Summary Report: Lead Scoring Strategy Model Analysis

Solution Summary

Step 1: Reading and Understanding Data:

- Read and inspected the data.

Step 2: Data Cleaning:

- First step to clean the dataset we chose was to drop the variables having unique values.
- Then, there were few columns with value ‘Select’ which means the leads did not choose any given option. We changed those values to Null values.
- We dropped the columns having NULL values greater than 30%.
- Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.
- All sales team generated variables were removed to avoid any ambiguity in final solution.

Step 3: Data Transformation:

- Changed the binary variables into ‘0’ and ‘1’

Step 4: Dummy Variables Creation:

- We created dummy variables for the categorical variables.
- Removed all the repeated and redundant variables

Step 5: Test Train Split:

- The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.

Step 6: Feature Rescaling:

- We used the Min Max Scaling to scale the original numerical variables.
- The, we plot the a heatmap to check the correlations among the variables.
- Dropped the highly correlated dummy variables.

Step7: Model Building:

- Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
- Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
- Finally, we arrived at the 11 most significant variables. The VIF's for these variables were also found to be good and all <5.
- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 86% which further solidified the of the model.
- Then, checked if 80% cases are correctly predicted based on the converted column.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.41 and area under ROC as 0.86
- Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 79.2%; Sensitivity= 79.944%; Specificity= 78.053% in train data and accuracy value to be 78.34%; Sensitivity= 78.8209~ 79%; Specificity= 77.91% in test data.

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 79% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.
- Features which contribute more towards the probability of a lead getting converted are:
    a. Total Time Spent on Website
    b. Last Activity - Had a Phone Conversation
    c. Lead Origin_Lead Add Form

Learnings:

- EDA is extremely important step prior to building model. Key insights from EDA helps in treating the data correctly.

- Data cleaning helps in building efficient model. Steps like missing value imputation, scaling, outlier treatment must be performed at minimum to ensure quality of data is not compromised.
    a. Missing value – Columns with higher percentage of missing value can be dropped whereas columns with lesser percentage can be imputed.
    b. Outlier treatment – Outlier can impact model and result in less effective model. Hence care should be taken to treat outlier effectively. Care should also be taken to ensure that this does not result in significant loss of data.
    c. Scaling – Quantitative columns should bee scaled to ensure that they are on same scale.
- RFE is efficient technique to identify key features to start building model. On other hand PCA is helpful in dimensionality reduction by building new principal components.
- Functions to perform repetitive steps can help in building a modular code. This also help in reusability of the code.
- Understanding trade-off between sensitivity and specificity is key in determining ideal optimal cutoff for the mode.
- Confusion metrics is good indicator to determine how model performs. Accuracy, sensitivity, specificity can be derived from confusion metrics.

-By,

Tanuja Modupalli

Mrinmayi Kunkaliencar

Mrunal Paunikar