LEAD SCORING

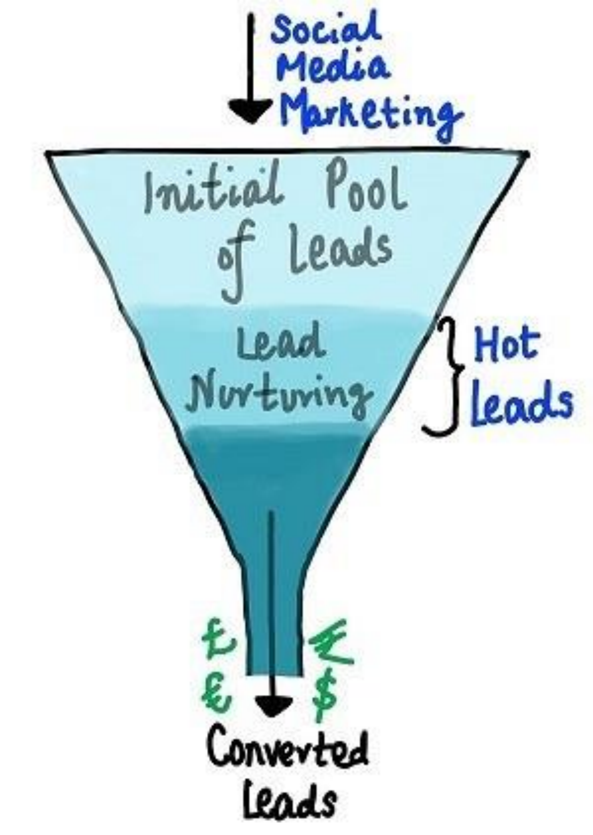
Satyavrat

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. A typical lead conversion process can be represented using the following funnel:



PROBLEM STATEMENT

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.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.The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data

You have been provided with a leads dataset from the past with around 9000 data points.This dataset consists of various attributes such as Lead Source,Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

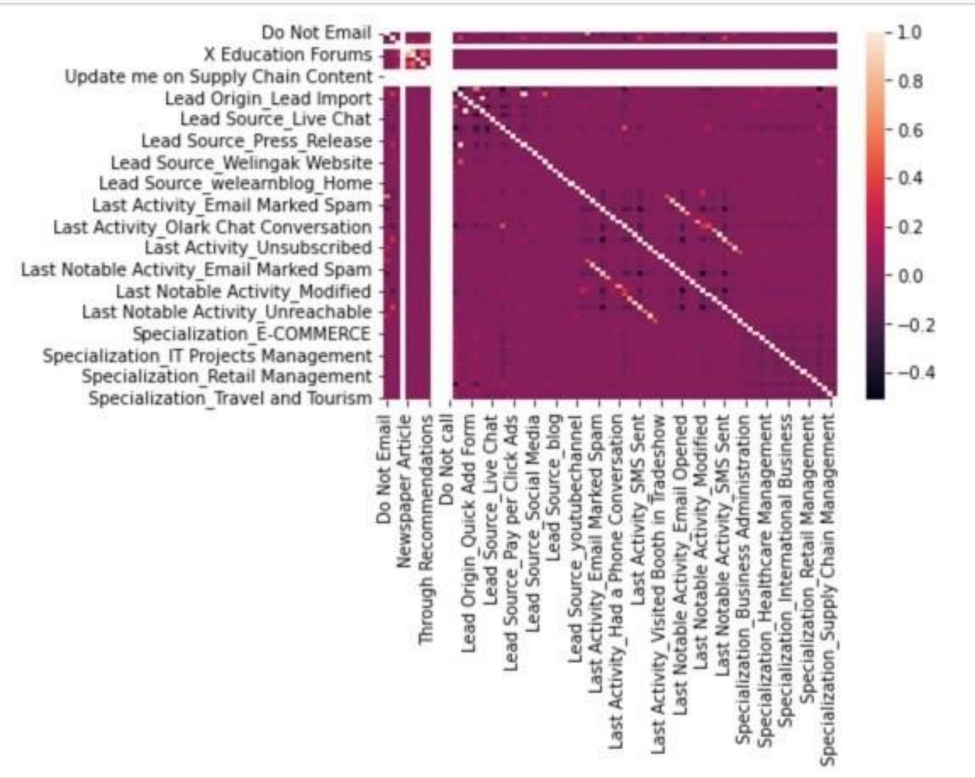
Analysis Approach

In order to analyse the data, we will first understand the basic structure of the data, handle the missing values, imbalance and correct the erroneous records and then finally we will start our analysis by doing the following:

- Exploratory data analysis
- Checking outliers
- Handling missing values
- Removing insignificant columns from data set

Correlation in the data

After performing the basic EDA operations, we will now proceed further and try to look at the correlations in the data. For this purpose, we will split data in to test and train data set, standardize the data and then with the help of a heatmap we will find out the correlation among different variables in the data set.

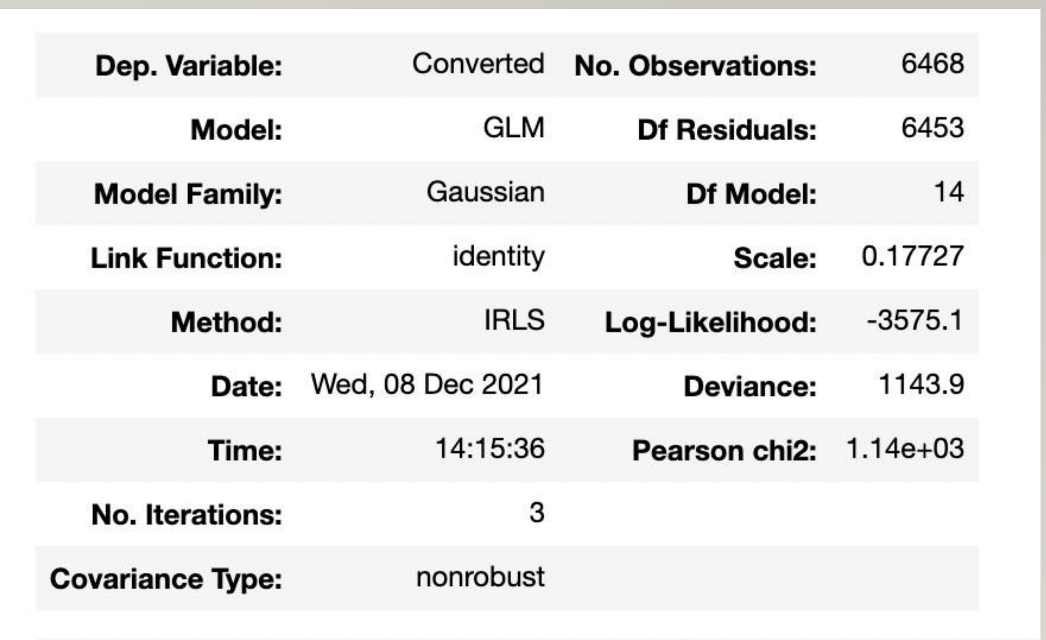


Building the Model

Once we build the model using logistic regression library from sklearn library, we found out that there are still insignificant columns which have xero to very minimum impact on the final output. Hence, removed those columns. We chose RFE method to remove those columns.

Once these insignificant columns deleted, we are left with a model that only have significant columns and low VIF values.

After finishing first two steps, we now evaluated our model.

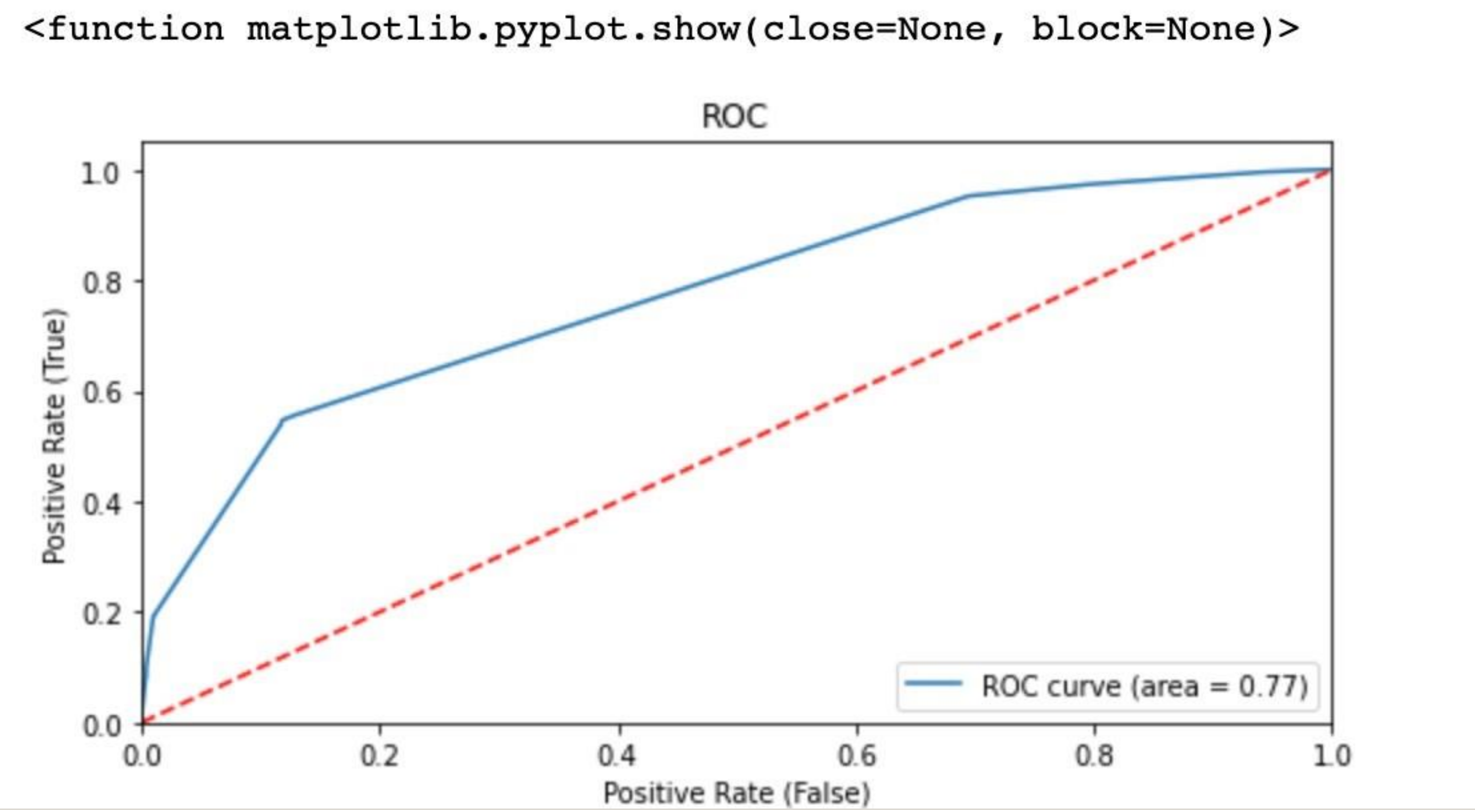


Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6453
Model Family:	Gaussian	**Df Model:**	14
Link Function:	identity	**Scale:**	0.17727
Method:	IRLS	**Log-Likelihood:**	-3575.1
Date:	Wed, 08 Dec 2021	**Deviance:**	1143.9
Time:	14:15:36	**Pearson chi2:**	1.14e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

Evaluating the Model

After making the final model, we then created the ROC on it to find the model stability. As depicted in the graph the area covered under the graph is 77%

Since our graph is towards the left-hand side of the red line which means that we have good accuracy.



```
<function matplotlib.pyplot.show(close=None, block=None)>
```

Finding the prediction on test set

There are various steps that we need to perform first.We will first standardize the test set then we will start predicting.

The accuracy score was found 75.06%

Sensitivity score was found 55.71 and Specificity was found 86.98.This indicates that the model we have built has high accuracy and good score.

This also shows that our model is stable.

Conclusion

There are 3 columns that are very useful:

- Last notable activity
- Lead origin and,
- Current occupation

The model we have built has high accuracy and good score.This also shows that our model is stable.

Thanks

Submitted by: Satyavrat