Analysis of the Seoul Bike Share

Northeastern University - ALY 6015.23546

Group Epsilon Final Project - February 22, 2021

Ryan Goebel, Nicole Castonguay, Brad Viles, Tim Przybylowicz

Table of Contents

|  |  |
| --- | --- |
| Table of Contents....................................................................................................... | 2 |
| Introduction and Research Questions........................................................................ | 3 |
| Data Set..................................................................................................................... | 4 |
| Descriptive Statistics.................................................................................................. | 6 |
| Hypothesis Testing..................................................................................................... | 9 |
| Regression Models.................................................................................................... | 11 |
| 1. Ordinary Least Squares Model.................................................................. | 11 |
| 1. Regularization Models............................................................................... | 13 |
| 1. Ridge Regression............................................................................ | 13 |
| 1. Lasso Regression........................................................................... | 14 |
| 1. Elastic Net Regression.................................................................... | 15 |
| 1. Regularization Results.................................................................... | 16 |
| 1. Data Mining: Random Forest Regularization............................................. | 18 |
| 1. Regression Model Comparisons................................................................ | 20 |
| Data Mining: Random Forest Classification............................................................... | 22 |
| Conclusions................................................................................................................ | 25 |
| Recommendations and Future Analysis.................................................................... | 26 |
| References................................................................................................................. | 28 |

Introduction and Research Questions

Group Epsilon decided to study data collected from the public Seoul bike share company between December 2017 and November 2018 on the demand of rental bikes along with various other factors. Our goal was to understand what peak drivers of demand for their bikes were, and how they should consider implementing a potential demand-based pricing model into their business plan. In order to implement this new pricing model, we will explore five different potential predictive models for demand. There are any number of ways to attach a pricing model once demand is predictable, so we believe that the initial demand model is the logical starting point for this development. In addition to predicting demand, we plan to explore the Other potential business questions to be solved by this analysis include identifying the best times to pull bikes out of circulation for routine maintenance, or seasonal changes in demand that would allow for segments of the fleet to be upgraded to new models. Research questions to be answered by this paper are as follows:

1. When is demand for bikes the highest?
2. Can we predict absolute bike demand?
3. Can we classify demand into relative groups?

By answering these questions, we will provide recommendations to improve the business model of the bike share and increase the revenue generation of the business. We will also provide recommendations in order to improve future data collection and utilization based on our findings. By refining the data collection to be aligned with the analyses that we present here, we will be able to create more accurate model iterations and further define areas for opportunity for the business. This data lifecycle is imperative to increasing the growth of the business and the efficiency of research allocation.

Data Set

The dataset used for this analysis was found from the UCI Machine Learning Repository, which contains over eight-thousand time-stamped records between December 2017 and November 2018. Each record – taken at hourly intervals – includes the count of bikes rented, current weather factors, and categorical descriptions. The structure of the dataset is summarized in Figure 1 below.

|  |
| --- |
|  |
| *Figure 1 – Dataset structure prior to cleaning* |

After looking at this pre-processed data, the following cleaning operations were performed on the dataset:

* Each variable name was converted to shorthand to facilitate programming operations.
* The ‘Date’ variable was converted to a categorical value representing the day of the week.
* Each categorial variable was converted to a factor to facilitate analysis.
* Numeric variables were converted from metric to imperial units.

Since our study centered around identifying drivers for bike count, the ‘count’ variable was assigned the dependent variable, and the remaining numeric and categorical data were explanatory variables. The pre-processed data included a variable called “Functioning.Day” designating days when bikes were available for rent. For the records marked as “non-functioning”, bike count defaulted to zero, but independent measurements were maintained. As including these records in the dataset would introduce errors in prediction, these records were removed and the variable “Functioning.Day” was deleted. The table and figure below show the structure and summary of the data once these cleaning operations were performed.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Variable | Description | Units |
| **Categorical** | season | Season of the Year | N/A - Factor |
| day | Day of the Week | N/A - Factor |
| hour | Hour of the Day | N/A - Factor |
| holiday | Holiday or Not Holiday | N/A - Factor |
| **Numeric** | count | Number of Rented Bikes | bikes |
| temp | Temperature | F |
| humidity | Relative Humidity | percent |
| wind | Wind Speed | mph |
| visibility | Visibility | miles |
| dewpoint | Dewpoint Temperature | F |
| solarrad | Solar Radiation | kWh/m2 |
| rain | Rainfall | inches |
| snow | Snowfall | inches |

|  |
| --- |
|  |
| *Figure 2 – Dataset summary after cleaning* |

Descriptive Statistics

Once initial cleaning of the data was complete, basic descriptive statistics was completed on all variables in the data set to understand what relationships were evident (Figure 3). A histogram of bike rentals per hour shows that very rarely did demand reach the peak capacity of the fleet (Figure 4). Additionally, variables like snow and rain logically had a negative relationship to the demand for rental bikes. Demand was markedly lower in winter months following the same logical explanation. There appeared to be interesting trends in both hourly and weekly demand, which we decided to pursue further analysis on. The bike share rarely reached its maximum available supply, indicating that there exists great opportunity for better utilization of the fleet to increase peak rider numbers.

|  |
| --- |
|  |
| *Figure 3 – Scatterplots of numeric variables* |

|  |
| --- |
|  |
| *Figure 4 – Histogram of bike rentals per hour* |

More thoroughly examining seasonal demand, it becomes clear based on the plot in Figure 5 that winter is the off-peak season for the bike share. By this figure alone, we recommend transitioning scheduled preventative bike maintenance to winter months. Implementing this change, the bike share is less likely to be caught by surprise with abnormally high demand when the fleet is depleted for maintenance. This also allows for the greatest number of bikes to receive basic maintenance and preventative care before there are functional issues with the bikes. Taking advantage of this window for maintenance will become more important as the number of peak riders increases as well as the total mileage of the fleet.

|  |
| --- |
|  |
| *Figure 5 – Boxplots of categorical variables with mean bikes counts per hour indicated by the gray dashed lines* |

Hypothesis Testing

A hypothesis test was conducted to determine if weekday demand was consistent with weekend demand despite the difference in range evident from the descriptive statistics. As the ratio of variances was determined to be not equal and the sample passed a Shapiro-Wilk normality test, a two-sample Welch t-test for equal means was performed.

|  |
| --- |
|  |
| *Figure 6 – Variance and normality checks for t-test conditions.* |

The null hypothesis of this test was that the mean number of riders on weekdays is equal to the mean number of riders on weekends, and the alternative hypothesis was that the mean riders on these days differed. As is shown in the figure below, the null hypothesis was rejected at a confidence interval of 95%. This is consistent with two observations from the descriptive statistics: a drop in mean demand on Sundays, and a drop in maximum demand on weekends in general. Based on this result we can say with confidence that there is a statistically significant difference in ridership between weekends and weekdays.

The obvious spike in demand at the beginning and end of working hours alongside this fact leads us to believe that many of the bike passengers are using the service for their daily commute. Based on this knowledge, we recommend the creation of a commuter program. By implementing commuter passes, the bike share would be able to offer discounts on evening rides to all commuters that also rode the bike share that morning. This change is intended to further create demand in the morning when there is already a small surge in demand. Additionally, the bike share has the option to apply certain privileges to riders holding a commuter pass such as protection for price surges or a small flat rate discount. Once it is established, there exists a great opportunity to take advantage of this new program using any number of other incentive programs.

|  |
| --- |
|  |
| *Figure 7 – Hypothesis test with:*  *Null hypothesis: Mean # of riders on weekdays = Mean # of riders on weekends*  *Alt. hypothesis: Mean # of riders on weekdays ≠ Mean # of riders on weekends* |

Regression Models

In order to predict demand for the bike share, Group Epsilon created five models for predicting demand for the Seoul bike share. With this predictive model, the bike share could implement a demand-based pricing structure in order to maximize profits when demand is at its highest, and implement discounts to increase volume when demand is at its lowest. A model of this nature introduces a proactive cost lever into the operation of the business. Using the lever created by our model, the bike share will be able to maximize profits at both ends of the demand curve.

The regression models were created to predict rental demand based on the provided numeric and categorical predictors in the dataset to determine the best option. Each of the models described in this section were created using the same high-level methodology:

1. Split the dataset into training and testing data.
2. Develop the model using the training dataset.
3. Apply the model to the test dataset to create predicted values.
4. Compare the predicted values to the actual values in the test dataset.
5. Calculate the performance metric (RSME) based on the difference between predicted and actual values.

Ordinary Least Squares (OLS) Model:

As a means of establishing a performance baseline for modelling activity, Group Epsilon first developed a multivariate linear regression model using R’s native linear model package. Gauging by the significance level of the coefficients (Figure 8), the relationships between certain explanatory variables and bike counts established during hypothesis testing were validated – such as certain days of the week and hours of the day.

|  |
| --- |
|  |
| *Figure 8 – OLS model summary* |

Due to the number of interrelated weather variables used in this study, there existed a high likelihood of multicollinearity within the OLS model. These relationships were visualized using the correlogram shown in Figure 9.

In this plot, the color indicates the type of relationship and the size of the shape indicates the magnitude of the relationship. As shown, there exists potentially significant relationships between ‘Temperature’ and ‘Dewpoint Temperature’, ‘Humidity’ and ‘Visibility’, and ‘Humidity’ and ‘Dewpoint Temperature’. Since multicollinearity tends to inflate the values of the coefficients within regression models, ridge, lasso and elastic net regularization techniques were employed to limit the impact of this collinearity.

|  |
| --- |
|  |
| *Figure 9 – OLS weather variables correlogram* |

Regularization Models:

*Ridge Regression Model*

The first regularization model that Group Epsilon developed was a Ridge Model. This type of model uses each of the predictors but constrains their coefficients through L2 norm penalization. By constraining the coefficients, the multicollinearity between the different weather predictors is minimized. Cross-validation techniques within R’s glmnet package were used to predict optimal values for the tuning parameter, lambda – which controls the weight of the penalization.

|  |
| --- |
|  |
| *Figure 10 – Cross-Validation Plot to identify optimized lambda values* |

The two vertical lines on the cross-validation plot indicate two optimal values for the tuning parameter: lambda.min and lambda.1se (Figure 10). To provide the most regularized model, lambda.1se is used in the results of this study.

*Lasso Model*

The second regularization model that was developed for this study is a Lasso Model. Similar to ridge regression, this model contains a tuning parameter which controls a penalty term (L1 norm). In this type of model, the penalty term shrinks some coefficients to zero based on their predictive value. To visualize the relationship between the tuning parameter and the number of non-zero coefficients, the lasso trace plot shown in Figure 11 was created.

|  |
| --- |
|  |
| *Figure 11 – Lasso trace plot showing coefficient paths* |

For consistency between models, labmda.1se was also selected for this study – providing the most regularized model and reducing the most coefficients to zero. By identifying which features had non-zero coefficients, this model identifies which predictors were the most important. As further validation of the hypothesis testing results, most of the non-zero coefficients were either specific hours or days.

*Elastic Net Model*

The last regularization technique used to model bike demand was Elastic Net. This model is a mix of the L1 norm and L2 norm penalties found in ridge and lasso regression models. To optimize the mixing parameter alpha, we ran a loop from 0 to 1 in increments of 0.01, returning the alpha and lambda values producing the lowest RMSE. These models were compared against the baseline OLS model as percent improvement. For model comparison, the optimized elastic net model was selected, coinciding with an alpha of 0.02 and a lambda of 89.58.

|  |
| --- |
| Inserting image... |
| *Figure 12 – Optimized Elastic Net Model Parameters* |

*Regularization Results*

The RMSE results for each of the regularization models is shown in Figure 13.

|  |
| --- |
|  |
| *Figure 13 – RMSE for each of the regularization models* |

After applying each model to the test dataset, the OLS model outperformed all three regularization models. This makes sense intuitively, as regularization models are generally employed when the number of predictors outweighs the number of records. In our models, we had 41 explanatory variables and over 8000 records, negating the major benefit of these techniques. The multicollinearity that we were worried about did not seem to have a major impact on the accuracy of the OLS model. To better visualize how each model treated the different parameters, a variable importance plot was generated (Figure 14).

|  |
| --- |
|  |
| *Figure 14 – Variable Importance Plots Comparing OLS to Regularization Models* |

This plot displays the absolute value of the coefficients for each variable, showing the change in demand for a unit increase for each predictor. Since the variance within the variables differ significantly, this plot is used more for identifying how the model treated each parameter. Variables that were consistently ‘important’ in each model were:

* Rain
* Hours (specifically 18, 19, and 8)

Variables that were consistently not as important, based on these plots were:

* Visibility, dewpoint, days, and humidity.

Data Mining: Random Forest Regression

The final regression model that Group Epsilon explored was a random forest model. Random forest is a data mining technique in which the training data is bootstrapped, or randomly sampled with replacement, and then only a random subset of the predictor variables is available for use by the model when splitting at each node. Then, all the trees’ predictions are averaged together to yield the final prediction.

In a random forest model, there are two primary variables that we adjusted to optimize the model’s fit to the training data. The first of these variables was *mtry*, which is the number of randomly selected variables that are available at each split of a tree node. The other variable that was optimized was *ntree*, which is the number of decision trees used in the model. To optimize *mtry*, we ran a loop with a small number of trees (*ntree*=100) to determine which value gave the lowest out-of-bag (OOB) error estimate. The OOB data is the data that is not contained in the bootstrapped (bagged) data. We found that an *mtry* valueof 9 gave the lowest OOB error (see Figure 15). Then, we re-ran the model using an *mtry* of 9 and *ntree* of 500 and verified that the OOB error had been minimized by the time it reached 500 trees.

|  |
| --- |
|  |
| *Figure 15 – Random forest regression optimization* |

The variable importance plot from our random forest regression model can be used to gauge which input explanatory variables have the largest impact on predicting the number of bikes that are rented each hour. Random forest has two types of variable importance plots: “%IncMSE” and “IncNodePurity” (Figure 16). The “%IncMSE” plot is interpreted as the percentage increase in MSE if that particular variable were removed, which in our case shows that the hour of the day is overwhelmingly the most critical variable for predicting the number of bikes, followed by day of the week and the current temperature. The “IncNodePurity” plot, on the other hand, is the increase in node purity (related to the reduction in error) whenever that variable is chosen at a node split. In our case this shows that the hour of the day and current temperature have the most significant impact on demand.

|  |
| --- |
|  |
| *Figure 16 – Random forest regression model variable importance* |

Our final random forest regression model (Figure 17) was able to explain 91.75% of the variance in the out-of-bag data, with a MSE of 33827.42 (equivalent RMSE = 183.92). Applying this model to the testing data, resulted in an RMSE of 188.03, which was significantly better than any of our previous models.

|  |
| --- |
|  |
| *Figure 17 – Final random forest regression model* |

Regression Model Comparisons

By comparing the calculated error rates of these models, there are four with remarkably similar RMSE, between 369 and 377. The outlier of the five models is the random forest model, which outperforms all the others with an RMSE that is almost half of any of the other options at 188.03 (Figure 18). A comparison of predicted bike counts per hour from the models versus actual bike counts per hour from the test data set can be seen in Figure 19.

|  |
| --- |
|  |
| *Figure 18 – RMSE comparison of regression models* |

|  |
| --- |
|  |
| *Figure 19 – Predicted vs. actual values for the five regression models* |

Data Mining: Random Forest Classification

Although sometimes exact bike count predictions may be desired as given in the regression models, for operational purposes it is often necessary only to predict specific categorical divisions. For instance, you may want to predict when a particular hour on a particular day with specified weather conditions will have high demand, low demand, or something in the middle. In our case, we have defined “high demand” as being above the 85th percentile (1424.4 bikes per hour), “low demand” as being below the 15th percentile (137 bikes per hour), and “mid demand” as being anything in between.

|  |
| --- |
|  |
| *Figure 20 – Random Forest Classification Model Optimization* |

Just as in the random forest regression model, the training data is bootstrapped in the random forest classification model. Instead of averaging the numeric result of each tree, a “vote” is tallied from each tree and the classification is given to whichever has the most “votes.” Similarly to the random forest regression model, we programmed a loop to determine the value of *mtry* which would yield the lowest error rate, which was *mtry* = 10. (Figure 20). We then verified that the estimated out-of-bag error and the estimated errors for each of the classification were stabilized near a minimum using our model’s *ntree* value of 500 (Figure 20).

|  |
| --- |
|  |
| *Figure 21 – Random forest classification model variable importance* |

Similar to our random forest regression model, we can look at a variable importance plot for the classification model to visualize the relative effects that each predictor variable has on the model (Figure 21). The “Mean Decrease Accuracy” can be interpreted as the proportion of observations that would be classified incorrectly if that particular variable were removed. “Mean Decrease Gini” refers to the Gini impurity, which is a measure of uniformity at each node. After a split is made at the node, the Gini impurity can be thought of as a weighted proportion of incorrect instances of either side of the split. In our classification model, the hour of the day and the current temperature are shown as having the most significant impact to the model’s predictions.

|  |
| --- |
|  |
| *Figure 22 – Final random forest classification* |

The final random forest classification model’s output (Figure 22) shows that the out-of-bag estimate of the error rate is 9.49%, and the confusion matrix breaks this down further with the error of predicting individual classifications. The model correctly classified 790 low-demand hours as “low”, incorrectly classified 227 low-demand hours as “mid”, and incorrectly classified 1 low-demand hour as “high” for an error rate of 22.40%. 4468 mid-demand hours were predicted correctly, 148 were incorrectly classified as “low”, and 128 were incorrectly classified as “high” for an error rate of 5.82%. For high-demand hours, 871 were predicted correctly, zero incorrectly classified as “low”, and 139 incorrectly classified as “mid” for an error rate of 13.76%. It should be noted that these error rates are considered estimates since they are based on the out-of-bag data from the training set.

The model is then applied to the test data. As is shown in Figure 23, the model has a 90.29% accuracy of predicting what category/class the bike count will be each hour, with an error rate of 24.24% on the low demand hours, 5.30% on the mid demand hours, and 14.62% on the high demand hours.

|  |
| --- |
|  |
| *Figure 23 – Accuracy of random forest classification model* |

Conclusions

In conclusion, our exploratory analysis revealed some trends in the data such as lower bike demand during the winter, higher demand during commuter hours, and lower demand on weekends. Based on a histogram of the demand for bikes, it was rare for demand to reach the maximum supply available. Using a hypothesis test, we were able to confirm with 95% confidence that mean weekend demand is not equal to mean weekday demand. We created five different predictive regression models and were able to determine the variable that has the greatest impact on demand – hour of the day. We also uncovered which variables were decidedly unimportant when predicting demand – visibility, dewpoint, day, and humidity. Of our five regression models, the random forest model had the lowest RSME and was by far our most accurate model in predicting exact bike demand. Alternatively, if exact bike counts aren’t required, we developed a random forest classification model that has a 90.29% accuracy at predicting demand into high, mid, or low categories.

Recommendations and Future Analysis

Our final business recommendations based upon the bike share data center around three main ideas. Implementation of a commuter pass, seasonal maintenance, and introduction of a flexible pricing model. A commuter pass could offer different incentives such as a discounted afternoon ride for a morning rider, or immunity to surge prices. These passes could follow either a monthly or an annual pricing model. By implementing an annual pricing model, the bike share could recoup some of the revenue stream that is lost in the winter months, when passes may not be bought under a monthly purchasing model. Seasonal maintenance allows for the fleet to remain operational and ensures that demand can be met in the peak months.

The final and largest recommendation that we make for the bike share is the implementation of a flexible pricing model based on the random forest classification model highlighted above. By introducing an additional cost lever into the system, the bike share could institute demand-based pricing during times of high demand, and rider discounts during times of low demand. Included in this pricing model is the inclusion of the commuter pass, which would shield regular customers from an increase in transportation fees through this model.

An area for opportunity is in increased capabilities in support of location tracking. This would allow the bike share to understand what areas consistently reached peak supply of bikes and may be able to support a greater fleet presence. Because maximum supply is rarely reached, it would allow for areas of low demand to have the fleet density lowered to better allocate resources to where they will be used. Additionally, if a commuter program is established it will be crucial to track behavior patterns within the two subsets created within the population. This would allow the bike share to fine-tune pricing models and target them toward specific parts of the rider population.

By continuing to monitor this data over the coming years, we feel that accuracy of our model will improve. Greater refinement could better tune the demand regression model. Several of the explanatory variables are not independent of each other - creating multicollinearity in the linear regression model. This allows for inaccuracies in the model that could be prevented with additional time and data collection. Due to the fact that our data only covers one year, time series analysis comparing the seasonality of demand will also improve our accuracy. There is a very small sample size on the behavior of demand on holidays that would improve with additional years of data collection.

References

Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.