A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered

Summary Report of the Group Assignment

The most important part of this exercise was cleaning the data. It involved following steps

1. Identifying duplicate fields (for example Prospect ID & Lead Number have same information from our assignment perspective so keep only one)
2. Following fields has 99+ % have same information repeating for all the leads. So Remove them, they are not going to help us in this assignment.
    a. 'Digital Advertisement',
    b. 'Do Not Call',
    c. 'Get updates on DM Content',
    d. 'I agree to pay the amount through cheque',
    e. 'Magazine',
    f. 'Newspaper',
    g. 'Newspaper Article',
    h. 'Receive More Updates About Our Courses',
    i. 'Search',
    j. 'Through Recommendations',
    k. 'Update me on Supply Chain Content',
    l. 'X Education Forums'
3. Two Index fields (categorical variables) and 2 Score fields (numerical variables) are correlated. We cannot impute numerical fields using mean in this situation. We need to ensure the values in score corresponds to “high”, “medium”,”low” of index variables. For this we need to do imputation of Index variable first and then impute numerical variable using these index values.
4. There is huge percentage of null values in categorical variables which cannot be imputed using mode so we need to use logistic regression for prediction purpose.

Second important aspects of this assignment is to develop different model using various algorithms and perform the prediction using each. Compare which model is serving best our need. For that purpose we used ROC curve.

In our case, the best model is that which is coming from StatsModel algorithm.

After a best model is identified then we need to select a probability cut-off which will optimize the results and give us the accuracy of 80+%. In our case 65% probability is cut-off, it gives us the desired 80+% accuracy at the same time we get reasonable other metrics like recall, precision, error-rate.

- Accuracy: 0.91
- Recall: 0.92

- Precision: 0.86
- Specificity: 0.90
- Error Rate: 0.09
- FPR 0.10
- FNR: 0.08

Finally get the lead_score of each lead. We computed it by multiplying 100 with conversion probability of the lead.

Learnings Gathered

- Before modeling data cleaning is huge work
- We need to be creative in null value imputation. There are many ways of doing imputation, depending on variables type (date, number, bool, categorical, text), business importance of variable, number of null values we can decide imputation technique
- If there are long text columns we need to do feature engineering, so we need to create a new field from the text column. In this exercise we didn't feel the need of this.
- Choose which algorithm can be suitable for which purpose in our work. Every algorithm has specific purpose if we use algorithms properly we can develop a great model
- Compare results of multiple models using ROC and compare the result always on test data
- Always perform test/train split on the given dataset.
- Never scale the dataset before split
- Sometimes when concating, merging dataset there is no common field so use index.
- Visualize the distribution to know the data imbalance. Distplot and boxplot are the best tools for this.
- Write functions as much as possible, it will help in optimizing/improving code. It will improve modularity of the code and reusability of the code.
- Understanding and application of confusion matrix is super critical for interpreting the model results.
- Document the code as clearly as possible. It should not happen that we ourself not able to read the code after sometime.