Leads Scoring Case Study

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

Answer:

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

1. **Data Cleaning:**

    a. We found that some columns are having label as ‘Select’ which means the customer has chosen not to answer this question. The ideal value to replace this label would be null value as the customer has not opted any option. Hence, we changed those labels from ‘Select’ to null values.

    b. After removing the redundant columns, we choose to remove the redundant variables/features.

    c. We removed categorical columns containing only single variables.

    d. Removed columns having more than 40% null values.

    e. We removed categorical columns having frequency for a single variable greater than 95% or highly skewed data.

    f. For remaining missing values, we have imputed values with maximum number of occurrences for a column.

2. **Data Transformation:**

    a. Changed the multicategory labels into dummy variables and binary variables into ‘0’ and ‘1’.

    b. Removed all the redundant and repeated columns.

3. **Data Preparation:**

    a. Split the dataset into train and test dataset and scaled the dataset.

    b. After this, we plot a heatmap to check the correlations among the variables.

    c. Found some correlations and they were dropped.

4. **Model Building:**
    a. We created our first model with rfe count 15 variables and proceeded form there.
    b. Our did manual tuning on the models using VIF and p values.
    c. Our final model is having a total of 9 predictive variables and all of them are significant.
    d. For our final model we chose the probability cutoff to 0.5 and check the accuracy, sensitivity and specificity for the training set. It gave us unacceptable range for sensitivity and specificity.
    e. Using Sensitivity vs Specificity graph, we found one convergent points and we chose that point for cut-off and predicted our final outcomes.
    f. We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs on the train set**.**
    g. Prediction made now in test set and predicted value was recoded.
    h. We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is
    i. We found the score of accuracy and sensitivity from our final test model is in acceptable range.
    j. We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.

5. **Conclusion:**

Learning gathered are below:

- Test set is having accuracy, recall/sensitivity in an acceptable range.
- In business terms, our model is having stability an accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.
- Top features for good conversion rate:
    1. **Total time Spent on the Website.**
    2. **Total Visits to the website.**
    3. **Lead source_Lead Add Form.**