Forecasting Bikesharing Usage for DC's Capital Bikshare System

Table of Contents

Introduction

Bike sharing systems are an increasingly popular solution in major urban areas to increase the usage of bicycles as a mode of transport. Due to the bike being used in place of the car in some situations, the riding of bikes helps to improve the lives of both users. It also help non-users since it reduces the number of cars on the road and CO_2 emissions. 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 usage, then it would allow the company to plan for the best route to increase their fleet as well as expanding the available stations that are offered. A station is a location where the bikes are stored and can be rented.

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 the business model?

We made an initial hypothesis that we would see a higher usage during the summer months and when the weather was nice (sunny no rain). With our initial exploratory data analysis, we could see that the usage data contains a trend and seasonality. 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 impact?

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. This is 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 exploratory data analysis (EDA) performed on the dataset indicated we did not have sufficient data to fit any models. We found that we could scrape the Capital Bikeshare website in order to acquire more usage data. The only issue with this is that it didn't include weather data. Therefore, we decided to also scrape weather data.

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's website directly, then we ran `join_data.py` (which imports from `noaa.py`) to join the data with weather data from the NOAA API. All these code files can be found in our `Other Resources` directory.

Upon scraping the additional data, we combined this with our initial data set to create a new dataset that spanned from 2011 to 2017. We also added a few new feature variables such as `season`, which was missing from the new scraped dataset.

Off of this dataset, we also created another dataset that was used for time series models. Instead of forecasting on a daily, which is how the new dataset was created, we created one that was weekly. By creating a weekly data set, we are taking the mean of the usage. This will help smooth over any outliers, but they wouldn't be removed. It allowed for a cleaner more presentable dataset. This also makes sense to have it on a weekly, since you don't want your employees to have to constantly change bike quantities on a daily basis, but at least a weekly timeframe.

Therefore, we had one dataset that we performed EDA and performed linear regression and LightGBM on. All time series models were run against the weekly dataset.

The Weather data included:

- `TMAX`: high temperature (in tenths of degree Celsius),

- `PRCP`: precipitation (tenths of mm), and
- `AWND`: average daily wind speed (km/h).

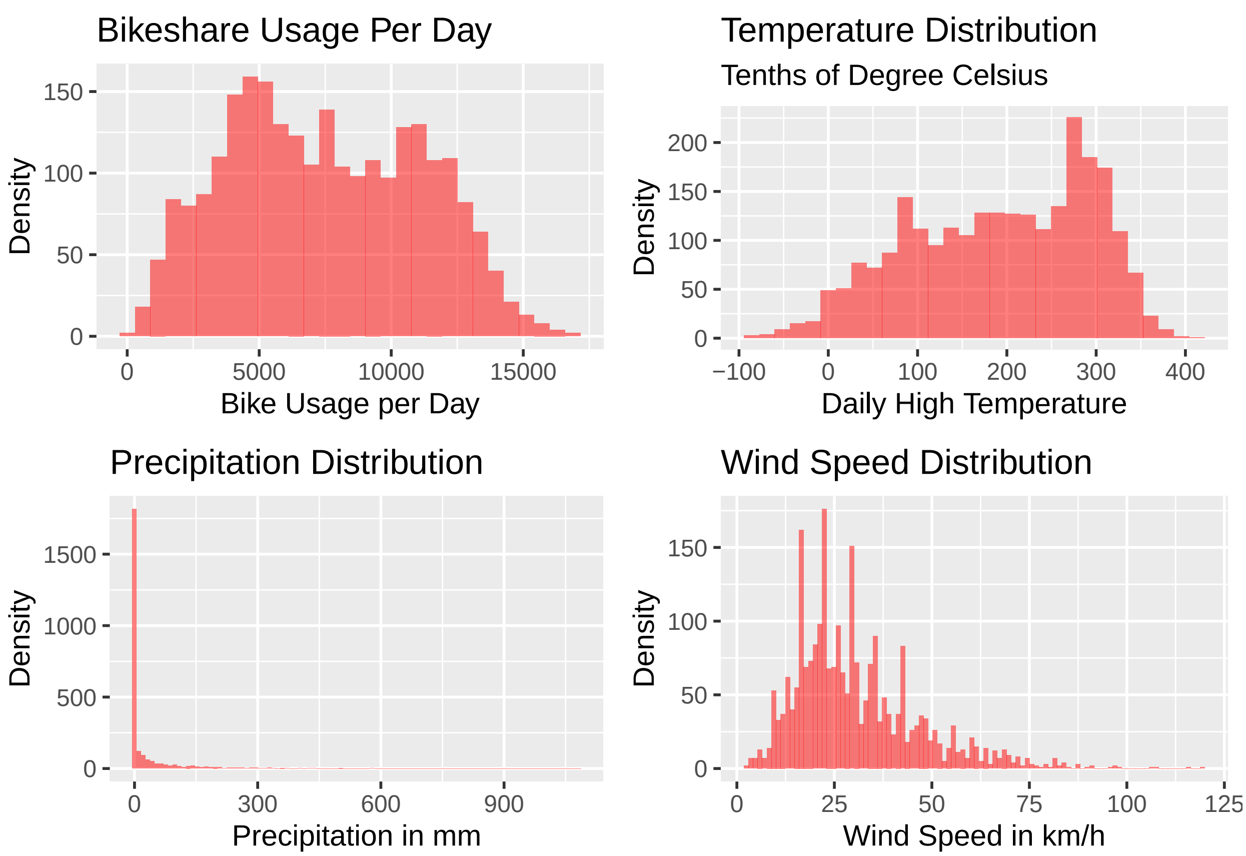
The final merged dataset used for analysis is `final_df.csv` located in the `Data` folder.

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)

Exploratory Data Analysis

We can see that fortunately, Bikeshare usage per day is not heavily skewed, but has closer to a normal distribution.



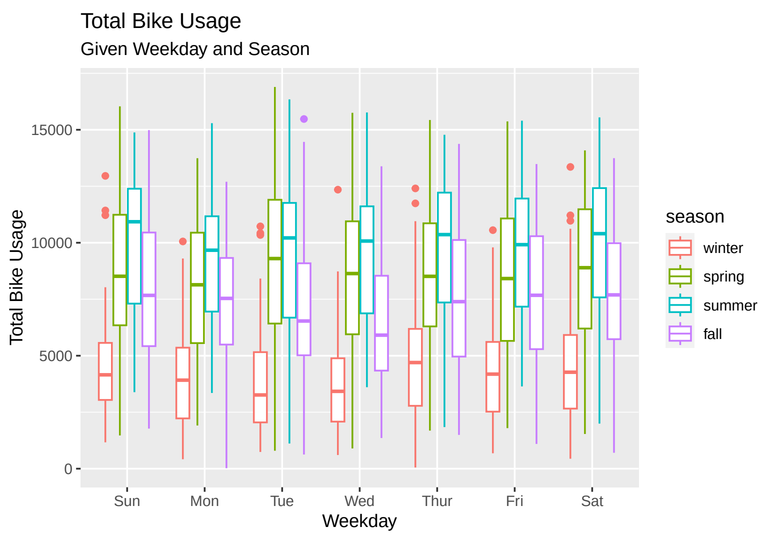
We can see that the temperature distribution has a leftward skew. This makes intuitive since because there would be fewer cold days.

We can also see that wind speed and precipitation exhibit rightward skew, which makes sense:

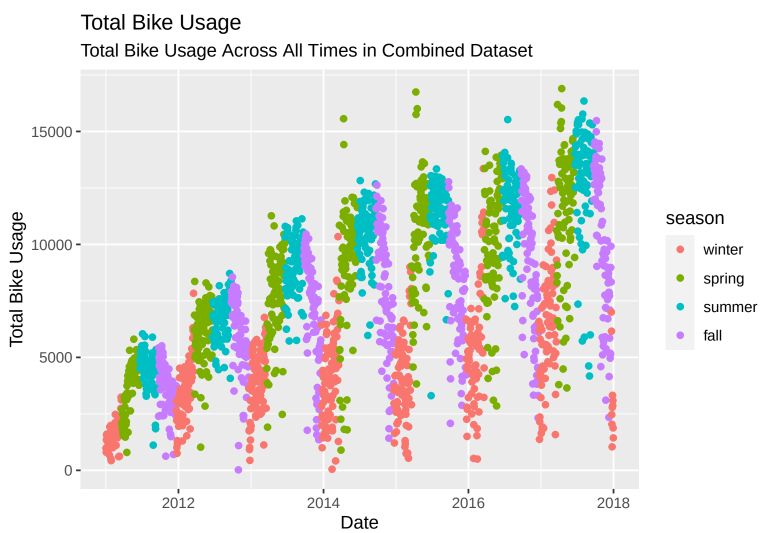
- Most days have no precipitation

- Most days have moderate wind speed, while a few have very high wind speed

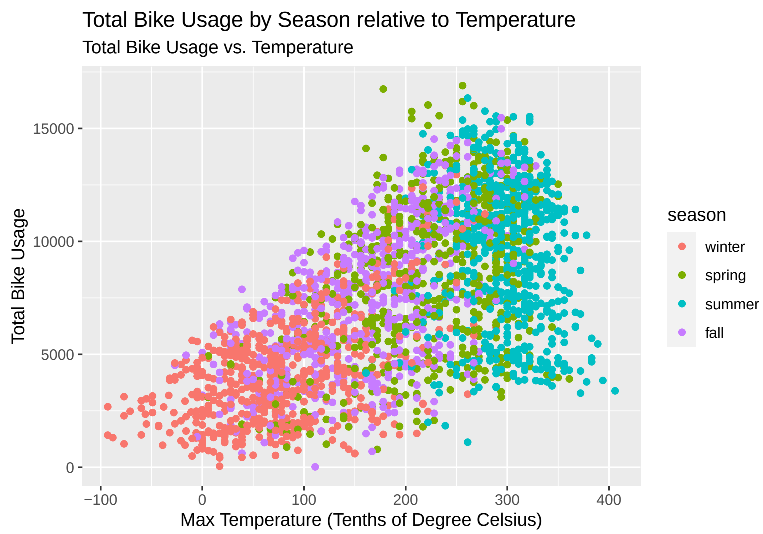
The data depicts that there is a "steady" state for each of the distributions that the data is trying to form around. We expect that in general we would have warmer days with less rain and low wind speeds.



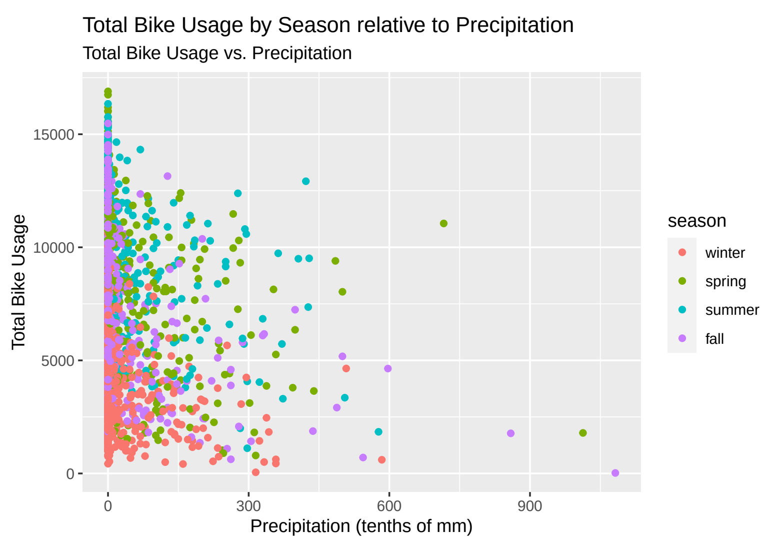
We can see in the above chart the effect of season on the bike share usage. Usage is lowest in the winter, highest in the summer, with spring and fall in between. For the most part, the usage per weekday seems to be consistent with minor fluctuation (correlating seasons). Fortunately, there is not a huge amount of variation in usage between weekdays and weekends.



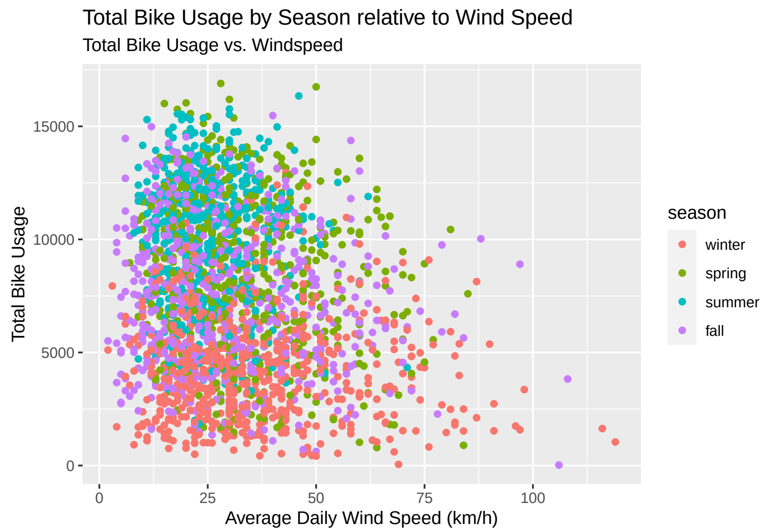
The most challenging aspect of our dataset is that usage was not static. It grew from 2011 through 2017, which shows an upward trend. While this was obviously good for Capital Bikeshare, it meant that models would need to take into account not just variation within a year but an upward trend.



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



The relationship between precipitation and usage appears somewhat weaker than temperature. Still, the highest precipitation days tend not to have high usage, consistent with an overall negative effect of precipitation on usage. This is also somewhat hard to definitively define, unlike temperature, since rain isn't as consistent. You could almost see higher precipitation as an outlier in the data.



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

A correlation matrix comparing our temperature, precipitation, wind speed and usage variables validates our graphical EDA:

- There is a positive correlation between temperature and count
- There are negative correlations between precipitation and count, as well as wind speed and count

	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

We can also see the negative correlation of temperature with wind speed. Fortunately, precipitation has minimal correlation with temperature (-0.006) and relatively low correlation with wind speed (0.093). You would expect wind speed and precipitation to be somewhat correlated since there is often a breeze when it is raining.

Key Predictors

The key predictors for time series will be time, which is made up of a trend and seasonality component.

Overview of Modeling

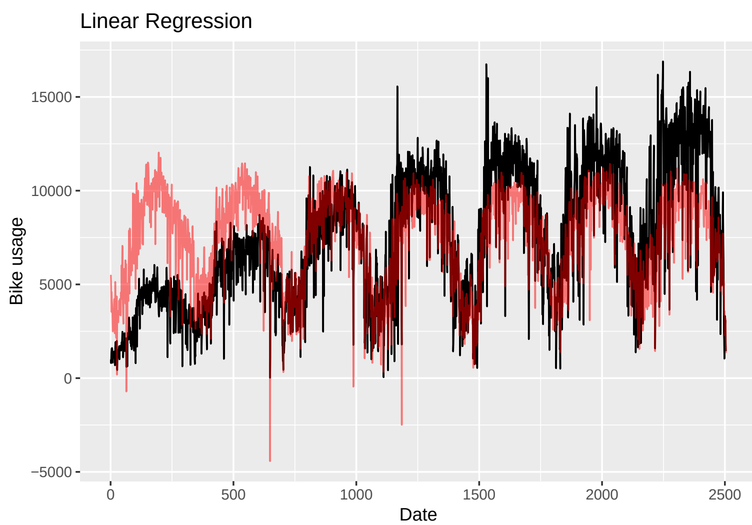
For our modeling, we used the `final_df.csv` file in the `Data` directory, which is described in the accompanying readme on GitHub page. We also used an edited data set that converted the above dataset into weekly instead of daily for the time series models. We wanted to run models that were able to forecast in the future as well as the capability of classifying on a daily basis. We decided to use time series models to forecast into the future. We used different trees, gradient boosting and regression models to be able to classify the bike usage given some inputs.

Model Types and Comparison

We ran multiple different models to try to 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 Linear Regression (with and without time series), ARIMA and LightGBM (gradient boosting). Other models were run and can be found in our `Code/Models` directory.

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.

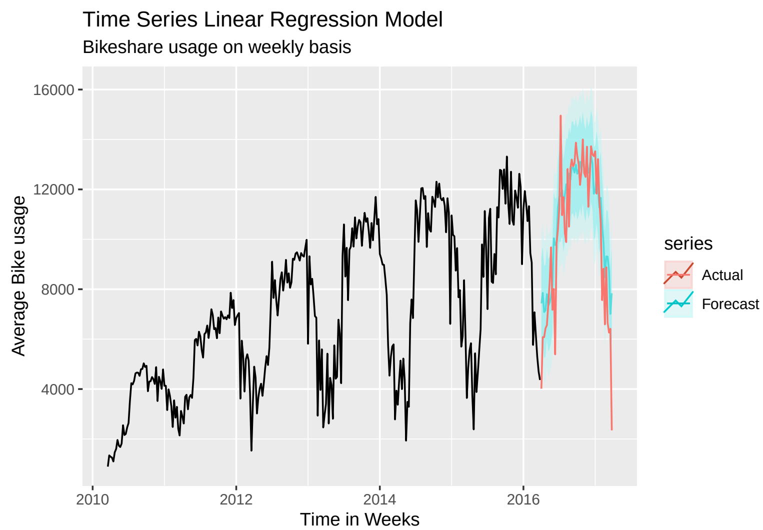


A naive linear regression does not incorporate any time series factors such as trend. It has an adjusted R-squared of 0.45. We can see that the linear regression over-predicts usage in early years, and under-predicts usage in later years, due to lacking a trend/time series component.

Linear Regression with Time Series

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

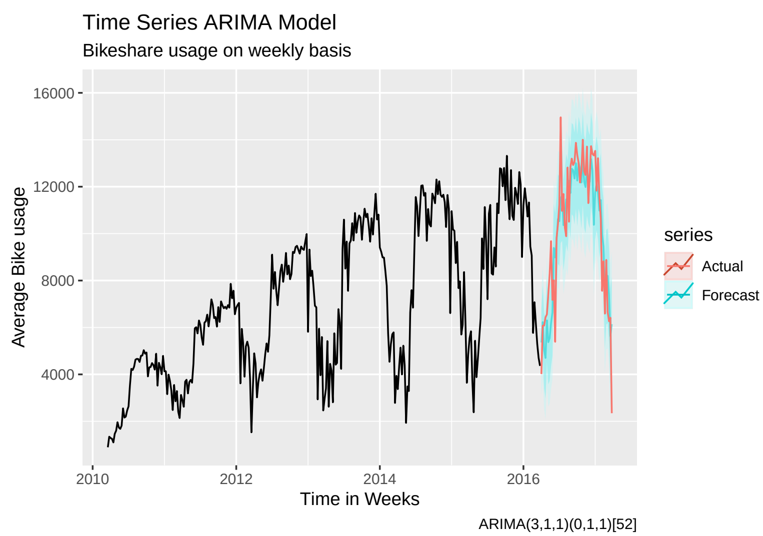
Therefore, the data was changed for the time series models to use weekly data instead of daily. The reason is for cleaner plots and reduced time in calculating some of the time series models. The total bike usage was determined by taking the average over a week's time.



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 using a test set, we have a RMSE of **1659.65**. The test set is the entire year of 2017. It was with held from the training data.

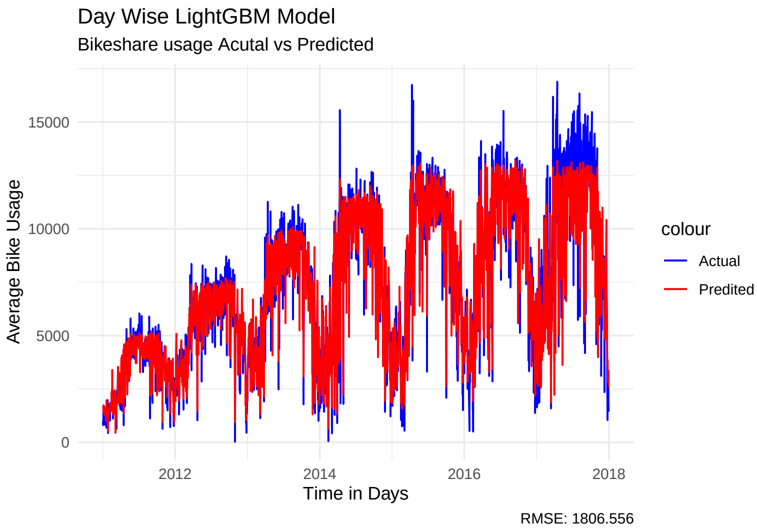
ARIMA

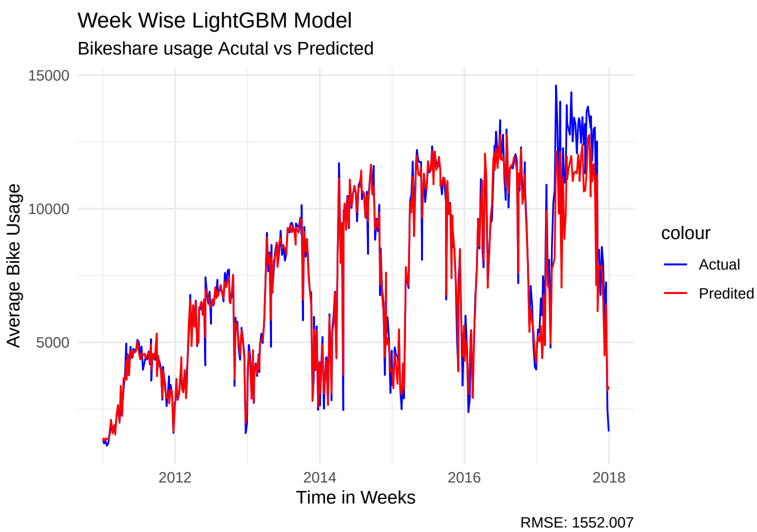
When plotting the data as well as seeing the above Linear Regression with Time Series, we know we have a time series problem. One of the best models for dealing with Time Series is ARIMA and its derivatives.



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





We implemented a powerful and efficient gradient boosting framework to predict/forecast the usage for the later years. The Predicted year of 2017 is out of sample i.e the test data set. Pre 2017, the predicted line describes the fitting on the training data set itself. However, despite the overall success in capturing trends, there's a notable limitation in predicting the increase in usage observed in 2017. The RMSE (Root Mean Squared Error) of weekly forecast is lower than daily forecast as it is able to remove the noise in the data and capture it well. Nevertheless, the model still struggles to accurately predict the surge in usage during 2017.

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 time series models trained on 2011-2012 data and performed poorly. They were unable to learn the trend and seasonality with so little time.

Literature Review Summary

We reviewed a few papers modeling bikeshare usage 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

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