Capital Bikeshare In DC

Final Report

By Team 5

Arfa Tahir, Kirill Kim, Matthew Ng, Riwaaz B. Sijapati, Sangmo Lama, Tugba Yildirim

Table of Contents

Introduction and Background

Capital bikeshare is the DC based bikeshare company also known as CaBi that was launched in september 2010 where they had 49 stations and 400 bikes. The regional system now has grown across 7 Jurisdictions (Washington, DC.; Arlington, VA; Alexandria, VA; Montgomery, MD; Prince George's County, MD; Fairfax County, VA; and the City of Falls Church, VA) and it has more than 500 stations and 4300 bikes. With the bike share system, passengers can easily rent a bike from one of the stations and return the bike to any other stations in the system. Bikesharing has become a popular way to explore cities for tourists to visit the museums, parks and memorials located in DC. Capital Bikeshare has made the ride easily accessible with their bike sharing app. Riders can use the Capital BIkeshare app to locate the nearest station and they can purchase hourly, daily or yearly membership at any station. Anyone can get a membership for capital bike at the station using the capital bike share app.

Bike sharing is booming at an increasing rate, largely due to the relatively low cost of the schemes and the ease with which they can be implemented compared to other transport infrastructure. Bike share is mostly used in dense areas of urban core. Just like everyone else, capital bike might need work because of flat tire of any other mechanical problems. If the bike needs service, a rider can press a button which locks the bike until a capital bikeshare personnel unlocks the bike. With its minimal membership fees, Cabi has become helpful for low-income commuters. A $75 purchase provides a year membership. The Capital Bike Share also provides various deals and discounts for tourists and residents.

Capital Bike Sharing system releases a large amount of their data to the public. The data is recorded every time a rider rides a bike and the dataset is collected quarterly. For the analysis in this project, dataset from from year 2011-2012 is being used. This dataset is collated by Hadi Fanee-T and Joao Gama and it was uploaded by Mark Kaghazarian to Kaggle.

Motivation of our research question

Our motivation for this project came from our collective dislike of the MTA. We all agreed that the MTA breaks down a lot and can be very slow and unreliable, so we thought rental bikes would be a great alternative to public transportation. An article was recently published that claimed the MTA's 24 hour service may no longer be available, so we cannot rely on this type of transportation forever. We considered the success for large companies like Capital Bike Share, Citi Bike, Hubway and Divvy who operate in cities with large populations and we noticed the growth in bike sharing across the United States. The number of bikes operating through a bike share company in the United States has increased from 42,500 in 2016 to over 100,000 by the end of 2017. We also considered the cost of transportation for everyday people(MTA fare, gas prices, etc.) and wanted to challenge the stereotype that Americans are overweight.

Bike-share is still a relatively new idea, and as it spreads to new cities and areas where it already operates, we are only beginning to understand the many ways in which cities and residents will profit from it. We found that bike sharing is an inexpensive, flexible and accessible way to commute from one place to the other, especially in big cities like Washington D.C, New York, etc. Bike riding can be a fun and easy way for people to travel while also improving the passenger's health. It is an active way of transportation for short trips rather than just sitting in a bus or a train. It has some socioeconomic and environmental benefits such as less traffic and noise pollution and a decrease in CO2 emissions. In congested areas, such as; New York, DC, etc, riders spend less time in traffic and finding parking in bike sharing than other forms of transportation. Bike sharing also encourages the riders who wouldn't

normally ride a bike start using bikes as a means of transportation. An article by David Cranor says that there is a secret station for capital bike share that does not show in the capital bike share app. It says that the station inside the White House is for people who can get inside the Whit House Security Perimeter only. Commute in DC could be very hectic because of traffic and tourists and Bike share can be best and reliable alternative to get to the destination on time.

Dataset description and variable introduction

Our analysis based on rental bike information. This data is collected by the Capital Bikeshare company from 2011-2012 years. It consists of 731 observations and 15 variables. We found the data in the open-source Kaggle.com. The variables of the dataset can be split into three descriptive sections:

1. Time Metrics: Contains Date, Year, Month, whether it's a holiday, day of the week, and whether it is a working day. Information such as whether it's a holiday was sourced from the dc.gov.
2. Weather Records: Contains the season, what type of day it was (bright, misty, snowy, etc.), the actual and Feels-Like temperature normalized in Celsius, normalized humidity, and normalized wind-speed. This information was sourced from freemeteo.com.
3. Rider type: Contains count of casual and registered users. This information was sourced from Capital Bikes.

After we categorized all variables, decided to establish their types. The dataset includes nine quantitative (instance, dteday, casual, registered, count, temp, atemp, humidity, windspeed) and seven qualitative variables(season, year, month, holiday, weekday, workingday, weathertype). Upon closer examination of our dataset, we removed dteday, casual, and registered variables from our consideration. Variable dteday is the date of the observation, and it had no relevant information for our research question. Considering casual and registered variables, we decided to use only one dependable variable. We focused on count as our predicted variable. Numerically count is equal to the sum of casual and registered users. In our case, the use of count reduced the unnecessary complexity of the multilinear regression.

As we mentioned above, our data includes seven Categorical variables (also known as a factor or qualitative variables). Those variables cannot be entered directly into a regression model and be meaningfully interpreted. This issue required a different type of code. This code is known as "dummy coding." Likely for us, the dummy coding is done automatically by the statistical software R. For instance, in our dataset categorical variable season has four levels (Winter, Spring, Summer, and Fall), R automatically chose Fall as a baseline and construct three dichotomous variables that contain the same information as the single categorical variable season. To interpret our regression results correctly, we must remember that the default option in R is to use the first level of the factor as a reference and interpret the remaining levels relative to

this level. In other words, the coefficients of the Winter, Spring, and Summer show only the relationship to their baseline (Fall), but not the relationship to the predicted variable count. The same principle is applied to our other qualitative variables where the baseline for variable month is Apr, and the baseline Holiday is used for the variable holiday. We will discuss them more in our multilinear regression analysis.

Data summary statistics

We found that the bike-sharing rental process is highly correlated to the environmental and seasonal settings. Bike rentals were the highest when the weather was good, which was pretty self explanatory. We also found that the humidity and wind speed had little to no effect on the number of bike rentals for the day. Out of the 731 observations, we found that 68.40% of those were working days and 31.60% were not working days. Of those that weren't working days, 90.91% were not a holiday(weekend) and 9.09% of those were holidays. We were not surprised to find that the number of bike rentals were higher on work days and non-holidays. When we plotted the temperature-season box plot, we had to double check if the data was correct because we were confused when we saw that Fall actually had the highest temperature and Spring had the lowest. It took awhile for us to wrap our head around this, but then we recalled that the data was collected from the Capital Bikeshare system in Washington D.C and it all made sense. We compared the number of bike rentals from 2011 to 2012 and we noticed that there was about a 70% increase. The highest amount of bike rentals were in the months of June through October which were synonymous with our seasonal charts that found that Fall and Summer had the most bike rentals.

Methodology

For our project, we will be using supervised learning approach. Supervised learning is the branch of machine learning that focuses on learning from labeled data, the examples from which the algorithm learns also contain the desired output. We have two types of supervised learning techniques

- Classification: The algorithm assigns inputs to a class from a set of different classes.
- Regression: The algorithm outputs a continuous value.
    - Regression with Interaction Terms: The algorithm outputs a continuous value as a result of interaction between variables.

Predicting the bike rental for capital bike share system based on time and weather metrics requires of a regression algorithm. The regression algorithms are flexible, as no distribution on the data is assumed. We will be using different regression algorithms to find out the best model to predict the bike rentals. For this project, we used Multilinear regression model, as there are multiple independent variables that could affect the number of bikes rented. For instance, weather conditions, precipitation, day of the week, season and other factors can affect the rental behaviors. We expect to see multilinear relationship between the total number of bike rental

users and the variables of interest. We used the multiple linear regression model, then evaluate the adequacy of the fitted model. We then fine tuned our model by applying Interaction terms because we noticed that certain variables affect the outcome of other variables.
The second algorithm we will use to predict the bike rentals is the Regression tree model. We used the regression tree model as the model is easily interpretable. Making Predictions is fast, as there are no complicated calculations. With regression tree, it is easy to understand what variables are important in making the predictions. For instance, the internal nodes are those variables that most largely reduced the deviance.
to assess the performance of Multilinear regression model and regression tree model, observations will be divided into a training set and a test set. The training set is used together with the machine learning algorithms to teach the computer model, the test set is used to evaluate how well that model can generalize by comparing the predictions of the trained algorithm on the test set with the known results through a loss function. To reduce the variance, this can be done multiple times using different partitions and averaging their results to obtain the validation results. this process is called cross validation and can be used to compare the Multilinear regression model and the Regression tree model, to determine which model is the best to predict the bike rentals.

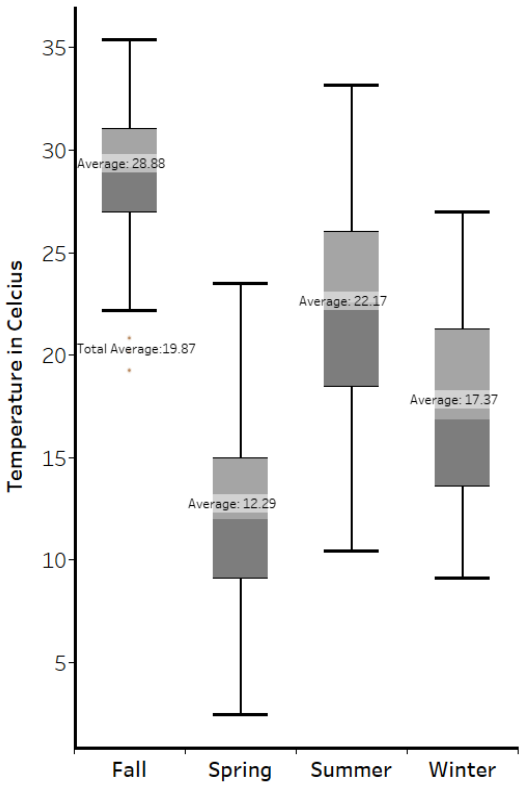
Models and Evaluation

Multi Linear Regression Model

The Multi linear regression model we applied employs the variables from both the time and weather metrics to predict total bike rentals. The original model contained In order to do this we started off with a model that included all independent variable. Due to the fact that some of the variables had insignificant p-values we decided to tune the model by using stepwise regression with a backwards elimination approach. The backwards elimination approach starts with all the variables provided, then removes statistically non significant variables. In doing so we remove the variables workingday and atemp from the original model. With the new model we noticed that the year variable has a huge effect on the models performance, however including the variable in a model that is meant to be used for predictive purposes did not seem significant.

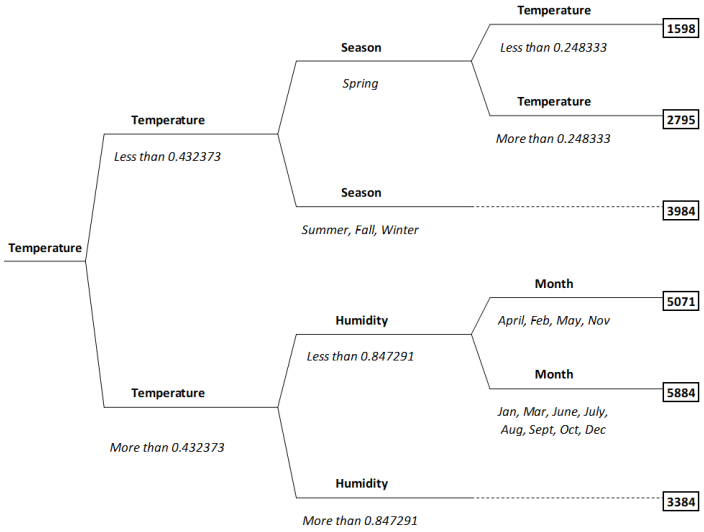
Original Model	Predicted count = season + month + year + holiday + weekday + workingday + weathertype + temp + atemp + humidity + windspeed
New Model	Predicted count = season + month + weekday + year + holiday + weathertype + temp + humidity + windspeed

After conducting a stepwise regression on the original model without the year variable we removed variables workingday, weekday, and atemp from the equation. After conducting an anova test between the two models we concluded that the p-value is larger than .05 which gives credence to the fact that the prediction capability of the variables that we removed are limited. However the new model reduced the



adjusted coefficient of determination by .04, however it increased the residual mean squared value from 1266 to 1267.

It became clear that a multi linear regression model was not enough to form a strong predictive model so we took a closer look at our data set. The seasons fall and summer on average had more rentals, so we went about finding out if there are any variables that could have caused this effect. As shown in the box plot to the left, we noticed that during the fall and summer season the average temperature was greater compared to the total average temperature. So we decided to refine our regression model with an interactive model. In order to check which variables had the most effect in interaction we can take a look at the regression tree from our regression tree model as shown below:



The plot shows that temperature is the most important variable. When temp is lower than .43 when its is spring the rental count is lower. This shows us that there is an interaction between temperature and seasons. To test this hypothesis we applied this interaction to our model and the coefficient of determination increased from 58.4 to 60.8 and reduced the rmse value from 1267 to 1232. Due to this

improvement we included tested the model with the interaction seen in the second branch between temperature, humidity and month. Compared to model with no interaction terms the final interaction model increased our coefficient of determination from 58.4 to 69.1 and decreased the residual mean squared error from 1267 to 1122. In addition to the base lines established in the beginning of this report, the model considers interaction terms: April*Humidity, Fall*Temp, April* Temp, and April*Humidity*Temperature as additional baselines. Our final model have the following variables and respective coefficient:

Section description	Variable	Coefficient
Season	*Spring*	463.5
	Summer	-372.2
	Winter	1817.3
Month	*August*	27014.6
	December	-7843.6
	February	-5533.4
	January	-6572.4
	July	9395
	June	3642.4
	March	-6632.8
	May	-3038.1
	November	-11396.8
	October	-2352.9
	September	-568.8
Type of Weather	*Cloudy:Adequet*	-79.4
	LightRain:Bad	-1211.3
Standalone Variables	*Humidity*	-9474.9
	Temperature	2073.4
	Not a Holiday	487.3
	Windspeed	-3726.5
	(Intercept)	7702.8
Interaction between Month And Humidity	*August*Humidity*	-42102.4
	*December*Humidity*	8594.1
	*February*Humidity*	6341.2
	*January*Humidity*	6806
	*July*Humidity*	5794.1
	*June*Humidity*	9930.7
	*March*Humidity*	5109.4
	*May*Humidity*	6517.6
	*November*Humidity*	19227.1
	*October*Humidity*	-497.6
	*September*Humidity*	1472.5
Interaction between Season and Temperature	*Spring*Temp*	-3427.6
	*Summer*Temp*	40.1
	*Winter*Temp*	-2431.2
Interaction between Month And Temperature	*August*Temp*	-41340.3
	*December*Temp*	14035.2
	*February*Temp*	5780.7
	*January*Temp*	12264.5
	*July*Temp*	-14943.8
	*June*Temp*	-4955.3
	*March*Temp*	9411.8
	*May*Temp*	6706.7
	*November*Temp*	23026.3
	*October*Temp*	8716
	*September*Temp*	9085.6
Interaction between Month, Temperature, and Humidity	*August*Humidity*Temperature*	62629.8
	*December*Humidity*Temperature*	-13670
	*February*Humidity*Temperature*	-2926.9
	*January*Humidity*Temperature*	-8533.3
	*July*Humidity*Temperature*	-6713.9
	*June*Humidity*Temperature*	-16648
	*March*Humidity*Temperature*	-302.8
	*May*Humidity*Temperature*	-12635.7
	*November*Humidity*Temperature*	-42107.3
	*October*Humidity*Temperature*	-4403.5
	*September*Humidity*Temperature*	-13365.5
Interaction between Humidity and Temperature	*Humidity*Temperature*	7982.2

Observations (n)	731
Coefficient of Determination (R^2)	69.10%
Coefficient of Determination (R^2)	66.50%
Residual Standard Error	1120

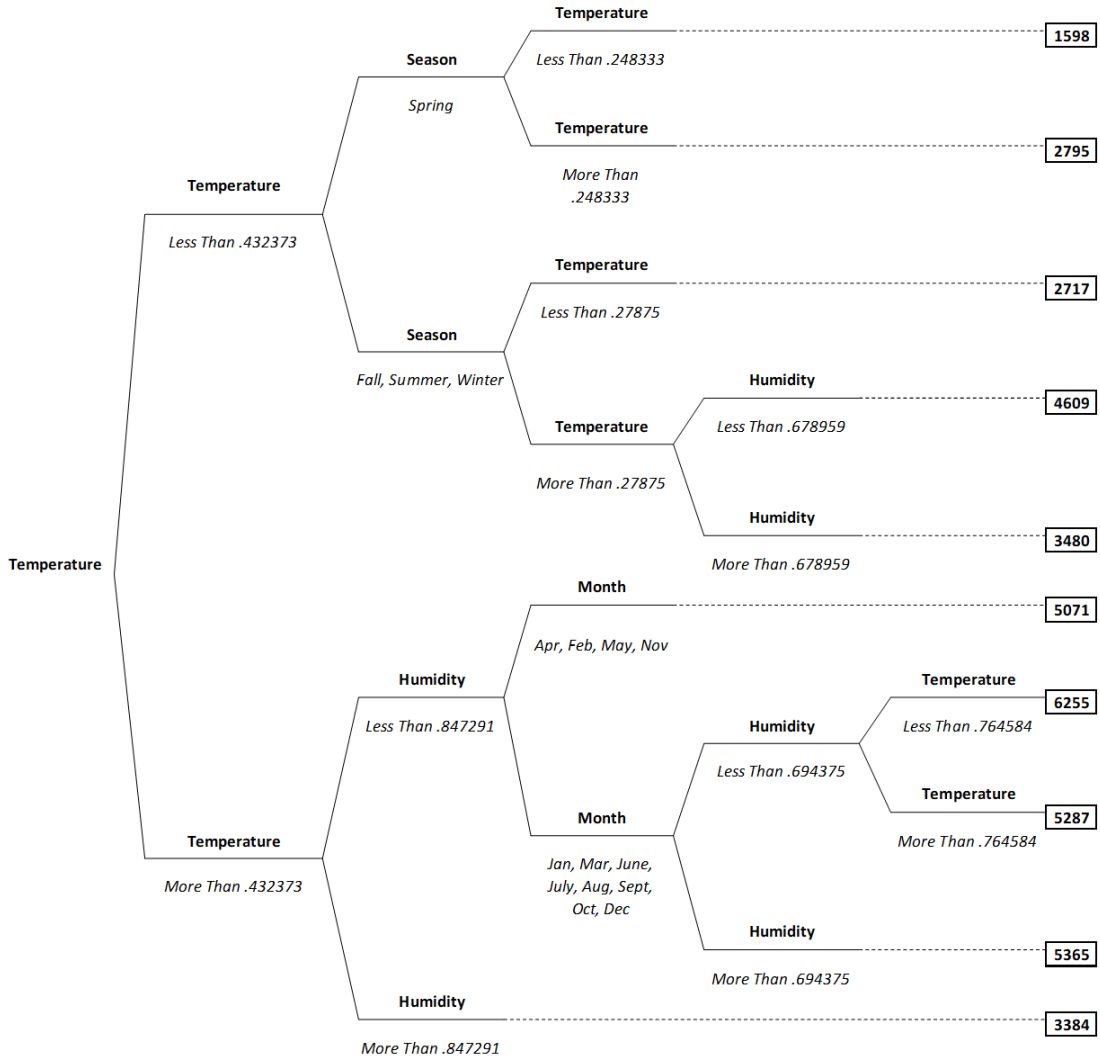
To explain the model we will consider a datapoint, in this case instance number 472. This instance occurs in the Summer month of April, it is a holiday and the weather is clear with a temperature of .664, humidity of .562, and wind speed at .2848. By excluding any variables that does not fit the criteria of the data point our equation is as follows:
*Predicted Count = 7702.8 - 372.2*Season:Summer + 2073.4*Temperature - 9474.9*humidity - 3726.5*Windspeed + (40.1*Season:Summer*Temperature) + (7982*Humidity*Temperature)*

This in turn gives us the predicted rental of 5326.39. Ultimately this model can explain roughly 69.10% of the variation in rental count based on the variables used. We conducted a 10 fold cross validation to measure the robustness of our model it acquired an RMSE value of 1168

Regression Tree Model

How do weather metrics and temperature metrics play a part in number of bikes rental? Visualizing the data by Regression trees may give us some insights. Using the rpart and tree package we first created a regression tree on the normalized data set. We produced the Regression Tree by setting half of the observations as the training dataset and finding the best split that maximizes the difference in between-group sum-of-squares. The actual variable used for the Regression Tree are “Temp”, “Season”, “Humidity” and “Month”. We find the total terminal node to be 10. The branches are annotated with the conditions of each split. Based on this tree, Temp is the most important variable in determining number of bikes rental. We will further analyze the tree

What does the tree mean?



Plot 1: Regression tree for the training data set

- The tree represents a series of splits starting at the top of the tree.
- The top split assigns observations having temp < 0.432373 to the left branch with season: spring and temp > 0.432373the right branch with humidity < 0.847291
- The left branch with season: spring, when temp > 0.248333 moves into right with number of bikes =2795 and when temp<0.248333 it moves left with number of bikes=1598
- Branch with season: Fall, Summer Winter, when temp<0.27875= 2717 bikes

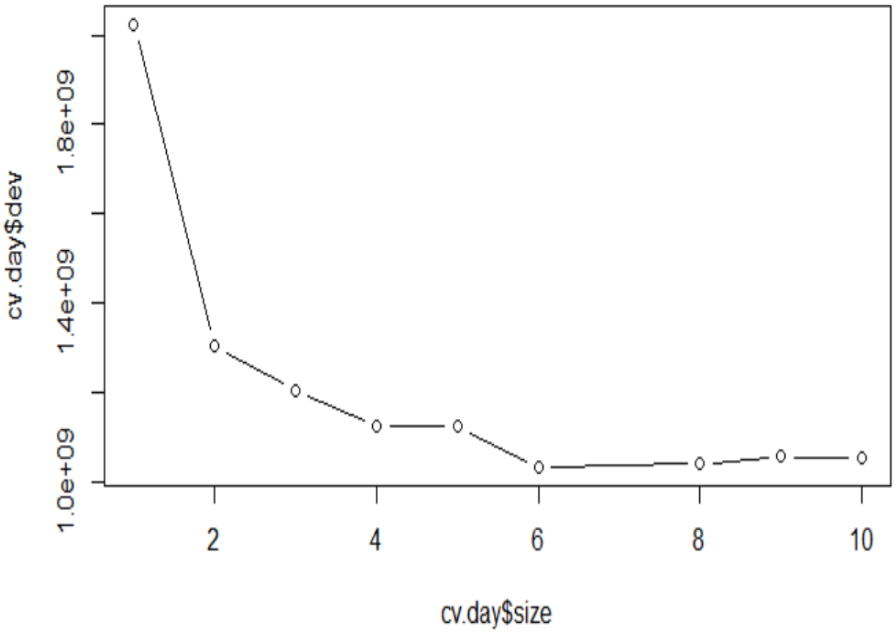
- Branch with season: Fall, Summer Winter, when temp>0.27875 and humidity<0.678959 = 4609 bikes .When humidity>0.678959 the number of bikes= 3480
- The top split assigns observations having temp > 0.432373, humidity < 0.847291 and month: April,Feb,May,Nov the number of bikes = 5071
- To the right branch with humidity<0.847291 moves to right with number of bikes = 3384 and left with month: April,Feb,May,Nov
- Observations having month: Jan, Mar, June, July, Aug, Sept, Oct, Dec and humidity <0.694345, with temp<0.764584= 6255 bikes and if the temp>0.764584 =5287
- Observations having month: Jan, Mar, June, July, Aug, Sept, Oct, Dec with humidity> 0.694345 = 5365 bikes
- With temp>0.4323 and humidity>0.847291 has number of bikes = 3384

Cross Validation and Pruning

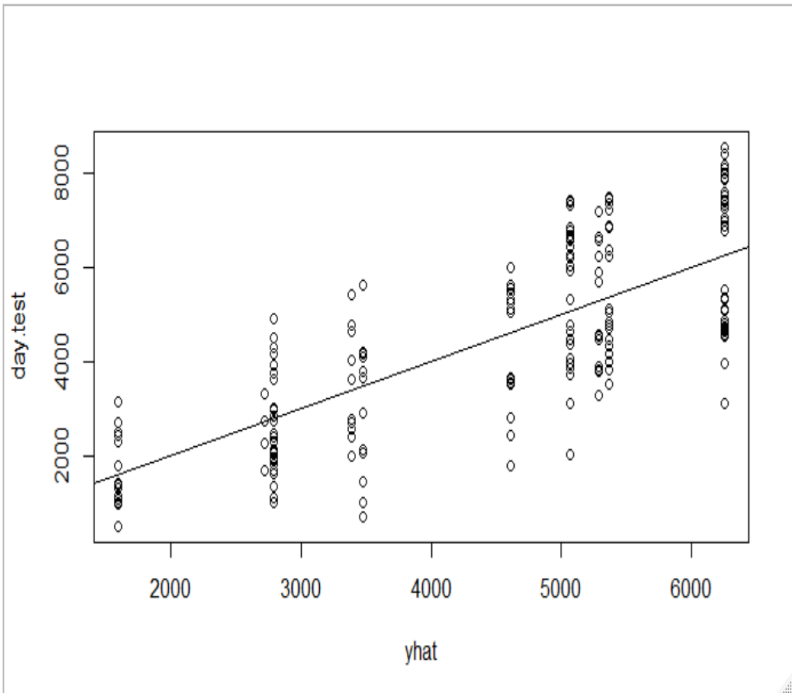
The tree package contains function cv.tree to perform cross-validation.We use the default method="deviance", which fits by minimizing the mean squared error (for continuous responses). The cv.tree() function reports the number of terminal nodes of each tree considered (size) as well as the corresponding error rate and the value of the cost complexity parameter used (k, which corresponds to α in (8.4)). We use cross validation function and produce the plot to find the best terminal node. We found the best terminal node = 6 with the lowest corresponding error rate of 1032866712. We calculate the predicted value of the Dependant variable. on the test data using the trained model and calculate the MSE. We find the MSE = 1688770 and RMSE = 1299.527 The following table and plot shows the number of terminal nodes with their corresponding deviance.

Tree Size	10	9	8	6	5	4	3	2	1
Deviation	1052948462	1056352417	1040701822	1032866712	1124146219	1125837833	1204490720	1304394808	2024344102
k	-∞	21010711	24080592	24982326	39475615	40401816	101983737	142059517	769463796

Table 2: Size of each terminal node with corresponding deviance and k



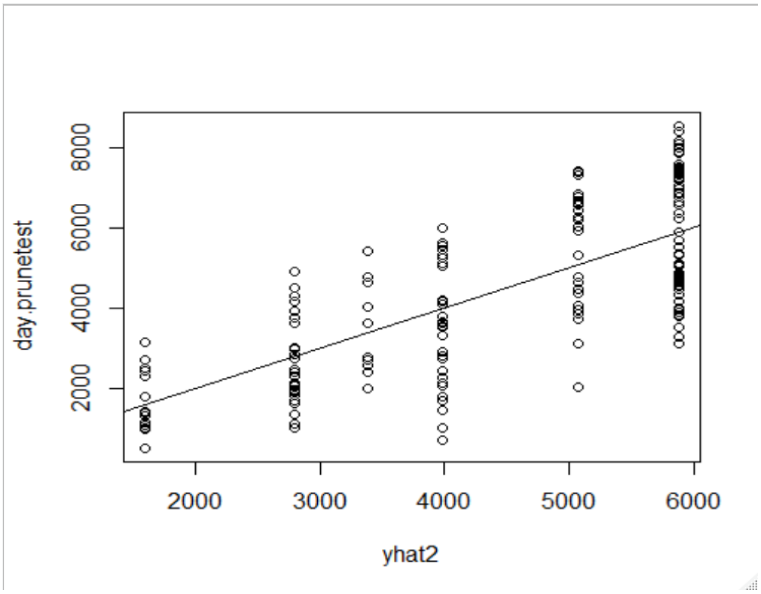
Plot 2 : Plot for Cross validation



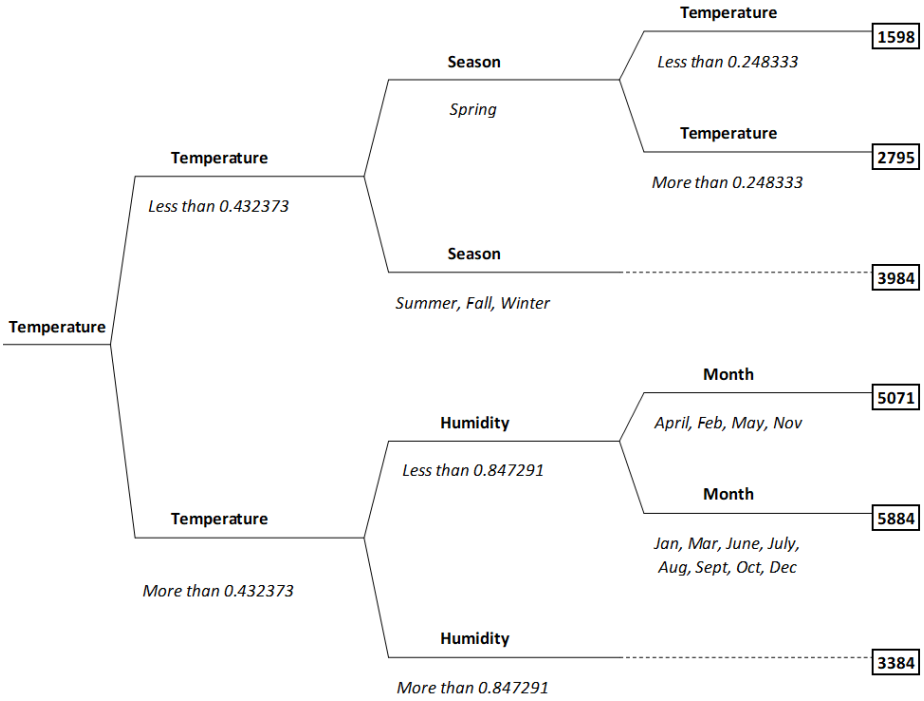
Plot 3: Prediction values for the DV on the test data set using the training model

Prune Tree

We use the function prune.tree, which takes a tree you fit by tree, and evaluates the error of the tree and various prunings of the tree, all the way down to the stump. We will prune the tree to see if it improves the performance. We will prune the tree on the training data(the default). We will use size 6 to prune the tree, as its has the lowest deviance. Once we prune and plot the tree. We calculate the predicted value of the dependent variable on the test data using the prune tree. We then calculated the MSE for the prune tree. The MSE is =1889280 and RMSE = 1374.511. Which is higher than the MSE for the training model. Therefore, we use unpruned tree to make predictions on the test data.



Plot 4: Prediction values for the DV on the test data set using prune tree



Plot 5:Prune tree

What does the Prune tree mean?

- The tree represents a series of splits starting at the top of the tree.
- The top split assigns observations having temp < 0.432373 to the left branch with season: spring and temp >0.432373 move to the right branch with humidity < 0.847291
- The left branch with season: spring, when temp < 0.248333 moves to the right with number of bikes= 2795 and left with number of bikes = 1598
- The right branch with season: Summer, Fall and winter and temp <0.432373 move left with number of bikes = 3984
- Temp>0.432373, humidity>0.847291 moves to right with number of bikes = 3384
- Temp> 0.432373, Humidity<0.847291moves to left and splits in two branches right with month: Jan, Mar, June, July, Aug, Sept, Oct, Dec = 5884 bikes and left with month:April, Feb, May, Nov =5071 bikes

Results and Conclusion

Below are the Algorithms with their corresponding RMSE(Residual mean square error) values and ranking based on their performance.

Algorithm	RMSE	Ranking
Interaction Model	1122	1
Multi Linear Regression	1267	2
Regression Tree	1299	3

Table 4: Residual mean square error results and ranking based on the RMSE

Ultimately we decided on the interaction model as our choice for predicting bike rentals based on the lower residual mean squared error value of this model compared to the other models. The interaction model explains that roughly 69.10% of the variation in rental count can be explained by the effect of time and weather metrics provided. Though it is not as robust a model as we wanted, this might be due to the fact that the variables provided and the number of datapoints may not have been enough to construct a perfect model. If we were to, for instance, consider the hourly metrics along side the weather metrics we could probably form a stronger model.

Practical Implications

All methods emphasized that number of using bikes is changing based on the season, holiday or working days. Important usage of the rental bikes has the biggest value mostly in fall and these numbers are getting to increase by the year. So, the major usage of this project would be to forecast bike rentals for the coming years based on independent variables we found in our research ‘Time Metrics’ and ‘Weather Records’. With our measurement, Capital BikeShare can decide whether they should expand the number of bike rental stations or adjust their pricing policy. Or if they want to do some innovation on their old bikes or check their system, such as optimizing battery replacement for stations and developing models for future system expansion they can see times with the lowest values from our statistics to utilize that time within the smallest portion of bike usage.

While people's interest in sharing bicycles leads to an increase in the number of bicycles, this increase can lead to new business opportunities. this analysis can be used in helping a company interested in starting a bike sharing service in bike friendly cities with similar weather conditions as D.C. Based on the weather record, 59 degrees is the average weather in

Washington D.C. for the last 30 years. In any city with similar weather conditions, this analysis can be helpful to find the basic information to start this business, through our analysis.

Bike sharing in China, with dozens of bike-share companies had city streets with millions of rental bicycles. However, the rapid growth, immediate customer demand became a problem in the country. Basically, riders would park bikes anywhere, or just abandon them, resulting in bicycles piling up and blocking already-crowded streets and pathways. Even most of the companies tried to make a provision against to increasing numbers, the bike sharing cities abandoned and broken bicycles have become a familiar sight in many big cities. Bike sharing remains very popular in China, and will likely continue to grow, just probably at a more sustainable rate. As we saw in China, following, the idea of sharing bicycles will benefit the country as well as adversely affect its economy. In order to overcome such problems, it is useful for bicycle companies to know the total number of bicycles in the country or the amount of use of bicycles owned. With the help of any analysis, it is possible to take measures before bike sharing becomes a problem. We believe that our analysis can guide current bike sharing companies to avoid such problems.

Our project can help with the idea of sharing bicycles in many areas. We used two different independent variables in our project. In these variables, the weather conditions are more effective on the cities with the similar conditions, while the time-dependent variable can be obtained in any city or for any company. Apart from these reasons, our research and the statistics we find may also refer to environment-friendly research. People who do research on this subject can see how many eco-friendly bikes are available or when they contribute more to the environment.

Bibliography

Capital Bikeshare. (2019, October 12). Retrieved from https://en.wikipedia.org/wiki/Capital_Bikeshare

Cranor, D. (2010, November 26). White House has a "secret" CaBi station. Retrieved from https://ggwash.org/view/7400/white-house-has-a-secret-cabi-station

China Is Still Sorting Through Its Colorful Bike-share Graveyards, Alan Taylor, https://www.theatlantic.com/photo/2018/08/china-abandoned-bike-share-graveyards/566576/

National Weather Service, https://www.weather.gov

Cosp Arqué, Oriol. "Demand forecast model for a bicycle sharing service." (2015).

STUDYING URBAN MOBILITY THROUGH MODELING DEMAND IN A BIKE SHARING SYSTEM, https://faculty.nps.edu/