Great Learning

AIML | MODULE PROJECT

Ensemble Techniques

TOTAL SCORE 60

General Instructions:

1. Submission of all the parts is expected in 1 notebook only

2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only

3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.

4. If output for any code cell is missing, 50% marks will be deducted.

5. Any kind of Plagiarism will lead to 0 (zero) Marks.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and

2. '.html' (Jupyter Notebook > File > Download as > HTML)

5 Marks will be deducted if submission in any of the formats is missing.

Part A - 30 Marks

- **DOMAIN:** Telecom
- **CONTEXT:** A telecom company wants to use their historical customer data to predict behaviour to retain customers. You can analyse all relevant customer data and 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:** To Build a model that will help to identify the potential customers who have a higher probability to churn. This helps the company to understand the pinpoints and patterns of customer churn and will increase the focus on strategizing customer retention.

- **STEPS AND TASK [30 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: [5 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 Improvement: [10 Marks]**

   A. Train a model using XGBoost. Also print best performing parameters along with train and test performance. [5 Marks]

   B. Improve performance of the XGBoost as much as possible. Also print best performing parameters along with train and test performance. [5 Marks]

Part B - 30 Marks

- **DOMAIN:** IT
- **CONTEXT:** The purpose is to build a machine learning pipeline that will work autonomously irrespective of Data and users can save efforts involved in building pipelines for each dataset.

- **PROJECT OBJECTIVE:** Build a machine learning pipeline that will run autonomously with the csv file and return best performing model.
- **STEPS AND TASK [30 Marks]:**
  1. Build a simple ML pipeline which will accept a single '.csv' file as input and return a trained base model that can be used for predictions. You can use 1 Dataset from Part 1 (single/merged).
  2. Create separate functions for various purposes.
  3. Various base models should be trained to select the best performing model.
  4. Pickle file should be saved for the best performing model.

     Include best coding practices in the code:
     - Modularization
     - Maintainability
     - Well commented code etc.

Please Note:

Here, if you need to perform some research to build a pipeline. If you could, very well done! If not, please follow below instructions:

1. Create separate function fo every step individually.
   For Example: Separate function to remove null values, separate function for normalization etc.
   On top of it, if you could build some rule based logic, you'll gain better experience.
2. Once you are done with building smaller functions, you can group similar functions into another function to proceed with.
   For Example: create a function 'preprocessing_' and call all the preprocessing related functions within that function.
3. Once done with this, Stack all the functions sequentially within 'main' function to conclude.
4. Here, knowledge and skills required are of Supervised Learning and Python module only.
5. By building function modules in pipelines, you will start gaining industry best practices as you go further in the AIML program else only marks are gained with traditional approach of programming.
6. If this project is solved by traditional approach, evaluation will be done out of 20 Marks. And if industry approach is followed successfully, bonus of 10 marks will awarded and evaluation will be done out of 30 Marks.