Summary of Lead scoring Assignment
Name-Vikas Bahrtiya
Batch- DS-43

Steps Followed to complete the Assignment

1. Importing Libraries
2. Loading and reading the data
3. Understanding and cleaning the data
4. EDA Univariate and Bivariate analysis
5. Preparing data for modelling
6. Model Building
7. Evaluation of the Final model
8. Evaluation of the final Model
9. Checking other accuracy beyond simple accuracy
10. Plotting the ROc curve
11. Finding the otimal cutoff
12. Calculating the Precesion and recall
13. Making the prediction on test set
14. Final Conclussion.

1. **Importing Libraries**
   - All necessary libraries are loaded for numerical computation, visualization and plotting, for suppressing the warning, feature selection, model building and evaluation of the model.

2. **Loading and reading the data**

- Loaded the dataset
- Checked the shape and info
- Checked the stats of the numerical column using the describe function.

3. **Understanding and cleaning the data**

- Null value checked and removed the column having less than 20 percent null value.

- Select value also as part of the null value that also been considered.
- Using the value count checked the spreads of the column data showing single value has been removed as this will not train the model

4. EDA Univariate and Bivariate analysis

- Carried out univariate and Bi variate analysis.
- Completed the outliers treatment using the box plot.
- Checked the correlation using the heatmap

5. Preparing data for modelling

- Categorical column having the less value has been replaced with the others.
- Created the dummy variable of the categorical column.
- Splitted the dataset in train and test set.
- Applied standard scaler to the train dataset (fit_transform)
- -Applied standard scaler to test data(transform)

6. Model Building

- Added the constant in X train and fitted with the y train
- Applied this first model to whole dataset
- Checked the statistics of the model.

7. Feature Selection using RFE

- Used the logistic regression to checked the feature selection.
- Selected 15 features from the data.
- Applied (fitted) the stats model to these selected 15 features.
- Checked the VIF and P value from the summary.
- Repeated the feature until p value less than 0.05 and VIF less than 5.
- Finally got our fourth model good as per the stats value.

8. Evaluation of the final Model

- Used the default value 0.5 for the cut off and converted the predicted value in 0 and 1.
- Overall accuracy 0.79 and also printed the confusion matrix.

9. Checking other accuracy beyond simple accuracy

- Checked the other accuracy
- Sensitivity → 0.6625
- Specificity → 0.8747
- false positive rate → 0.125
- positive predictive value → 0.7634
- Negative predictive value → 0.8094

10. Plotting the ROc curve

- Plotted the Roc curve using the sklearn class metrics.roc_auc_score.
- Final curve between the converted value 0,1 and probability in float between 0 and 1.

11. Finding the optimal cutoff

- Created the column using the different dataframe at the interval of 0.1 from 0 to 0.9.
- Calculated the accuracy, sensitivity and specificity
- Plotted and taken the optimal cutoff value
- Optimal cutoff value obtained 0.35.
- Again checked the metrics as per the new cutoff value.
- Accuracy → 0.7914
- Senstivity → 0.7932
- Specificity → 0.7903
- false postive rate → 0.2096
- Positive predictive value → 0.6977
- Negative predictive value → 0.8623

12. Calculating the Precesion and recall

- From Sklearn.metrics imported precision_score, recall_score
- precision_score → 0.6977
- recall_score → 0.7932

13. Making the prediction on test set

- Added constant in the test data set.
- Predicted X test
- Mapped the prob value in 0 and 1 as per the cut off point 0.35
- Accuracy of the test data set → 0.7837
- Sensitivity of the test set → 0.7740
- Specificity for the test set → 0.7897

14. Final Conclusion.

- Final Model (res) res = logm4.fit()
- Can be converted to pickle file for further use
- Cut off probability 0.35
- Greater than 0.35 converted as lead
- Less than 0.35 will not converted as lead
- Accuracy of the train data 0.791
- Accuracy of the test data 0.784
- For increasing or decreasing number of the Lead Cut off prob can be adjusted
- We can also follow the lead score targeting from the top.