# IE 7300 PROJECT PROPOSAL

## *Topic:*

Predicting Bike Share Due to Weather Forecast

## Group 12:

Vincenzo Coppola

Varun Vyas

Abhijeeta Gupta

## PROBLEM DEFINITION:

Bike-sharing programs have become increasingly popular in recent years as they provide an affordable and convenient way for people to rent bikes without the need for face-to-face interaction. These systems automate the entire rental process, enabling users to borrow a bike from one location and return it to a different location in the same city or even miles away. As of 2021, there are approximately 3000 bike-sharing systems globally, offering over 9 billion bikes. This represents a significant growth since the creation of this dataset in 2011-2012, where there were only around 500 systems and 500,000 bicycles.

The dataset highlights the importance of bike-sharing programs, even in cities like Boston, where the use of Blue Bike rides is becoming more common. With concerns around global warming and the downsides of driving in large cities, such as heavy traffic and health issues, the growth of these programs makes sense. The increasing number of users proves that people want to ride bikes, and bike share programs must keep up with the demand. Therefore, the ability of data scientists to predict this demand using machine learning techniques is crucial.

This project aims to predict bike rental counts on an hourly or daily basis by utilizing weather and seasonal data from 2011-2012. The data provided by the Laboratory of Artificial Intelligence and Decision Support at the University of Porto in Portugal will aid in achieving this goal. The objective is to create a model that can be implemented on a larger scale, thus providing a more accurate prediction of bike rental demand. The implementation of several machine learning techniques and algorithms will be necessary to accomplish this objective.

## 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 17 feature attributes describing the hourly count of rental shares for bikes in the Capital bike-share 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:

| Attribute Name | Attribute Description |
|---|---|
| 1. instant | record index |
| 2. dteday | date |
| 3. season | season (1: winter, 2: spring, 3: summer, 4: fall) |
| 4. yr | year (0: 2011, 1:2012) |
| 5. mnth | month (1 to 12) |
| 6. hr | hour (0 to 23) |
| 7. holiday | Whether the day is a holiday |
| 8. weekday | Day of the week |
| 9. workingday | If the day is neither weekend nor a holiday is 1, otherwise is 0. |
| 10. weathersit | Description of weather:<br>1. Clear, Few clouds, partly cloudy, partly cloudy.<br>2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist<br>3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds<br>4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| 11. temp | Normalized temperature in Celsius:<br>1. The values are derived via (t-t_min)/(t_maxt_min), t_min=-8, t_max=+39 (only in hourly scale) |
| 12. atemp | Normalized feeling temperature in Celsius:<br>1. The values are derived via (tt_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale) |
| 13. hum | Normalized humidity. The values are divided into 100 (max) |
| 14. windspeed | Normalized wind speed. The values are divided into 67 (max) |
| 15. casual | Count of casual users |
| 16. registered | Count of registered users |
| 17. cnt | Count of total rental bikes including both casual and registered |

## SOLVING METHODS:

This project's 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 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