Summary

Lead Scoring Case Study involved building a logistic regression model that predicts lead score for potential hot leads for an education company named X Education that sells online courses to industry professionals. We started the case study with 'Reading and Understanding the Dataset' of past customers with both sales and technical features. This step was done for a basic overview and inspection for further steps. We started with 'Data Pre-processing' step where the main focus was to clean, impute and drop data so as to build a dataframe with optimal information and less garbage values. We started by dropping sales team generated columns as these were extra information not relevant for our model, then we took care of missing values and skewed columns. The columns with so many categories were grouped into 'others' category to reduce extra variables. Lastly, outlier treatment was done by capping the necessary columns with percentile values. Then, 'Exploratory Data Analysis' was done to understand the data in detail by plotting various countplots, boxplots and pairplots. Some useful insights from this step were that our Leads originated from landing page submission followed by API and Google, Direct Traffic were two major sources of leads. Also, working professionals had significant conversion rate as well those customers with reference had higher chances of conversion.

*In the 'Data Preparation' step, tasks like Dummy variables for categorical columns, Train-test split and Scaling data using Standard Scaler were performed. With this step we proceeded to the 'Model Building' where we first started with 'Feature Selection' using RFE to select top 11 features. Once we got these desired features, we performed manual feature selection using P-value and VIF where any feature with greater and 0.05 P-value and/or greater than 5 VIF value was dropped. Since logistic regression gives its output as probability, it is required to set an optimal threshold in order to predict each customer as converted or not converted from the lead score (probability score * 100). So, in order to set the optimal threshold, we calculated Accuracy, Sensitivity and Specificity on a range of threshold from 0.0 to 0.9. An optimal threshold of 0.4 was found out from the line plot.*

Once our model was ready with 9 features: 'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Lead Origin_Lead Add Form', 'Lead Source_Google', 'Lead Source_Olark Chat', 'Lead Source_Organic Search', 'Do Not Email_Yes', 'Occupation_Working Professional' and an sub-optimal threshold of 0.36, we proceeded with model evaluation both training and testing data. We got an Accuracy, Recall and Precision of 77%, 78% and 75% respectively on training data and 77%, 77% and 78% on testing data. Finally, we multiplied 100 with the probabilities of each customer that our model predicted for obtaining a lead score on a scale of 0-100. From the distribution plot of lead score we could observe that there were roughly two groups with lead score such as 10-36 and 70-100 that we can label as cold and hot leads.