Summary Report for Lead Scoring Case Study

The Lead Scoring Case Study is based on an Education Company(X Education) that deals in selling online courses to industry professionals. The company collects a lot of data about their customers at different phases and has provided the same for model building.

The company currently faces a situation where the conversion rate is found to be at 30% or in simple words. The sales team is only able to convert 3 in 10 people, the reason to which lies in the lack of understanding of which customers have a higher probability of converting and which customers don't.

Thus, the model here is expected to assign a Lead Score to every customer so the sales team could look at the Lead Score and get an idea on weather a customer may or may not convert in the first place. This additional piece of information could help them create better strategies to convert people by differentiating a hot lead from a cold lead.

The process for building the model is inclusive of multiple steps that start from Data Understanding. Here, we analyse the data to look for any discrepancies like outliers, null-values, incorrect information and un-labelled information.

These operations are performed by making use of inbuilt functions like df.info(), df.describe(), df.head() and df.isnull().

Post analysis the rows and columns with discrepancies in data are either imputed or dropped depending on the statistical and business importance. In this case we have dropped the columns with more than 40% missing data as they will not be able to return any significant insights in our analysis. Some other columns with a lot of missing data have also been imputed with their modal values (most frequent values). Except for the above set of operations, there also exist some columns that have been provided by the sales team, these columns of data are not always available thus cannot be used in building the model.

The Skewness/Distribution test along with EDA is also performed to get an idea of all the operations that need to be performed for cleaning the data.

Once all the analysis steps have been completed the test_train model is built with a train size of 70% and test size of 30%. The features are scaled for an even distribution of data and are later selected by making use of feature_selection using RFE. The model is built using 15 features and run over multiple iterations after which the p-values of individual columns are compared to decide whether or not a column needs to be present in the final model. The Variance Inflation Factor(VIF) is also used to analyse the impact of different columns on the model.

Once all the operations have been performed a new predicted column is generated to predict if or not a customer would convert. On further confusion matrix based analysis it is noticed that the accuracy of the model comes out to be around 77.8% which is a decent score for making accurate predictions. The ROC Curve also helps understand the performance of the classification model. The optimal cut-off point is calculated by plotting accuracy, specificity and sensitivity. The cut-off point post plotting is found to be around 0.3 with an accuracy of around 79%. All the other factors like precision score and recall score are calculated before finally producing the same set of results on the test data. The resultant of which is a column that mentions the Lead Score of every lead in the data.