Leads Scoring Case Study

A brief summary report explaining how we proceeded with the assignment and the learnings that we gathered.

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

1. **Data Cleaning:.**
    a. First, 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. Removed columns having more than 35% null values.
    c. For Columns with missing values more than 25%, we have imputed values with a new category called as "not available", as if we had done a imputation with mode/median, etc. the column could have become biased.
    d. For rest of the missing values, we did mode imputation for categorical variables and median for numerical variables.
    e. Finally for the country column we reduced the categories to three categories: "India", "Foreign" and "Not Available"
2. **Data Transformation:**
    a. Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.
    b. Checked the outliers for numerical variables and caped the outliers.
3. **Data Preparation:**
    a. Split the dataset into train and test dataset into 70:30 ratio
    b. Scaled the dataset using Normalization.

4. **Model Building:**
    a. We created our model with RFE using variable count of 25.
    b. Then based on the p-value and VIF score, we dropped some columns one by one.
    c. Finally we had a model with 17 feature variables.
    d. For our final model we checked the optimal probability cutoff by checking the accuracy, sensitivity and specificity tradeoff.
    e. We found one convergent points and we chose that point for cutoff and predicted our final outcomes.
    f. We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs**.**
    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 performing.
    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:

i. Test set is having accuracy, recall/sensitivity in an acceptable range.
ii. In business terms, our model is having stability in accuracy, sensitivity and specificity, meaning it will adjust with the company's requirement changes made in coming future.
iii. Top features for good conversion rate:

- **Total Time Spent on Website**
- **LastNotableActivity_had a phone conversation**
- **CurrentOccupation_working professional**