Summary

The model building and prediction are being done for company X Education to find ways to convert potential users. We will further understand and validate the data to conclude to target the correct group and increase the conversion rate. Let us discuss the steps followed:

1. **EDA:**

- Index columns are dropped.
- Dataset has 9240 rows and 37 columns, 7 variables are continuous, rest are categorical.
- Data does not belong to same scale.
- Dataset has 41039 null values in total.
- Quick check was done on % of null value and we dropped columns with more than 30% missing values.
- We also saw that few of the variables with nulls would cost us a lot of data and they seemed important in prediction, therefore, we replaced the nulls values with 'not provided' which was later encoded as class 0 in each of the instances.
- In other instances, for categorical variables, the nulls were replaced with the modes of the variables.
- Some variables had only one type of entry in them of 99% of the entries were of a single type, these variables are also dropped.
- Outliers are treated for the continuous variables.
- Few categorical variables are encoded manually, others are encoded using Label Encoder.
- Variables are scaled using standard scaler.
- Most variables have a multi-modal distribution as they were categorical variables transformed into ordinal.
- No significant collinearity/ multi-collinearity present (checked from correlation plot).

2. **Train-Test split & Scaling :**

- The split was done at 70% and 30% for train and test data respectively.

3. **Model Building:**

- VIf scores were checked, no significant scores were found, so the final set of variables were used in the final model.
- After fitting the initial model, few variables were found to not be significant in prediction (high p values), therefore, these variables were dropped to prevent clutter.
- Probability cut-off were chosen by plotting the accuracy, sensitivity and specificity graphs and choosing the cut-off as the point where these three curves met.

- Overall accuracy was checked for different probability cut-offs and after analysing all the factors, a cut-off of 50% was chosen.

4. Model Evaluation

Sensitivity – Specificity

The optimum cut-off value was found using the ROC curve. The area under the ROC curve was 0.88. After Plotting we found that the optimum cut-off was 0.35 which gave:

Train data scores:

- Accuracy 80.91%
- Sensitivity 79.94%
- Specificity 81.50%.

Test data scores:

- Accuracy 79%
- Sensitivity 79.23%
- Specificity 79.50%

Precision – Recall:

With the cut-off of 0.35 we get the Precision & Recall of 79.29% & 70.22% respectively. Therefore, to increase the above percentage we need to change the cut-off value. After plotting we found the optimum cut-off value of 0.44 which gave:

Train data scores:

- Accuracy 81.80%
- Precision 75.71%
- Recall 76.32%

Test data scores:

- Accuracy 80.57%
- Precision 74.87%
- Recall 73.26%

5. CONCLUSION

- Performance is stable across train and test sets proving that the model is a good fit.
- Final Regression equation is:

  Converted (y) = **0.44 *** (lead origin)
  + **0.32 *** (source)
  - **1.76 *** (do not email)
  + **0.17 *** (total visits)
  + **0.99 *** (total time spent on website)
  - **0.59 *** (page views per visit)
  + **0.24 *** (last activity)
  - **0.09 *** (specialization)
  + **0.62 *** (current occupation)
  + **0.58 *** (tags)
  + **0.16 *** (lead quality)
  + **0.40 *** (lead profile)
  + **0.27 *** (city)
  - **0.48** * (a free copy of mastering the interview)
  + **0.41 *** (last notable activity)
  - **0.30**

- Class 1 in "Converted" can be considered as **"Hot Leads"**
- "Churn_Prob" can be used to get a continuous score of all the leads higher churn_prob means higher chances of the lead becoming a paid customer.
- CEO's target is around 80% lead conversion, this model predicts with an accuracy of 75%, which is an indication that it will be a good fit in this scenario.