Problem Statement

X Education sells online courses to industry professionals based on the leads they receive via different channels the company has used to market. If a user fills up a form, the company treats that as a lead and tries to convert the Lead into a customer via various channels.

With the current strategy, the company associates get in touch with the Leads and convert only 38-39% of the Lead population, which is low.

Therefore, as part of the case study, the company has requested an ML model that can successfully determine the possible customers with > 80% accuracy and possibly save the company from contacting every Lead via different channels.

The objective of Case Study

- Understand the relationship between the variables the company have captured via different channels
- Build an ML model to assign a Lead Score between 0 – 100. One hundred would mean that the Lead is hot.
- The ML model should identify users with 80% accuracy who are willing to take up any course.
- Identify the top variables which contribute to the model

Approach

Following are the steps which we performed to build a Logistic regression model

1) Data Cleaning
    - Many columns contained ‘Select’ as values. Therefore, we have replaced the ‘Select’ with Null values
    - Dropped features with a very high percentage of null values.

        - How did you hear about X Education
        - Lead Profile
        - Lead Quality
        - Asymmetrique Activity Index
        - Asymmetrique Profile Index
        - Asymmetrique Activity Score
        - Asymmetrique Profile Score

    - Dropped Tags – ‘Tags’ is a columns which is filled by the company employee after they spoke with the Lead. But as per the case study we should be able to determine whom the company employee should call. Hence the ML model should not depend on ‘Tags’. Hence dropping ‘Tags’
    - Imputed missing values for below features
        - Lead Source -All null values were replaced with Other
        - TotalVisits - Since the number null values is less therefore, imputed the vales with median of the column
        - Page Views Per Visit- Since the number null values is less therefore, imputed the vales with median of the column

- Specialization - All Specialization Null Values was replaced with ‘Not Sure’
- City - There is a ‘Other Cities’ Category in the City column therefore all City Null values was replaced with ‘Other City’
- For the current occupation we have replaced null values with ‘Other’
- Country value can be derived from the City values. Also, if the city value is ‘Other Cities’ we have replaced with ‘Other Country’

2) Univariate and Bivariate Analysis

Univariate Analysis

We analyzed all columns in the dataset and identified that there was not much variance in the data for a few of the categorical variables. Below are the columns which we dropped:

- What matters most to you in choosing a course
- Do Not Call
- 'Search'
- 'Magazine'
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Through Recommendations
- Receive More Updates About Our Courses
- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque

Bivariate Analysis

Below are a few of our observations:

- There is no significant difference in the Total Visits per profession
- There seems to be a linear relationship between the Total Visits Vs. Page Views per visit
- The ‘Total Time Spent‘ median for Converted Leads via channels direct traffic / Google / Organic Search seems to be almost double compared to a Non-Converted Leads user

3) Preparing data for model
   - Yes/No columns were converted to 1/0.
   - Categorical columns were converted to dummy variables.
   - Numerical values were scaled using MinMaxscaler

4) Splitting the data into train and test

   Dataset was split into 70% training data and 30% test data

5) Build the model

   We used Recursive Feature Elimination (RFE) with 15 columns to find out the most significant columns.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6367
Model:                            GLM   Df Residuals:                     6351
Model Family:                Binomial   Df Model:                           15
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2534.6
Date:                Mon, 11 Oct 2021   Deviance:                       5069.1
Time:                        02:46:24   Pearson chi2:                 6.56e+03
No. Iterations:                     6
Covariance Type:            nonrobust
==============================================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                         -3.2836      0.186    -17.626      0.000      -3.649      -2.918
Specialization_Not Sure                       -0.9855      0.127     -7.736      0.000      -1.235      -0.736
Last Activity_Email Opened                     1.1006      0.094     11.660      0.000       0.916       1.286
Last Activity_Others                           2.5011      0.637      3.925      0.000       1.252       3.750
Last Activity_SMS Sent                         1.3345      0.163      8.208      0.000       1.016       1.653
Last Notable Activity_Others                   2.6162      0.503      5.201      0.000       1.630       3.602
Last Notable Activity_SMS Sent                 1.0827      0.157      6.888      0.000       0.775       1.391
What is your current occupation_Student        1.1013      0.239      4.602      0.000       0.632       1.570
What is your current occupation_Unemployed     1.0501      0.089     11.773      0.000       0.875       1.225
What is your current occupation_Working Professional     3.5713      0.208     17.153      0.000       3.163       3.979
Lead Source_Olark Chat                         1.1490      0.137      8.360      0.000       0.880       1.418
Lead Origin_Landing Page Submission           -0.9641      0.131     -7.379      0.000      -1.220      -0.708
Lead Origin_Lead Add Form                      3.7264      0.234     15.947      0.000       3.268       4.184
TotalVisits                                    2.1977      0.336      6.548      0.000       1.540       2.855
Total Time Spent on Website                    4.6105      0.171     26.883      0.000       4.274       4.947
Page Views Per Visit                          -1.9360      0.425     -4.552      0.000      -2.770      -1.103
==============================================================================================================


**********Variance Inflation Factor of the Model**********

                                      Features    VIF
0                                        const  23.45
4                          Last Activity_SMS Sent   4.31
6                      Last Notable Activity_SMS Sent   3.92
11                Lead Origin_Landing Page Submission   3.36
1                            Specialization_Not Sure   2.90
15                              Page Views Per Visit   2.51
13                                       TotalVisits   2.17
10                           Lead Source_Olark Chat   2.15
12                          Lead Origin_Lead Add Form   1.68
9   What is your current occupation_Working Profes...   1.43
2                           Last Activity_Email Opened   1.42
8          What is your current occupation_Unemployed   1.35
14                        Total Time Spent on Website   1.33
3                               Last Activity_Others   1.23
5                       Last Notable Activity_Others   1.23
7             What is your current occupation_Student   1.06
```

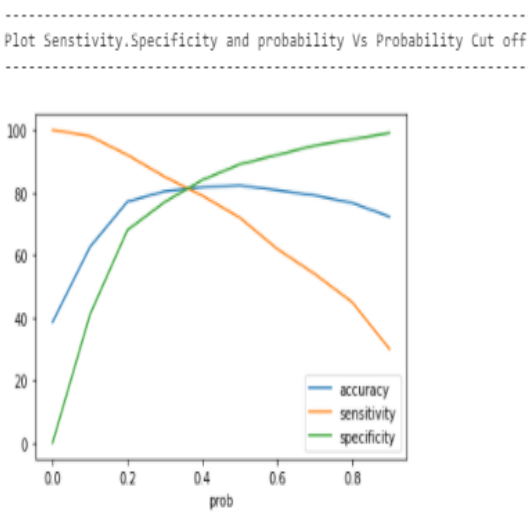
   Below are the top 3 features which we got after model building:

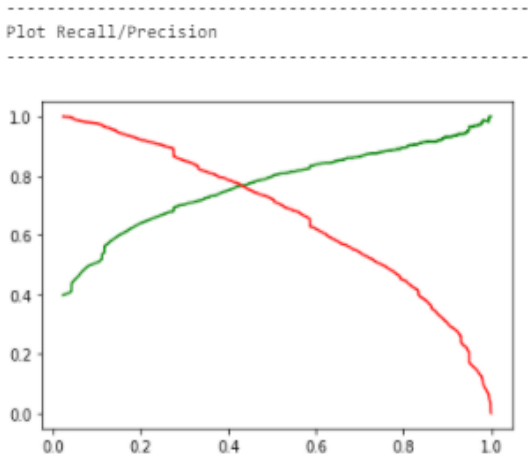
   - Total Time Spent on Website
   - Lead Origin_Lead Add Form
   - What is your current occupation_Working Professional

6) Evaluate the model on the training dataset

   We evaluated the model on the training dataset by calculating Accuracy, Sensitivity, Specificity.

   We also plotted the ROC curve, Sensitivity, Specificity, and probability Vs. Probability Cut off and Precision-Recall.





We got the optimum cut off for the model as 0.36. Below are the model paramters for training data set.

- Accuracy Score - 81.12
- Sensitivity - 81.0
- Specificity - 81.0
- False Positive - 19.0
- Positive Predictive Value - 73.0
- Negative Predictive Value - 87.0
- F1 Score - 81.27

7) Predict the Lead Score on Test dataset

The lead score was predicted on the test data set with a cut-off of 0.36. Below are the results which we got for test dataset

- Accuracy Score - 80.54
- Sensitivity - 80.0
- Specificity - 81.0
- False Positive - 19.0
- Positive Predictive Value - 71.0
- Negative Predictive Value - 87.0
- F1 Score - 80.75

Summary:

Below are the top features which we identified after a model building that would help in getting a lead converted:

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Last Notable Activity_Others
- Last Activity_Others
- Total Visits
- Page Views per visit