SUMMARY

The summary of the lead scoring case study is explained below based on our model:

1.Data cleaning:

As we can observe in our Problem statement as well as in our data there are select values for many column.This is because customer did not select any option from the list, hence it shows select.Select values are as good as NULL.So we are replacing them with Null values.we also dropped the columns having more than 70% NA values.Likewise,the remaining missing values are imputed with their modes and NA values accordingly.

2.EDA:

To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin, olark chat, organic search, direct traffic, and google leads at the same time focus on generating more leads from reference and welingak website.Most of the leads generated by Email opened as their last activity and most leads are unemployed with less conversion rate.Most leads are from mumbai with around considerable conversion rate.

3.Dummy variables:

Created a dummy variable for some of the categorical variables and added to the main dataframe

4.Train-Test split:

We splitted the data into train data and test data and performed feature scaling to check the churn rate.You saw that the data has almost 38% churn rate. Checking the churn rate is important since you usually want your data to have a balance between the 0s and 1s (in this case churn and not-churn).Here

the data is neither balanced nor heavily imbalanced so we'll not have to do any special treatment for this data.

5.Model Building:

1.After running the training dataset,we performed feature selection using RFE.
2.Built model using stats model for detailed statistics and calculated the VIF value.Dropped the variable with high VIF.
3.Created a dataframe with the actual churn flag and the predicted probabilities.
4.For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
5.We found one convergent point and we chose that point for cutoff and predicted our final outcomes.
6.We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs.
7.Prediction made now in the test set and the predicted value was recorded.
8.We did model evaluation on the test set like checking the accuracy, recall/sensitivity
9.We found the score of accuracy and sensitivity from our final test model is in acceptable range.

6.Observations:

After running the model on the Test Data these are the figures we obtain:

Accuracy : 90.78%

Sensitivity : 84.14%

Specificity : 94.57%

Final Observation:
Let us compare the values obtained for Train & Test:

Train Data:

Accuracy : 91.11%

Sensitivity : 85.73%

Specificity : 94.49%

Test Data:

Accuracy : 90.78%

Sensitivity : 84.14%

Specificity : 94.57%

—----------

Submitted by,
Krithiga K
Martin robert durai