Lead Score Case Study Using Logistic Regression

Problem Statement

- The problem statement is that we have Lead conversion data generated from the source and the company current has around 30% conversion rate and we need to identify parameters that increase the conversion rate to 80% from this data using logistic regression - by identifying hot leads and then assigning proper score to them based on probability percentage

- Using this we can reduce the effort of sales team by simply focusing on these hot leads and increase conversion rate

Data Sources

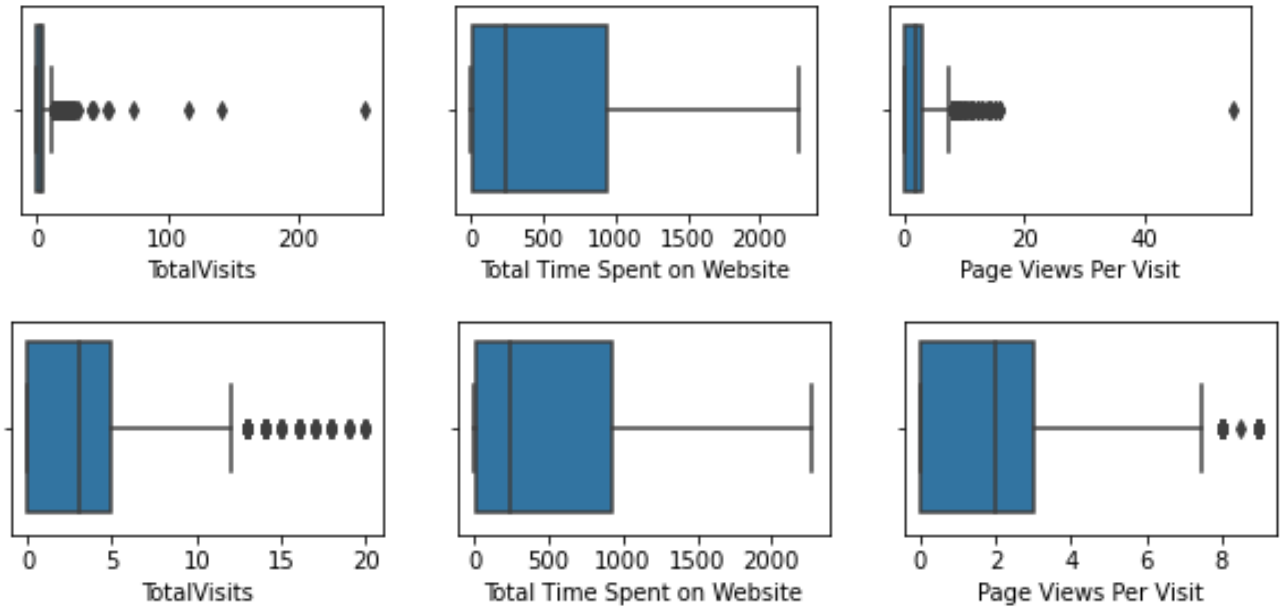
- Lead.csv
- Data Dictionary

Data Cleaning Steps

- Appropriately replaces “Select” values with np.nan as these values were not chosen by the lead
- Calculated null value columns and removed all columns with null value > 39%
- Removed all columns with skewed data as they are of no significance in the model and we cannot identify conversion trends with them - '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‘}
- Impute the remainder null value columns with mode – it is applied for even the numerical values as the outliers with high data gives a skewed average
- Columns with skewed data - values which appear minimal times are removed
- Converted the binary vars - 'A free copy of Mastering The Interview','Do Not Email','Do Not Call‘ to 0/1

Logistic Regression Steps

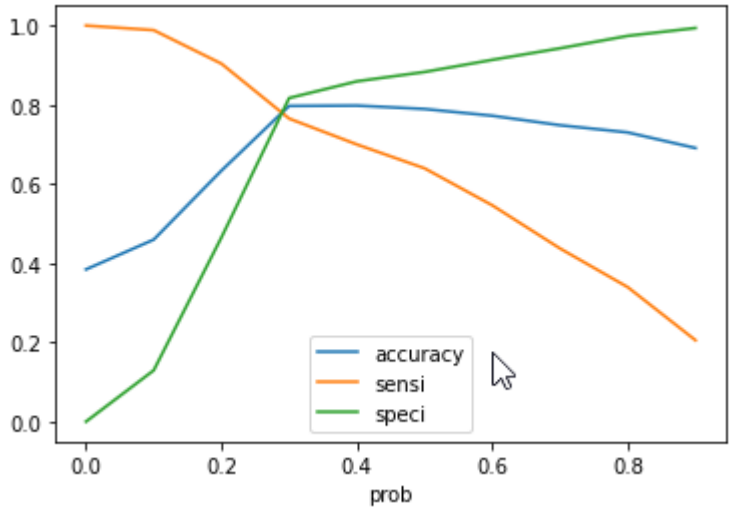
- Creating dummy variables for categorical columns and dropping the original columns
- Identifying outliers in the numerical values and removing them



Logistic Regression Steps

- Split the data into train – test
- Did feature scaling using Standard Scaler
- The current conversion Rate was calculated to 38.42
- Dropped columns that were highly correlated to each other
- Created the model using stats model
- 0.3% probability was calculated to be the most appropriate threshold to assign the yes probability and classify them as yes based on ROC model

Plotting Accuracy, Sensitivity and Specificity for different probabilities to identify optimum threshold



Confusion Matrix Achieved

```
[[3389,  759],
 [ 608, 1981]]
```

Final Observation:

- **Train Data:**
- Accuracy : 79.70%
- Sensitivity : 76.51%
- Specificity : 81.70%

- **Test Data:**
- Accuracy : 79.61%
- Sensitivity : 70.68%
- Specificity : 85.17%

- The model seems to predict the conversion rate very well

ThankYou