Lead Score Summary

The lead score case study has been a holistic case study in the terms of o learning as we needed to put all the things together that we have learned so far.

So the very first step was to download the dataset from upgrad portal and the data dictionary to understand the features that were present in the dataset. The next thing was to start working on the dataset in the jupyter notebook and understand the data thoroughly. We loaded the dataset in the jupyter notebook and tried to see it, the shape, size and the number of records etc. The Data set had around 24 features and 92000 records. Once regular checks were done we went a bit deeper to understand out data well and checked for null values and their percentage in each columns. All the columns having more than 45% of missing values were drooped and the rest of the columns were given the required treatment as per the analysis. The next thing was to do the EDA and understand the data on more granular level so univariate analysis were performed to understand the features and their impact on the target that is the “Score” variable which signifies the probability of the individuals getting converted to a customer for the company. After doing the EDA and understanding the data the next step was to prepare the data for the model building. Since the score variable was binary in nature we decided to use logistic Regression model. We tried to convert all the variables to numeric type as categorical variable won’t be supported by the logistic model, so dummies were created for the categorical variables having more than two categories.

Once Data Preparation was done we moved to model development and imported the necessary libraries. The first thing in model development was to separate the target variable from the reaming variable and then split the data into test and train for model building and evaluation. 70:30 split was used to split the data into training and testing dataset. Next we build the model using the training data. For feature selection we used RFE with 20 variables as output to select the most important 20 variables from the variable pool and then manual elimination was used basis the P values and VIF values to reach the final 16 variables used in the final model. After variable selection optimal cut-off of 0.35 was chosen for the model by plotting the different metrics like specificity sensitivity and accuracy for different thresholds.

Finally, the Model was evaluated on both training and test data sets and showed similar results for different metrics that indicated our model is performing well and is stable. The Accuracy of model in training was 82% and 83% for test datasets. Other metrics were also evaluated and they showed a satisfactory result.

The case study helped in learning the concepts practically I learned so far theoretically. It also made me learn how to overcome Unseen challenges.