AIML | MODULE PROJECT

Ensemble Techniques

TOTAL SCORE 60

- **DOMAIN:** Telecom
- **CONTEXT:** A telecom company wants to use their historical customer data and leverage machine learning to predict behaviour in an attempt to retain customers. The end goal is to develop focused customer retention programs
- **DATA DESCRIPTION:** Each row represents a customer, each column contains customer's attributes described on the column Metadata. The data set includes information about:
  - Customers who left within the last month – the column is called Churn
  - Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
  - Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
  - Demographic info about customers – gender, age range, and if they have partners and dependents

- **PROJECT OBJECTIVE:** The objective, as a data scientist hired by the telecom company, is to build a model that will help to identify the potential customers who have a higher probability to churn. This will help the company to understand the pain points and patterns of customer churn and will increase the focus on strategising customer retention.

- **STEPS AND TASK [60 Marks]:**

1. **Data Understanding & Exploration: [5 Marks]**

   A. Read 'TelcomCustomer-Churn_1.csv' as a DataFrame and assign it to a variable. [1 Mark]

   B. Read 'TelcomCustomer-Churn_2.csv' as a DataFrame and assign it to a variable. [1 Mark]

   C. Merge both the DataFrames on key 'customerID' to form a single DataFrame [2 Mark]

   D. Verify if all the columns are incorporated in the merged DataFrame by using simple comparison Operator in Python. [1 Marks]

2. **Data Cleaning & Analysis: [15 Marks]**

   A. Impute missing/unexpected values in the DataFrame. [2 Marks]

   B. Make sure all the variables with continuous values are of 'Float' type. [2 Marks]
   [For Example: MonthlyCharges, TotalCharges]

   C. Create a function that will accept a DataFrame as input and return pie-charts for all the appropriate Categorical features. Clearly show percentage distribution in the pie-chart. [4 Marks]

   D. Share insights for Q2.c. [2 Marks]

   E. Encode all the appropriate Categorical features with the best suitable approach. [2 Marks]

   F. Split the data into 80% train and 20% test. [1 Marks]

   G. Normalize/Standardize the data with the best suitable approach. [2 Marks]

3. **Model building and performance improvement : [40 Marks]**

   A. Train a model using Decision tree and check the performance of the model on train and test data (4 marks)

   B. Use grid search and improve the performance of the Decision tree model , check the performance of the model on train and test data , provide the differences observed in performance in Q3.a and Q3.b (5 marks)

   C. Train a model using Random forest and check the performance of the model on train and test data (4 marks)

   D. Use grid search and improve the performance of the Random tree model , check the performance of the model on train and test data , provide the differences observed in performance in Q3.c and Q3.d (5 marks)

   E. Train a model using Adaboost and check the performance of the model on train and test data (4 marks)

   F. Use grid search and improve the performance of the Adaboost model , check the performance of the model on train and test data , provide the differences observed in performance in Q3.e and Q3.f (5 marks)

   G. Train a model using GradientBoost and check the performance of the model on train and test data (4 marks)

Continued.....(Please turn to next page)

H. Use grid search and improve the performance of the GradientBoost model , check the performance of the model on train and test data , provide the differences observed in performance in Q3.g and Q3.h (5 marks)

I. Provide detailed analysis of the below steps (4 marks) :

(1) Compare the performance of each model in train stage and test stage

(2) Provide your observation on which model performed the best

(3) Provide your reasoning on why the model performed best

(4) Provide your final conclusion on your observation