ML Assignment Problem - learning purpose

M Tech AIML WIPL (Birla Institute of Technology and Science, Pilani)

Scan to open on Studocu

Problem Statement

Predicting customer churn is a critical task for telecommunications companies aiming to retain their customers and reduce revenue loss. Customer churn refers to the phenomenon where customers discontinue using the company's services. By accurately predicting which customers are likely to churn, the company can take proactive measures to improve customer satisfaction and retention.

For this assignment, you are provided with a dataset containing various features related to customer demographics, account information, and service usage patterns. The dataset also includes random missing values to simulate real-world data issues.

Objective:

The objective is to build a machine learning classification model to predict whether a customer will churn (i.e., stop using the company's services). You will need to preprocess the data, handle missing values, perform feature engineering, build and evaluate classification models, and provide insights based on the model's performance.

1. **Import Libraries/Dataset**

    a. Download the dataset.
    b. Import the required libraries.

2. **Data Visualization and Exploration [1M]**

    a. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
    b. Provide appropriate data visualizations to get an insight about the dataset.
    c. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**

3. **Data Pre-processing and cleaning [2M]**

    a. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell.
    b. Apply appropriate feature engineering techniques. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.

4. **Model Building [5M]**

a. Split the dataset into training and test sets. **Answer without justification will not be awarded marks.** [1M]
    i. Train = 80 % Test = 20%
    ii. Also, try to split the dataset with different ratios of your choice.
b. Build models [4 M]
    i. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.

5. Performance Evaluation [2M]

a. Compare the performance of the Logistic Regression and Decision Tree models using appropriate evaluation metrics.
b. Provide insights into which model performs better and why. **Answer without justification will not be awarded marks.**