Lead Scoring Assignment – Summary

The problem statement was to identify the hot leads of the company X education from a given pool of leads data, so as to better channelize the efforts of the sales team and concentrate on high probable hot leads.

This business problem can be considered as Machine Learning classification problem, where the model predicts the probability of conversion for the lead (i.e. Lead Score) and also to help business identify the features contributing to lead conversion.

We have decided to use Logistic Regression model for the predicting the classes as it has good interpretability.

We initially identified the short comings in the data, performed the data cleaning, EDA Analysis and prepared the data such that it can be processed by the M/L Model. We have proceeded with feature elimination using automated approach using RFE and further eliminated features manually using p-value, and also ensure there is no multi collinearity using VIF metric. The model performance was further evaluated using Cross Validation and found the model performance was equally good on unseen data.

As the business expectation was to have lead conversion of 80% from the leads identified by the model, we ensured that the model recall score to be more than 80% for the optimal cutoff value using Precision_Recall_Curve.

Finally arrived at the features affecting the model prediction and playing vital role in Lead Score, and provided recommendation for the business to act up on based on the different business situations