**QUESTION 1**

First, we import the necessary libraries that are required for the assignment and completion of task

Background pattern

Description automatically generated with low confidence

We then set the working directory



Reading the dataset and then checking whether the dataset has been read correctly or not. Also finding basic information about the dataset like the column names, number of rows and columns, information about each column, and checking if any row has a null value or not.

Graphical user interface, text, application

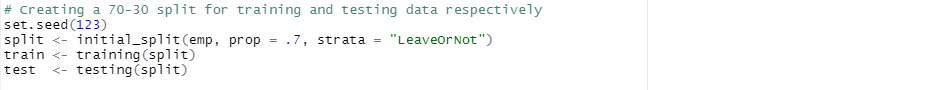
Description automatically generated

Converting the categorical string variables into categorical numerical variables

Text

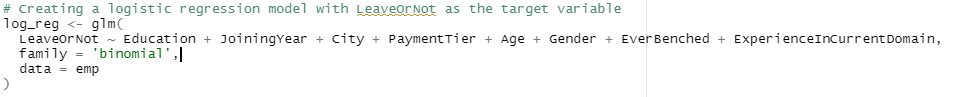
Description automatically generated

Creating a 70-30 train test split for further model building



**QUESTION 1(a) - Create a logistic regression model and explain the significant odds ratios in terms of LeaveOrNot.**

Creating a logistic regression model with LeaveOrNot as the target variable



Getting the summary of the logistic regression model



Text

Description automatically generated with medium confidence

Coefficients in exponential form



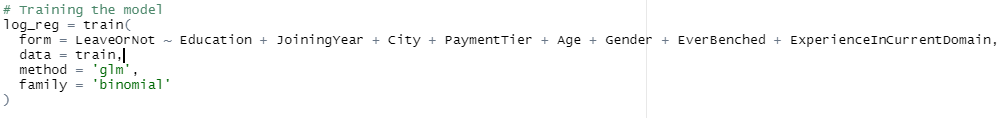
Table

Description automatically generated

We can clearly see the odds ratio of all the variables. The basic concept of odds ratio is that when the variable is greater than 1, it promotes the happening of outcome. When the odds ratio is less than 1, it prevents the happening of outcome. In this case Education, JoiningYear, Gener, EverBenched are promoting leaving of a candidate. On the other hand, City, PaymentTier, Age, ExperienceInCurrentDomain are preventing the leaving of a candidate.

**QUESTION 1(b) - Create a confusion matrix and explain how well the model is classifying the Leaves Company in 2 years events.**

Training the model



Predicting the dependent variable



Confusion matrix



Text

Description automatically generated with medium confidence

From the confusion matrix, we can conclude that accuracy of the model is 68.41%. This accuracy is low, and the model has further scope of improvement.

**QUESTION 1(c) - Create an ROC curve and calculate the c-statistic (auc). What does this mean about the model?**

Plotting the ROC curve

Text

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

Getting the AUC scores



Chart, line chart

Description automatically generated

The ROC curve talks about the trade off between sensitivity and specificity. In this case, the ROC curve is neither far from the 45-degree diagonal nor is near to it. Hence, from ROC curve we can conclude that the performance of the model is somewhat average.

**QUESTION 1(d) - What are the differences between the information in part a and part b?**

We create a binary classifier model that is a logistic regression model. This model helps in correctly classifying the binary outcome variable based on one or more independent variables. Part A has its outcome in the form of Odds Ratio. Odds ratios tell us how variables affect the outcome. If the odds ratio of a variable is greater than 1, the variable contributes to the likelihood of making the outcome happen. On the other hand, if the odds ratio of a variable is less than 1, then contributes to the likelihood of preventing the outcome from happening. In Part B, we create a confusion matrix. A confusion matrix is a matrix that describes the classification of the outcomes predicted by our model. The classifications are mainly classified as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The ratio of these in various combination gives us the performance metrics of the model. The metrics talks about precision, recall, f1-score, support, etc.

**QUESTION 1(e) - How does this model differ from the linear discriminate analysis you ran in Assignment 7?**

In Assignment we performed a LDA which stands for Linear Discriminant Analysis. Since LDA is linear, it is used in prediction of continuous or linear variables or classification of variables linearly. LDA looked for a particular linear combination of feature variables that best explained the target variable. After performing the LDA, we performed a cross validation.

Whereas in this assignment we performed and learned about logistic regression. Logistic regression is generally used when the outcome variable or a target variable is binary and categorical. Logistic regression predicted the binary outcome based on numerous feature variables with a regression at its core. Logistic regression being a machine learning algorithm required a large amount of data to train the regression model. No cross validation was performed in logistic regression.