Leads Scoring Case Study

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

Below are the steps how we have proceeded with our assignment:

1. Data Cleaning:

We took the following actions to clean the data.

a. Converted the 'Select' values to null values.

b. Checked the percentage of missing values in each column.

c. Removed columns having more than 30% null values.

d. Dropped the categorical columns which were heavily skewed.

e. For columns that had less percentage of missing values we dropped the rows having missing values.

After Data Cleaning we retained 9074 out of the 9240 rows and 24 out of the 37 columns that we started with.

2. Exploratory Data Analysis:

a. Checked the number of unique values in each column.

b. Dropped the columns which had only one unique value because they are not helpful in our analysis.

c. For categorical features with high cardinality we grouped categories with low count into 'Others' category to reduce the cardinality.
d. Visualized the numerical features with box plots and discretized them accordingly.
e. Visualized the categorical features and the rate of conversion using bar plots.
f. Made some inferences based on the data visualization – which variables are good predictors of lead conversion.

3. Data Preparation:

a. Dropped the columns which were not useful based on the inferences.

b. Performed One-Hot Encoding of all the categorical features.

c. Plotted a heatmap to check the correlations between the predictor variables and the target variable.

d. Separated the predictor variables and target variable.

e. Created the train-test split with test data size being 25% of the complete dataset.
f. Used Standard Scaler to scale the dataset to zero mean and unit variance.

4. Model Building:

a. Used Recursive Feature Elimination to trim down the number of variables to 15.

b. Fitted a Logistic Regression Model with the 15 variables.

c. Removed features with high VIF (VIF > 5) iteratively.

d. Removed features with high p-value (p-value >=0.05) iteratively.

e. Plotted the ROC Curve for the training dataset and found the Area Under Curve to be 0.85.

f. Plotted the Accuracy, Sensitivity and Specificity for different values of probability cut off and found the optimal probability cut off point**.**

g. Calculated the precision and recall for the training dataset at the optimal probability cut off point.

h. Made predictions on the test set.

i. Found the sensitivity & specificity score on the test data and made sure they are in acceptable range.

5. Assigned a score to each lead in the dataset.

Conclusion:

a. Test set is having sensitivity and specificity in acceptable range.

b. In business terms, our model is having stability in predicting whether a lead will be converted or not.

c. Also it can adjust with company’s changing requirement in the coming future.

Top features for good conversion rate:

1. Lead Origin_Lead Add Form
2. Total Time Spent on Website_1000+
3. Last Activity_SMS Sent