SUMMARY

As we all know by now, this is X education data and they sell online courses to industry professionals. People interested in the course land on the website and browse courses. They get a huge number of leads but are mostly irrelevant which eventually waste sales team time. The conversion ratio is 45.5% as per the data processing done by us. We tried our best to delete the least number of rows possible to generate better inferences.

Commence with, data cleaning after having the good understanding of the data because huge number of missing values and irrelevant columns/ rows are present in the data. We started by removing the columns having more than 40% missing value since they are incompetent to make conclusions. Then, synchronizing, imputing missing values was the next step we followed.

After finishing the data cleaning step, started with exploring data analysis by visualizing different variables and their impacts, Understanding Correlations, checking Outliers, treating outliers, Count Plots and histograms helped us visualize each variable and its impact in a better way.We used box plot to fix the outliers.

Post completing EDA, we have created dummy variables and convert binary variables into numeric values. Then, dropped the original variables by which we created dummies.

Then, we did splitting of the data into train-test split and scaling the variables by using standard scaler so that all the variables should be on the same scale. Then, started with Recursive Feature Elimination (RFE) approach keeping the 15 variables, checking the VIF and p- values and rebuilding the model until we get an adequate P-value and decent VIF score. We made 7 models again and again to reach the conclusion at last model all VIF were below 3 and p value of all the variable were 0.

After building the final model, we started calculating the accuracy matrix, specificity and sensitivity and optimal cutoff that is 0.4 to decide the convert probabilities. Plotted a ROC curve graph to measure the area (0.87) and accuracy was a mandatory thing to follow.

We plotted sensitivity, specificity and probability to calculate the optimal cutoff that is 0.4. We concluded the model by creating precision and recall.

Now, we have to make predictions on the test set on the same columns as the training model. Same calculation happened for the test set as well, measuring important matrix and accuracy, precision, recall, sensitivity, specificity.

Making comparison with the final training model and finding the difference between them to ensure that it is the correct model and ready to share with the sales team.

- The steps can be easily seen in the python notebook we have submitted.
- You may refer to PPT for visuals and references made during model building.
- The detailed inferences can be referred to in the comment section of the notebook.
- Subjective questions were answered in different PDF file.

Important columns from the data- Source of Lead, Lead origin, current occupation, Last activity,Time spend on the website, choosing_Course_Better Career Prospects, Last Notable activity and Specialization are the key variables. The conversion ratio can be increased by following the model observation.

Final output:

Observations on Test Set:

- Accuracy = 80%

- Sensitivity= 81.5%

- Specificity = 79.1%

Observations on Train Set:

- Accuracy=78.9%

- Sensitivity=71.9%

- Specificity=85.6%