Forecasting Bikesharing Usage for DC's Capital Bikshare System

Table of Contents

Introduction

Bikesharing systems are an increasingly popular solution in major urban areas to increase the usage of bicycles as a mode of transport. The riding of bikes helps to improve the lives of both users, as well as non-users, as each bike trip potentially represents a trip that would otherwise have required a car. We used data from the DC Capital Bikeshare from 2011 to 2017.

The purpose of this analysis is to determine variables/factors that help estimate bike usage and develop a model that forecasts the usage based on certain predictor variables.

Overview of Project

The idea of the project was to use different data acquisition, cleaning and modeling techniques to determine if we could forecast the bike usage. If we are able to model and forecast the bike usage, then it would allow the company to plan for the best route to increasing our fleet as well as expanding the available stations that are offered.

Some of the questions were: can we use the data to know when our usage is lower to repair bikes? When should we start increasing our fleet to best meet demand? Does weather and seasons have an impact on our business model?

We made an initial hypothesis that we would see a higher usage during the summer months and when the weather was nicer. With our initial exploratory data analysis, we could see that there is a trend and seasonality to the usage. The question then became, can we model this through a time series model. Also could we see what features were key factors in determining bike usage. Does the season or weather have an affect?

Literature Review Summary

We reviewed a few papers modeling bikeshare using in different cities across the globe (see Works Cited). These papers generally shared the same findings, which largely align with our Capital Bikeshare data:

- Usage increases as temperature increases, then starts to decrease as temperatures go into the 90s (Fahrenheit)
- Precipitation of any amount discourages cycling
- High winds can have a negative effect on cycling
- Usage is often higher in spring and summer, and lowest in winter

Overview of Data

Initial Data Set:

We started by just looking at two years of Capital Bikeshare usage from this dataset: https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

The dataset contains 2011 and 2012 historical usage data from Washington, DC's public Capital Bikeshare program, one of the first large scale bikeshare programs in the nation. Usage data is broken out by day and by hour. Additional data included a variety of information on weather, season, and whether a day was a holiday.

Data Cleaning Process

The dataset had required minimal cleaning. We had to convert several variables into factor variables (`season`, `holiday`, `weekday`, `workingday`, `weather`). Additionally we noted that the key for our dataset mislabeled the season variable, which was trivial to correct. Fortunately, there was no missing data.

Additional Scraping, Cleaning

Initial EDA on the dataset we began with indicated we did not have sufficient data to fit any models that could improve upon linear regression. So, we scraped and combined two different datasets. We used python scripts located in the Other Resources directory to do this.

Specifically, we ran get_bikeshare_data.py to get bikeshare data from Capital Bikeshare directly, then we ran join_data.py (which imports from noaa.py) to join the data with weather data from the NOAA API.

We used the same process to add the additional weather columns to the 2011 and 2012 as well.

Sources for additional datasets:

- Capital Bikeshare usage data from 01-01-2013 through 12-31-2017, from Capital Bikeshare (https://ride.capitalbikeshare.com/system-data)
- Weather data for DC for the same time period, from NOAA (https://www.ncdc.noaa.gov/cdo-web/webservices/v2)

The Bikeshare data only counted trips taken, and did not distinguish between registered or casual users.

The Weather data included:

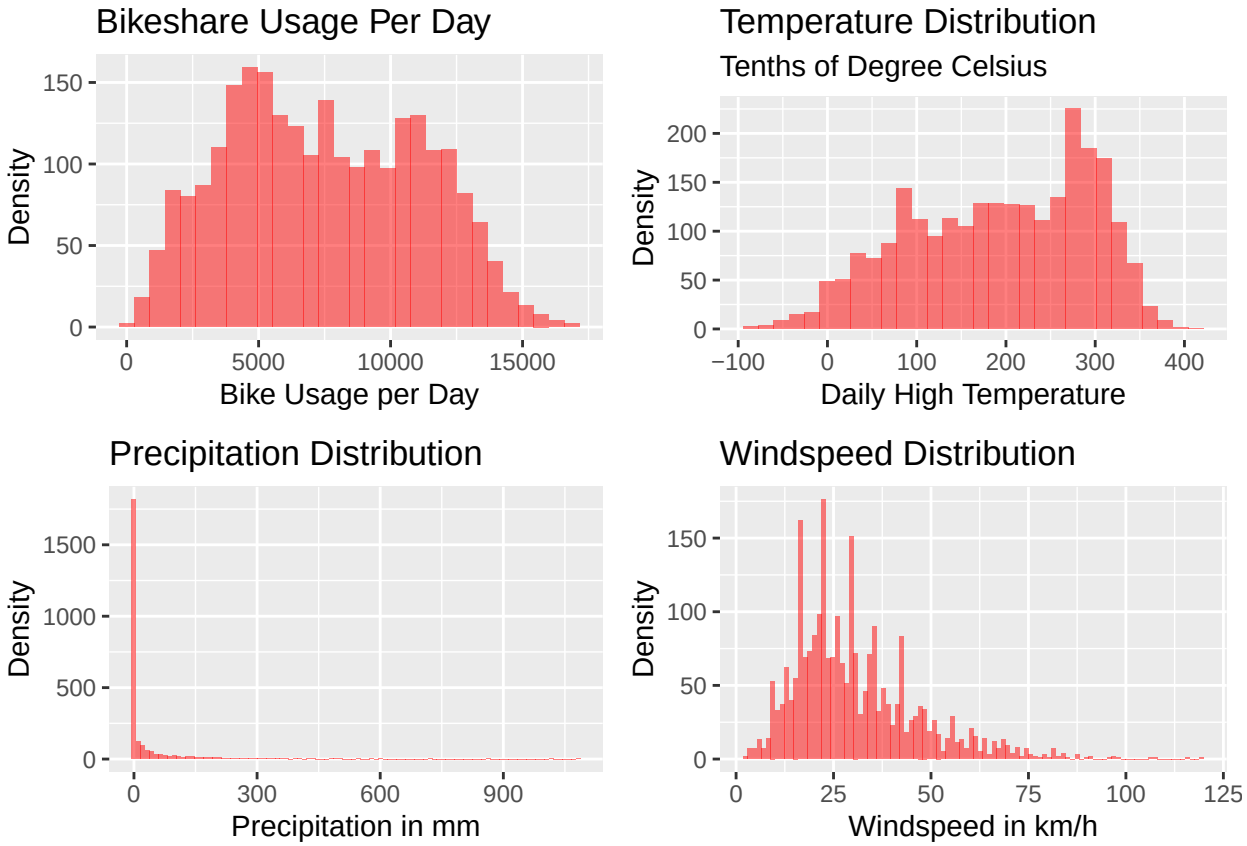
- TMAX: high temperature (in tenths of degree Celsius),
- PRCP: precipitation (tenths of mm), and
- AWND: average daily windspeed (km/h).

The final merged dataset used for analysis is final_df.csv located in the Data folder.

Our R code applies a season variable to the dataset, based on the date.

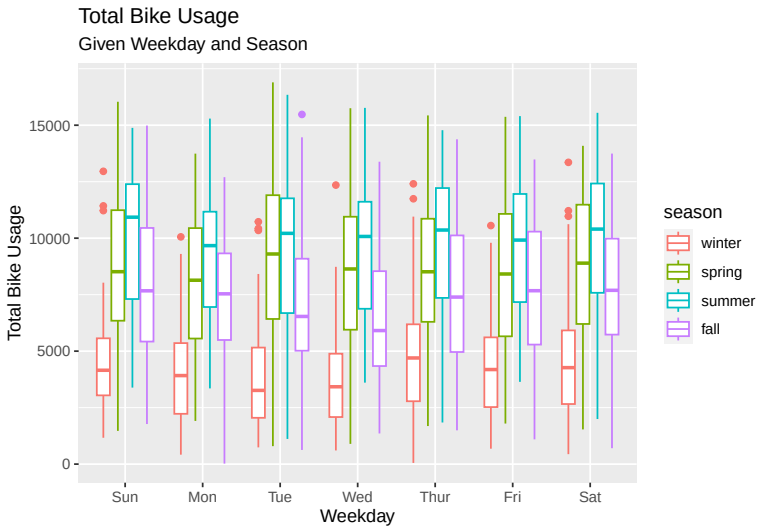
Exploratory Data Analysis

We can see that fortunately, Bikeshare usage per day is not heavily skewed.



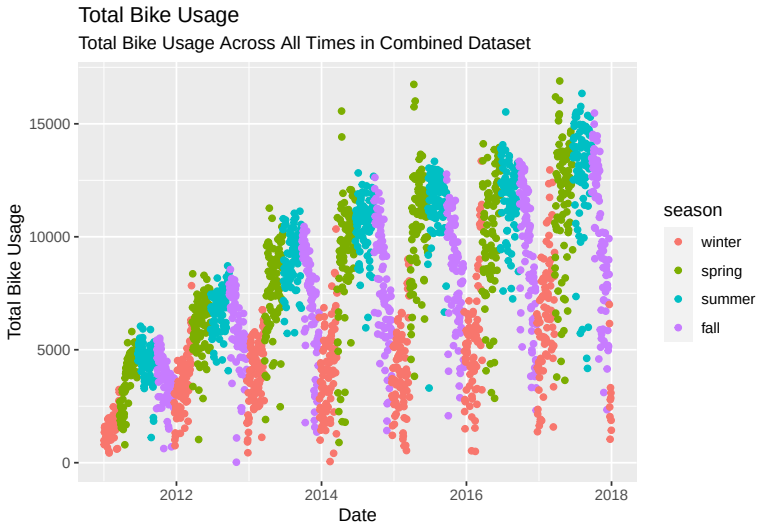
We see that windspeed and precipitation exhibit rightward skew, which makes sense: * Most days have no precipitation * Most days have moderate windspeed, while a few have very high windspeed

By contrast, temperature has leftward skew, with a smaller number of colder days

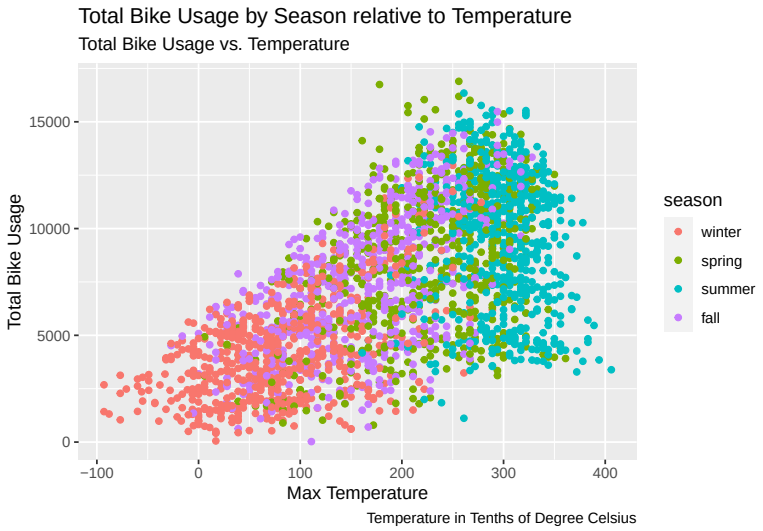


We can see in the above chart the effect of season as well as the effects of the day of the week on total bikeshare usage. Usage is lowest in the winter, highest in the summer, with spring and fall in between.

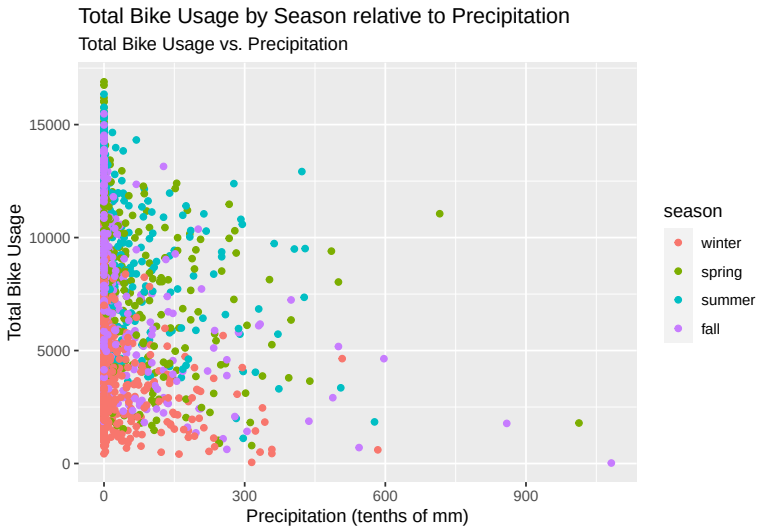
Fortunately, there is not a huge amount of variation in usage between weekdays and weekends.



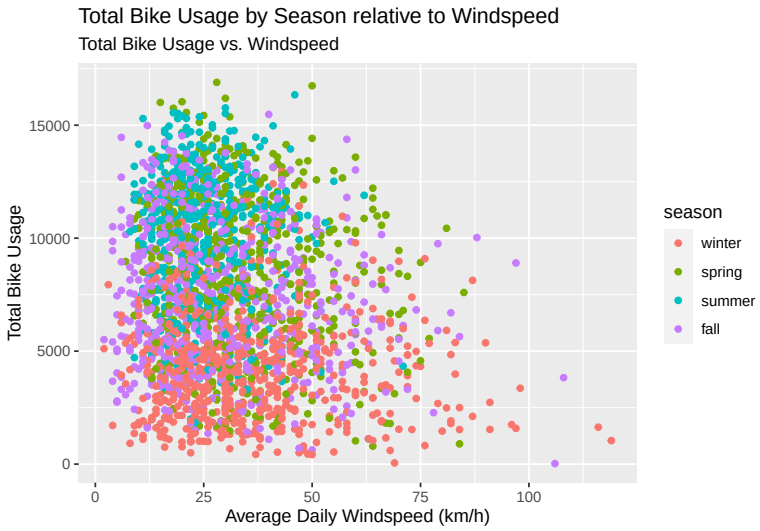
The most challenging aspect of our dataset is that usage was not static - it grew from 2011 through 2017. While this was obviously good for Capital Bikeshare, it meant that models would need to take into account not just variation within a year, but overall increasing usage over time.



Consistent with our literature review, see that as temperature increases, usage tends to increase, until around 30 degrees Celsius, after which usage tends to decrease.



The relationship between precepitation and usage appears somewhat weaker than temperature. Still, the highest precipitation days tend not to have high usage, consitent with an overall negative effect of preciptiation on usage.



We can also see a slightly negative relationship between windspeed and usage.

A correlation matrix comparing our numeric variables validates our graphical EDA:

- There is a positive correlation between temperature and count
- There are negative correlations between precipitation and count, and windspeed and count

We can also see the negative correlation of temperature with windspeed. Fortunately, precipitation has minimal correlation with temperature (-0.006) and relatively low correlation with windspeed (0.093).

	TMAX	PRCP	AWND	cnt
TMAX	1.0000000	-0.0065320	-0.2517998	0.6112456
PRCP	-0.0065320	1.0000000	0.0934646	-0.2114366
AWND	-0.2517998	0.0934646	1.0000000	-0.1837203
cnt	0.6112456	-0.2114366	-0.1837203	1.0000000

Key Predictors

Overview of Modeling

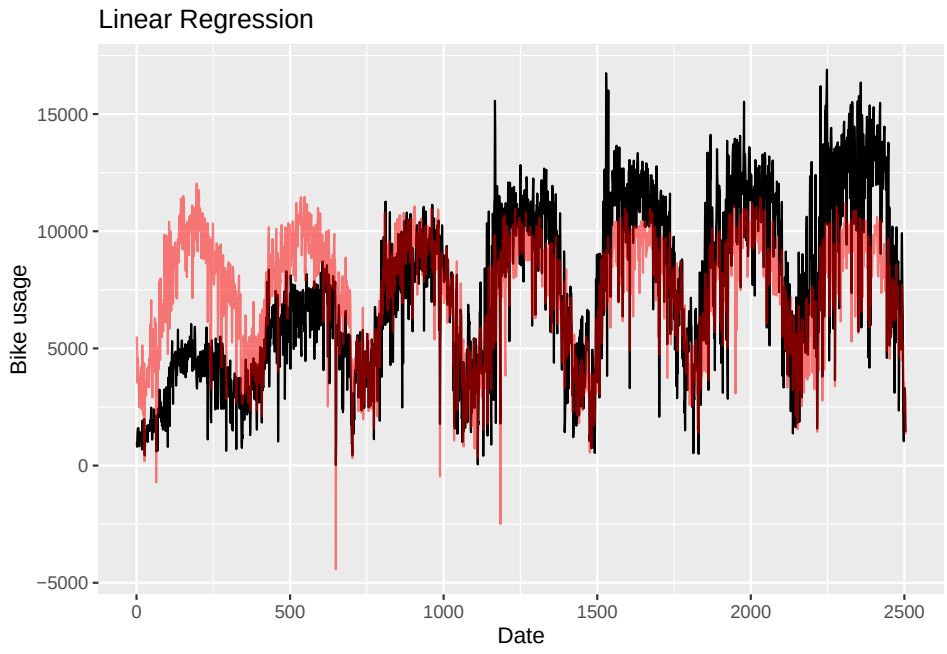
For our modeling, we used the `final_df.csv` file in the `Data` directory, which is described in the accompanying readme.

Model Types and Comparison

We ran multiple different models to try and best determine what fits the data best. Ideally since we are dealing with time series data, we are expecting that it will be the best at predicting. We decided to only showcase three of the models in this paper. The three are LASSO, ARIMA and LightGBM (gradient boosting).

Linear Regression

This is the very first model that we used as a base case. We also ran other basic linear regression models (LASSO, Ridge Regression, etc.) but they all seemed to perform similarly. The issue was they were unable to calculate a trend.

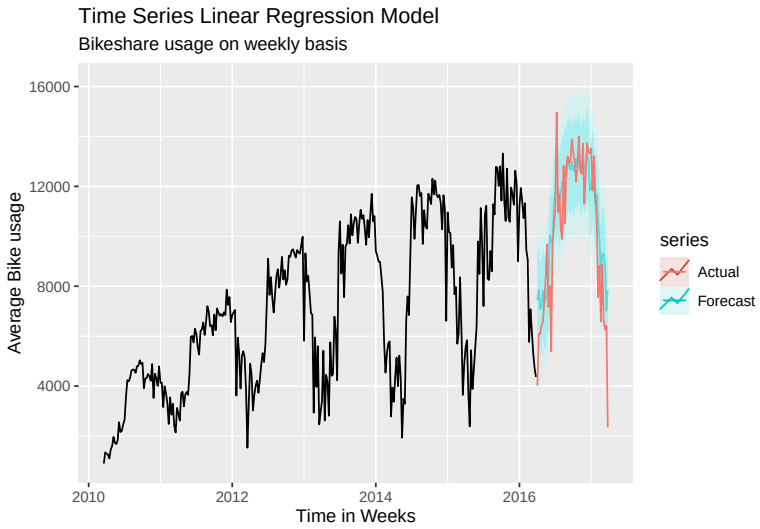


A naive linear regression does not incorporate any time series factors, has an adjusted R-squared of 0.45. We can see that the linear regression over-predicts usage in early years, and under-predicts usage in later years, due to lacking a trend/time series component.

Linear Regression with Time Series

Since we can see above that we are unable to predict without time series, what if we create a model that attempts to predict with trend and seasonality.

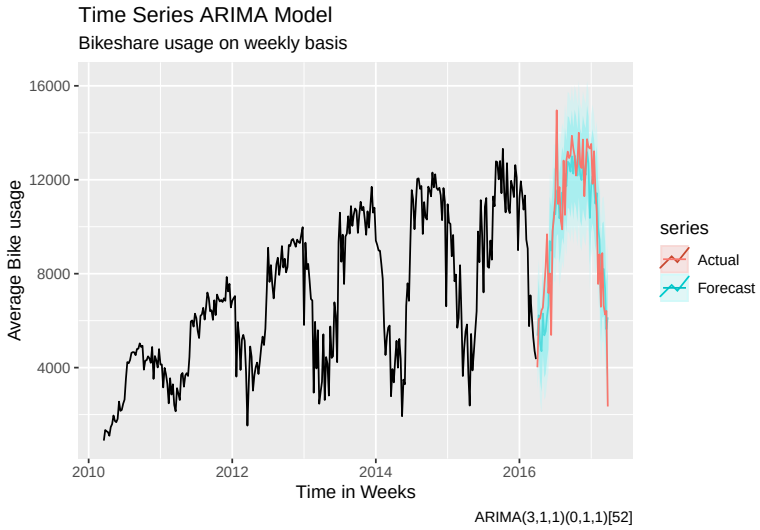
The data was changed for the time series models to use weekly data instead of daily. The reason is for cleaner plots and reduced time in calculating some of the time series models. The total count was determined by taking the average over a week's time.



By changing the linear regression model to take in the trend and seasonality, we can see that we are doing fairly well at accurately forecasting the year of 2017. For this model we have a RMSE of **1659.65**.

ARIMA

When plotting the data as well as seeing the above Linear Regression with Time Series, one of the best models for dealing with Time Series is the ARIMA and its derivatives.



It was determined that the best ARIMA model for our data was `order = (3, 1, 1)` and `seasonal = (0, 1, 1)`. This model had an RMSE of **1503.639**. It was better able to forecast/predict the final year as compared to the linear regression model.

LightGBM

Model Selection and Optimization

Model Performance

Detailed evaluation and interpretation of results

Conclusion

One finding was that forecasting effectively requires more data than two years of usage. Our initial models trained just on 2011-2012 data did not greatly outperform linear regression.

Works Cited

Bean, R., Pojani, D., & Corcoran, J. (2021). How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones. *Journal of Transport Geography*, 95. https://doi.org/10.1016/j.jtrangeo.2021.103155.

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, 54. https://doi.org/10.1016/j.scs.2019.101882

Ashgar, H. I., Elhenawy, M., & Rakha, H. A. (2019). Modeling bike counts in a bike-sharing system considering the effect of weather conditions. *Case Studies on Transport Policy*, 7(2), 261-268. https://doi.org/10.1016/j.cstp.2019.02.011