**Problem Statement**

**Telco Customer Churn:**

The objective of this assignment is to develop a predictive model to accurately determine the customer churn risk for a telecommunications company, leveraging a dataset containing information on over 7,000 customer accounts. The dataset includes attributes such as customer demographics (gender, age, family status), service usage (phone, internet, streaming), contract details, payment methods, and monthly charges. This project aims to apply advanced data analysis and visualization techniques to uncover patterns and relationships among these customer attributes, ultimately enabling the prediction of customer churn probability. The insights gained from this analysis will help improve customer retention strategies, allowing the telecommunications company to proactively identify and address the needs of at-risk customers, ultimately reducing churn rates and increasing customer loyalty.

**Dataset:** TelcoDataset

**Metadata:**

* **customerID -** Customer ID
* **gender -** Whether the customer is a male or a female
* **SeniorCitizen -** Whether the customer is a senior citizen or not (1, 0)
* **Partner -** Whether the customer has a partner or not (Yes, No)
* **Dependents -** Whether the customer has dependents or not (Yes, No)
* **Tenure -** Number of months the customer has stayed with the company
* **PhoneService -** Whether the customer has a phone service or not (Yes, No)
* **MultipleLines -** Whether the customer has multiple lines or not (Yes, No, No phone service)
* **InternetService -** Customer’s internet service provider (DSL, Fiber optic, No)
* **OnlineSecurity -** Whether the customer has online security or not (Yes, No, No internet service)
* **OnlineBackup -** Whether the customer has online backup or not (Yes, No, No internet service)
* **DeviceProtection -** Whether the customer has device protection or not (Yes, No, No internet service)
* **TechSupport -** Whether the customer has tech support or not (Yes, No, No internet service)
* **StreamingTV -** Whether the customer has streaming TV or not (Yes, No, No internet service)
* **StreamingMovies -** Whether the customer has streaming movies or not (Yes, No, No internet service)
* **Contract -**The contract term of the customer (Month-to-month, One year, Two year)
* **PaperlessBilling -** Whether the customer has paperless billing or not (Yes, No)
* **PaymentMethod -** The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* **MonthlyCharges -** The amount charged to the customer monthly
* **TotalCharges -** The total amount charged to the customer
* **Churn -** Whether the customer churned or not (Yes or No)

**1. Import Libraries/Dataset**

a. Download the dataset.

b. Import the required libraries.

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

a. Print at least 5 rows for sanity check to identify all the features present in the dataset and if the target matches with them.

b. Print the description and shape of the dataset.

c. Provide appropriate visualization to get an insight about the dataset.

**3.** 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.**

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

1. 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.
2. 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.

**5. 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 model using Logistic regression and decision tree [4 M]

i. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.

**6. 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.**

**Instructions for Assignment Evaluation**

1. Organize your code in separate sections for each task. Add comments to make the code readable.
2. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behavior.
3. Notebooks without output shall not be considered for evaluation.

***For clarifications, contact Vinayaka SP <spvinayaka@wilp.bits-pilani.ac.in>***