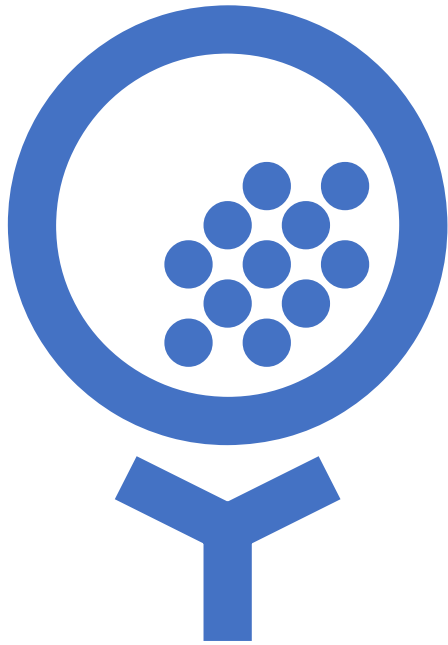


# Lead Scoring – Case Study

## **Problem Statement :**

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. 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%.



# Business Goal & Our Case Study Goal

## **Business Goal :**

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model where in you 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%.

## **Our Case Study Goal :**

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.

To adjust to if the company's requirement changes in the future so you will need to handle these as well.

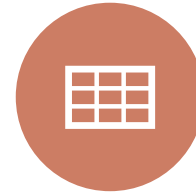
# Steps Involved



1. Read and understand the data



2. Clean the data



3. Prepare the data for Model Building



4. Model Building



5. Model Evaluation



6. Making Predictions on the Test Set

# Read and Understanding on Data

- We have imported the Lead.csv file using the pandas library and we have examined the data frame using the below commands
- `lead_info.head()`  
`,lead_info.info(),lead_info.shape` and `lead_info.describe()`
- We found the data set contains null values and there can be outliers and thus need to be handled appropriately .

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                     Non-Null Count  Dtype
---  -
 0   Prospect ID                               9240 non-null   object
 1   Lead Number                               9240 non-null   int64
 2   Lead Origin                               9240 non-null   object
 3   Lead Source                               9284 non-null   object
 4   Do Not Email                             9240 non-null   object
 5   Do Not Call                              9240 non-null   object
 6   Converted                                9240 non-null   int64
 7   TotalVisits                              9183 non-null   float64
 8   Total Time Spent on Website              9240 non-null   int64
 9   Page Views Per Visit                    9183 non-null   float64
10   Last Activity                            9137 non-null   object
11   Country                                  8779 non-null   object
12   Specialization                           7882 non-null   object
13   How did you hear about X Education       7893 non-null   object
14   What is your current occupation          8550 non-null   object
15   What matters most to you in choosing a course  6531 non-null   object
16   Search                                   9240 non-null   object
17   PageLine                                 9240 non-null   object
18   Newspaper Article                       9240 non-null   object
19   X Education Forums                      9240 non-null   object
20   Newspaper                               9240 non-null   object
21   Digital Advertisement                   9240 non-null   object
22   Through Recommendations                 9240 non-null   object
23   Receive More Updates About Our Courses  9240 non-null   object
24   Tags                                    5887 non-null   object
25   Lead Quality                             4473 non-null   object
26   Update me on Supply Chain Content       9240 non-null   object
27   Get updates on DM Content               9240 non-null   object
...
```

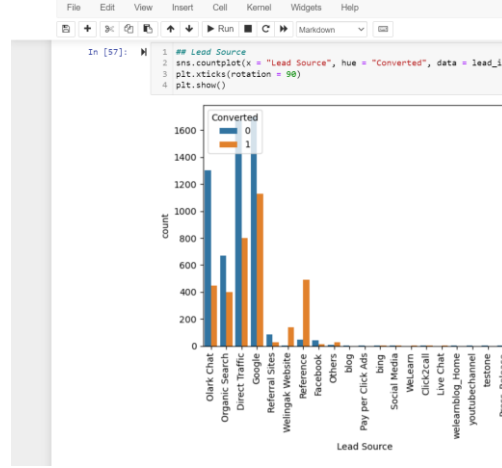
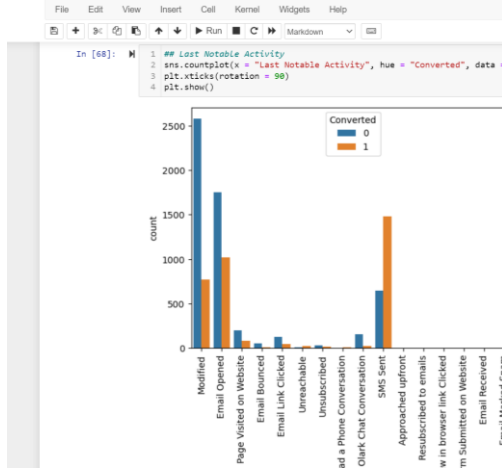
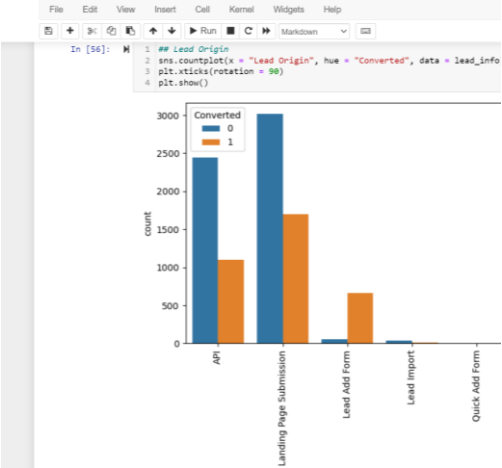
```
In [9]: 1 #checking for the null values as percentage
        2 percent_null = round(100*(lead_info.isnull().sum()/len(lead_info.index)), 2).sort_values(ascending=False)
        3 print(percent_null)

Lead Quality                                51.59
Asymmetrique Activity Index                 45.65
Asymmetrique Profile Score                  45.65
Asymmetrique Activity Score                  45.65
Asymmetrique Profile Index                  45.65
Tags                                         36.29
Lead Profile                               29.32
What matters most to you in choosing a course 29.32
What is your current occupation              29.11
Country                                     26.63
How did you hear about X Education           23.89
Specialization                              15.54
City                                         15.37
Page Views Per Visit                        1.48
TotalVisits                                1.48
Last Activity                              1.11
Lead Source                                 0.39
Receive More Updates About Our Courses       0.00
I agree to pay the amount through cheque    0.00
Get updates on DM Content                   0.00
Update me on Supply Chain Content           0.00
A free copy of Mastering The Interview      0.00
Prospect ID                                0.00
Newspaper Article                           0.00
Through Recommendations                     0.00
Digital Advertisement                       0.00
Newspaper                                   0.00
X Education Forums                          0.00
...
```



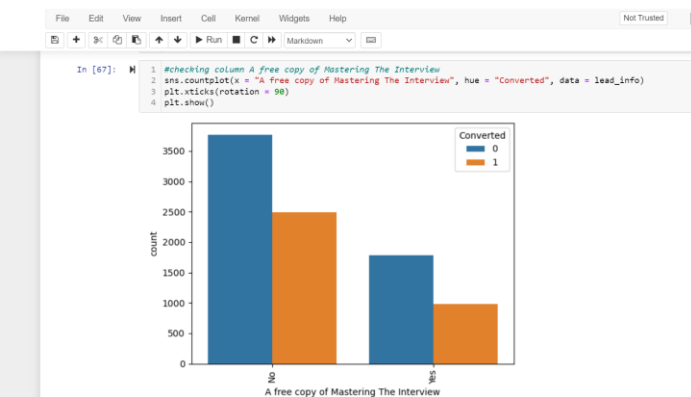
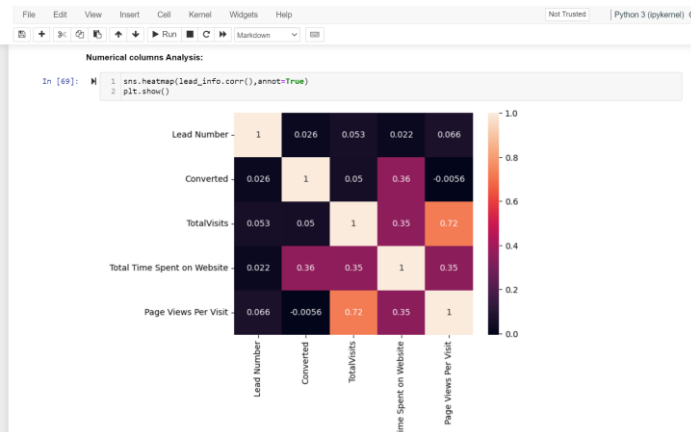
# Data Cleaning

- Initially we checked for duplicates and the percentage of the null values.
- We also know based on the case study few columns have the value 'Select' and we converted it to NAN.
- We have dropped null value columns > 30 %.
- Also, a few columns were dropped that have max data under one category 'NO'.
- We have treated the outliers for numerical column .

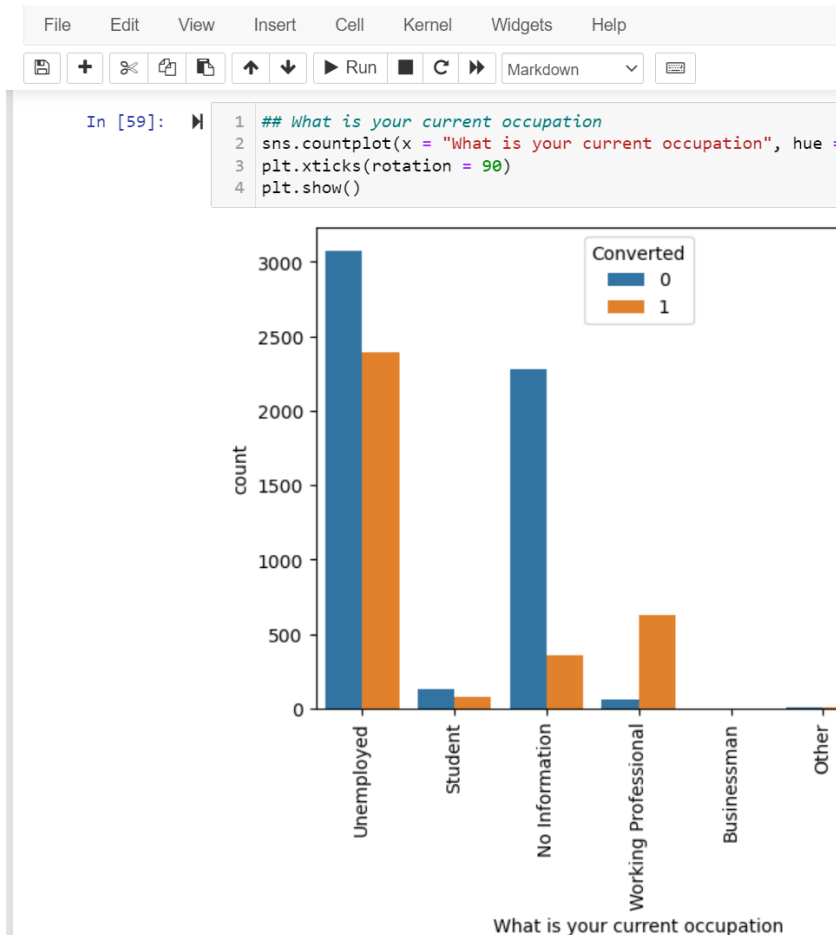


# Data Analysis

- We did a Uni-variate/Bi-variate analysis of data using EDA for few columns and made the inferences on the same .



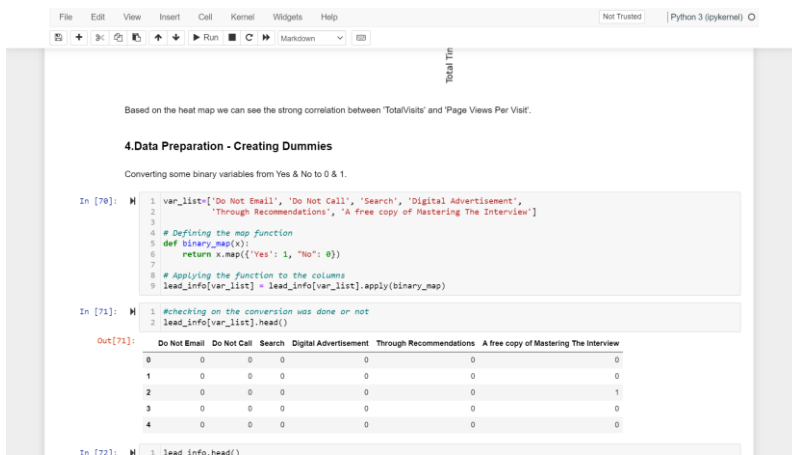
Based on the result we can infer as more leads are from those who do not ask for free copy of Mastering Interviews. Can be focused for conver



What is your current occupation

# Data Preparation

- We have converted the columns that are with values as Yes/No to 0 and 1.
- We also created dummies for columns with more than one category.



The screenshot shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar. The notebook content includes a text block about a heat map correlation, a section header '4.Data Preparation - Creating Dummies', and a code cell. The code cell contains Python code to convert binary variables to 0 and 1. Below the code cell, the output shows a table of converted data.

Based on the heat map we can see the strong correlation between 'TotalVisits' and 'Page Views Per Visit'.

#### 4.Data Preparation - Creating Dummies

Converting some binary variables from Yes & No to 0 & 1.

```
In [70]: 1 var_list = ['Do Not Email', 'Do Not Call', 'Search', 'Digital Advertisement',  
2               'Through Recommendations', 'A free copy of Mastering The Interview']  
3  
4 # Defining the map function  
5 def binary_map(x):  
6     return x.map({'Yes': 1, 'No': 0})  
7  
8 # Applying the function to the columns  
9 lead_info[var_list] = lead_info[var_list].apply(binary_map)
```

```
In [71]: 1 #checking on the conversion was done or not  
2 lead_info[var_list].head()
```

Out[71]:

	Do Not Email	Do Not Call	Search	Digital Advertisement	Through Recommendations	A free copy of Mastering The Interview
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	1
3	0	0	0	0	0	0
4	0	0	0	0	0	0

```
In [72]: 1 lead_info.head()
```

Generated Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6320
Model:	GLM	DF Residuals:	6305
Model Family:	Binomial	DF Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2522.2
Date:	Wed, 17 May 2023	Deviance:	5044.4
Time:	20:35:15	Pearson chi2:	6.30e+03
No. Iterations:	7	Pseudo R-sq. (CS):	0.4166
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-2.6532	0.148	-18.185	0.000	-2.939	-2.367
Do Not Email	-1.1622	0.175	-6.649	0.000	-1.505	-0.820
TotalVisits	1.2264	0.247	4.954	0.000	0.741	1.710
Total Time Spent on Website	4.5066	0.168	26.758	0.000	4.176	4.837
LeadOrigin_Lead Add Form	3.7825	0.202	18.683	0.000	3.386	4.179
LeadSource_Olark Chat	1.6450	0.124	13.263	0.000	1.402	1.888
LeadSource_WeinGak Website	2.0756	0.742	2.799	0.005	0.622	3.529
LastActivity_Email Opened	0.4606	0.115	4.008	0.000	0.235	0.686
LastActivity_Olark Chat Conversation	-0.6504	0.188	-3.453	0.001	-1.020	-0.281
LastActivity_SMS Sent	1.6433	0.116	14.213	0.000	1.417	1.870
CurrentOccupation_No Information	-1.2327	0.090	-13.734	0.000	-1.409	-1.057

```

In [189]: 1. #rechecking the VIF
2
3 vif = pd.DataFrame()
4 vif['features'] = X_train.columns
5 vif['vif'] = [variance_inflation_factor(X_train[col].values, i) for i in range(X_train[col].shape[1])]
6 vif['vif'] = round(vif['vif'], 2)
7 vif = vif.sort_values(by = "vif", ascending = False)
8 vif

Out[189]:
```

	Features	VIF
1	TotalVisits	2.63
2	Total Time Spent on Website	2.11
6	LastActivity_Email Opened	2.06
4	LeadSource_Olark Chat	1.94
8	LastActivity_SMS Sent	1.86
7	LastActivity_Olark Chat Conversation	1.75
12	LastNotableActivity_Modified	1.68
3	LeadOrigin_Lead Add Form	1.57
9	CurrentOccupation_No Information	1.57
5	LeadSource_WeinGak Website	1.30
10	CurrentOccupation_Working Professional	1.18
13	LastNotableActivity_Unreachable	1.02
11	Do Not Email	1.13

# Model Building

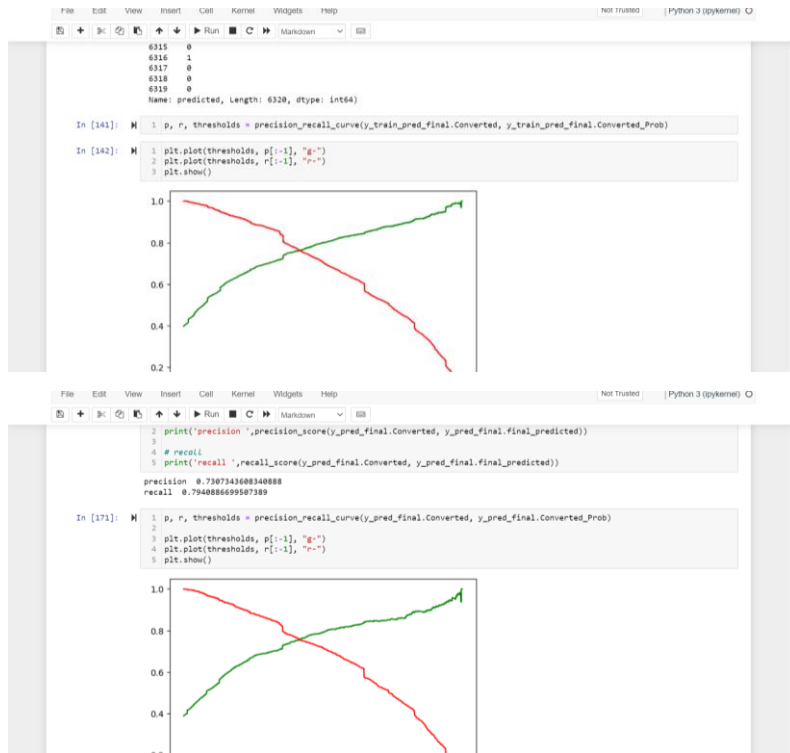
- We have created test-train-split using the ski-learn libraries for logistic regression model building.
- We have also scaled the data using MinMaxScaler.
- We also used RFE for the selection of the top 20 columns for model building.
- We have used the stats model to build a logistic regression model.
- We rebuild the model multiple models until we had the stable p-values and VIF <5.
- Finally, we arrived at the model 8 with all stable p-values and VIF less than 5.





# Model Predictions

- After building the model we have to predict the model based on the metrics to see its overall performance on the train data set.
- We used the confusion matrix and calculated the accuracy score.
- We also didn't stick with the accuracy score and tried to calculate the sensitivity, specificity, and false positive rate.
- Also, we checked by plotting the ROC curve.
- We also found the optimal cut-off point based on the probability of balanced sensitivity and specificity.



## Model Prediction-Cont'd

- we will need to now predict with the obtained cut-off on the test set.
- We continue the overall metrics to make predictions on the test set with accuracy, sensitivity, specificity, and false positive rate.
- We also made a confusion matrix.
- We also calculated the Precision and Recall for both the train and test data sets.



# Conclusion

- As we have checked Sensitivity, Specificity, Precision, and Recall as Metrics, we have considered the optimal cutoff limit as 0.37 for calculating the final prediction.
- Accuracy, Sensitivity, and Specificity values of the test set are around 81%, 79%, and 82% which are approximately closer to the respective values calculated using the trained set.
- Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.
- Firstly, need to sort out the best prospects from the leads you have generated. 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
- Focus on converted leads.
- Hold question-answer sessions with leads to extract the right information you need about them.
- Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.