Summary

In this assignment we aimed to find the right parameters that can efficiently predict whether a Prospective customer can be converted or not. We preceded by reading the dataset and listing the variables according to their data types, then checking the number of null values in each column. The aim was to precisely predict the conversions, so we deterred from any kind of missing value imputation, and dropped all the columns with more than 15% missing values. There were some unselected columns which essentially meant null values, hence some more cleaning was necessary. There were some variables that had only one level, so we dropped those as well.

The next step was to select only the necessary features. So we factorized all the categorical variables, and mapped the binary variable, so we could fit them into a mutual information regressor, we did this to check for non-linear correlations. Next we plotted the respective mi scores and set an arbitrary cut-off of 0.3. Variable having mi scores of 0.3 or above were selected. Next we ran Recursive Feature Elimination on the numeric variables, and rejected the Total Visits column. This might seem counter intuitive, but visualizations showed that most of the information available in this variable were captured in the remaining two numeric type variables (this analysis was not shown in the notebook).

Next we visualized the data based on the remaining variables. There were some interesting interplays among these variables that intuitively explained the behaviour of the Target people. We found that people who took the time to read the contents of the webpages had a higher chance of opting for the course, rather than those who skimmed through the whole website. We also noticed that most of the customers are coming through organic searches, or they are specifically searching for the website(Direct Traffic). Direct traffic can be a result of advertisements, but we could not be sure. We found that the remaining Lead sources are not leading to almost any traffic. We did the same for the other categorical columns, the detailed interplay of the variables have been presented in the ppt.

We factorized the categorical variables with more than 10 levels, and dummy encoded the remaining categorical types. At this point we split the data into train and test sets. We used statsmodels.api to perform the logistic regression. Next we calculated two different cut-offs for the model, one being the intersection for specificity, sensitivity, and accuracy for different cut-offs ranging from 0 to 1. The other one was found similarly based on precision and recall. We evaluated the model based on both these cut-offs. The model demonstrated 78% accuracy for both the cut-offs on test set.

We recommend using the second cut-off for most of the year, as it has 83% specificity (77% for the other cut-off), and a very low False Positive Rate. So the second cut-off greatly captures the customers who are not going to opt for the courses, and classifies very few targets as positives who are actually not going to opt for the course. In that sense this model is more conservative. Hence, the model saves a lot of time and effort for the Business Development team. During the period when the company will have the interns it is recommended to use the more aggressive cut-off.