

Assignment : Regression

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.

Problem Statement

In this project, you are asked to combine historical usage patterns with weather data in order to forecast hourly bike rental demand.

Data

You are provided with following files:

1. train.csv : Use this dataset to train the model. This file contains all the weather related features as well as the target variable “count”. Train dataset is comprised of first 18 months.
2. test.csv : Use the trained model to predict the count of total rentals for each hour during the next 6 months.

Data Dictionary

Here is the description of all the variables :

Variable	Definition
datetime	hourly date + timestamp
season	Type of season (1 = spring, 2 = summer, 3 = fall, 4 = winter)
holiday	whether the day is considered a holiday
workingday	whether the day is neither a weekend nor holiday
weather	weather
temp	temperature in Celsius
atemp	"feels like" temperature in Celsius
humidity	relative humidity
windspeed	wind speed
casual	number of non-registered user rentals initiated
registered	number of registered user rentals initiated
count	number of total rentals

How good are your predictions?

Evaluation Metric

The Evaluation metric for this project is Root Mean Squared Logarithmic Error (RMSLE). The RMSLE is calculated as:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$

Where:

- n is the number of hours in the test set
- p_i is your predicted count
- a_i is the actual count
- $\log(x)$ is the natural logarithm.

Solution Checker

You can use solution_checker.xlsx to generate score (RMSLE) of your predictions.

This is an excel sheet where you are provided with the timestamp and you have to submit your predictions in the count column. Below are the steps to submit your predictions and generate score:

- a. Save the predictions on test.csv file in a new csv file.
- b. Open the generated csv file, copy the predictions and paste them in the count column of solution_checker.xlsx file.
- c. Your score will be generated automatically and will be shown in **Your Score** column

You can also check out the baseline Python Notebook provided to get started.