Seoul Bike Sharing Demand: An Analysis on the Effects of Weather Data and Holiday Information on Bike Sharing Demands in Urban Cities using Data from Seoul, South Korea December 2017 – November 2018

Amy Lovas

Hannah Abraham

Kyle Smith

Sagar Goswami

OR-568, Applied Predictive Analytics
George Mason University
Dr. Vadim Sokolov

Abstract

Bike sharing has become a popular form of ground transportation in urban environments. In recent years, society has advocated for less traffic congestion and reduced vehicle emissions, especially in metropolitan areas and densely populated cities. During the COVID-19 Pandemic, the demand for bike sharing skyrocketed as people were less inclined to use public transportation, such as taxis or buses, due to the higher risk of infection in enclosed spaces. People have turned to bike sharing as the most viable option for ground transportation as the safest option to get to their destination. This is especially for travelers and sightseers that do not have a personal vehicle and do not want to waste precious time walking from site to site. Hence, for bike-sharing companies to meet the supply of bikes to the public, it is crucial to be able to predict bike demands within the operating town or city to meet demands and optimize company profit. By studying the data collected in Seoul, South Korea, a densely populated city where bike-sharing was already a well-established form of transportation before COVID-19, we may be able to analyze the demand trends and the relationship with weather conditions and holidays to provide insights to other bike-sharing companies. This insight will help companies ensure the proper number of bikes are available during peak demand times, decreasing wait times or walking distance to the nearest bike. Additionally, bike-sharing companies may use this information to determine a scaled price rate throughout the day, week, or season dependent on demand and bike availability.

Keywords: Bike-sharing, Bike Count, Weather, Season, Holiday, Hour

Introduction

With the information provided in the *Bike Sharing Demand, 2017-2018* dataset, we will use the given information to determine the relationship between weather conditions, seasons, holidays, and other contributing factors and bike demand in urban and metropolitan cities.

The dataset, sourced from the UCI Machine Learning Repository, contains 8760 observations of bike sharing demand from November 2017 through December 2018 in Seoul, South Korea. The dataset contains 16 independent variables and one dependent variable, bike count, provided below:

Variable	Data Type
Date: DD/MM/YYYY	Date
Bike Count by Hour	Integer
Hour of Day: 24 Hours	Integer
Temperature: Celsius	Float
Humidity: Percent	Integer
Windspeed: m/s	Float
Visibility: 10m	Integer
Dew Point: Celsius	Float
Solar Radiation: MJ/m2	Integer
Rainfall: mm	Float
Snowfall: cm	Float
Season	Categorical
Holiday	Binary
Functional Day	Binary
Day of Month	Integer
Day of Year	Integer
Month	Integer

Table 1: Table of Dataset Variables.

Logically, we predict that the demand for bike sharing would be greater when the weather is more desirable, that seasons have an effect on bike-sharing demand, and that there is a relationship between demand and holidays.

To test our intuitions and meet our project's objective to predict anticipated demand and plan a resource allocation strategy, we formulated four key questions to assist bike-sharing companies in demand predictions.

1) What features can bike share companies use to create a dynamic bike demand model and, potentially, a dynamic price rate that fits with the predicted bike-sharing demands.
2) What are the bike-sharing behaviors among the public in Seoul?
3) Is bike sharing more popular during the holidays?
4) Does weather effect the demand for bike sharing?

Methods

Data Cleaning

To ensure the dataset was suitable for fitting predictive models and hypothesis testing, it first needed to be cleaned. The dataset does not have any empty rows or missing categories, making the entire dataset complete. Seasons, originally entered as text, were transformed into categorical or factored variables. Holidays and days, referred to as functional days in the dataset, were transformed into binary variables. Finally, in anticipation of future transformations necessary for model fitting, any bike counts that were zero were changed to a very small non-zero number, 0.000001. This change to the bike count variable does not change the outcome of the predictive models in that it is small and incomprehensible in terms of "bikes rented per hour". However, this minuscule change will allow for future transformations that are possibly required for predictive modeling that we cannot perform with a dependent variable equal to zero.

Initial Variable Analysis

Initial review of summary statistics and data correlation shows that several of our variables are correlated, which makes sense in terms of weather variables.

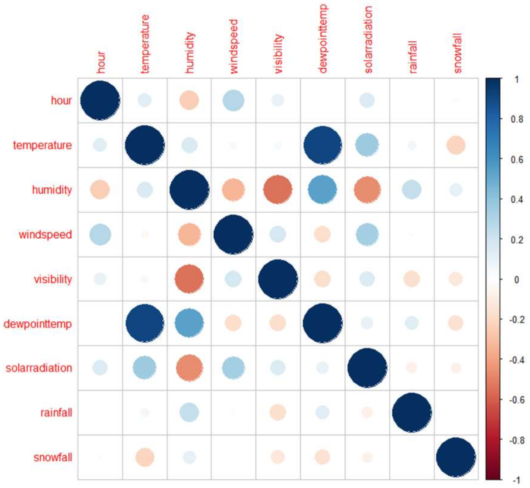


Figure 1: Correlation Plot of Weather Variables.

We see that there is a high correlation between dewpoint temperature and temperature as well as dewpoint temperature and humidity. There is also a large negative correlation between visibility and humidity as well as solar radiation and humidity.

Bike rentals by month shows that bike rental counts are lowest during the months of December, January, and February and then steadily increase to the peak demand in June. Demands then decrease and level off between July and October before rapidly declining again in November. A quick visual comparison to a daily temperature graph throughout the year shows that the bike rental count appears to correlate and follow the same trends as daily temperature.

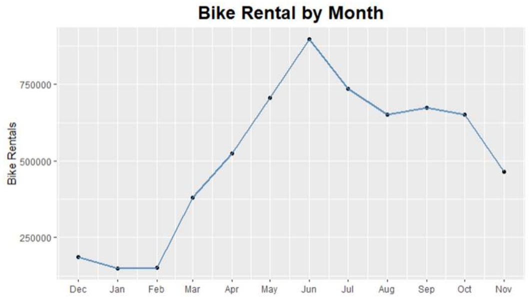


Figure 2: Bike Rental Count by Month December 2017 – November 2018.

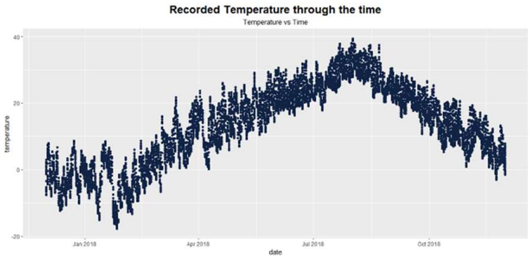


Figure 3: Temperature Recordings December 2017 – November 2018.

Using the seasonal categories provided by the dataset obviously reflects this same trend with winter having the lowest bike-sharing demand and summer having the largest demand.

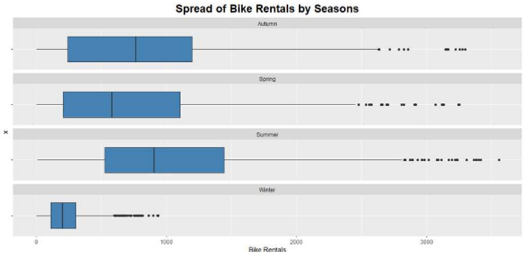


Figure 4: Bike-Sharing Demand by Seasons 2017 – 2018.

Predictive Modeling

Linear Regression

In our preliminary attempts to determine which features can be used to build a predictive model for bike demands, we created a linear regression model using all provided variables. This initial model using raw data provided a low adjusted R-squared value of only 0.5545. Assessing this initial model for our data showed that transformation was required to meet the assumption of normality of error for linear regression modeling.

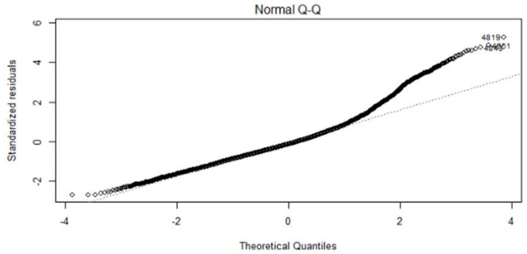


Figure 5: Quantile Plot of Raw Data Showing Lack of Error Normality.

Applying a log transformation to our data improved the normality of error as well as the adjusted R-squared value, improving to 0.9631.

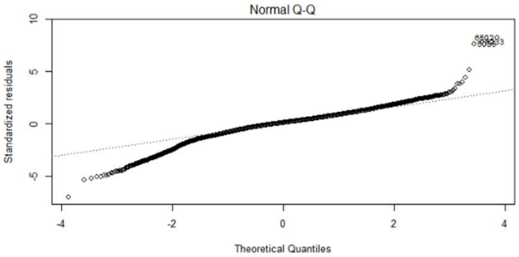


Figure 6: Quantile Plot of the Logarithmic Model Residuals Showing Normal Distribution of Errors.

Following the initial linear regression analysis, we continued with two feature selection methods to determine if any variables could be eliminated to simplify or improve the model's prediction capabilities. The best subset selection using both Mallow's Cp and Bayesian Information Criteria shows that a model containing 13 variables will produce the best prediction results. The adjusted R-squared variable selection method recommends including 17 variables in the model. We can further analyze the best number of variables suggested above using cross validation. With k-folds equal to five in the cross-validation method, we find that 15 variables is the best number of variables to include in our prediction model, this includes the categorical variables for seasons. We will remove the variables for visibility and dewpoint. Additionally, as predicted due to the correlation with other variables, the variables month, day of the year, and day of the month are removed from the model because of redundancy. Snowfall showed a high p-value of 0.197, meaning it is insignificant in the model and it was also removed. This leaves us with the following 14 variables and their respective coefficients and an adjusted R-squared value of 0.9628.

Predictor Variable	β_i	
Intercept	β_0	17.8
Date	β_1	-0.00185
Hour	β_2	0.0413
Temperature	β_3	0.0418
Humidity	β_4	-0.0171
Windspeed	β_5	-0.0160
Solar Radiation	β_6	-0.0413
Rainfall	β_7	-0.249
Season2 (Summer)	β_8	1.329
Season3 (Spring)	β_9	0.898
Season4 (Winter)	β_{10}	0.655
Holiday	β_{11}	-0.338
Work Hour	β_{12}	20.305
Weekday	β_{13}	0.0315

Table 2: Table of Coefficients for the Best Fit Model Predicting log(bikecount).

Splitting the data into a training set and a test set and then reapplying the same variables to the model showed that the logistic regression produced an in-sample R-squared value of 0.592 and an out of sample R-squared value of 0.539 and an RMSE of 415.7 and 453.8, respectively.

Ensemble Models

We then moved on to modeling using Random Forest, Boosting, and Bagging methods using the same training and testing data split used in the linear regression model.

The Random Forest data showed that the hour of the day and the temperature are the most important predicting factors for the models. These variables are closely followed by humidity and season. These variables have the most effect on the node purity, of the measure of how the model error increases when variables are randomly shuffled. The Random Forest model produced an RMSE of 189.1 for the testing data.

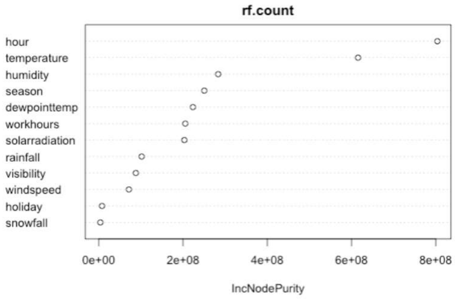


Figure 7: Random Forest Inclusion MSE for Node Purity.

The Boosting model using a Gaussian distribution and 1000 trees also shows that hour and temperature have the highest influence on the model with temperature having a relative influence of 35.89 and hour's relative influence of 29.41. Work hours was the next most influential variable with a relative influence of 9.14 followed by season at 8.31. This shows a drastic decrease in relative influence between variables and is reflective of our observations in the Random Forest model. The boosting model produced a RMSE of 388.31 for the testing data.

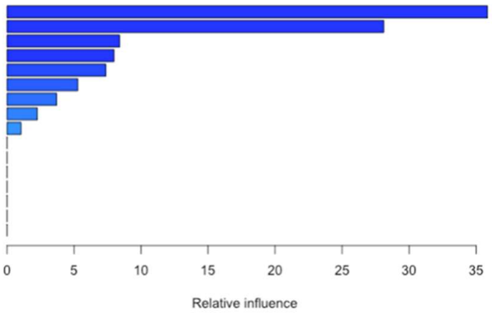


Figure 8: Boosting Relative Influence of Variables.

Our Bagging Model with 150 bootstrap replications of the data, which is necessary because our dataset is relatively small with only one year worth of data collection, produces an out of bag estimated RMSE of 178.37, the best error so far of all of our models. Applying the Bagging model to our testing data, we get a RMSE of 175.52, which is reflective of the predicted error initially given by the model formulation. For the Bagging model, the variables with the most importance are temperature, humidity, windspeed, and dew point. This is generally reflective of the variables with most effect on the predictive model for both the Random Forest and Boosting methods, however, the Bagging model shows that the top four predicting variables are all close in their level of importance in Bagging, whereas only the first two variables were significantly more important than the other variables in the Random Forest and Boosting models.

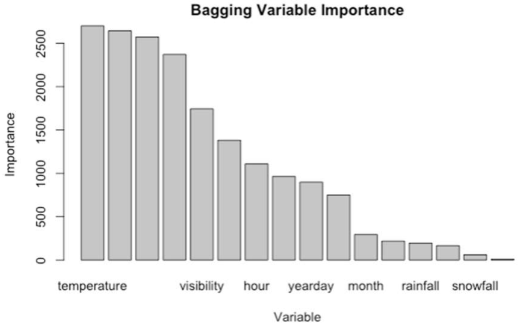


Figure 9: Bagging Variable Importance for Bike Demand Prediction.

Combination Model

Using the linear regression, random forest, boosting, and bagging models we created a combination prediction model. The models were weighted by their general performance on their own and then combined to produce a full model. The Linear Regression model was weighted at 10%, Random Forest at 15%, Boosting at 15%, and Bagging at 60% due to their independent model prediction performance. This combined model resulted in an RMSE of 156.55, predictably the best model performance so far, but only slightly better than the bagging prediction with an RMSE of 175.52.

Deep Learning

Deep Learning Procedure

Deep learning is a much more involved and lengthy process for predictive modeling, but it can have great predictive capabilities when allowed the amount of time needed for training the model.

The first step in Deep Learning is the data preparation. The model must first go through type conversion, data cleaning and data extraction so it can be properly implemented into the next steps of the computer algorithms. This is similar to the data cleaning and processing conducted in our previous models.

Data preprocessing is the next step where the data is encoded for the deep learning process. Our model used the scikit-learn package, which is also applied in the data vectorization. The data was sparsely vectorized for feature extraction. Because deep learning requires a lot of memory space and time, sparse vectors are used to save space by only storing the non-zero values. The data was then split, using scikit-learn.

The model assembly step determines the number of layers and nodes that should be included in the model. This is followed by model activation using ReLU and Linear libraries and then model optimization using adam, RMSprop, and SGD. This is the lengthiest part of the deep learning process, but it is also the most important to produce the best predictive model. The loss function is then applied to calculate the MSE, MAE, MAPE, and MSLE. The loss function is a part of the deep learning decision theory and calculates the relative cost of each event in the learning process for each variable.

Finally, the deep learning model's evaluation step tests the loses and validation of the model's curves to make sure it correctly assessed the data while interpreting data and producing the algorithm for the predictive model.

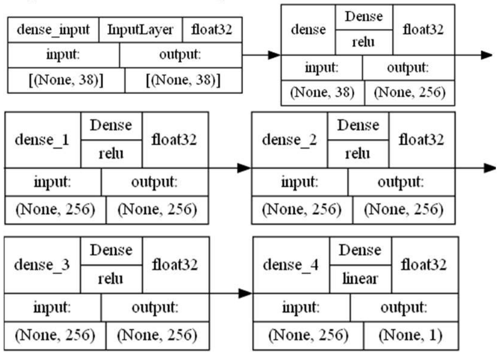


Figure 10: Evolution of the Deep Learning Best Model, Resulting in 4 Layers and 256 Neurons.

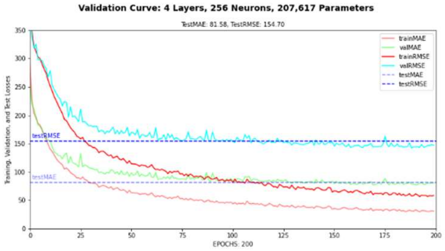


Figure 11: Validation Curve of the Best Deep Learning Model.

Our best Deep Learning model using Hyperparameter Tuning resulted in a Test/Validation RMSE of about 150/75. This indicates that our model may be overfitting and is biased toward our training data.

The Deep Learning model has the best predictive capability of our models, but it comes at a cost. Deep learning takes a lot of time. Depending on the cost-benefit analysis of the companies, either the previously discussed models or the combined models could be better. Or, if time is not a factor, the Deep Learning model would be best. The speed of the model could be improved for a faster training method by creating a TensorFlow-pipeline and dataset classes. This, again, take a lot of time and computer data. We could also test different combinations of layers while verifying the number of nodes, activations, and data dropouts. Removing outliers, attempting different repartions, and retraining the models could improve the bias effect of the training data. We could also train for higher Epochs by exploring possible convergences in further iterations or by conducting more iterations of the model. Finally, we could experiment with different and more unique combinations of node layers. All of this, however, would take more time and computer power than our team collectively has. So, we will continue with our current Deep Learning Model, recognizing that even without these improvements it provides the best predictive model for our data.

Hypothesis Testing

The Status Quo or the Alternative?

Hypothesis testing is used to test if there is a difference between the presence of a variable versus the lack of the presence of a variable. In our case, we will be using the Welch's Two Sample T-Test to test whether holidays have an impact on bike-sharing demand and whether hot versus cold weather has an impact on bike-sharing demand.

In our first hypothesis test, we will test the null hypothesis that there is no difference between bike-sharing demand when it is a holiday versus when it is not a holiday. The alternative hypothesis is that there is a difference in demand between holidays and not holidays. Our Two Sample T-Test shows that the test statistic is 7.6 on 490 degrees of freedom and a p-value 1.52e^{-13}. The mean bike count for holidays is 499.8 and the mean bike count for non-holidays is 715.3. The 95% confidence interval for our sample is (159.8,271.3). This information tells us that with 95% confidence, there is a difference of 160 and 271 additional bike rentals on non-holiday days than there are on holidays. The p-value for our T-Test also proves that there is a significant difference between the two samples. Therefore, we reject the null hypothesis in favor of the alternative hypothesis and conclude that bike renting is more popular on non-holidays.

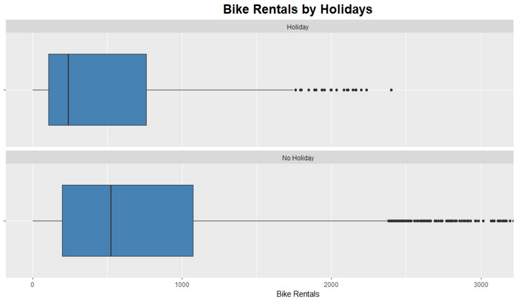


Figure 12: Box Plot of Bike-Sharing Demand on Holidays vs. Non-Holidays.

In our second hypothesis test, we test that there is no difference between hot days (represented by summer) and cold days (represented by winter) versus the alternative hypothesis that there is a difference between bike-sharing demand on hot days and cold days. Our Two Sample T-Test shows that the test statistic is 53.8 on 2420 degrees of freedom with a p-value of 2.2e^{-16}. The mean bike count for hot weather days (summer) is 1034.1 and the mean bike count for cold weather days (winter) is 225.5. The 95% confidence interval is (779, 838), meaning there is a difference of between 779 and 838 more bike rentals on hot days than there are on cold days. The small p-value for our T-Test, again, proves that there is a significant difference between our hot and cold weather samples. Therefore, we reject the null hypothesis in favor of the alternative hypothesis and conclude that bike-sharing demand is greater on hot weather days than on cold weather days.

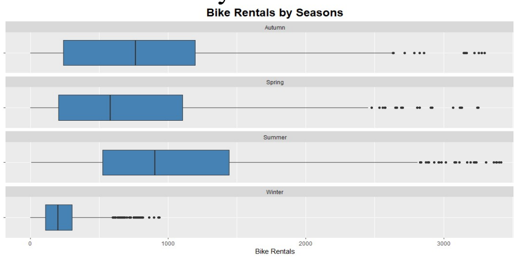
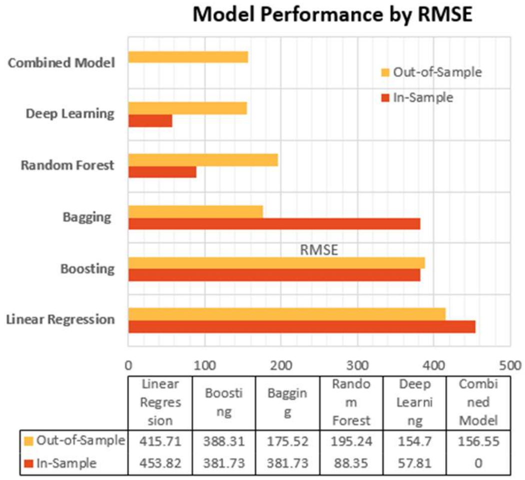


Figure 13: Box Plot of Bike-Sharing Demand on Hot (Summer) Days vs. Cold (Winter) Days.

Findings

Model Performance

Our data exploration and modeling has produced several possible predictive models for bike-sharing demand using the Seoul, South Korea dataset. The Deep Learning model and the Combined Model produce similar results with an out-of-sample RMSE of 154.7 and 156.6, respectively. The Deep Learning model, though it produces the best results on its own, takes a lot of time and memory to produce and predict requirements. If the company is willing to invest the time and memory space in their predictive model, the Deep Learning model is the best option and will provide the best overall predictive capability for bike-sharing demand. If the company is not willing to invest these types of resources, the combined model is best. However, the combined model relies on the accuracy of several models to properly predict demand. The best model that is included in the combined model is the Bagging predictive model, followed by Random Forest. Boosting and Linear Regression are also included in the combined model and produce the most error or all of our models. It is important to include these models, even with their larger error rates, in the combined model because they assist in representing the outliers and random errors in the data caused by variables that are either not included or out of the researchers' control.



	Linear Regression	Boosting	Bagging	Random Forest	Deep Learning	Combined Model
🟨 Out-of-Sample	415.71	388.31	175.52	195.24	154.7	156.55
🟥 In-Sample	453.82	381.73	381.73	88.35	57.81	0

Figure 14: Comparison of Model Performance and Predictive Capability by RMSE

Results

In our analysis, we attempted to answer four key questions for the bike-sharing companies. To find the features that best predict bike-sharing demand, question 1, we developed several models: Linear Regression, Random Forest, Boosting, and Bagging. In these models, we found that the variables with the most predictive power were hour, temperature, and humidity. Depending on the model, the order of importance of these variables varied. Other variables that regularly contributed to the model's predictive capability were season and workhours. Though these variables have high importance, they do have a correlation with the four variables that were identified as the best predictive variables. Our deep learning model addresses our second question, if we can predict the bike-sharing behavior of the public in Seoul. The Deep Learning Model had very good results in its predictive power for the bike-sharing demand. On its own, the Deep Learning model was able to produce a similar RMSE to the combined predictive models from question 1. This is very promising for the future if we can gather more data for future Deep Learning models and continue to make model adjustments for more accurate predictions. Finally, for our third and fourth questions, the difference between holidays and non-holidays and cold and hot weather seasons, we found that there is a larger bike-sharing demand on non-holidays and during warm weather months. Questions 1 and 2 can be used to assess how many bikes need to be available throughout the city given the combination of the variables included in the dataset. Questions 3 and 4 might assist the company in considering if there should be a scaled price for holidays and cold weather seasons to increase use during lower demand times, hopefully increasing the company's profit.

Future Research

Though the dataset included 8,760 entries, this is a very small amount of data considering it only spanned over the course of one year. To be able to better predict bike-sharing demand and customer behavior, several years of data should be applied to the analysis to account for differences from year to year and control for outliers that may have affected our analysis and results. Also, the data collected did not include anything regarding customer behavior such as how far the bike-sharing customer traveled, how long they used the bike, or what were the most common locations for bike-sharing started and terminated. This information would provide good insight into customer behavior and prediction as well as better assessments on not only how many bikes need to be available on a given day, but where the bikes need to be to maximize availability and bike-sharing demand.

References

"Seoul Bike sharing Demand Data Set." UCI Machine Learning Repository. Center for Machine Learning and Intelligent Systems, March 3, 2020. https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand.