Assignment Summary Report

1. Importing Dataset

- Import necessary libraries which are required for analysis and load the from .csv to a data frame.
- Take a backup copy of original data frame.

2. Data Understanding

- Verify various properties of data frame like Shape, Info, Summary Statistics etc.

3. Data Cleaning

- Replace “Select” values with np.NaN.
- Calculate the Column wise and Row wise null count.
- Calculate Column wise null value percentages and drop the columns which have high null value percentages and skewed columns.
- Drop the rows which have 5 or more null values.
- Impute the missing values of Categorical and Continuous variables with Mode and Median respectively.
- Bucket the levels of Categorical variables which have high number of levels.
- Calculate the percentage of rows retained after cleaning the data.
- Treat the outlies of Continuous variables using capping to make sure that we are not losing any information.

4. Exploratory Data Analysis

- Drop “Prospect ID” and “Lead Number” from the data set as they are not useful in modelling.
- Split the data into “conv” and “non_conv” where “Conv” will have records where Converted column value is 1 and “non_conv” will have records where Converted column value is 0.
- Perform Univariate and Bivariate analysis on Categorical and Continuous using both data sets to draw valuable insights.

5. Data Preparation

- Map the variable which have only “Yes” and “No” with 1’s and 0’s respectively.
- Create dummies and attach dummy variables to main data frame.
- Split the data frame in to train and test.

- Scale the Continuous variables in train data frame and create X_train and y_train.

6. Building a Model

- Build a model with all the variables.
- Perform feature selection using RFE to start model building with 15 variables.
- Calculate Confusion matrix and accuracy with an arbitrary cut-off using the model.
- Check the VIF's and p-values of the variables used in model.
- Drop 1 variable at a time which have high VIF or p-values and rebuild the model with remaining variables until we get VIF's and p-values of all variables under acceptable limits.

7. Model Evaluation

- Construct Confusion matrix and calculate accuracy using an arbitrary cut-off.
- Calculate other matrix like Sensitivity, Specificity, Precision, etc.

8. ROC Curve

- Plot the ROC curve using fpr and tpr to check whether our final model is a random model or not.

9. Finding Optimal Cut-off Point

- Classify the leads to Converted and Not Converted using various probability cut off's ranging from 0.0 to 0.9.
- Calculate Accuracy, Sensitivity and Specificity for predictions made using the various probability cut off's ranging from 0.0 to 0.9.
- Plot a graph of Accuracy, Sensitivity and Specificity to find the optimal cut off probability at which all 3 curves will intersect.
- From the plot, the optimal cut off probability is at around 0.35. Precision recall trade off plot was also plotted and the optimal cut off probability is at around 0.4.
- To make final predictions, 0.35 was used as cut probability.

10. Model Re-evaluation

- Calculate accuracy from the predictions made using optimal cut off.
- Calculate other matrix like Sensitivity, Specificity, Precision, etc.

11. Making Predictions on Test data

- Scale the Continuous variables in test data frame and create X_test and y_test.
- Make predictions using the final model and variables used in final model.
- Create a data frame using “Lead Number, Converted and Probability scores” and classify the leads using optimal cut off.
- Calculate accuracy from the predictions made using optimal cut off.
- Calculate other matrix like Sensitivity, Specificity.

Learnings:

- “Select” value in our data set means that Lead has not selected any value for that particular field. So, it is equivalent no null value.
- There are columns in data set where 1 value’s is clearly dominating the other values. These are called Skewed columns. We have to drop such columns as they might give biased results.
- Initially we have calculated the Accuracy, Sensitivity, Specificity etc. with 0.5 probability as cut off value which is completely arbitrary. As we increase the cut off probability, Specificity increases and Sensitivity decreases. So, we need to find the balance between them. To find the optimal cut off, roc curve is used.