Lead Scoring Case Study

By: Ningareddy Modase

Rudin Bose

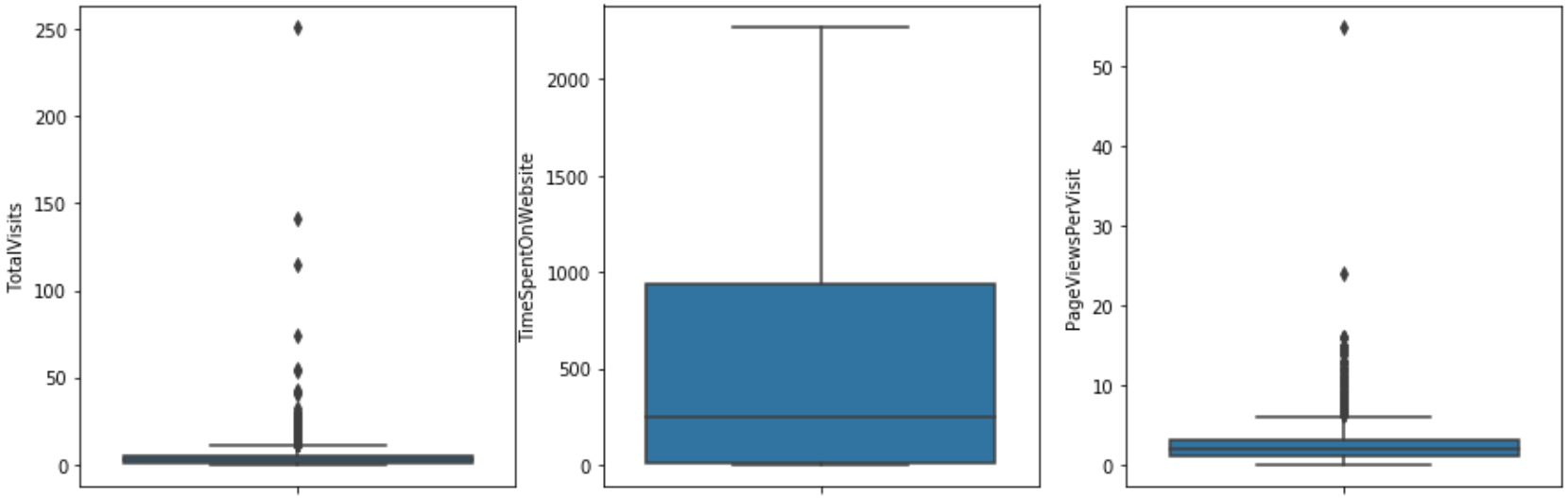
Objective

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

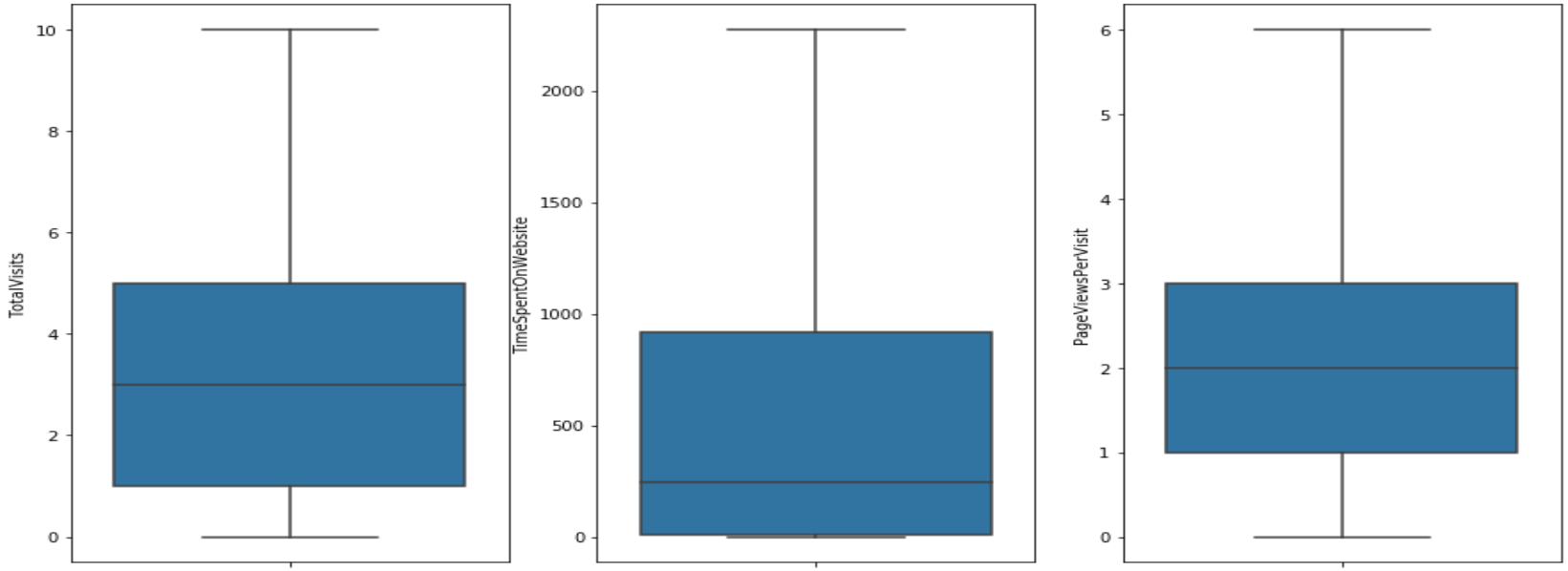
Steps

- Replaced all 'Select' value from each categorical variable with NaN value
- Missing values handling
  - Dropped features with high percentage of missing value (i.e. % > 30)
  - Dropped features 'Last Notable Activity’ and 'Lead Number’
  - Numerical columns were replaced by Median value
  - Categorical columns were replaced by Mode value
  - All rows having missing value percentage very small were dropped
  - Columns with Binary values (Yes/No) were replaced by 1/0
  - Outlier Treatment was performed for “TotalVisits” and “PageViewsPerVisit” by capping the values 0.95 percentile.
- Created dummy variables for the categorical features having multi level
- Test-Train Split (70:30)
- Feature Scaling
- Model Building
- Feature Selection using RFE
- Computed metrics (Sensitivity, Specificity etc)
- ROC curve and finding optimal cut-off point
- Model Prediction on Test set
- Recommendations

Before capping



Post capping

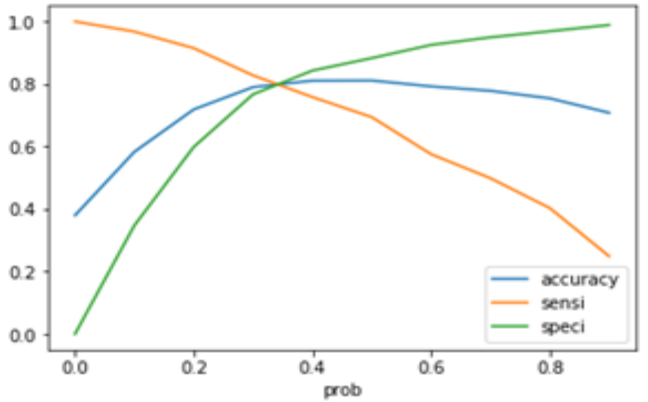


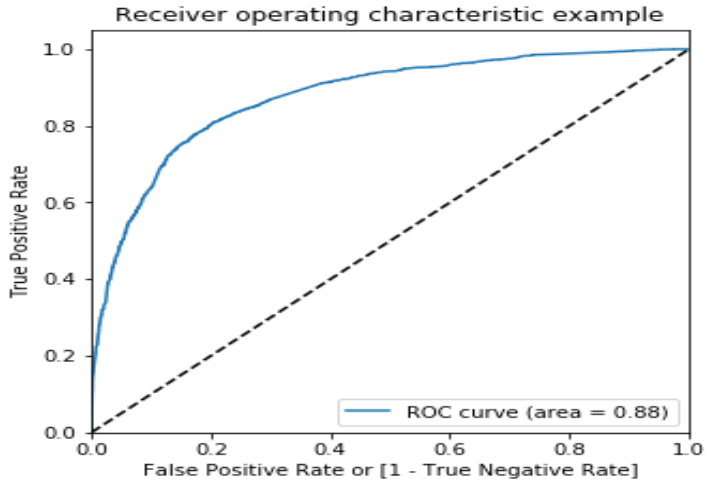
Trained Model Stats

- Top 20 features are selected using RFE, followed with manual elimination as per their P values and VIFs.
- Using cut off probability of 0.3 from accuracy, sensitivity and specificity graph, we got following stats,

Final LR Model Stats (Train set) is,

- Accuracy Score %= 80.84
- Sensitivity %= 77.18
- Specificity %= 83.08
- False Positive Rate %= 16.92
- Positive Predictive Rate %= 73.62
- Negative Predictive Rate %= 85.61
- Precision %= 73.62
- Recall %= 77.18



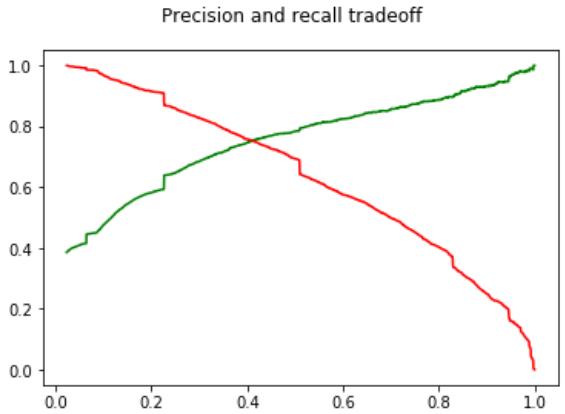


Predicted Model (on Test set) Stats

- Tradeoff value of 0.4 which we got from the graph shown was used to to predict on test set.

Predicted LR Model Stats (Test Set) is,

- Accuracy Score %= 80.96
- Sensitivity %= 76.1
- Specificity %= 83.96
- False Positive Rate %= 16.04
- Positive Predictive Rate %= 74.53
- Negative Predictive Rate %= 85.06



Conclusion

Final LR Model Stats (Train set) is,

- Accuracy Score %= 80.84
- Sensitivity %= 77.18
- Specificity %= 83.08
- False Positive Rate %= 16.92
- Positive Predictive Rate %= 73.62
- Negative Predictive Rate %= 85.61
- Precision %= 73.62
- Recall %= 77.18

Predicted LR Model Stats (Test Set) is,

- Accuracy Score %= 80.96
- Sensitivity %= 76.1
- Specificity %= 83.96
- False Positive Rate %= 16.04
- Positive Predictive Rate %= 74.53
- Negative Predictive Rate %= 85.06

Top 3 variables

1. Lead Origin
2. Occupation
3. Lead Source

Top 3 categorical/dummy variables

1. Lead Origin_Lead Add Form
2. Occupation_Working Professional
3. Lead Source_Welingak Website

Recommendations

Strategy 1:

Consider lead score **>= 75** as high score (Hot Score)

Target only those customers whose lead score is high **(>= 75)** and **model predicted value is 1**.

As we have a greater number of interns, let use divide these interns into 3 groups

Group 1: Contact customers whose lead score is **>=95**

Group 2: Contact customers whose lead score is **>=85 and < 95**

Group 3: Contact customers whose lead score is **>=75 and < 85**

This strategy will help more aggressively achieve high conversion rate.

Strategy 2:

Consider lead score **>= 90** as high score (Hot Score)

Target only those customers whose lead score is **high (>= 90)** and **model predicted value is 1**.

This strategy will help more aggressively achieve high conversion rate.