Predicting Rental Demand for Public Bike Sharing – A Case Study of the Seoul Bike Sharing System

ANA630: Advanced Analytic Applications

By Magnus Aghe

Abstract

Public bike sharing is becoming an increasing transportation feature of major cities across the world with over 3, 000 cities now offering the service as at 2022 (O'Sullivan, 2022). Because of the high demand for this service, there is a need to make the rental bike available and accessible to the public at the right time thereby reducing waiting time. This research paper attempts to predict the hourly rental demand for public bikes using the Seoul Bike Sharing System by evaluating various machine learning regression models namely Linear Regression, Polynomial Regression, Decision Tree Regression, Random Forest Regression, Bagging Regressor, and Stacking Regressor. The Seoul Bike Sharing Demand dataset covers 12 months of data from December 1, 2017, to November 30, 2018, and was obtained from the UC Irvine machine learning data repository website. Several evaluation indices such as R-squared (R^2), cross validation accuracy score and mean squared error (MSE) were used to measure the prediction performance of the regression models.

Introduction

Today, major cities across the world have adopted public bike sharing – a public bicycle scheme, to help reduce traffic congestion, air pollution, and alleviate the economic impact of rising oil prices, amongst others. The concept of bike sharing has roots from the mid-1960s Amsterdam where Provo, a counterculture movement launched a free bike sharing program called the White Bicycle Plan, to fight against traffic congestion, pedestrian deaths, and air pollution caused by vehicles on the roads of Amsterdam (O'Sullivan, 2022). The idea never went away with Amsterdam launching a more formal bike-sharing system in 1998, which introduced docks, three years after Copenhagen introduced theirs.

Seoul's public bike sharing system, called Ttareungyi, was set up in 2015, 15 years after the city's first public bicycle rental service started. Inspired by Canadian city, Montreal's modern bike rental called Bixi, this new generation bike sharing system took off. Ttareungyi is an unmanned rental system that can conveniently be used anywhere, anytime, by anyone. The system was designed to improve citizens' health by enabling the use of bicycles in daily life, achieve a national vision of "Low Carbon, Green Growth", and reduce CO_2 emissions by using bicycles as an alternative form of transportation (Seoul Metropolitan Government).

Seoul Bike has become one of the city's most popular public transport systems, garnering over 100 million trips since launching in December 2015. With over 25,000 bicycles in more than 2,000 rental bike stations located near substations and in heavily crowded areas, the demand for rental of Seoul Bike has soared, and availability and accessibility of rental bikes to the public at the right time whilst reducing waiting times, has become a major concern.

Problem Statement

It is necessary to be able to predict accurately the number of prospective renters of public bikes per time in any given area within the city where Seoul Public Bike operates. It can be disappointing for a prospective rider to try to rent a bike and find out that there is none available at the location of their choice at the time they want it. Knowing the bike count required each hour will help optimal allocation of bikes to locations where they are needed most.

Objective

The objective of this study is to build a predictive model that would predict the required bike count at each hour, for the stable supply of rental bikes to meet demand.

Description of the Dataset

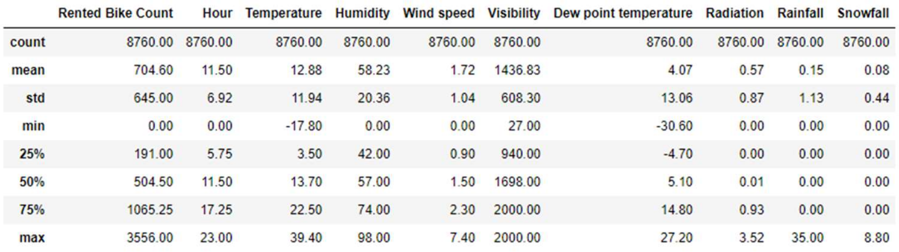
The Seoul Bike Sharing Demand dataset is used in this study. This dataset covers 12 months of data from December 1, 2017, to November 30, 2018, and was obtained from the UC Irvine machine learning data repository website. The dataset has 8760 rows and 14 attributes. It contains weather information – 8 attributes (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar Radiation, Snowfall, Rainfall), and other attributes like Date, Rented Bike Count, Hour of the day, Seasons of the year, Holiday, and Functioning Day. 10 attributes are numeric, 3 are categorical, and 1 is a date attribute.

Understanding the Data

Upon loading the data, we see that there are no missing or null values for all 14 attributes. Of the 10 numeric attributes, 6 are float data types and 4 are integer data types. Date is a date data type. The three categorical data are in object data type but converted to category data type. These are Seasons, Holiday, and Functioning Day. ‘Seasons’ has 4 categories namely Autumn, Winter, Spring and Summer. Holiday has 2 categories, Holiday/Non-holiday and Functional Day has 2 categories, Functional Day/Non-functional Day.

Summary Statistics

The mean Rented Bike Count per hour is 704.6 with a minimum of 0, a maximum of 3566 and a median of 504.5. The mean temperature is 12.88 °C, with a minimum of 0, a maximum of 39.4 °C and a median of 13.7 °C. Solar radiation has a mean of 0.57 MJ/m2, with a minimum of 0, a maximum of 3.52 MJ/m2 and a median of 0.01 MJ/m2. Below is a table of the summary statistics of the numerical attributes.

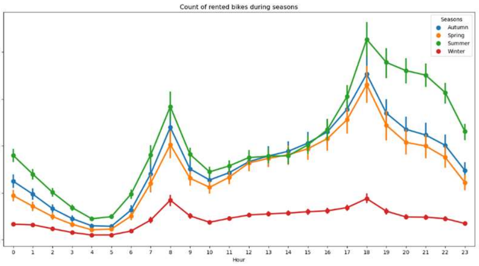


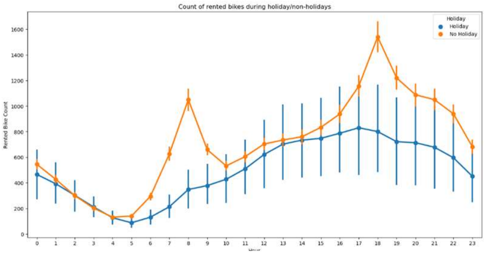
	Rented Bike Count	Hour	Temperature	Humidity	Wind speed	Visibility	Dew point temperature	Radiation	Rainfall	Snowfall
count	8760.00	8760.00	8760.00	8760.00	8760.00	8760.00	8760.00	8760.00	8760.00	8760.00
mean	704.60	11.50	12.88	58.23	1.72	1436.83	4.07	0.57	0.15	0.08
std	645.00	6.92	11.94	20.36	1.04	608.30	13.06	0.87	1.13	0.44
min	0.00	0.00	-17.80	0.00	0.00	27.00	-30.60	0.00	0.00	0.00
25%	191.00	5.75	3.50	42.00	0.90	940.00	-4.70	0.00	0.00	0.00
50%	504.50	11.50	13.70	57.00	1.50	1698.00	5.10	0.01	0.00	0.00
75%	1065.25	17.25	22.50	74.00	2.30	2000.00	14.80	0.93	0.00	0.00
max	3556.00	23.00	39.40	98.00	7.40	2000.00	27.20	3.52	35.00	8.80

Visualizing the Data

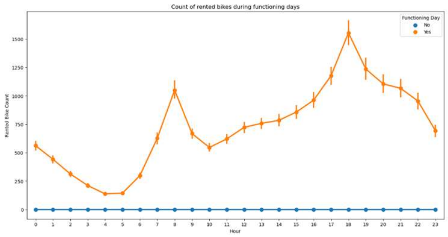
Categorical Variables

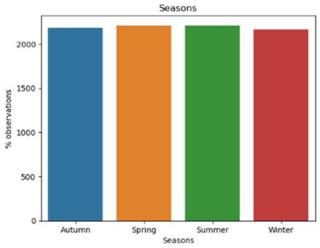
Across all four seasons we observe peak bike rental demands at 18:00 hours for any given day, with the demand much higher during the summer and lowest across peaks during the winter. Spring, Summer and Autumn show similar graph pattern unlike Winter. Between 5am and 8am there is a steady hourly increase in demand for rental bikes which peaks at 8am, suggesting that is the time most people go to work. This demand then steadily decreases till 10am, before gradually ascending from 2pm (14:00 hours) and reaching a peak at 6pm (18:00 hours), suggesting rush hour period with most people trying to commute home from work.





Also, there seems to be more demand for rented bikes during non-holidays than during holidays. There appears to be negligible demand for rented bikes across non-functioning days. Almost all demand is experienced during functioning days.

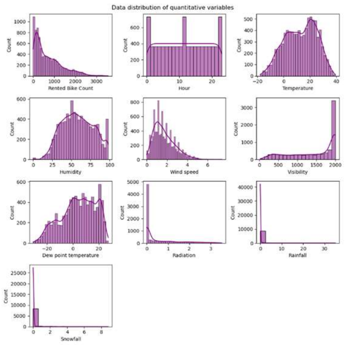


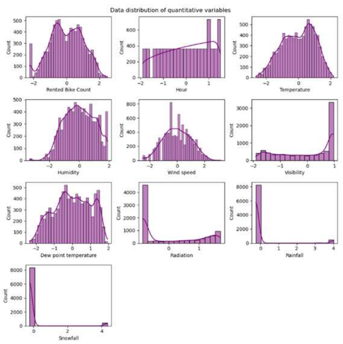


The spread of number of observations in the dataset is uniform across seasons. There were slightly more Spring and Summer days than Autumn and Winter.

Quantitative Variables

From the plotted histograms of the numeric attributes below, we can see that both positively and negatively skewed distributions exist in the dataset. Since most of the predictor variables are not normally distributed, we have to transform the variables.

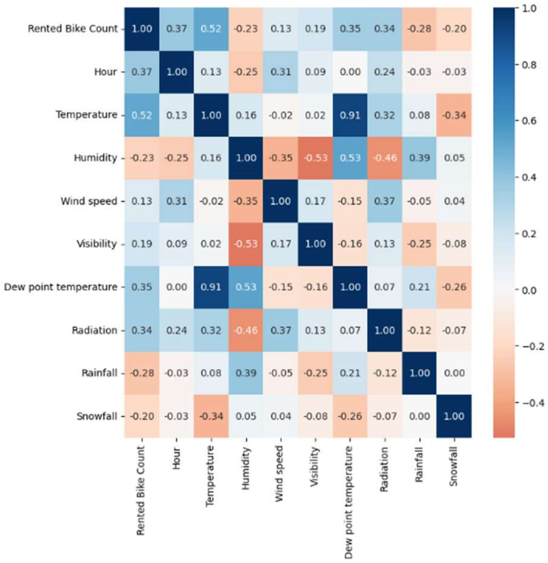




Upon applying power transformation, the distribution plots now look more symmetrical as seen below.

Correlation Analysis

From the correlation analysis, we see that Temperature is the most correlated with Rented Bike Count. Hour, Dew point temperature and Radiation are also quite strongly correlated with Rented Bike Count. Amongst the predictor variables, we see high collinearity between Temperature and Dew point temperature (0.91). This suggests we might remove Dew point temperature from the list of predictor variables.



Data Preparation

Coding Dummy Variables and Identifying Input Variables

Feature selection methods are intended to reduce the number of input variables to the most useful to a model in order to predict the target variable. Using variance inflation factor as our feature selection technique, we see that Dew point temperature is the most highly correlated, we remove it from our numerical input variables. The remaining numerical variables have variance inflation scores between 1 and 3, which shows that they are moderately correlated, and are thus good predictor variables. We then plot a feature importance graph of our predictor variables. Temperature is the most important predictor at 26.02% then Functioning Day (14.88%), Season's Winter (14.02%), Hour (11.2%), Solar Radiation (9.19%), Season's Summer (7.48%), Rainfall (6.08%), Humidity (3.79%). These 8 variables account for 92.66% of the total feature importance. In total, we have 13 predictor variables to be used in our predictive model.

Building the Model: Machine Learning Methods

First, we split our dataset into two parts 70% training set, and 30% test set. Then we normalize or scale our data. We then build models using Linear Regression, Polynomial Regression, Decision Tree Regression, Random Forest Regression, Bagging Regressor, and Stacking Regressor.

Linear Regression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

A **polynomial regression** model is a machine learning model that can capture non-linear relationships between variables by fitting a non-linear regression line, which may not be possible with simple linear regression. It is used when linear regression models may not adequately capture the complexity of the relationship.

Decision trees are used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.

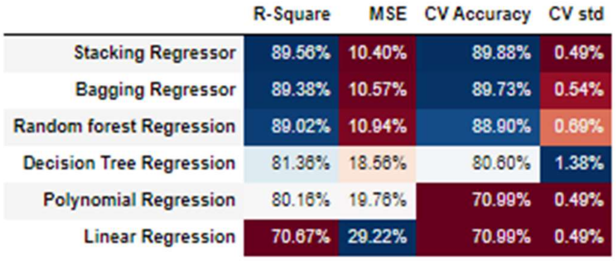
Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

Stacking regressions is a method for forming linear combinations of different predictors to give improved prediction accuracy. The idea is to use cross-validation data and least squares under non-negativity constraints to determine the coefficients in the combination.

Results: Model Evaluation

Performance measures for regression used in this study include R-square, Mean Square Error, and K-Cross Validation. Cross-validation is a technique in which we train our model; using the subset of the dataset and then evaluate using the complementary subset of the dataset.



	R-Square	MSE	CV Accuracy	CV std
Stacking Regressor	89.56%	10.40%	89.88%	0.49%
Bagging Regressor	89.38%	10.57%	89.73%	0.54%
Random forest Regression	89.02%	10.94%	88.90%	0.69%
Decision Tree Regression	81.36%	18.56%	80.60%	1.38%
Polynomial Regression	80.16%	19.76%	70.99%	0.49%
Linear Regression	70.67%	29.22%	70.99%	0.49%

From lowest to highest in terms of prediction accuracy, the linear regression model had an R-square of 70.67%, a mean square error of 29.22%, cross validation accuracy of 70.99% and CV standard deviation of 0.49%. Polynomial regression model had an R-square of 80.16%, MSE of 19.76%, CV accuracy of 70.99% and a CV deviation of 0.49%. The decision tree regression model had an R-square of 81.36%, MSE of 18.56%, cross validation accuracy of 80.60% and a CV deviation of 1.38%.

Random forest regression model had an R-square of 89.02%, MSE of 10.94%, CV accuracy of 88.90% and a CV deviation of 0.69%. Bagging regressor model had a better R-square of 89.38%, MSE of 10.57%, CV accuracy of 89.73% and a CV deviation of 0.54%. Stacking regressor, the best model, had an R-square of 89.56%, MSE of 10.40%, CV accuracy of 89.88% and a CV deviation of 0.49%.

Conclusion

This study focused on predicting the hourly rental demand for public bikes using the Seoul Bike Sharing System Dataset. Six machine learning regression methods were applied namely, Linear Regression, Polynomial Regression, Decision Tree Regression, Random Forest Regression, Bagging Regressor, and Stacking Regressor. In my findings, stacking regressor, bagging regressor and random forest regression were the three models that stood out with the best R-square and CV accuracy scores, as well as the least mean squared error and least CV standard deviation. Stacking regressor topped with the overall best model with r-square of 89.56%, CV accuracy of 89.88%, MSE of 10.4% and CV standard deviation of 0.49%.

References

UC Irvine Machine Learning Repository (2020). Seoul Bike Sharing Demand. http://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

Seoul Metropolitan Government (2021). 2019 Safe convenient people-centered Seoul Transportation., pp. 31-34. https://english.seoul.go.kr/wp-content/uploads/2021/05/2019-Safe-convenient-people-centered-Seoul-Transportation.pdf

Seoul Metropolitan Government (Undated). Seoul Public Bike. https://english.seoul.go.kr/service/movement/seoul-public-bike/1-seoul-public-bike/

O'Sullivan, F., (2022). The Radical Roots of Bikesharing. https://getpocket.com/explore/item/the-radical-roots-of-bikesharing?utm_source=pocket-newtab

Brianna, X., (2019). How to Use the Seoul Bike Rental System https://10mag.com/seoul-bike-rental/