Computer Science Department
CS675 – Introduction to Data Science (CRN: 74028)
Fall 2023

Project #1 / Due 10-Oct-2023

The goal of this assignment is to understand the logic and methods of exploratory data analysis (EDA). The mode of analysis concerned with discovery, exploration, and empirically detecting phenomena in data. EDA has become the default pre-modeling step for every Machine Learning project engagement. Exploratory Data Analysis (EDA) is a way to investigate datasets and find preliminary information, insights, or uncover underlying patterns in the data. Instead of making assumptions, data can be processed in a systematic method to gain insights and make informed decisions.

Prior commencing your efforts on coding, you must install the following libraries:

- Pandas Profiling
- SweetViz

https://pandas-profiling.github.io/pandas-profiling/docs/master/index.html
https://pypi.org/project/sweetviz/

Perform an **Exploratory Data Analysis** (**EDA**) on **Customer Churn** data within the **Telecommunication** industry. Although there will be no need to build a model based on the data provided, you are asked to look for issues in the data and find correlation among the various variables in order to improve/lower customer churn predictions.

Churn rate is a critical metric of customer satisfaction. Low churn rates mean happy customers; high churn rates mean customers are leaving you. A small rate of monthly/quarterly churn compounds over time. 1% monthly churn quickly translates to almost 12% yearly churn.

Investigating the data should be done two-fold:

1. Manually by utilizing the classic (legacy) EDA libraries: **NumPy**, **Pandas, graph** libraries (MatPlotlib, Seaborn, Plotly), and Python's **Statsmodel** modules.
2. Generate 'html' reports by integrating **Pandas Profiling** and **SweetViz** Python libraries.

The analysis of the data should be focus on predicting the customer churn rate.
The data (**telco-customer-churn.csv**) is available for you to download.

The dataset has 7043 rows and 21 columns.

There are 17 categorical features:
CustomerID: Customer ID unique for each customer
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)
Dependent: Whether the customer has dependents or not (Yes, No)
PhoneService: Whether the customer has a phone service or not (Yes, No)
MultipeLines: 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 an 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 years)
PaperlessBilling: The contract term of the customer (Month-to-month, One year, Two years)
PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

Next, there are 3 numerical features:
Tenure: Number of months the customer has stayed with the company
MonthlyCharges: The amount charged to the customer monthly
TotalCharges: The total amount charged to the customer

Finally, there's a prediction feature:
Churn: Whether the customer churned or not (Yes or No)

NOTE: You do NOT build/select a model, you only perform deep-dive analysis on the data.

Write **Python** scripts in order to complete the following tasks along with their output. All work should be done and submitted in a single **Jupyter Notebook.**

1- Prep the data in order to be ready to be fed to a model.
    Look for missing, null, NaN records.
    Find outliers.
    Transform data – all entries should be numeric.
2- List all types of data, numeric, categorical, text.
3- Perform EDA on data.
    Present dependencies and correlations among the various features in the data.
    List the most important variables (Feature Importance) that will affect the target label.
4- Split the dataset into training and test datasets (80/20 ratio). Using SweetViz's 'compare' command contrast the training vs test datasets on the target ('churn')
5- State limitations/issues (if any) with the given dataset.