LEAD SCORE CASE STUDY

Group Members

1.Priyanka

2.Joseph

Problem Statement

Case study is done for An education company named X Education sells online courses to industry professionals. They wanted to know who are the professionals visiting their site and based on the data they will have to shortlist and decide potential customers to whom they can reach to sell their courses. We were given data about the Customers on how they visit the site, time they spent in site, Kind of specialization they search for, their occupation, what matters most in choosing the course, their last activity, what is their tags (last status) based on which to select most promising leads, i.e. the leads that are most likely to convert into paying customers. To build a model wherein we 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.

Business objective:

- X education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- 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.

Analysis Approach

We have analyzed the problem using the below steps:

1. Data loading
2. Data analyzing/ handling/ cleansing
3. Data transformation using imputation
4. Data preparation using dummies
5. Train- Test dataset creations
6. Data scaling
7. Data correlation
8. Generalized Linear Model Regression to identify p values
9. Regression score

Reading and Cleaning dataset

- Imported required libraries and read the dataset.
- We understood the dataset by knowing shape - dataset has 9240 rows and 37 columns
- We have Dropped unwanted columns like 'Prospect ID', 'Lead Number', 'Country', 'City', 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview' since these were not much helpful in Analysis
- Handled 'Select' level for categorical columns - using np.nan because We need to do this as user has not selected any values for those fields in the form. Still we need to handle them for accurate analysis Fields like – Specialization.
- Checked for missing values. Identified number of columns we need to drop for having more than 30% of missing values and dropped them.

Data transformation using imputation

- Used Imputation - For columns having less % of missing values - Lead source, Total Visits, Page per Views Per Visit, Last activity, What is your current occupation and What matters most. and Imputed with - Max occurrence values.
- We made sure that there are no null values in our dataset.

Data preparation using dummies

- Converted Categorical to numeric - Do Not Email, Do Not Call, Magazine, Newspaper Article, Education Forums, Newspaper, Digital Advertisement
- # Yes : 1 , No : 0 - created dictionary for two categories
- Created dummy variables for 8 categories and dropping the first level and Added these dummies to our original dataset. After creating dummies we removed original columns.
- dropped redundant variables.
- Converted some categories to numerical as they are imported as an 'Object'. all our data types are numeric for further analysis.
- we had checked for outliers and excluded them

Train- Test dataset creations

- Separated target variable from dependent variable.putting target variable 'Converted' to a new series 'y' and adding dependent variable in a new dataset called 'X'.
- Splitted the dataset into train and test dataset
- Checked the conversion rate from 'converted' column as it denotes the target variable. We have conversion rate of almost 39%.

Data scaling and Data correlation

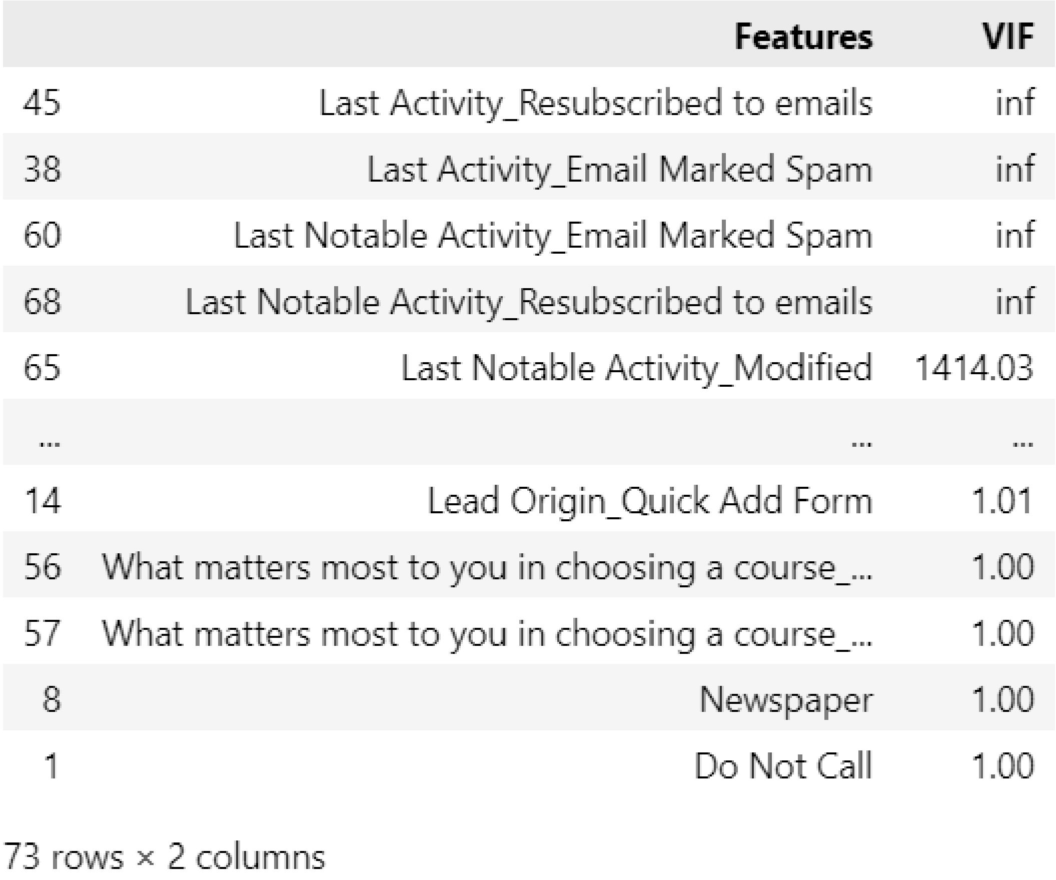
- Imported Standard Scaler method from sklearn - preprocessing library. Scaled the 'Total Time Spent on Website' variables with standard scaler and fitting - transforming the X - train dataset.Checked the conversion rate from 'converted' column as it denotes the target variable
- identifying correlation of the dataset, had set figure size and plotted heatmap. Dropped highly correlated dummy variable/categories.
- From above heatmap we got it was difficult to spot the highly correlated variables so we had proceeded with Model building based on the p-values and VIFs.

Generalized Linear Model Regression to identify p values

- Imported library for modelling. Performed Generalized Linear Model Regression using the library. we can see that there are many variables with insignificant p-values. So we had used RFE for feature selection as we have 70 variables and checking one by one is not an efficient way to do so.
- Recursive Feature Elimination, or RFE for short, is a popular feature selection algorithm.
- RFE is popular because it is easy to configure and use and because it is effective at selecting those features (columns) in a training dataset that are more or most relevant in predicting the target variable.

Regression score

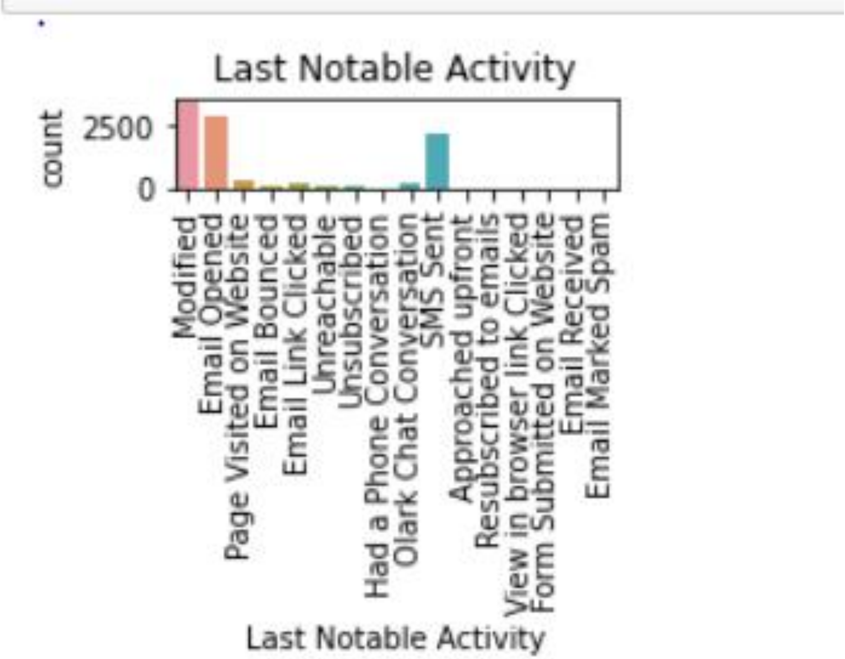
- We have calculated regression score - 0.43
- VIF



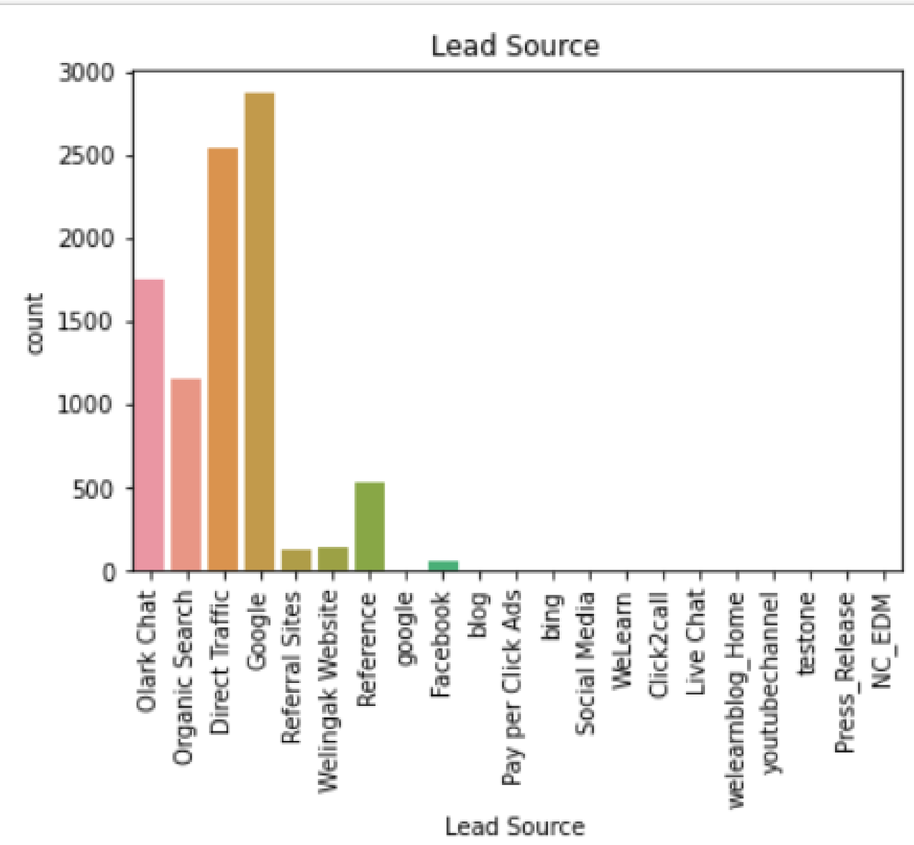
	Features	VIF
45	Last Activity_Resubscribed to emails	inf
38	Last Activity_Email Marked Spam	inf
60	Last Notable Activity_Email Marked Spam	inf
68	Last Notable Activity_Resubscribed to emails	inf
65	Last Notable Activity_Modified	1414.03
...	...	...
14	Lead Origin_Quick Add Form	1.01
56	What matters most to you in choosing a course_...	1.00
57	What matters most to you in choosing a course_...	1.00
8	Newspaper	1.00
1	Do Not Call	1.00

73 rows × 2 columns

Last Notable Activity



Lead Source


Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

THANK YOU