Lead scoring case study report

1. Data Cleaning:

- First step to clean the dataset we choose was to check the percentage of null values in each column and we dropped every column which have more then 32% of null values
- After removing the null values, we checked the value counts in each column to check the variance. The columns having negligible variance can be dropped directly as they won't affect our analysis.
- The null values in the remaining columns can be dropped directly so as to make the model building more precise.

2. Data Preparation:

- Split the dataset into train and test dataset in the ratio of 70%-30%.
- We have major number of categorical columns present in our dataset so we make dummy variables for them.
- The column Specialization has one value as select which means that lead didn't choose any specialization, so we make dummy variables for that column and dropped the variable Specialization_select.
- Scaling the numerical variables with minmax scaler.
- After this, we select the most important features that affects our target variable directly with the RFE (recursive feature elimination) method.

3. Model Building:

- We created our model with 15 number of variables.
- The columns (What is your current occupation_Housewife, Last Notable Activity_Had a Phone Conversation, Lead Source_Reference and What is your current occupation_Working Professional) were dropped one by one because their p-value was more than 0.05.
- The vif score for each remaining column was found to be decent as they were below 5.
- We predicted the values from our final model by assuming an arbitrary cut-off of 0.5 and calculated different evaluation metric like the accuracy score, specificity and sensitivity and our model performed quite well on these metrics.
- We plotted the ROC curve which finds the auc (area under the curve) which came out be 0.86 and is quite well.
- Then we assign the same columns to our test set to predict whether our model is working fine or not.
- We evaluated out model on test set with the help of different metrics like accuracy, specificity and sensitivity and they all were near to the metrics of the train set which means that our model performed good on the test set as well.
- We assigned a lead score to each of the leads which is nothing but the conversion probability multiplied and rounded up to 0 decimal places.
- The ROC curve on the test set has an auc of 0.85 which is also near to the ROC of train set.
- The precision and recall values are also decent on both the train and test dataset.

4. 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:
    - TotalVisits
    - Total Time Spent on Website
    - Lead Origin_Lead Add Form