Summary Report

We started our study by first importing all the necessary data & libraries required for the analysis. Primarily we looked at the metadata and checked the missing values.

After looking at metadata and before dropping missing values we first decided to drop the columns which have only one value in it as it carries no valuable information. After that we decided to change the values of the binary columns with (Yes/No) to (1/0).

We proceeded by replacing the 'Select' value with null value and then we decided to drop the columns which had more than 30% of the missing entries except City and Specialization as we considered this might give some good insight while building model.

Then we proceeded by missing value imputation after looking at the data types of columns & the distribution of values in the columns. We confirmed that there are no missing values in the data set and then continued.

Starting with univariate analysis we understood that treating the outliers might impact our model so we did not drop any outliers. In bivariate analysis we understood that the Country column is not adding much value and hence we dropped it. Multivariate analysis confirmed that there is strong relationship between "Total time spent on website" and getting the customer converted.

We proceeded ahead with dummy creation step & then we dropped the Lead Number and Prospect ID as they are just the unique identification keys and hence do not carry any useful information for creation of model.

Going ahead we did the train test split with 70/30 ratio. We decided to scale the continuous numerical variables with MinMaxScaler.

As the number of columns obtained after the dummy data creation was high, hence we decided to go ahead with hybrid approach for feature selection. In RFE we began with 15 features and then kept the feature which were significant and had a VIF score lesser than 2 to maintain. Which gave us total 11 number of features which were able to produce a decent model.

We checked the accuracy of the model which was good. However, confusion matrix showed that model was not showing good recall score at 50% of cut-off value.

In order to select a reasonable cut-off value where we will have a good trade-off between recall, precision and accuracy we used multiple methods like ROC, recall precision trade-off and decided to take 25% as the cut-off value.

To confirm that our model is working well on the test data set we used the model on test data to predict the results. And we were able to manage good recall, precision and accuracy score. Which confirmed that model was able to predict decently on both test & train data.

To conclude important learning, X Education should focus on the customers who pay more visits to platform, spend considerable time on the website & decide to have a phone conversation & should not spend much efforts on customers who do not want to be contacted by email.