# Factors that Influence the Use of Bike-Sharing Programs

Rachel Bellflowers

May 5, 2020

## Contents

# 1 Introduction

Bike-sharing programs allow individuals to rent a bike at one location and then leave it at another location. Often found in cities, these programs help citizens avoid traffic and travel around quicker than walking. With over 500 bike-sharing programs in existence, it is clear that this model has gained traction. Given that these bicycles have tracking software that records who is riding, for how long, and their final destination, data about the use of such programs is easily available.

Data from the Capital Bikeshare system in Washington, D.C. were used to determine what factors contribute to higher usage of rental bikes on a daily basis in the years 2011 and 2012. Started in 2008, the Capital Bikeshare program has over 500 stations spanned across the nation's capital, Virginia, and Maryland.


# 2 Method

## 2.1 Data Collection

This dataset was downloaded from the UCI Machine Learning Repository. The uploader of this dataset retrieved the bicycle data from Capital Bikeshare's website data, the weather data from i-weather.com, and data about holidays from the DC Department of Human Resources. All data were selected for the years 2011-2012. Further information is included in the researchers' article *Event Labeling Combining Ensemble Detectors and Background Knowledge*, Fanaee-T and Gama (2014).

## 2.2 Exploratory Analysis

I created a series of bar plots and scatterplots to explore the dataset. I removed variables one-by-one to see if any would lower the RMSE. However, as it appeared that all variables had some effect on the response variable `cnt`, none were removed. I then used the `summary()` function to examine the distribution of the variables.

### 2.3 Statistical Modeling

After trying two linear models (10-fold with cross-validation and 10-fold with cross-validation repeated 5 times) and three random tree models (10-fold cross-validation, 10-fold cross-validation repeated 5 times, and a model with bootstrapping), I found that the model with the lowest $RMSE$ and the highest $R^2$ was the 10-fold random forest model with 5 repetitions of cross-validation. The $RMSE$ for this model's predictions on the test set was 651.6911, and the $R^2$ was 0.8924. I removed the variables for the count of casual users (`casual`) and the count of registered users (`registered`) as I was only interested in examining the count of total rental bikes (`cnt`). Additionally, I removed the date column `dteday` as it increased the RMSE.

### 2.4 Reproducibility

All code used is included in the attached `writeup.Rmd` file. The R function `set.seed()` was used to ensure the reproducibility of the models and the partitioning of the original dataset. Additionally, links to the data sources are included in the "Data Collection" section of this report. Users will need the latest version of the `R` data analysis software, the `RStudio` GUI, and the following `R` packages installed on their computer: `corrplot`, Wei and Simko (2017); `caret`, Kuhn (2020); `dplyr`, Wickham, François, et al. (2020); `ggplot2`, Wickham, Chang, et al. (2020); `magrittr`, Bache and Wickham (2014); and `yardstick`, Kuhn and Vaughan (2020).

## 3 Results

The data I used to create this model on the number of bikes rented out (`cnt`) were `season`, the year (`yr`), the month (`mnth`), hour of the day (`hr`), whether it was a holiday or not (`holiday`), day of the week (`weekday`), if the day was a working day (`workingday`), weather conditions (`weathersit`), the normalized temperature in Celsius (`temp`), the normalized feeling temperature in Celsius (`atemp`), the normalized humidity (`hum`), and the normalized windspeed (`windspeed`). The graph comparing the predicted values to the residuals can be seen in Figure 1. A table of the correlations between all variables is located in Figure 2.

## 4 Conclusion

The results of this analysis provide some explanation for the difference in use of rental bikes in the Capital Bikeshare system over time. Interestingly, it seems that there is no significant difference for the number of bikes rented on different weekdays or whether it is a working day. I had predicted that people would be more likely to ride on weekdays when people need to travel to work.

It would be interesting to see how well this model predicts data from programs in other locations with different climates and other public transport infrastructure.

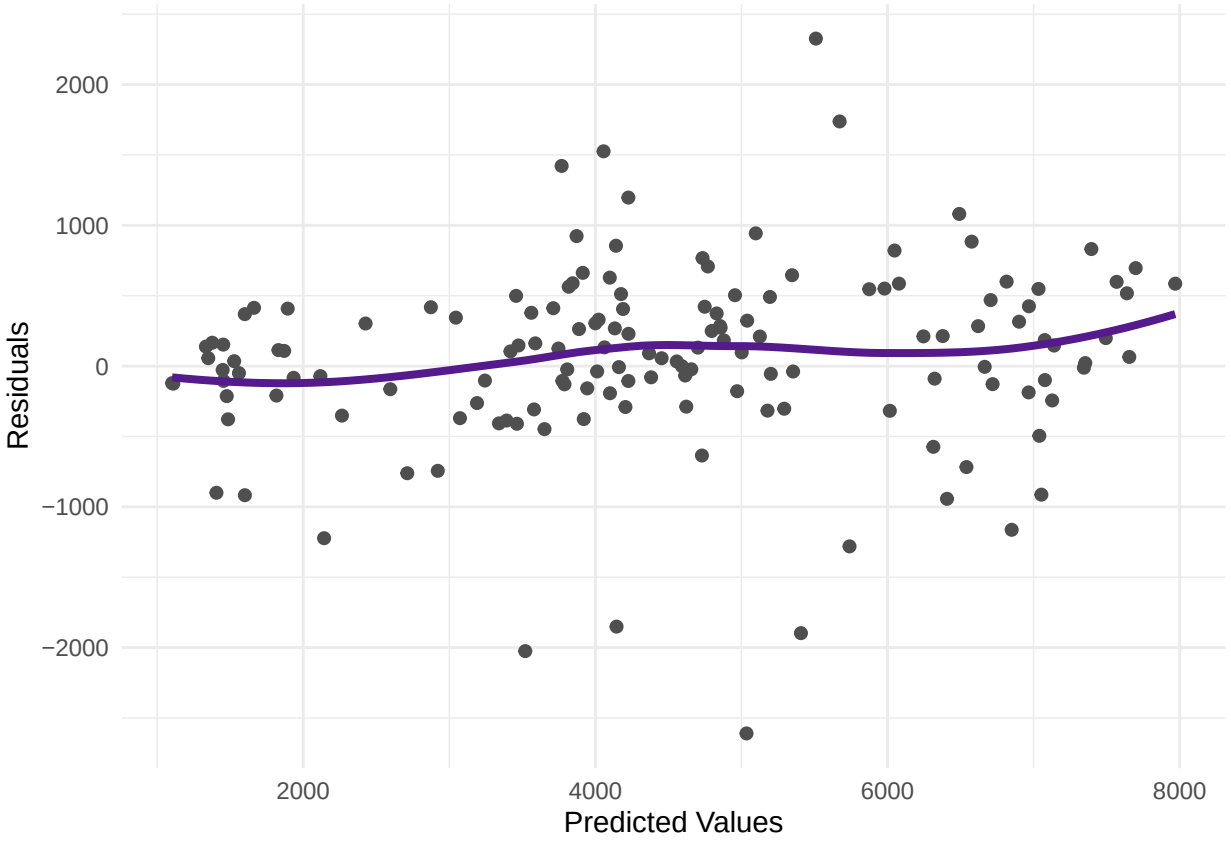


Figure 1: The model best fits the values around the midpoint of the dataset. On the lefthand side, the model is on average about 250 points off, while on the righthand side a more dramatic difference can be seen.

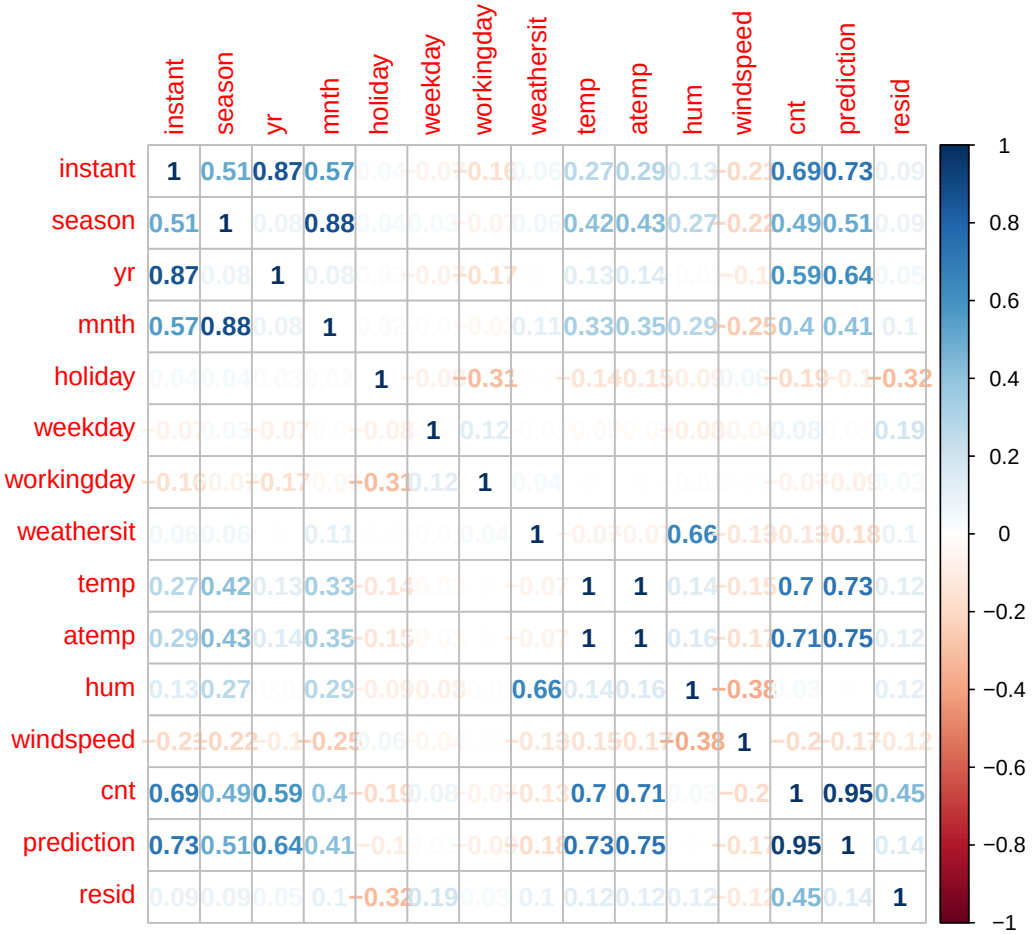


Figure 2: The season, year (2011 vs. 2012), and the temperature are strongly correlated with the predicted bike rental counts, while the month and the weather situation are moderately correlated with the predicted values. In contrast, the day of the week, whether it was a holiday, whether it was a working day, the level of humidity, and the windspeed had no statistically significant relationship with the predicted values.

# References


Bache, Stefan Milton, and Hadley Wickham. 2014. *Magrittr: A Forward-Pipe Operator for R.* https://CRAN.R-project.org/package=magrittr.

Fanaee-T, Hadi, and Joao Gama. 2014. "Event Labeling Combining Ensemble Detectors and Background Knowledge." *Progress in Artificial Intelligence* 2 (2): 113–27. https://doi.org/10.1007/s13748-013-0040-3.

Kuhn, Max. 2020. *Caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret.

Kuhn, Max, and Davis Vaughan. 2020. *Yardstick: Tidy Characterizations of Model Performance.* https://CRAN.R-project.org/package=yardstick.

Wei, Taiyun, and Viliam Simko. 2017. *Corrplot: Visualization of a Correlation Matrix.* https://CRAN.R-project.org/package=corrplot.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.