**Project Title:** Regression Analysis on bike sharing demands of Washington, D.C.

**Author:** Pradeep Sathyamurthy (DePaul University, Chicago)

**Guiding Professor:** Prof. Nandhini Gulasingam

**Dataset Name:** Kaggle's competition on Bike Sharing Dataset

- test.csv
- train.csv

**Dataset Description:**

- Capital Bike share program in Washington, D.C. is similar to the DIVVY program in City of Chicago
- Data generated by the IoT (network KIOSK) at each docking stations throughout the city attracts researchers because of the diversified information present as part of this dataset
- Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city
- This is one of the Univariate dataset used for Kaggle's competition

**URL Details:**

- We can obtain this dataset from Kaggle and as well from UCI data repository.
- Data present in UCI is more normalized and structured while dataset present in Kaggle are raw data as such generated from network KIOSK. We will use dataset shared by Kaggle:
    1. Kaggle: **https://www.kaggle.com/c/bike-sharing-demand/data**
    2. UCI: **https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset**

**Problem Description:**

- We are asked to *combine historical usage patterns with weather data in order to forecast bike rental demand* in the Capital Bike share program in Washington, D.C.

**Proposed Approach:**

| In – Scope Approaches | Shadow Scoping (will be tried to improve model if needed) |
|---|---|
| Data Split-up [in this case we might not need it] and Outlier Studies | Jacknife Regression |
| Exploratory Data Analysis | Logistic Regression |
| Hypothesis Testing to validate the data and its inference | Locally Estimated Scatter Plot Smoothing (LOESS) |
| Data Exploration, smoothing, transformation and data preparation | Multivariate Adaptive Regression Splines (MARS) |
| Linear Regression - Model Building, Model Assumption, Model Validation | Advance Data Analysis Methods if Data Demands |
| Residual, outlier and influential points Analysis | |
| Prediction with Test Data | |

*I will first try to build a model using linear regression methods studied as part of this course CSC-423. Model efficiency will be the primary focus; we will observe the maximum variance in rental demands of bicycle that will be explained by Linear Regression Model. If the Adjusted R-square of final model is less than 50%, we will study the limitations of Linear Regression in such case and try to improve the model with one advance data analysis method that is not in scope of CSC-423 to experience the need of Advance regression methods in model building*.

**Citation Request and Reference:**

*Fanaee-T, Hadi, and Gama, Joao,* Event labeling combining ensemble detectors and background knowledge, *Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg. (IEEE ref inside this are studied)*