Summary

This analysis aims to assist X Education in attracting more industry professionals to their courses.

1. Clean the data: Starting with convert some categorical variables from select to NaN. Then looking through all the labels to the correct format, either numerical or categorical. Then remove the missing value by using dropna.
2. When the data is well formated, then checking the distribution of the data and decide top drop columns where the distribution is super unbalanced can may have no impact on explaining the Converted status
3. Then create dummy variables for categorical variables, ready to split the data into parts for testing and modeling
4. Train-test split was done with 70% train and 30% test split.
5. The model is then builts with logistic regression, refined with RFE to reduce the number of features down to 15.
6. Model is then evaluated with using Statmodels. Confusion matrix is employed and accuracy, sensitivity, precision and specificity are calculated with different thresholds to find the optimal cut-off value, which is 0.3 in this case.
7. Re-assess the model with the new cut-off value and test set
8. Find the important features in the model using coefficients.