Lead Score Case Study –Summary

- Initially we thoroughly inspected the data for missing values, found several columns with >30% and went forward with dropping then as they didn't have any particular insight. During EDA, we found that several categorical columns had highly skewed values some with constant values only and dropped those too. While preforming Sanity checks we found that the entry 'Select' was present in many columns which is the default value in a drop-down, as the customer didn't make any selection we went ahead and treated it as a missing value which wherever appropriate was handled by mode/median and where it's dominance was >30%, the columns was dropped. Few columns also had the presence of outliers which was handled by capping at 99th percentile and categories with <5% counts was included into a category called 'Others'. We then went ahead with performing EDA both uni-variant and bi-variant analysis where an interesting insight that working professional with referrals had the highest conversion rate but was approached less. Then prepared the data for model building by performing scaling the variables, test-train split, adding constant, creating dummies for categorical variables.

- First, built a model with all the variables which seemed to performed well but was suspected of high over-fitting (variance) the data. So went ahead with Recursive Feature Elimination with variables. Then it was found that 1 variable `Last Notable Activity_Unreachable` was insignificant hence dropped it. After re-building the model we found that all the variables were significant and having Variance Inflation Factor (VIF) <2 which indicates that there was no multi-collinearity present in between the independent variables which were – 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Do Not Email_Yes', 'What is your current occupation_Working Professional', 'Last Notable Activity_Had a Phone Conversation', and 'Last Notable Activity_SMS Sent'.

- By performing a random thresholding (cut-off) we found that the model produced an accuracy score of 79.0 % and AUC of 0.77 which is better than a randomly guessed model. In order to find the optimum threshold we plotted a line plot between Accuracy, Sensitivity and Specificity Plot and found that the threshold value of 0.348 was where all the 3 met as is the optimum. Then plotted the confusion matrix to understand the sensitivity, specificity, precision and f1-score of the model which had a values of 0.82, 0.78, 69.94 and 0.76 respectively. Lastly, we went ahead with performing the evaluation by testing the model on the test data-set where the model's sensitivity decreased by quite a margin but rest remained virtually same or increased by 1% or so. The AUC on the train set was 0.8 while on test set it 0.81, suggesting that the model is also able to generalize well. Finally, we calculated the Lead Score by multiplying the prediction by 100 and retaining the integer part only and label them as hot leads/cold leads based on a optimum cut-off point. We then went forward with presenting our insights and findings as well as showcasing the model with the help of a PPT which can aid in improvising the conversion rate.