Lead Scoring Case Study: Summary

Data size

The data set contains 9240 rows & 37 columns

Data value “select” under many columns

“Select” values of all columns replaced with Null value

Missing values

- Columns with more than 40% missing values dropped from analysis
- Few Other columns have 27-36% missing values, these values are imputed with median values.

Columns with very few missing values

Columns TotalVisits, Page Views Per Visit, Last Activity” had 1% missing values, missing values dropped instead imputing

Columns with highly skewed data

Dropped columns with high skewed data

Columns with Most unique value

Some columns have 100% unique values. So, these columns dropped from analysis

Columns not necessary for the analysis

'Prospect ID', 'Lead Number' dropped

Final chosen features for analysis

- **Categorical:** Lead Origin, Lead Source, Do Not Email, Last Activity, Specialization, What is your current occupation, Tags, Last Notable Activity
- **Numerical:** TotalVisits, Total Time Spent on Website and Page Views Per Visit

Outliers handling for numerical features

Columns “TotalVisits” and “Page Views Per Visit” have lot of outliers.Data capped to 95% percentile.

Data Imbalance

“Converted” and “not-converted” data is 38 & 62% . Data is balanced

Exploratory data analysis

Lead Origin: value 'Lead Add form' has highest conversion rate

Lead Source: values 'Reference' and 'Wellingak Website' have high conversion rate followed by “Google”

Last Activity: "Head Phone Conversation" and 'SMS sent’ have high conversion rate

Specialization: No value has any significant higher conversion rate than others

What is your current occupation: “Working Professionals” have high conversion rate

TotalVisits: Increasing conversion rate as TotalVisits increased

Logistic regression model building

- **Dummy creation:** binary categorical values converted to “0/1”, One hot encoding done to data having more than two levels
- **Independent and response features:** “Converted” assigned to “y” & all other features assigned to “X”
- **Train-Test split:** Data split to Train and Test in 70-30% ratio
- **Feature scaling**:Rescaling done on numerical features
- **RFE:** top 20 features selected using RFE method

- **Feature elimination**: Manual features elimination carried out based on p values from statsmodes logistic regression fit models summary and VIF values (criteria p<0.05 VIF<5)
- **Prediction on train set:** prediction on train set carried out using defaults threshold probability value of 0.5
- **Train set performance metrics:** Final model created using probability cut-off value to 0.38, which is obtained by the intersection point of "Accuracy", "Sensitivity" and "Specificity" of the previous model

- **Train set performance metrics from final fine tuned model**

Accuracy	Sensitivity	Specificity
80.9	0.78	0.82

- **Prediction on the test set**

Accuracy	Sensitivity	Specificity
80.7	0.86	0.77

- **Assigning lead score to test set:** lead scores of 0 to 100 assigned to test data based on their probability values
- **Predicting hot leads:** hot leads, which have lead score of greater than or equal to 85

<u>Model summary</u>

- Accuracy on train set & test set is 80.9 and 80.7
- Sensitivity on train and test is 0.78 and 0.86 which reasonably good performance indicator