Summary

We had to build a logistic regression model to predict whether the lead for the online courses for the company X education will be successfully converted or not.

In order to build a good model, we went ahead with the following approach:

1. **Reading and understanding the data:** We started with reading, understanding, checking the shape and the variables in the dataset given to us.
2. **Data cleaning and manipulation:** In this step we came across many variables which were having high percentage of missing values.
   We removed the variables which were having greater than 60% missing values, we imputed few categorical variables with the mode function and dropped the rest of the rows having very few missing values.
3. **Check the data for duplicate values:** This step checked for any duplicated rows in the dataset – it was clean with no duplicates.
4. **EDA :** We analyzed important features using various plots like: Countplot, Boxplot, and heatmap of correlation coefficients. Also performed univariate analysis on the variables and found many of them were highly skewed and not useful for interpretation. We checked the variables with target variable as hue to see how variable's categories responded.
   We also capped the outliers at 95^{th} percentile so that it doesn't affect our model.
5. **Data preparation:** Create dummies of the variables containing different categories. Encode binary categorical variables to 0s and 1s.
6. **Train/Test split and Scaling:** Split the dataset into train (70%) and test (30%) sets. Before building the model we did scaling on the continuous numerical variables on the train set.
7. **Model Building**:
   a. Apply RFE to select the top 15 variables: Through RFE the model selected for us the best 15 variables and we proceeded with same to build our model.
   b. We build our model using GLM. We checked for any p-values which were higher as a large p-value indicates weak evidence against the null hypothesis, so we fail to reject the null hypothesis. Hence we drop the variables with higher p-value.
   c. Once we checked all the variables by constantly checking the higher p-value and dropping it we finally reached to the final model which had decent p-value for all variables and checked the VIF
8. **Model Evaluation:**
   a. Check various scores of the model like, accuracy, sensitivity, specificity, precision, recall etc.

   b. The sensitivity checks all the conversions have been classified which we needed for our purpose. Specificity is for all the conversion misclassified as non-converted leads which we wanted it to be lower.
9. **Threshold value(cutoff):** The higher the threshold cutoff the lower the sensitivity and thus the lower the conversion rate. To see if our model justifies our purpose, we tried different threshold values and decided on 0.35 which gave the decent conversion rate as needed for our business.
10. **Conclusion:** Once confirmed with the final model – we found a list of variables that could be suggested to the business to look further into to in order to increase the lead conversion rate.