Final Project on "customer_churn" Dataset

Problem Statement –

You are the Data Scientist at a telecom company "Neo" whose customers are churning out to its competitors. You have to analyse the data of your company and find insights and stop your customers from churning out to other telecom companies.

Customer_churn Dataset:

The details regarding this 'customer_churn' dataset are present in the data dictionary

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
7590-VHVEG	Female	0	Yes	No	1	No	No phone service
5575-GNVDE	Male	0	No	No	34	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No
7795-CFOCW	Male	0	No	No	45	No	No phone service
9237-HQITU	Female	0	No	No	2	Yes	No
9305-CDSKC	Female	0	No	No	8	Yes	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes
6713-OKOMC	Female	0	No	No	10	No	No phone service
7892-POOKP	Female	0	Yes	No	28	Yes	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No

Lab Environment: R-Studio

Domain – Telecom

Tasks to be done:

A) Data Manipulation:

a. Extract the 5th column & store it in 'customer_5'
b. Extract the 15th column & store it in 'customer_15'
c. Extract all the male senior citizens whose Payment Method is Electronic check & store the result in 'senior_male_electronic'
d. Extract all those customers whose tenure is greater than 70 months or their Monthly charges is more than 100$ & store the result in 'customer_total_tenure'
e. Extract all the customers whose Contract is of two years, payment method is Mailed check & the value of Churn is 'Yes' & store the result in 'two_mail_yes'
f. Extract 333 random records from the customer_churn dataframe & store the result in 'customer_333'
g. Get the count of different levels from the 'Churn' column

B) Data Visualization:

a. Build a bar-plot for the 'InternetService' column:
    i. Set x-axis label to 'Categories of Internet Service'
    ii. Set y-axis label to 'Count of Categories'
    iii. Set the title of plot to be 'Distribution of Internet Service'
    iv. Set the color of the bars to be 'orange'

b. Build a histogram for the 'tenure' column:
    i. Set the number of bins to be 30
    ii. Set the color of the bins to be 'green'
    iii. Assign the title 'Distribution of tenure'
c. Build a scatter-plot between 'MonthlyCharges' & 'tenure'. Map 'MonthlyCharges' to the y-axis & 'tenure' to the 'x-axis':
    i. Assign the points a color of 'brown'
    ii. Set the x-axis label to 'Tenure of customer'
    iii. Set the y-axis label to 'Monthly Charges of customer'
    iv. Set the title to 'Tenure vs Monthly Charges'
d. Build a box-plot between 'tenure' & 'Contract'. Map 'tenure' on the y-axis & 'Contract' on the x-axis.

C) Linear Regression:

a. Build a simple linear model where dependent variable is 'MonthlyCharges' and independent variable is 'tenure'
    i. Divide the dataset into train and test sets in 70:30 ratio.
    ii. Build the model on train set and predict the values on test set
    iii. After predicting the values, find the root mean square error
    iv. Find out the error in prediction & store the result in 'error'
    v. Find the root mean square error

D) Logistic Regression:

a. Build a simple logistic regression model where dependent variable is 'Churn' & independent variable is 'MonthlyCharges'
    i. Divide the dataset in 65:35 ratio
    ii. Build the model on train set and predict the values on test set
    iii. Build the confusion matrix and get the accuracy score

b. Build a multiple logistic regression model where dependent variable is 'Churn' & independent variables are 'tenure' & 'MonthlyCharges'
    i. Divide the dataset in 80:20 ratio
    ii. Build the model on train set and predict the values on test set
    iii. Build the confusion matrix and get the accuracy score

E) Decision Tree:

a. Build a decision tree model where dependent variable is 'Churn' & independent variable is 'tenure'
    i. Divide the dataset in 80:20 ratio
    ii. Build the model on train set and predict the values on test set
    iii. Build the confusion matrix and calculate the accuracy

F) Random Forest:

a. Build a Random Forest model where dependent variable is 'Churn' & independent variables are 'tenure' and 'MonthlyCharges'
    i. Divide the dataset in 70:30 ratio
    ii. Build the model on train set and predict the values on test set
    iii. Build the confusion matrix and calculate the accuracy