Lead Scoring Case Study

ML Assignment

Problem statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. (about 30%)
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- X Education wants help in identifying the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires to build a model wherein a lead score will be assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

High level Approach

Data Cleansing & Analysis

- Inspecting and cleaning data set
- Identify and handle null values
- Univariate, bivariate analysis
- Correlation analysis
- Multivariate analysis

Model Building

- Transformation-Dummy variable creation
- Train-test split
- Build logistic regression model
- Use RFE to reduce the features

Model Evaluation

- Accuracy, Precision metrics
- Optimum cutoff point
- ROC curve

Fixing Null values

- Identify all columns that have more than 45% null values
- Drop the columns from the dataset
- Remove unwanted columns from the dataset
- Impute the null values wherever possible

- Following columns are removed as don’t add to the modeling - *‘Prospect ID', 'Lead Number’*
- *‘Select’* is converted to Null for all columns
- *Country and City* columns are removed as they show high class imbalance
- *What matters most to you in choosing a course* is dropped as it has too many null values

Following columns are dropped as they show high class imbalance or have high number of null values

'Tags',

'Do Not Call',

'Search',

'Magazine',

'Newspaper Article',

'X Education Forums',

'Newspaper',

'Digital Advertisement',

'Through Recommendations',

'Receive More Updates About Our Courses',

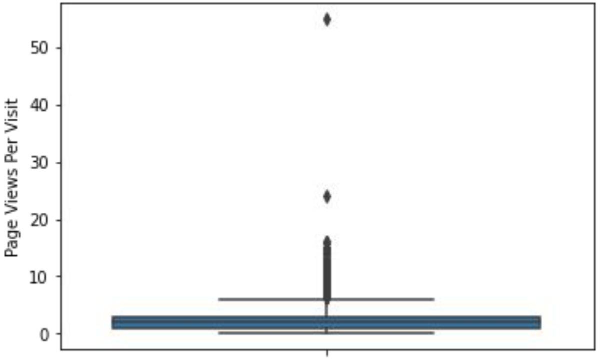
'Update me on Supply Chain Content',

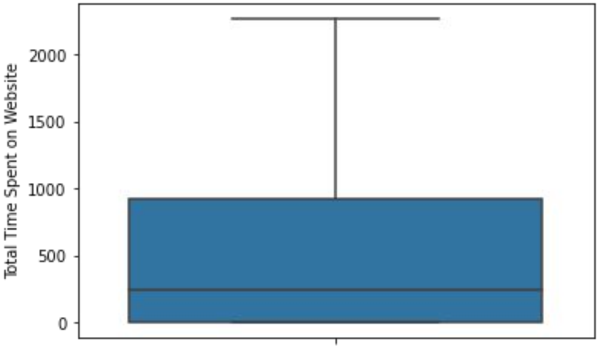
'Get updates on DM Content',

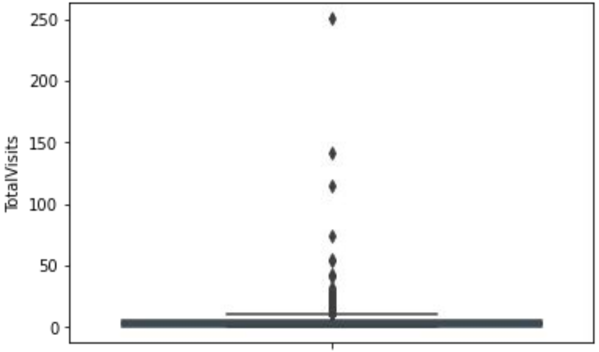
'I agree to pay the amount through cheque'

Data corrections

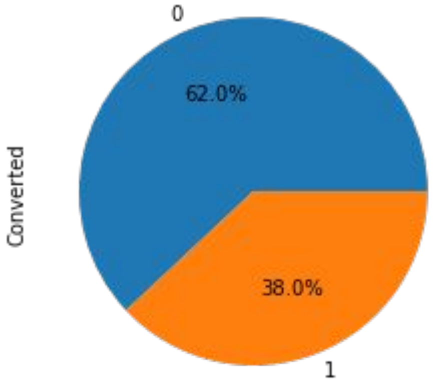
- Dropped the rows with null *Total visits*
- Replaced Null values and combining with low frequency values of *Last Activity*
- Wrong values of *Lead Source* are corrected
- No outliers seen with the variables *‘Page View Per Visit’ ,* *‘Total Time Spent on website’, and ‘Total Visits’*

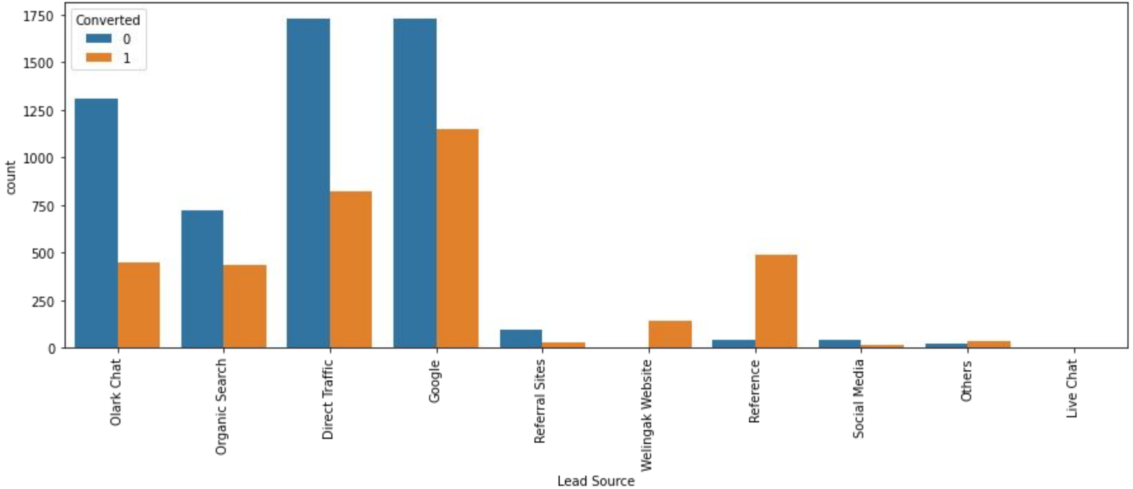






Data analysis





38% of the clients got converted, which is a good data set to train the model without major class imbalance

Google and Direct Traffic are the highest source of leads

Total columns at this stage : 12

Data Preparation and fitting the model

- Converted binary variables to dummy variables
- Dataset is split to Train and Test in the ratio of 70:30
- Total number of columns at this stage is 67
- Scaled the numerical variables using MinMax scaler
- Created the logistic regression model using statsmodels

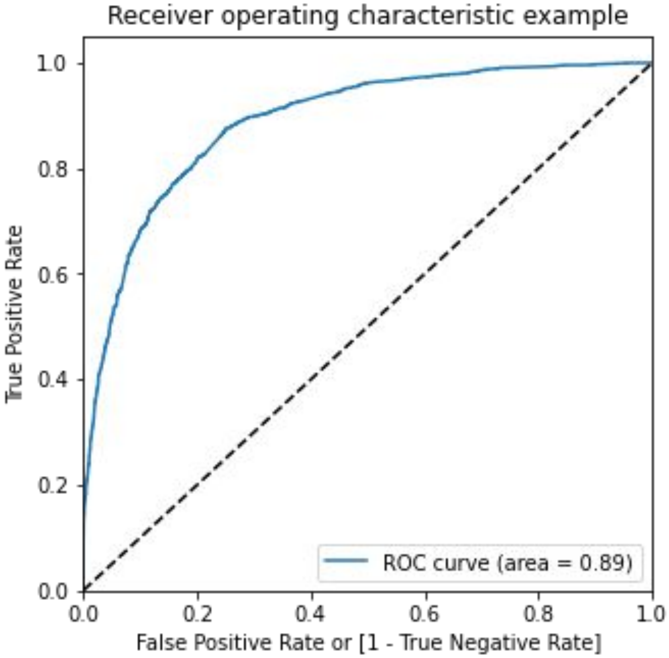
Model Building and RFE

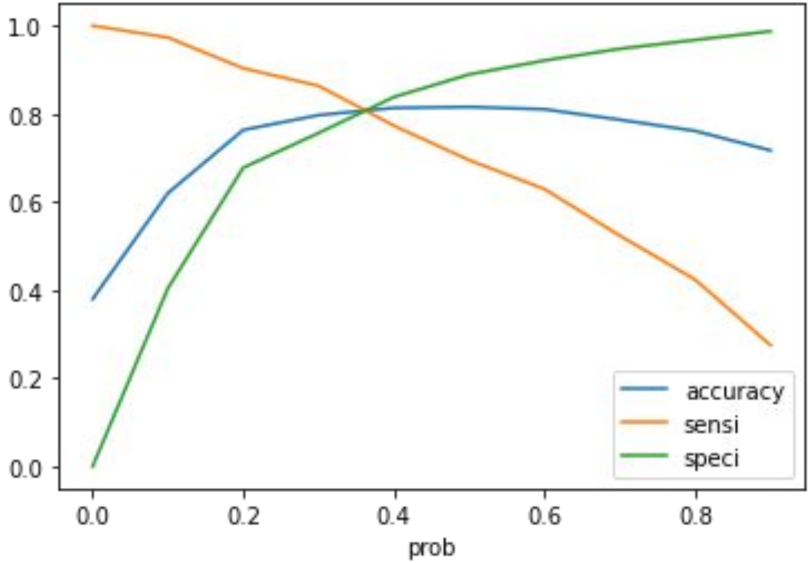
- RFE used to find the most significant 15 features
- Rebuilt the model based these features
- *What is your current occupation_Housewife* variable is dropped as it was found to have high p-value
- Rebuilt the model after this
- All p-values of the variables are in the reasonable range (below 0.05%)
- VIF values are less than 5 for all variables
- This can be treated as the final model with 14 variables

Model Evaluation

- Initial Confusion matrix calculated for the trained values with 50% as cut off
  - **Accuracy : 81%, Sensitivity : 69%**
- Area under the ROC curve : 89%

- New cut off arrived based on accuracy plot : **38% - Accuracy : 81%, Sensitivity : 78%**
- Lead score is calculated with 38% conversion probability





Conclusion

- Customers who make more visits and spend more time spent on the website have higher chances of conversion
- Customers whose leads came via Olark chat and Welingak website have high conversion probability
- Working Professionals likely to have higher conversion chances
- People who opted NOT to be emailed have lower chances of conversion
- Last Activity of SMS or others have higher chances of conversion