Data 583: Project Proposal

Ricky Heinrich & Vimaljeet Singh

2023-02-27

Overview

- The data set is taken from UCI machine learning repository and related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available at http://capitalbikeshare.com/system-data. The data was aggregated on two hourly and daily basis. Then the corresponding weather and seasonal information was added on to this data. Weather information are extracted from http://www.freemeteo.com. The data set contains two csv files, one for hourly data and one for daily data. The variables (columns) in each data set are the same except the missing "hr" column in daily data.
- 'day' csv file is 731 x 16 and that of the 'hour' file is 17379 x 17.
- The dataset was checked for any missing values (NAs) and both the files had no missing data in them.

Description of variables

- **instant**: record index
- **dteday**: date
- **season**: season (1: springer, 2: summer, 3: fall, 4: winter)
- **y**: year (0: 2011, 1:2012)
- **mnth**: month (1 to 12)
- **hr**: hour (0 to 23)
- **holiday**: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- **weekday**: day of the week
- **workingday**: 1, if day is neither weekend nor holiday, 0 otherwise
- **weathersit**:
  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 divided to 41 (max)
- **atemp**: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- **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

The variables 'temp', 'hum', 'windspeed' are not normally distributed as will be shown in EDA.

Underlying scientific processes that may affect the data

Underlying scientific processes that might have affected the data would be the occurrence of natural events like Hurricane Sandy on Oct 30, 2012. So, at instances where data might seem aberrant, it might be because

the count of bike users was affected by these natural events or anomalies. There should be no reason why data should vary for any reason other than this.

Probable questions to answer

Number	Questions
Question 1	Can we detect any anomalies or events that affect the rental count, such as holidays or weather events, using the data set?
Question 2	Can we predict the future demand for bike rentals based on past data and environmental factors?
Question 3	How does the demand for bike rentals differ between registered and casual users?
Question 4	What is the impact of seasonal changes on bike rental demand?
Question 5	How does the rental count of bikes vary with seasonal and environmental factors such as weather conditions, temperature, humidity, and wind speed?
Question 6	Can we identify any patterns in the hourly rental counts, such as spikes during commuting hours?