Lead Scoring case study

By : Vidhata RAWAT

PROBLEM STATEMENT:

X Education is an organization which provides online courses for industry professionals. The company marks its courses on several websites like google.

X Education wants to select most promising leads that can be converted as promising.

Although, X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rates could go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE:

- X Education wants a model to be built for selecting most promising leads.
- Lead score to be given to each leads such that it indicates how promising the leads could be. The higher the lead score, the most promising the lead to get converted, the lower it is the lesser the chances of conversion.
- The model to be built in lead conversion rate around 80% or more.

Problem solving techniques

- Data cleaning and data manipulation.

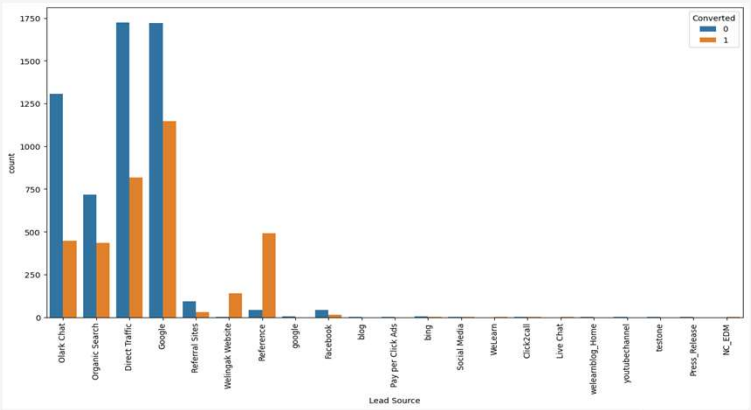
Check and handle duplicate data.
Check and handle NA values and missing values.
Drop columns, if it contains large amount of missing values and not useful for the analysis.
Imputation of the values, if necessary.
Check and handle outliers in data.

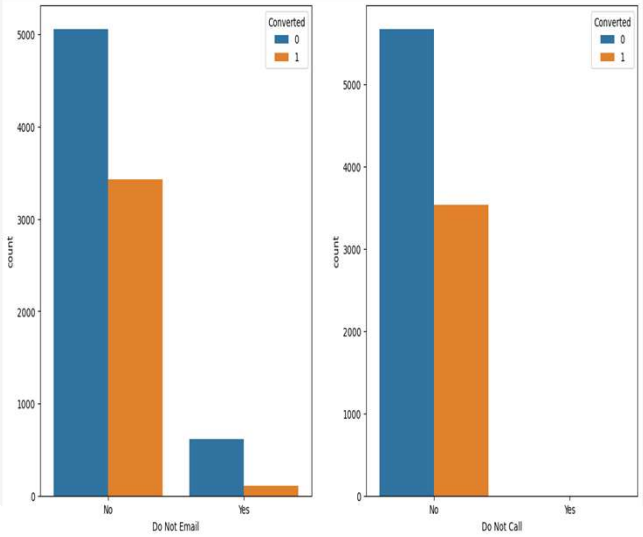
- EDA

Univariate data analysis: value count, distribution of variable etc.
Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the modelmaking and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

EXPLORATARY DATA ANALYSIS

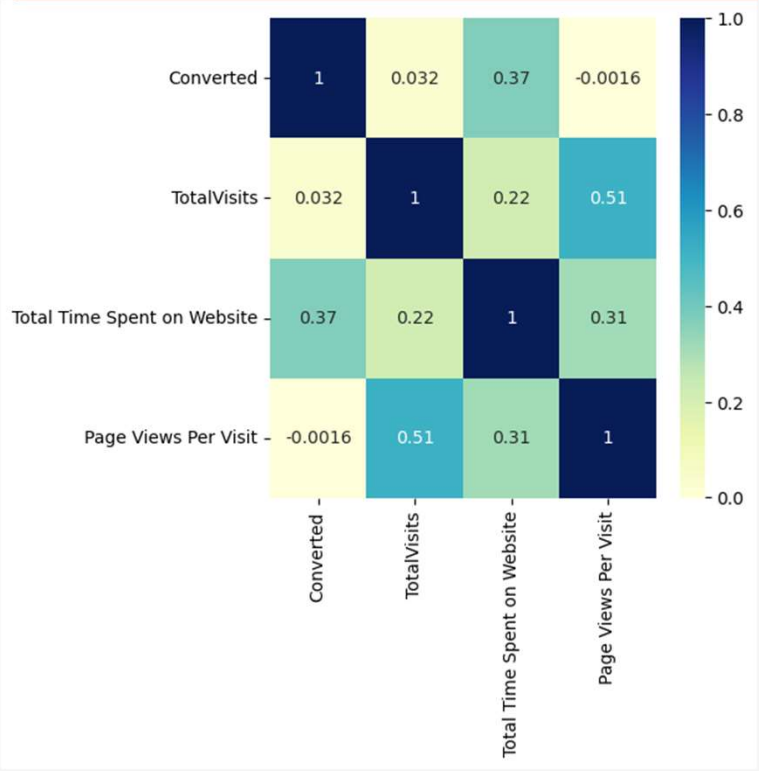




1.Google and Direct traffic generates maximum number of leads.
2.Conversion Rate of reference leads and leads through welingak website is high.

Many customers choose the option for Do Not Emails and Do Not Call

EXPLORATARY DATA ANALYSIS

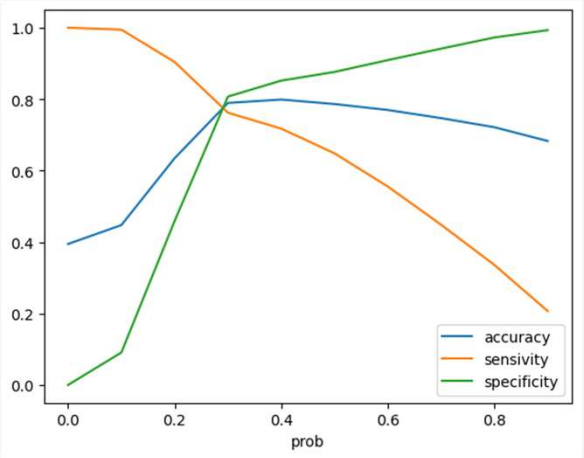


There is positive correlation between Total Time Spent on Website and Conversion 2.There is almost no correlation in Page Views Per Visit and Total Visits with Conversion.

MODEL BUILDING

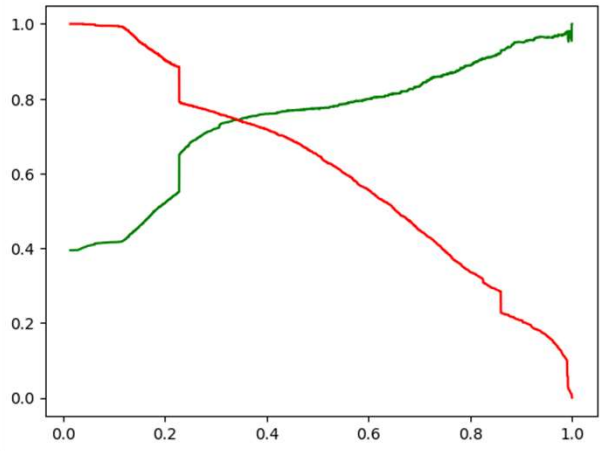
- Splitting into train and test set.
- Scale variables in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p-values.
- Check VIF values for all the existing columns.
- Predict using train set.
- Evaluate accuracy and other metric.
- Predict using test set.
- Precision and recall analysis on test predictions.

MODEL EVALUTION (TRAIN)



ACCURACY SENSITIVITY AND SPECIFICITY

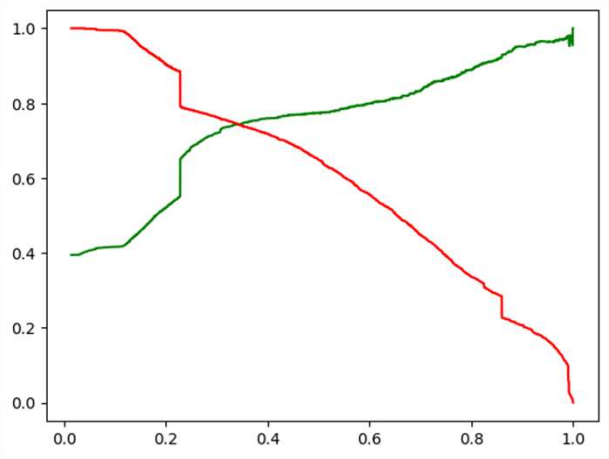
- 80.9% Accuracy
- 77.6% Sensitivity
- 82.9% Specificity



PRECISION AND RECALL

- 73.4% Precision
- 77.6% Recall

MODEL EVALUTION (TEST)



PRECSION AND RECALL

- 74.4% Precision
- 75.5% Recall

Test set threshold has been set as 0.41

CONCLUSION:

EDA:

- People spending higher than average time are promising leads, so targeting them and approaching them can be helpful in conversion.
- SMS messages can have higher impact on lead conversion.
- References and offers for referring a lead can be a good source of higher conversion.
- An alert message or information has seen to have high lead conversion rate.

Logistic regression model:

- This model shows high close to 80% accuracy.
- The threshold has been selected from Accuracy, Sensitivity, Specificity measures and precision , recall curves.
- The model shows 76% sensitivity and 83% specificity.
- The model finds correct promising leads and leads that have less chances of getting converted .
- Overall this model proves to be accurate.

Thank you !