**IE 7300 PROJECT PROPOSAL**

***Topic:***

Predicting Bike Share Due to Weather Forecast

**Group 12:**

Vincenzo Coppola

Varun Vyas

Abhijeeta Gupta

**PROBLEM DEFINITION:**

Bike sharing systems have been implemented over the last few decades to provide a convenient and affordable way for residents to rent bikes. The system automates the entire bike rental process by allowing members or others to pick-up, rent, and return the bike with no person-to-person contact. Users can borrow a bike from one location in a city and return it to a different location miles away. There are currently (as of 2021) around 3000 bike-sharing systems in place globally with over 9 billion bikes offered [1]. At the time of this dataset’s creation (2011-2012) the numbers were only 500 systems offering roughly 500,000 bicycles [2].

This dataset shows its importance even in a city like Boston where one can see Blue Bike rides being utilized on a daily basis. The growth of bicycle share programs can truly be seen from the growth in these numbers and programs as aforementioned. With global warming a rising concern, this growth makes sense. Similarly, driving in large cities has many downfalls in addition to environmental concerns, including heavy traffic and causing health issues from a lack of physical exercise. Whatever the case, people want to ride bikes and the growing numbers prove that. That being said, it is important for bike share programs to keep up with demand which is clearly growing and to do so it is the help of data scientists’ ability to predict this demand.

This project aims to do just that. The data provided by the Laboratory of Artificial Intelligence and Decision Support (LIAAD) at the University of Porto in Portugal provides information regarding Capital Bike sharing programs which will aid in the project’s purpose as follows. The objective outlined herein is to predict bike rental count on an hourly or daily basis utilizing the weather and seasonal forecast (2011-2012) which will provide a model that can be implemented on a larger scale moving forward. To do so, several Machine learning techniques and algorithms including dimensionality reduction and multiple types of normalized and unpenalized regressions will be implemented.

**DATA SOURCES:**

Dataset: <https://archive-beta.ics.uci.edu/dataset/275/bike+sharing+dataset>

The provided dataset is split into two csv files one with hourly data and one with daily. The hourly data which is intended to be used contains 17379 instances and 16 feature attributes describing the hourly count of rental shares for bikes in the Capital bikeshare system for 2011 and 2012 utilizing weather and holiday information pulled from freemeteo [3] and the dchr govt [4].

**DATA DESCRIPTION:**

The following variables contain bike sharing attributes information:

* instant: record index
* dteday: date
* season: season (1:winter, 2:spring, 3:summer, 4:fall)
* yr: year (0: 2011, 1:2012)
* mnth: month (1 to 12)
* hr: hour (0 to 23)
* holiday: Whether the day is a holiday or not (extracted from [Web Link])
* weekday: Day of the week
* workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
* weathersit: Description of weather
  + 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  + 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  + 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  + 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
* temp: Normalized temperature in Celsius.
  + The values are derived via (t-t\_min)/(t\_maxt\_min), t\_min=-8, t\_max=+39 (only in hourly scale)
* atemp: Normalized feeling temperature in Celsius.
  + The values are derived via (tt\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale)
* hum: Normalized humidity. The values are divided to 100 (max)
* windspeed: Normalized wind speed. The values are divided to 67 (max)
* casual: Count of casual users
* registered: Count of registered users
* cnt: Count of total rental bikes including both casual and registered

**SOLVING METHODS:**

This project goal is to achieve significant prediction results through various regressions. Therefore, the data will be cleaned if need be, and dimensionality reduction techniques will be used including feature engineering and potentially code to implement PCA. Then, using the final prediction input x the model(s) will be trained and evaluated using a variety of metrics including rmse on a test sample of the data. For this project, at a minimum - the following regression techniques will be implemented to aid in predicting bike rental count hourly (potentially daily):

* Linear Regression
* Lasso Regression
* Ridge Regression

If time permits, the following techniques will also be coded:

* Principal Component Analysis (PCA)
* Decision Tree Regression

Using these a combination of these methods and evaluating upon best performance. The concept of regression in data science will be properly understood and bike shares for 2011-2012 Capital Bike Programs will be successfully predicted with future large-scale implementation possible.

**SOURCES:**

[1] <https://www.pbsc.com/blog/2021/10/the-meddin-bike-sharing-world-map>

[2] <https://archive-beta.ics.uci.edu/dataset/275/bike+sharing+dataset>

[3] <http://www.freemeteo.com>

[4] <http://dchr.dc.gov/page/holiday-schedule>