Group 22: Anushka Tak and Chandan Manjunath

PROBLEM

- **Background**
- Bike Sharing Systems is one of the most popular commercial businesses, to rent bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city.
- **Problem Statement**
- The absolute utility of this commercial business needs methods to anticipate the demand in advance and be able to provide service to customer to maximize commercial profits in the market.
- **Objective**
- We aim to use data mining techniques to predict the total count of bike rentals on an hourly basis, learning from the market behavior.

SOLUTION DESIGN

- **Data Preprocessing/Exploration, Variable Selection**
- Has many missing values, which we replace with average values.
- We plotted Correlation Matrix to study predictor behavior, plotted several histograms and boxplots to further gain insights about the data behavior(outlier analysis, data distribution) for initial analysis.
- We will use PCA, AIC ,(Stepwise Regression) for variable selection in the final model.
- **Prediction/Classification/Time Series Forecasting/Unsupervised Learning**
- We intend to use various prediction algorithms and develop a model to fulfill our objective. We will implement Linear Regression, Random Forest and Gradient Boosting Regression and compare all of their performance.
- **Predictors/Outcomes**
- Our outcome variable is a count of bikes likely to be rented in a particular hour.

DATA

- **Data Origin**
- 2 year data from Bike rental Service in Washington D.C. called Capital Bike Share
- Approximately 11.4k records, 12 variables
- Source - *https://www.capitalbikeshare.com/system-data*
- **Key Attributes**
- From the Initital EDA, we find these to be most influential- Month, Season(1-spring,2- summer,3-fall,4-winter), Day, Hour, Temperature(in degree Celsius), Windspeed
- **Data Quality**
- There are a few variables that give redundant information(temp, atemp, holiday, workingday) and have missing values(windspeed) but overall the data quality is good enough for modelling.
- Drop casual and registered variable since they provide no relevant information.

DATA MINING

- **Techniques**
- Regression Analysis using three algorithms for the task- Linear Regression, Random Forest and Gradient Boosting Regression.

- **Validation**
- We will split the entire dataset into training dataset, validation dataset and test dataset **a**nd use tuning to come up with good models.

- **Performance**
- We will evaluate our models with the measure of RMSE values for each model. The minimum RMSE is the desired characteristic.