Lead Scoring Case Study

A brief summary report in 500 words explaining how you proceeded with the assignments and learnings that you gathered.

These are the various steps we have proceed with our assignment.

1. **Data Cleaning:**

   Since Data cleaning is the one of the most important Step to eliminate unnecessary columns and unnecessary rows. We performed data cleaning into five stages.

Mapping few categories into missing values:

- We Observed that few columns having label as 'Select' which means customer did not chosen any answer related to the question. So, we replaced this label as null values.

Eliminate unnecessary/redundant columns:

- Removed columns having more than 40% null values.
- We found that there are few columns, which contains 90% single category data, so we want to eliminate these columns as well from the dataset.

Eliminate unnecessary rows from the dataset:

- Removed rows having less than or equal to 5% null values.

Missing Values Imputation:

- For remaining categorical columns, we replaced missed values with the most frequent columns.

Categorical Variables Sanity Checks:

- For Categorical columns, we replaced low frequency of categories in each column into Others.

2. **Data Transformation:**

- Changed the categorical labels into dummy variables and binary variables into 0 and 1.
- Removed low frequency of categories in each dummy variable and eliminated unnecessary/ repeated columns from the dataset.

3. **Data Preparation for Modelling:**
   - We split the dataset into train (70%) and test dataset (30%) and performed standard scalar on the dataset.
   - After this, we want to eliminate the multi collinearity, so for that we plot a heatmap and eliminated highly correlated variables.
4. **Model Building:**

- We Created Logistic regression model with top 19 rfe value count variables and based on the p-value summary, we found few variables are insignificant. We eliminated highest p-value variable from the model summary and rebuild the model again, until p-value should be less than 0.05.
- After that we found, VIF value is greater than 5 for few columns. We eliminated highest VIF value variable and rebuild the model again, until all VIF values should be less than 5.
- Finally, we got the model for 15 variables and we checked the optimal propability cutoff by finding points and checking the accuracy, sensitivity and specificity. Based on the optimal point, we decided to optimal point is 0.4.
- Predicted the probabilities for target variable in test dataset and labelled whether a person will be converted to lead or not based on the optimal point.

5. **Model Accuracy Metrics and Results:**

- Finally, we got the best model, we calculated the accuracy, sensitivity, specificity, Precision, Recall and false positive rate.

Train-Set Results:

Metric Name	Accuracy
Accuracy	93.16%
Sensitivity	90.47%
Specificity	94.85%
Precision	91.67%
Recall	90.47%

Test-Set Results:

Metric Name	Accuracy
Accuracy	91.73%
Sensitivity	87.76%
Specificity	94.00%
Precision	89.30%
Recall	87.76%

6. **Summary:**

- Since Test set and train set having very good accuracy, model trained well and predicted well for lead conversion. Both Recall/Sensitivity are in acceptable range, we should be able to give the CEO confidence in making good calls based on this model.
- Top three Features for Lead conversion rate:
  1. Total Time spent on Website.
  2. Tags Closed by Horizons.
  3. Tag Lost to EINS.