Module 6 Live Coding Assignment : Logistic Regression and Applying GLM

Throughout this assignment, please remember to import all the necessary libraries at the beginning of every exercise.

Question 1

Read the dataset "churn.csv" and save it in a variable called "df". See the basic information about your data using the functions "head()", "info()", and "describe()", and the methods "dtypes" and "shape".

Question 2

Cast the variable "TotalCharges" to a float. Make sure to use the argument errors equal to 'coerce'. Delete rows and columns including null values, and drop the "CustomerID" column.

Replace relevant columns to integers. For "Gender", use 1 for "Male" and 0 for "Female". For the others, use 1 for "Yes" and 0 for "No".

Question 3

Perform an univariate analysis for all categorical and continuous variables. Plot their distributions using histograms in seaborn. Then use seaborn to create a "pairplot" for the columns "tenure", "MonthlyCharges" and "TotalCharges" with the hue equal to "Churn".

Question 4

Get and plot the correlation values between "Churn" and the remaining columns.

Question 5

Separate our dataframe into "X" and "y" values. "y" should be the "Churn" column in our dataset. Assign to "X" the remaining columns.

Don't forget to normalize to "X" in order to scale all values between 0 and 1.

Question 6

Use the RFE method to get the best 8 variables for predicting the Churn. Finally split the data set as train and test with %80-%20 ratio.

Question 7

Build a logistic regression model using the selected variables from above. Fit it with your training data and then see the predictions in your test data. Get the accuracy for your model and plot the ROC curve for it.