CASE STUDY: CDAC PROJECT

X EDUCATION COMPANY MARKET ANALYSIS

INTRODUCTION:

This is the part of CDAC ML Case study , to reach the conclusion and answer key to the business question we will go through following steps :-

1. *ASK – Problem Statement*
2. *Data Preparation*
3. *Data Analysis*
4. *Train Model*
5. *Test Model*
6. *Answering Problem statement*

SCENARIO:

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.

About The Company:

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.

Objective:

The objective for this analysis is to draw insights from the available data and answer key business question.

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%.

Steps:

1. Ask:

There are quite a few goals for this case study.

1. 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.*

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

2. *Data Preparation:*

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.

o ***EDA***

i. *For this we have loaded the dataset first on google Collab*

ii. *Then analysed the columns and rows in dataset*

iii. *To check the data integrity, we have to check the the consistency in column and data type.*

Feature Selection

- *Checked the null values in the columns, on the basis of null values percentage we have dropped sum columns.*
- *Analysed the categorical feature first, Plotted the count graph for the categorical features, on the basis of biasness in the data we have taken a call to drop that columns.*
- *Imputed the missing values with mode in the categorical feature.*
- *Analysed the Numerical features, Imputed the missing values with median , on the basis of skewness and kutosis imputation taken place.*

- *Handled the outliers, replaced the outliers with the median, some outliers have been dropped.*
- *After these steps our data is clean, now it is ready for further steps.*

3. *Data Analysis*

- Plotted the co-relation matrix for the features.

4. *Model Train*

- *Created the dummies for the categorical features.*
- *Created the logistic regression model*

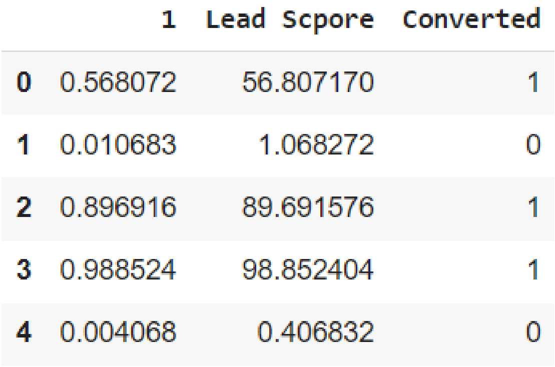
5. *Model Test*

- *Tested the model on test dataset.*

- *Then checked for the accuracy by plotting confusion matrix*

6. *Answering Problem statement*

1. **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.**



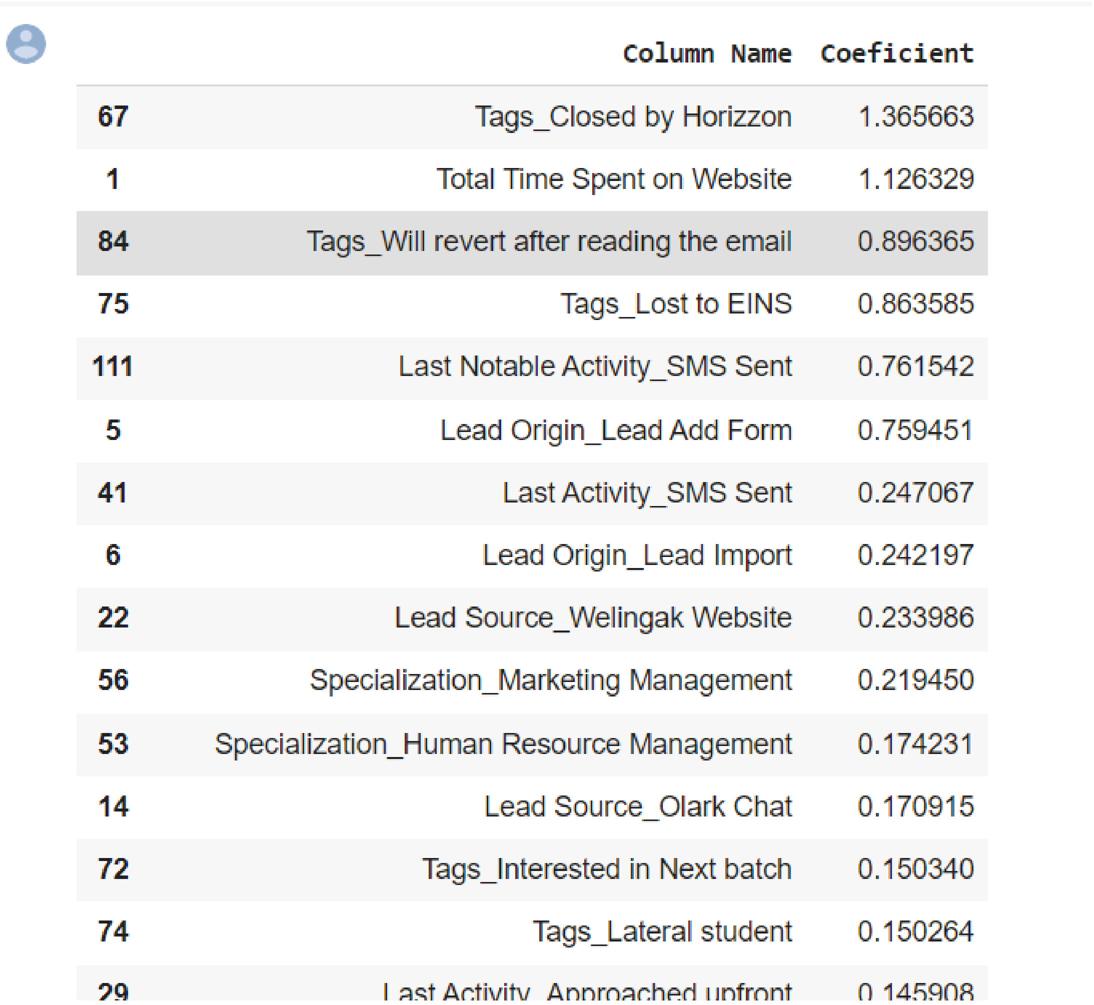
	1	Lead Scpore	Converted
0	0.568072	56.807170	1
1	0.010683	1.068272	0
2	0.896916	89.691576	1
3	0.988524	98.852404	1
4	0.004068	0.406832	0

2. **Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?**

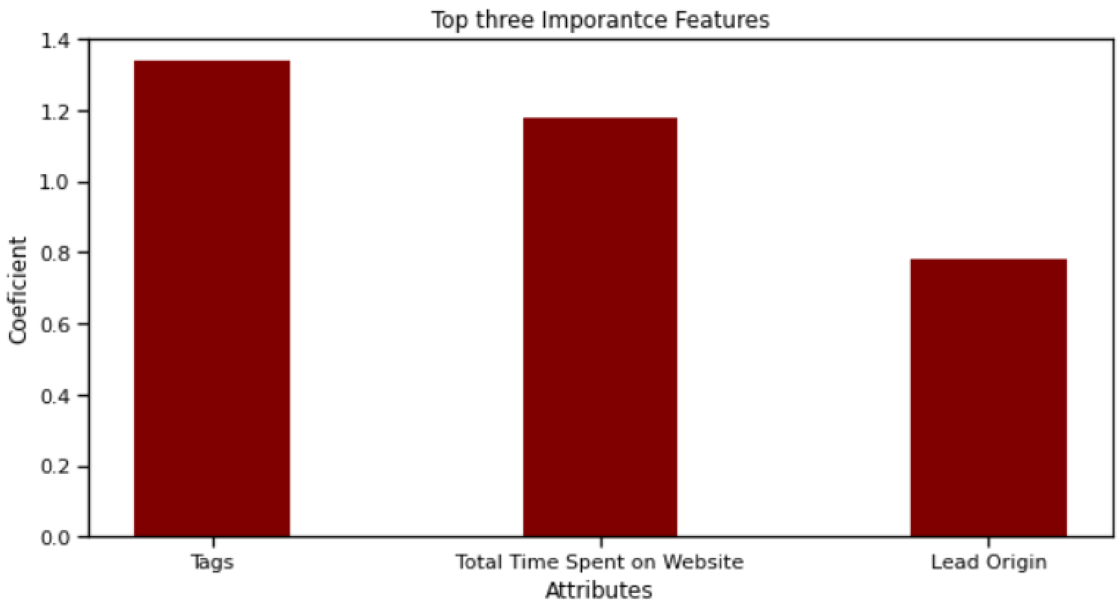
For this we have taken the coefficient of each feature and find the important features.

Top three important features are :

- *Tags*
- *Total Time Spent on website*
- *Lead origin*

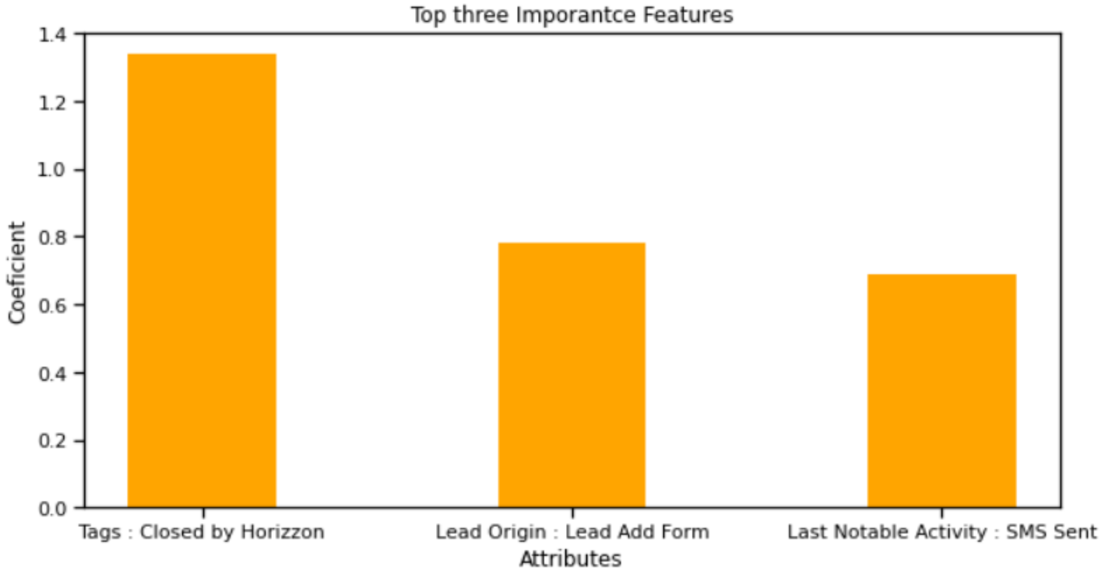


	Column Name	Coeficient
67	Tags_Closed by Horizzon	1.365663
1	Total Time Spent on Website	1.126329
84	Tags_Will revert after reading the email	0.896365
75	Tags_Lost to EINS	0.863585
111	Last Notable Activity_SMS Sent	0.761542
5	Lead Origin_Lead Add Form	0.759451
41	Last Activity_SMS Sent	0.247067
6	Lead Origin_Lead Import	0.242197
22	Lead Source_Welingak Website	0.233986
56	Specialization_Marketing Management	0.219450
53	Specialization_Human Resource Management	0.174231
14	Lead Source_Olark Chat	0.170915
72	Tags_Interested in Next batch	0.150340
74	Tags_Lateral student	0.150264
29	Last Activity_Approached upfront	0.145908

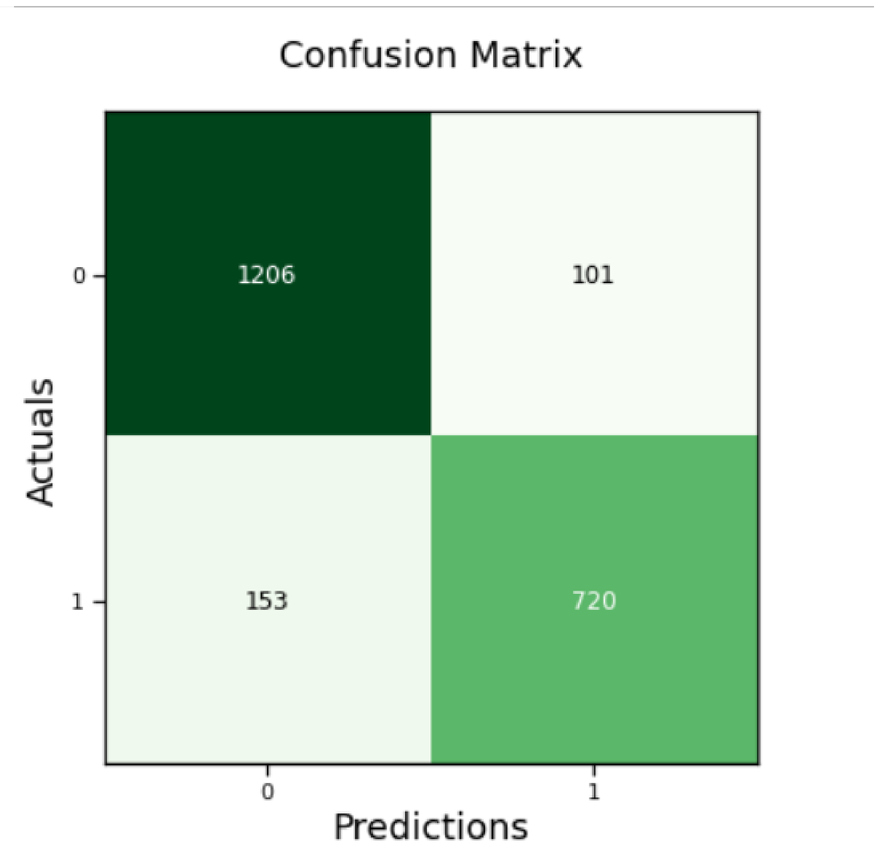


3. **What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?**

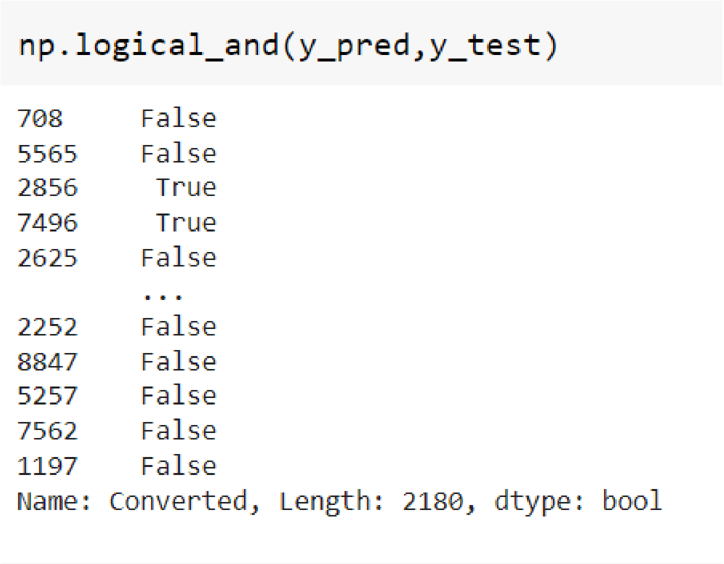
- *Tags : closed by horizon*
- *Lead Origin : Lead Add Form*
- *Last Notable Activity : SMS Sent*



4. **X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.**



From Confusion Matrix we can see there are total 101 false positive cases are there, so we need to focus on these false positive to convert into true positive cases.



```
np.logical_and(y_pred,y_test)
```

```
708     False
5565    False
2856     True
7496     True
2625    False
        ...
2252    False
8847    False
5257    False
7562    False
1197    False
Name: Converted, Length: 2180, dtype: bool
```