Diabetes Diagnostics Analysis

Approach Taken :-

Have followed the below steps while analysis of this project :-

Step 1 - Data Collection

In the this step we will be loading our input diabetes data in csv format into our system. As the provided dataset has no columns in the input datase, we will provide the column names separately.

Step 2 - Data Exploration & Data Preparation.

Observation from EDA

- Plasma Glucose Concentration for Diabetic Patients is significantly higher as compared to non diabetic ones. High levels of Q1,Q2, Q3 and upper whiskers along with IQR for diabetic patients signifies the same.

- No. of Times Pregnant & BMI is also high for Diabetic Patients.

- Age also plays quite relative role , as majority of the patients who are diabetic are relatively older. Q1, Q2, Q3 and upper whiskers for diabetic ones are higher than the not diabetic one, which signifies the roles of Age here.

- To some extent, skin Fold Thickness & blood pressure is also marginally high for Diabetic Patients.

- Lucky enough to have no missing values in the input dataset.

Step 3 – Creating the Train and Test Data sets for training the model on train data set and then test the model on Test Dataset.

Step 4 - Selection of Models and Model building.

Have used various ML algorithms using 10-fold cross validation type.

Step5 - Use models build in above Step to predict the outcome of employees left or not on Test dataset.

Conclusion:- One can see from the above stats that model performance that either of GBM (Accuracy 78.34 Kappa 49.68) Or LDA (Accuracy 77.42 Kappa 47.54) looks close the best models for our business. To decide further we have ploted the ROC for both these models and could see that the Area Under the ROC (AUC) is slightly larger for GBM as compared to LDA. This is quite obvious as the Accuracy, Sensitvity and Specificity for GBM is slightly greater than the LDA. hence we will consider GBM model for predicting the Diabetic conditions for the provided input dataset.