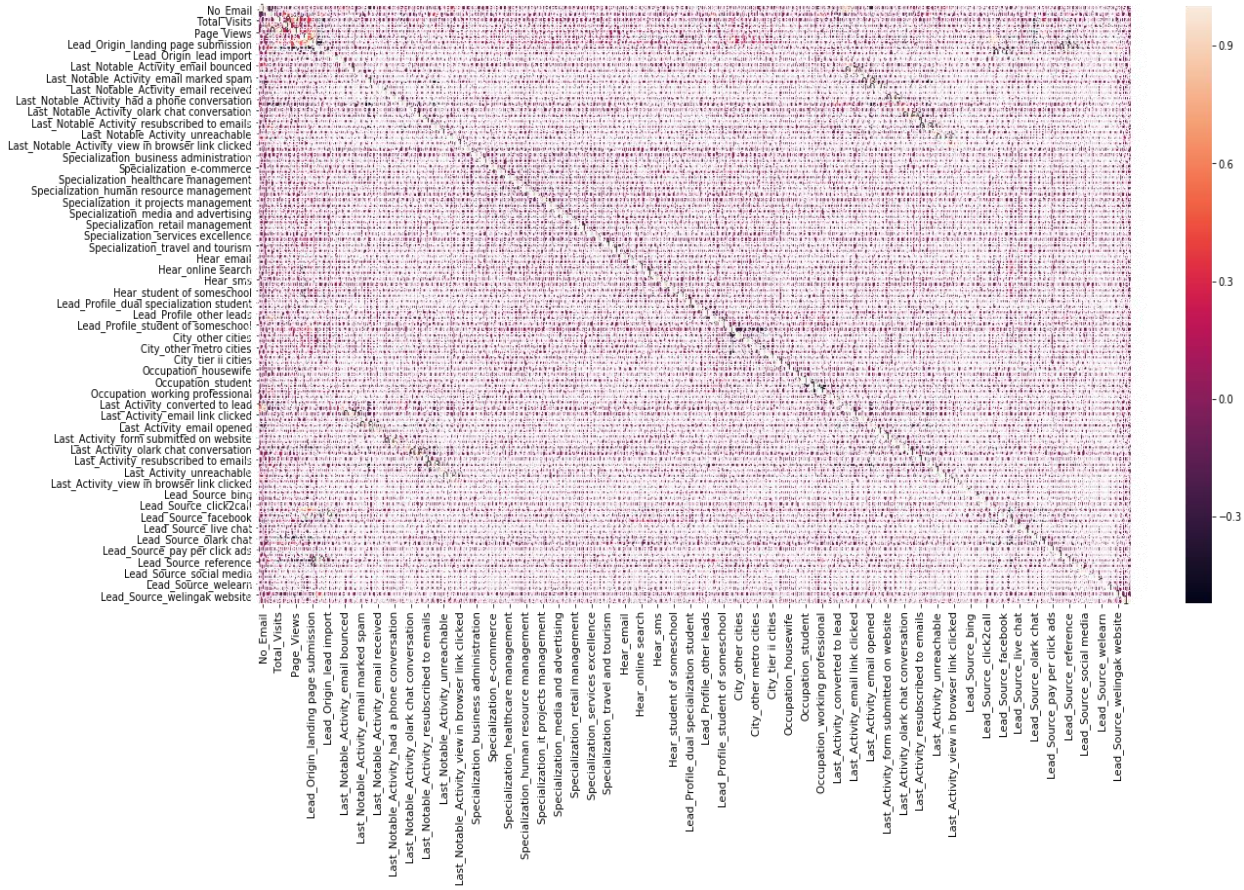


Columns 'Total_Visits' and 'Page_Views' have outliers in their data and will be dealt with the common capping method where we assign the 'Q3+1.5*IQR' value to the values greater than that.

Model Creation:

- **Test-Train Split**: Split the data into train (70%) and test (30%) using sklearn library
- **Feature Scaling**: Using “StandardScaler”, standardize all the dummy variables.

- **Checked for “Multi Collinearity” using heat maps**: Since it is not clear with the heat map, we will let RFE deal with dropping the variables with high collinearity and subsequently using manual elimination based on VIF and p-values.

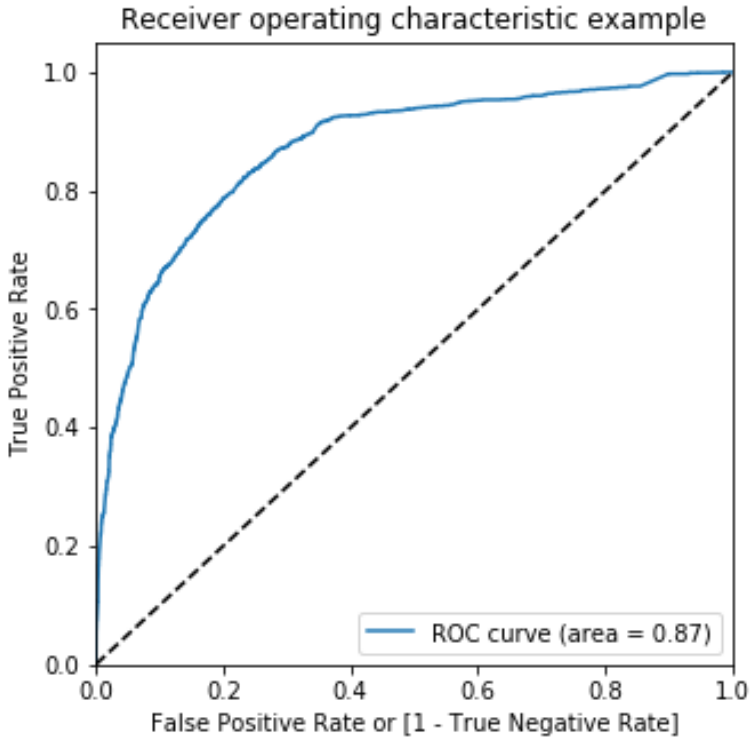


- **Feature Selection Using RFE:** Build the regression model using RFE with 15 variables and successively keep dropping variables using VIF and P-values until all p-value are < 0.05. Keep checking the model accuracy after dropping any variable and rebuilding the model.
  We are left with below features which are used by our model to predict the output.

Features	VIF
Lead_Profile_potential lead	1.3
Lead_Origin_lead add form	1.2
Occupation_working professional	1.2
Last_Activity_sms sent	1.2
Time_On_Website	1.1
No_Email	1
Lead_Profile_student of someschool	1

Plotting the ROC Curve:

To check the performance of our model, we will plot a ROC curve.

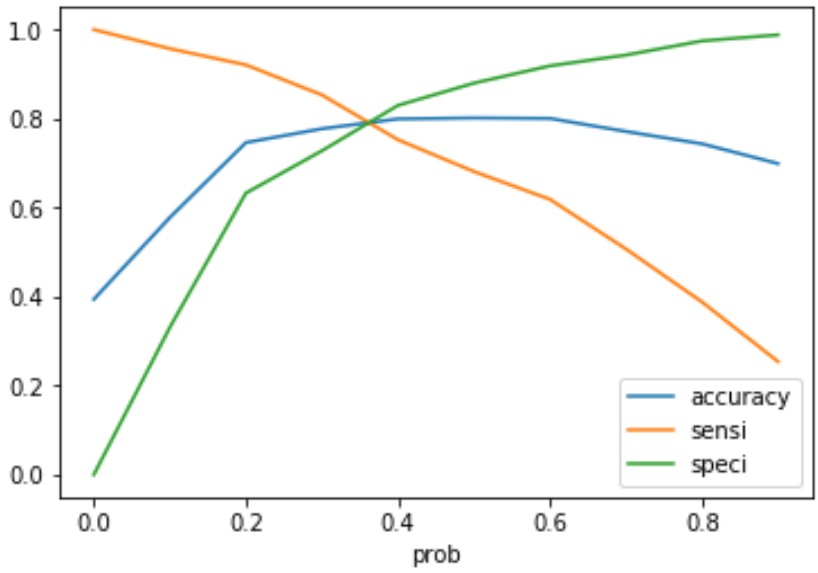


The AUC is 0.87 which is decent and the curve is not close to the diagonal.

Finding Optimal Cutoff Point:

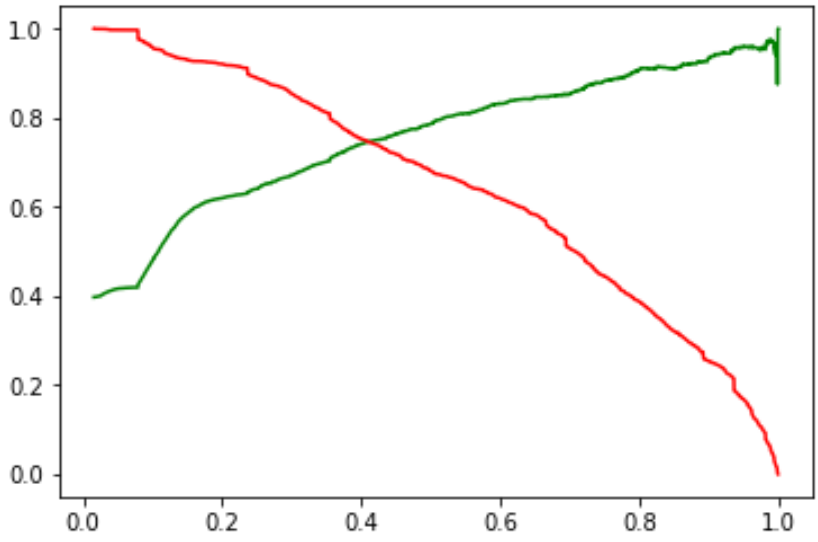
Now we have to decide a probability threshold value. If the probability is greater than the threshold value, we will predict the lead as “converted” otherwise, we will predict the lead as “not converted”.

Let's plot accuracy sensitivity and specificity for various probabilities.



From the plot above, 0.35 is the optimum point to take it as a cutoff probability.

The false positive rate is high with cut-off of 0.35. To further improve the model, we will also plot a “Precision” and “Recall” Tradeoff graph.



From the curve above, 0.42 is the optimum point to take it as a cutoff probability. This leads to good performance measurements of our model.

Accuracy - ~80%

Sensitivity - 74%

Specificity – 83%

false positive rate – 16%

We can tweak the cut-off value to meet the company's requirement changes in the future. Suppose, if the company wants to make lead conversion more aggressive then the cut-off threshold for conversion probability need to be brought down so that the **Sensitivity** (True Positive Rate) value increases thereby providing a higher number of Hot Leads to be followed up on.

Making predictions on the test set and populating lead_score column:

Here's the predicted values of test data in "final_predicted" column which is matching with "converted" column. Lead score (lies between 1 to 100) can be calculated easily using the probability obtained from the model. Higher the value of "lead_score", more are the chances of lead getting converted.

CustID	Converted	Converted_Prob	final_predicted	lead_score
3833	0	0.339469	0	33.946856
2508	1	0.70686	1	70.685997
5291	1	0.735638	1	73.563838
3892	0	0.014144	0	1.414365
3254	0	0.113834	0	11.383396

Conclusions:

We conclude by this analysis that:

1. **Lead Origin, Lead Profile & Occupation** are the top the 3 variables that contribute most towards the probability of a lead getting converted as these featured in all the 3 models built and have high positive beta co-efficient values

3) Below are the variables of the model we chose to go with along with their beta co-efficient values:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4914	0.052	-28.543	0.000	-1.594	-1.389
No_Email	-1.1550	0.157	-7.374	0.000	-1.462	-0.848
Time_On_Website	0.9408	0.037	25.277	0.000	0.868	1.014
Lead_Origin_lead add form	3.2830	0.205	16.053	0.000	2.882	3.684
Lead_Profile_potential lead	1.8635	0.098	19.075	0.000	1.672	2.055
Lead_Profile_student of someschool	-1.7822	0.434	-4.109	0.000	-2.632	-0.932
Occupation_working professional	2.3915	0.189	12.631	0.000	2.020	2.763
Last_Activity_sms sent	1.2874	0.076	16.903	0.000	1.138	1.437

The above variables make total business sense due to the below reasoning:

- Lead Profile “Potential lead” as assigned by the person allotted to a particular lead itself says we should consider him/her as hot lead.
- Occupation “Working professional” since the working professionals are the ones that look out for online courses than the students due to time constraints.