Seoul Bike Rental

Hanin Lutfi Falatah

Abstract

The goal of this project was to use regression models to predict the Rental bikes in Seoul in order to have the bike count required at each hour for the stable supply of rental bikes. I worked with data provided by [UCI](https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand), leveraging plotting and with regression model to achieve promising results for predicting the rental count.

Design

This Data originates from the [Public data for all Seoul citizens](http://data.seoul.go.kr/)

Data

The dataset contains 8,760 bike rental record with 14 features for each. A few feature highlights include Date, Hour which represent hour of Day, Rented Bike count number of bikes rented hourly, Hourly Weather Conditions which include (Temperature, Humidity, Windspeed, Visibility, Dew point temperature, Solar radiation, Rainfall and Snowfall), Seasons, and Holiday. Functional.

Algorithms

DATA ENGINEERING

1. Renaming the columns for easy Data handling
2. Sampling the data
    a. The holiday column to be 1 for Holiday and 0 for No Holiday.
    b. Functioning Day column, no to be 0 and 1 for yes.
    c. Seasons column, to be ‘Spring' to be 1, ‘Summer' to be 2, 'Autumn' to be 3, and ‘Winter' to be 4.

MODEL

Support Vector Regression, Linear Regression, Random Forest Regressor, Kernel Ridge Regression, Gradient Boosting Regression, and Elastic Net Regression. Random Forest Regressor is with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice in model.

MODEL EVALUATION

Training set has 7008 samples bike rental record which means that the data is spilt 80/20.

and all scores reported below were calculated with 5-fold cross validation on the training portion only. R squared on the 20% holdout were limited to the very end, so this split was only used, and scores seen just once. The official metric for the Data was R squared.

THE RANDOM FOREST 5-FOLD CV SCORES:

R Squared: 0.8864458102890593

The random forest final training score:

R Squared: 0.9857087992640947

RMSE: 77.0867597510915

The random forest testing score:

R Squared: 0.8769467374613737

RMSE: 226.4280876447692

Tools

- NumPy

- Pandas

- Scikit-learn

- Matplotlib

- Seaborn

Communication

In addition to the slides and visuals presented, [Seoul Bike Rental]() .