SUMMARY

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.

Below are the steps followed:

1. Data loading

Imported required libraries and read the dataset. We understood the dataset by knowing shape - dataset has 9240 rows and 37 columns. Datatypes are object,int,float. There are 7 numeric columns are rest are categorical columns. statistical info about the dataset like outliers, mean, std etc.

2. Data analyzing/ handling/ cleansing

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'.

Handled 'Select' level for categorical columns - using np.nan

We need to do this as user has not selected any values for those fields in form. Still we need to handle them for accurate analysis Fields like - Specialization

Created a for loop and the required columns are appended into a new list. Copied the required columns in separate DataFrame. Replaced 'Select' label with nan values. After replacing in temporary dataframe, we update our original dataset with new set of data from dataset df_1.

Checked for missing values. Identified number of columns we need to drop for having more than 30% of missing values and dropped them.

3. 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.

4. 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

5. 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%.

6. Data scaling

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

7. Data correlation

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.

8. 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.

9. Regression score

We have calculated regression score and VIF.