SUMMARY REPORT ON LEAD SCORE CASE STUDY

The data used in the lead scoring consists of 37 features in total. Although a large number of features most of them were not so useful for model creation. First Stage of building a ML model is the Data Preparation and Data cleaning process. First of all all the redundant columns Prospect ID, Lead Number were deleted, as they provide no useful insights regarding customers leads. Then features only having a Single value were deleted, those were Receive More Updates About Our Courses(only no), Update me on Supply Chain Content,I agree to pay the amount through cheque ,Magazine,Get updates on DM Content. Then within data preparation process, we took care of the features with null values, as they features with higher percentage of null values heavily skew the model building process. First columns with null values greater than 45% were deleted. Second, categorical and numerical features, which were filled with default values present in the form were binned together in a special column or were merged with median or mode values of each column. Third, features with least number of null values, their rows were deleted. After this, we proceeded with data visualisation and gathered insightful observations from the plots. Those observations were categorically added in the notebook. Fourth, for the model building process features having only Yes/No values in their rows were converted to 0/1. After this we checked for correlation coefficients. Highly correlated variables were then dropped from the model building process. Then we moved towards the dummy variable creation step, In this process various categorical columns were converted to individual value columns, with each subsection of the column representing a category of the feature. After this initial process, we moved towards implementing logistic regression t4echnique for model building. First, the data were separated into test data and train data. Selected numerical columns in the leads dataset were standardized as a pre-requisite for model building. Then using sklearn library we implemented logistic regression on the train dataset. Then, we moved towards the feature selection process using RFE. RFE helped us in screening through a large number of features to a least number of independent features that best explained the behaviour of the train data set. Again the logistic regression technique was applied on the train dataset, but with only selected features given through RFE. After this a manual selection of features, took place where we rooted out the features based on their VIF rating and their p-values. This process was repeated until and unless we got a mixture of independent features which best explained the model.We moved on to finding the optimal threshold point, which gave us the boundary below above which a lead is taken to be converted or not. For this, we looked at the performance metrics of the model, where we got info regarding the model accuracy, specificity, sensitivity, precision and recall values. When we were satisfied with the evaluation metrics of the model. We tested the model on test data and assigned a lead score for each value.