Boulder Bicycle Traffic Forecasting

Jacob Munoz
University of Colorado Boulder
jamu0075@colorado.edu

Abstract

Boulder, Colorado, has consistently been recognized as one of the best cities for cyclists. This is based on factors such as the number of riders, safety, path network, and availability of the paths. Boulder has a dedicated riding community that may one day require traffic monitoring to improve safety and commuting time while providing the city of Boulder valuable information on the trends of their riders. The driving hypothesis of this research is that cyclists ride in predictable patterns based on their environment with an end goal of predicting bicycle traffic.

I. Introduction

Boulder, Colorado, has consistently been recognized as one of the best cities for bicyclists in the country[1]. The judging criteria for the ranking includes features such as ridership, safety, network, reach, and acceleration. Network being how well the bike network connects people to destinations, how long it takes to navigate Boulder. Reach is how well the bike network serves everyone equally. Acceleration is the city's commitment to growing bicycling quickly, maintaining and updating infrastructure. In 2019, Boulder was named one of the best cities in the U.S. for bikes out of 500 communities[2].

The goal of this research is to help the city of Boulder understand its cyclists' trends and to forecast bicycle traffic using recent observations and weather data. Cycling is a fantastic environmentally sustainable mode of transportation that should be constantly improving to encourage more riders. There may one day be a need for bicycle traffic mapping similar to that of current car traffic to promote safe and timely commuting.

While the results of this research provide some insight to cyclist patterns, there is plenty of room for forecasting improvements and expansion on location tracking. This is the first step towards real-time bicycle traffic modeling.

II. Data

Two data sets were used throughout this project. The first being bicycle counts at various intersections throughout Boulder and the second being daily weather data. The bicycle data is obtained from the City of Boulder website that has publicly available data that is updated regularly[3]. The weather data is obtained from the National Oceanic and Atmospheric Administration (NOAA) website and is also updated regularly[4].

The bicycle data includes the count of bikes observed at each intersection every 15 minutes. The intersections being observed are highly trafficked and capture many common routes in Boulder. Every intersection has data up until the current day and begins at various dates. Some sets begin early 2015 and others early 2016 but the longest, Folsom & Boulder Creek Path, begins 8/8/2011. It is uncertain how the data is collected but it is assumed to be a simple count of objects passing through each intersection's bicycle lane. This would capture other modes of transport such as skateboards or scooters but the vast majority of traffic in these lanes are assumed to be bicycles.

The weather data includes the daily temperature minimum and maximum in degrees Fahrenheit, snow cover in inches, and precipitation in inches. NOAA has records of weather readings beginning in 1897, although most features where not recorded until early 1900's.

This information is updated on a monthly basis and comes directly from NOAA's observations here in Boulder, Colorado.

While Boulder provides good coverage of data with their bike counts, this research will focus on one. Folsom & Boulder Creek Path is an important intersection given its central location and longevity of data collection. More importantly, this is the only intersection that is connected to the multi-use bike lanes that navigate Boulder. This is an important feature because these multi-use paths are more popular and practical than bike lanes, however Boulder is not yet collecting data on these paths. This intersection will capture some of the traffic that uses multi-use lanes to get into the heart of Boulder.

	date	total	tmax	tmin	precip	snow	snowcover	dayofweek
0	2015-01-03	47	35	15	0.03	0.70	8.0	5
1	2015-01-04	978	25	0	0.01	0.40	8.0	6
2	2015-01-05	813	56	2	0.00	0.00	5.0	0

Figure 1: *Folsom & Boulder Creek Path count with features*

III. Methods

To begin, a monthly count plot was created to view yearly seasonality. Figure 2 shows a clear seasonal correlation that repeats each year. The monthly counts remained mostly constant year after year. This supports the idea that bicycle traffic is a predictable variable.

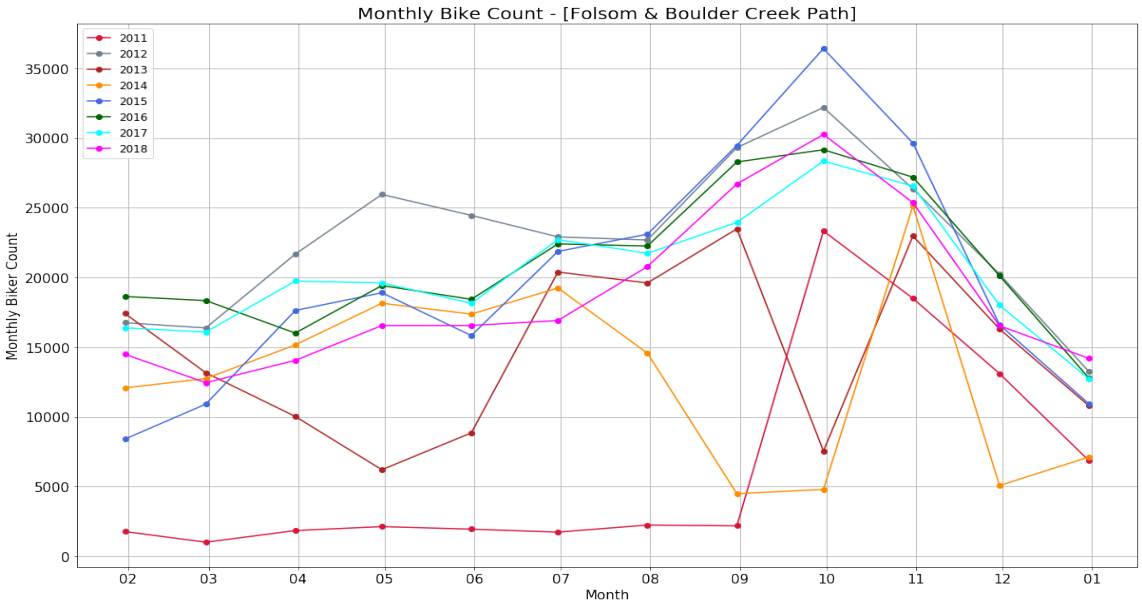

Figure 2: *Monthly bike count at Folsom & Boulder Creek Path 2011-2018*

Figure 3 shows daily temperature and monthly means that follows a very consistent climate pattern. For the last 10 years the monthly means for low and high temperature has remained very consistent with only a few degrees of variation. There is a clear correlation between the temperature and bike count throughout the year that follows the seasons.

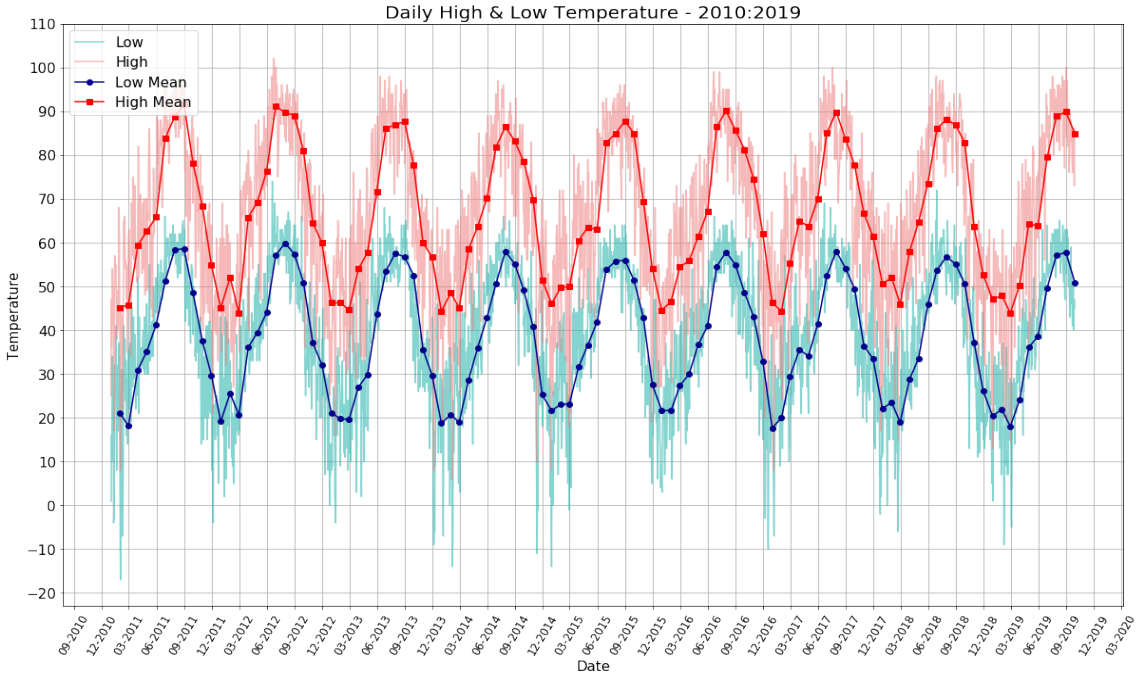

Figure 3: *Monthly average temperature(degrees Fahrenheit) with daily observations 2010-2019*

Before a model could be trained the bicycle count data had to be cleaned. Folsom & Boulder Creek Path's data longevity was appealing, however, after cleaning the data it became comparable to the other intersections (2015-2019). The years 2011-2014 were removed due to wildly different monthly counts compared to more recent years, as shown in figure 2. This may be as a result of physical changes to the bike lane or something else early on but for the purpose of forecasting, that will be saved for later exploration.

The number of daily zero-counts was 41 (2.25% of 2015-2018). When looking at the data the zero-counts could be observed to be grouped by three or more days in a row, and repeated monthly. This is highly unlikely in reality and may suggest the counting mechanism was offline for some scheduled maintenance. For this reason all zero-counts (within the daily grouping) were removed. Furthermore the data had only two outliers and were three or more standard deviations out. These dates did not appear to be any known holiday or event (i.e., Boulder bike-to-work day) and thus were removed. After these changes the skew for the Folsom & Boulder Creek Path count was 0.53. The daily counts was significantly right skewed and the data was standardized in an attempt to mitigate this skew. The data was standardized using sklearn[5] and resulted in a standardized skew of 0.49. The skew is difficult to interpret but it seems likely due to the fact that the counts vary drastically and within a wider range than initially thought. At this point regression was used on the Folsom & Boulder Creek Path data set to make simple predictions.

A simple linear regression model was built using the daily temperature high as the dependent variable with sklearn LinearRegressor[6]. Daily high was used after observing the correlation table, figure 4, for the total and features: daily low, daily high, precipitation, snow, snow cover, and day of the week. Daily max temperature outperformed all other features.

	total	tmax	tmin	precip	snow	snowcover	dayofweek
total	1.000000	0.392215	0.358899	-0.170514	-0.205226	-0.249324	-0.222049
tmax	0.392215	1.000000	0.890617	-0.197196	-0.319635	-0.476324	0.004508
tmin	0.358899	0.890617	1.000000	-0.044057	-0.252350	-0.461541	-0.018285
precip	-0.170514	-0.197196	-0.044057	1.000000	0.541321	0.268217	-0.013865
snow	-0.205226	-0.319635	-0.252350	0.541321	1.000000	0.633644	0.010305
snowcover	-0.249324	-0.476324	-0.461541	0.268217	0.633644	1.000000	0.003632
dayofweek	-0.222049	0.004508	-0.018285	-0.013865	0.010305	0.003632	1.000000

Figure 4: *Folsom & Boulder Creek Path correlation amongst features*

After examining each feature individually, multiple linear regression was used to complicate the model slightly. Features where chosen based on correlation to the total count and observed model performance. The best model is most accurate when using maximum temperature, and day of the week as features.

Ultimately, given the observed behaviour of the data, time series analysis became the focus[7]. When observing the auto correlation and partial auto correlation in figure 5 we can see clear seasonal trends amongst the daily groupings.

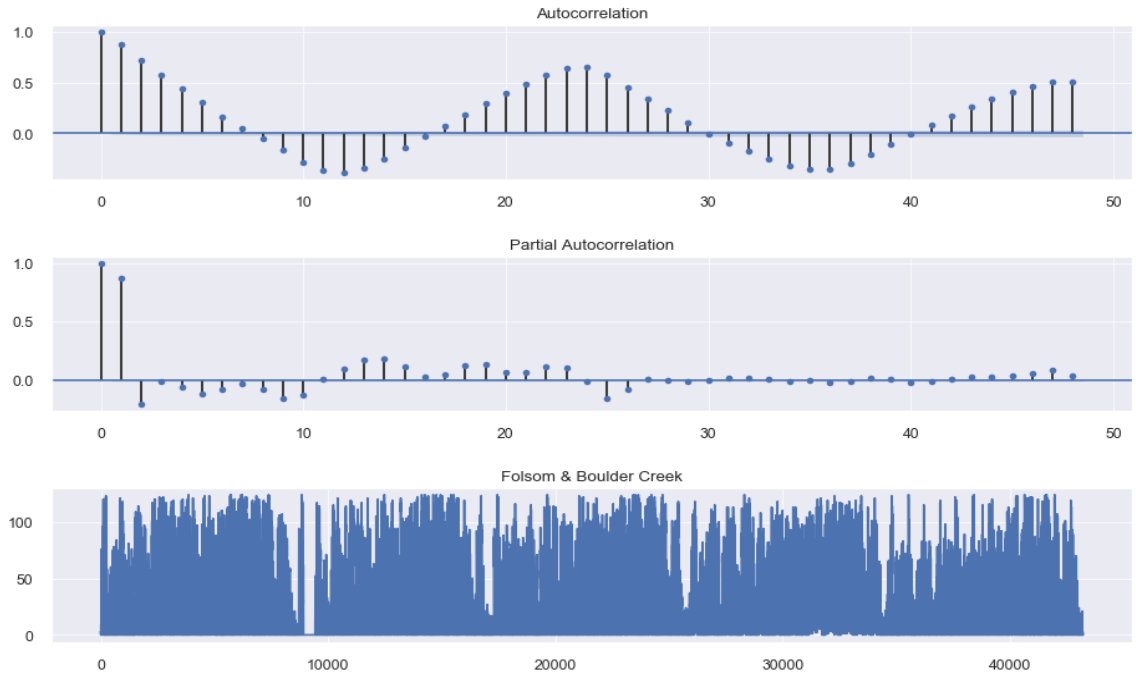

Figure 5: *Autocorrelation and Partial Aurocorrelation plots for Daily counts at Folsom & Boulder Creek Path*

FBprohpet[8] was used for time series forecasting due to its versatility and customization options. It is capable of capturing multiple seasonalities at once including sub-daily, daily, weekly, and yearly. It is also robust in its ability to handle outliers and trend change over time. Sub-daily forecasting was explored with some levels of success but the majority of the modeling was spent working with daily forecasting. That being said, sub-daily forecasting has great potential for success and practical use.

IV. Results

The results of univariate linear regression were not great but showed promise, specifically for maximum daily temperature as a feature. A p-value test was conducted using python StatsModels[9] to confirm correlation between bike count and features. The null hypothesis, there is no correlation between weather and bicycle count, was safely rejected after receiving a p-value of less than 0.01 with each weather feature.

The assumptions of linear regression that were acknowledged include a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity. Figure 6 shows the correlation between maximum daily temperature and bike count and there appears to be some amount of linear relationship, though not overwhelmingly. As for multivariate normality, the bike count and daily maximum temperature were both standardized to be on the same scale. Multicollinearity was not an issue for univariate analysis but was considered when exploring multivariate regression. Auto-correlation is however present in the data. As observed previously in figure 5, there is clear auto-correlation which negatively affected the performance of linear regression and led to time series analysis. As for homoscedasticity, the residuals in figure 7 appear to have a skew similar to that of the observed values which indicates room for improvement. The best univariate linear model had a root mean squared error of 358.12 and an R-squared value of 0.15. Thus, the model has room for improvement and a multivariate linear regression model was the next step.

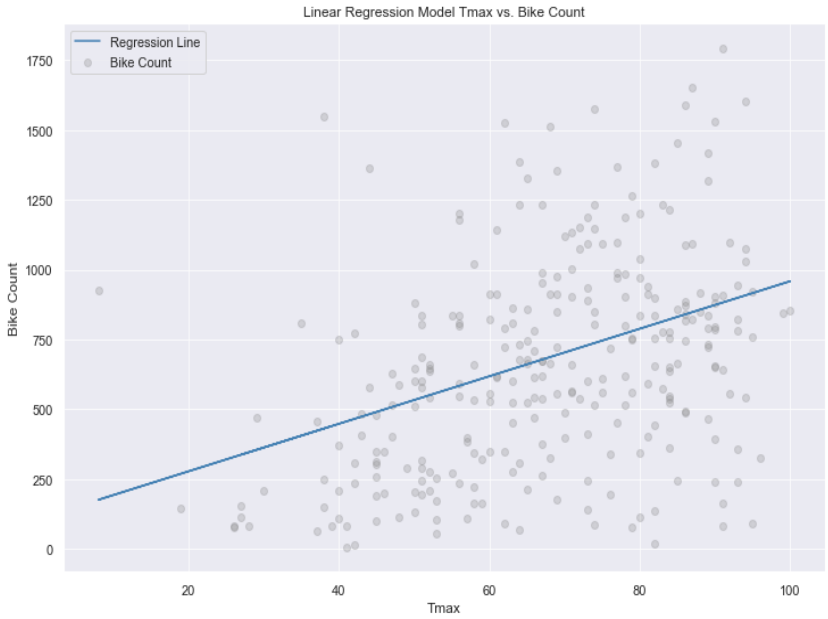

Figure 6: *Daily Count vs. Daily maximum temperature with a univariate regression line*

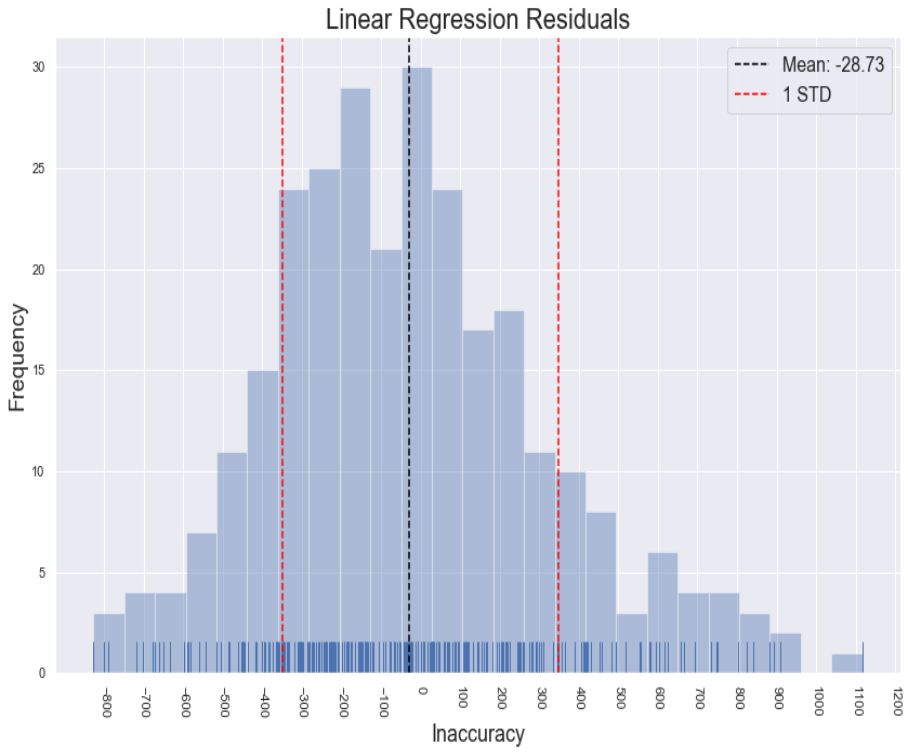

Figure 7: *Residuals for univariate regression predictions*

Many of the features were explored, but the best model was produced with the maximum temperature alongside day of the week as features. Surprisingly, the number of cyclists drops 25% on Saturday and Sunday. This suggests that there are a large number of cyclist who are commuter only, but there are other possible explanations. The root mean squared error got down to 338.85 with an R-squared value of 0.198. The residuals followed the same distribution of the univariate regression, confirming room for improvement. While still not an ideal model, there is significant improvement from the univariate linear model.

Although there is plenty of room for improvements with the regression models, focus shifted towards time series forecasting given the data auto-correlation.

The most success was found within daily forecasting. Using FBprophet, it was possible to forecast based not only on seasonalities but also extra regressors such as daily maximum temperature and day of the week. These extra regressors proved to provide a small increase in accuracy. The model is a multiplicative seasonality with a changepoint-prior-scale of 0.1 and daily, weekly, and yearly seasonality values of ten, eight, and fifteen respectively. This yielded a model that made the prediction in figure 8 below with the most success. Figure 8 shows the predicted values against the observed values. The model is able to capture the general trend quite well however there are several outliers. Daily bike counts certainly follow seasonalities however the unpredictable nature of humans results in dramatic shifts in bicycle count from one day to the next. This results in a model that fits for general trends but struggles with dramatic shifts in activity.

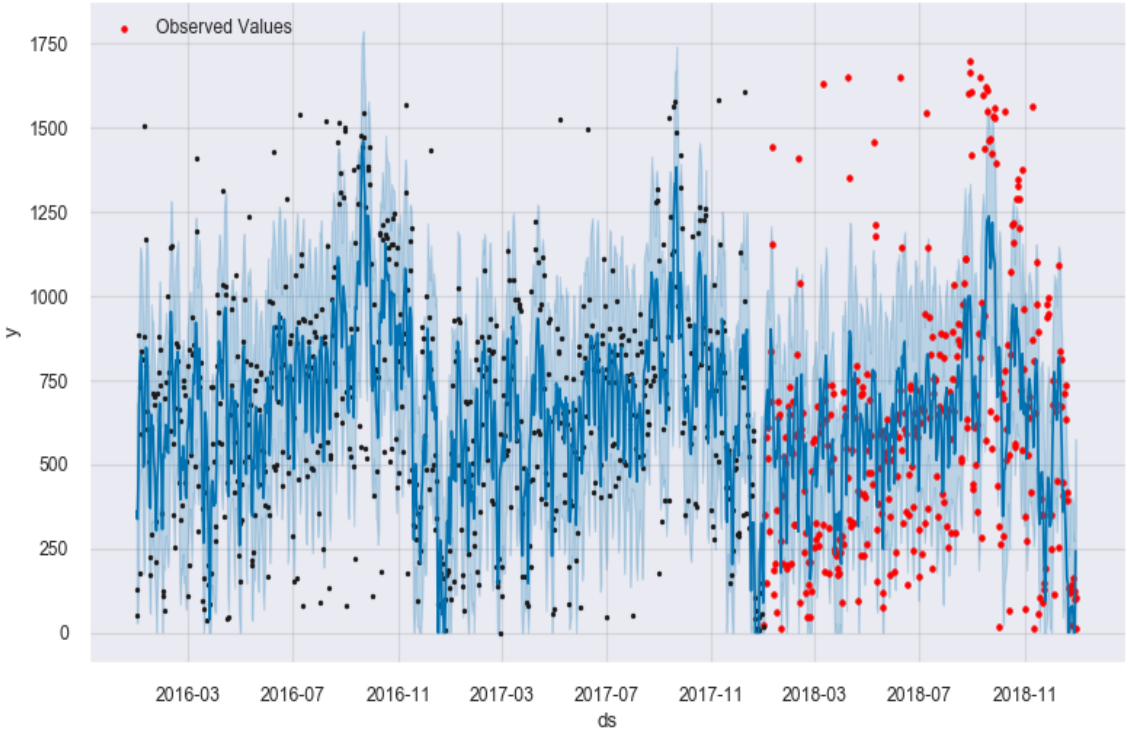

Figure 8: *The FBprophet model trained on 2016 and 2017 to predict 2018*

Figure 9, below, shows the observed values versus the predicted values with the range of possible predictions. It showcases the model's ability to capture the general trend while struggling to account for drastic peaks in activity. This could likely be improved with more time adjusting the models sensitivity to change. The residuals shown in figure 10 for the forecasting model show a much more normal distribution with a mean close to zero. This suggests the time series forecasting model more accurately captures the behaviour of cyclists with a more uniform error.

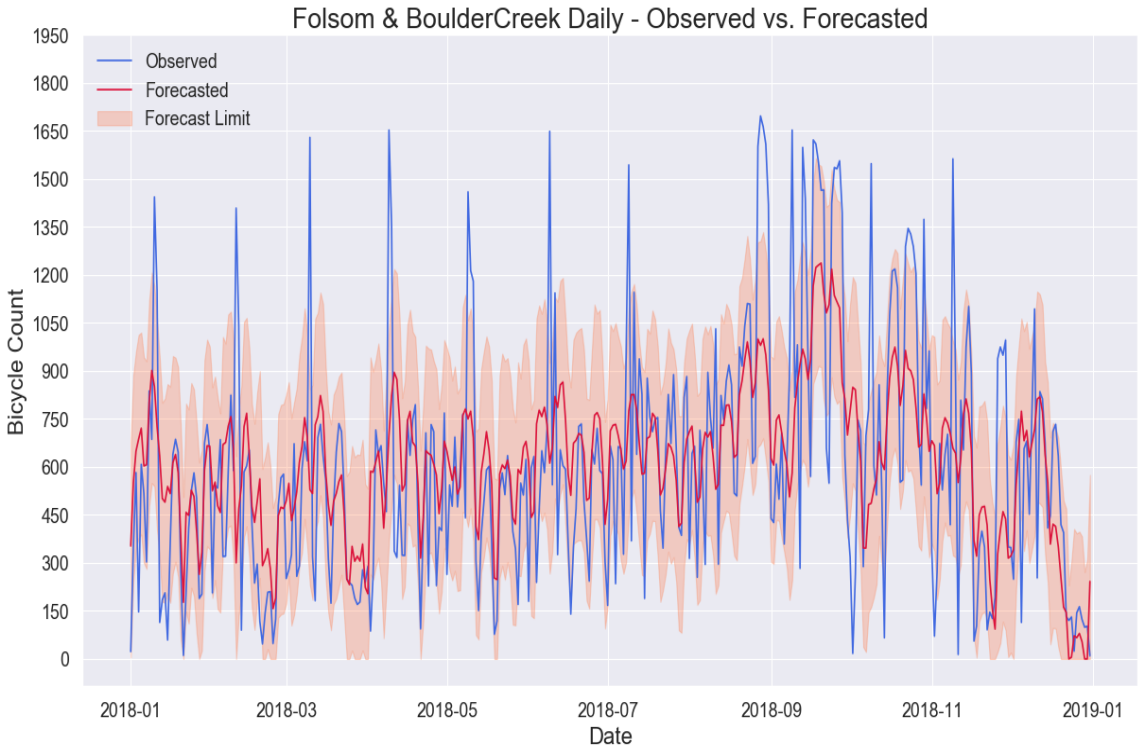

Figure 9: *The FBprophet model predicted values vs. observed values for 2018*

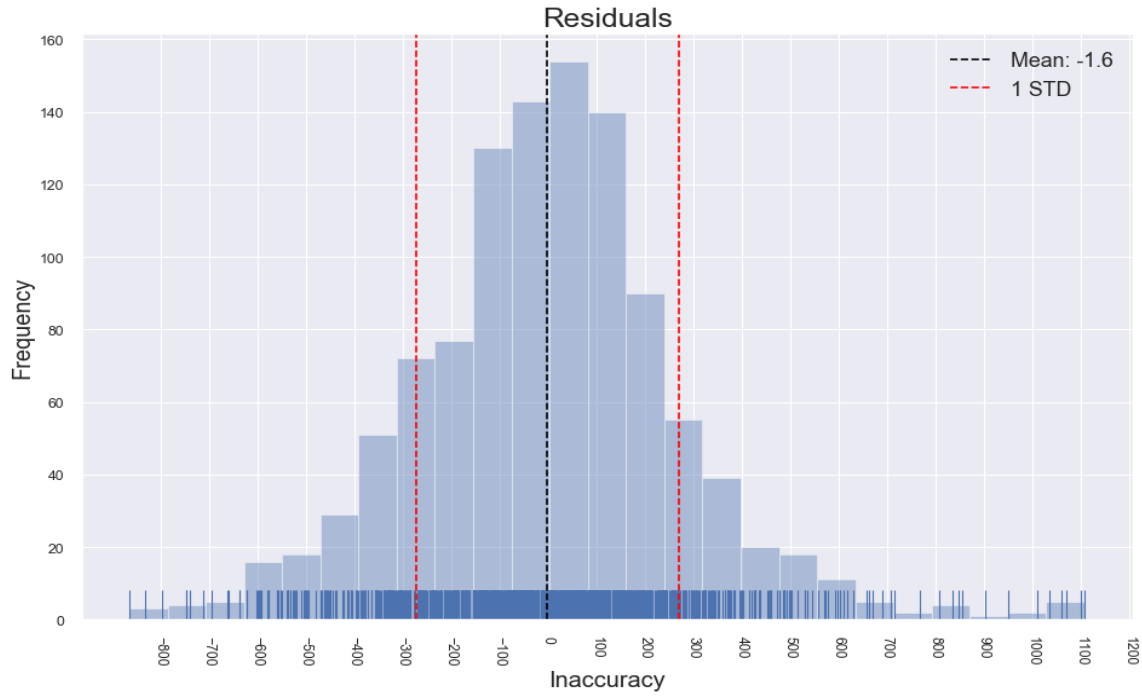

Figure 10: *The FBprophet model residuals*

When comparing the observed values to the predicted values they show a definite correlation, a calculated correlation coefficient[10] of 0.67. The model is able to more accurately predict counts between 450 and 700, possibly because the daily average falls in this range. Finally, the calculated root mean squared error of 270.90, a significant improvement upon the multivariate regression model.

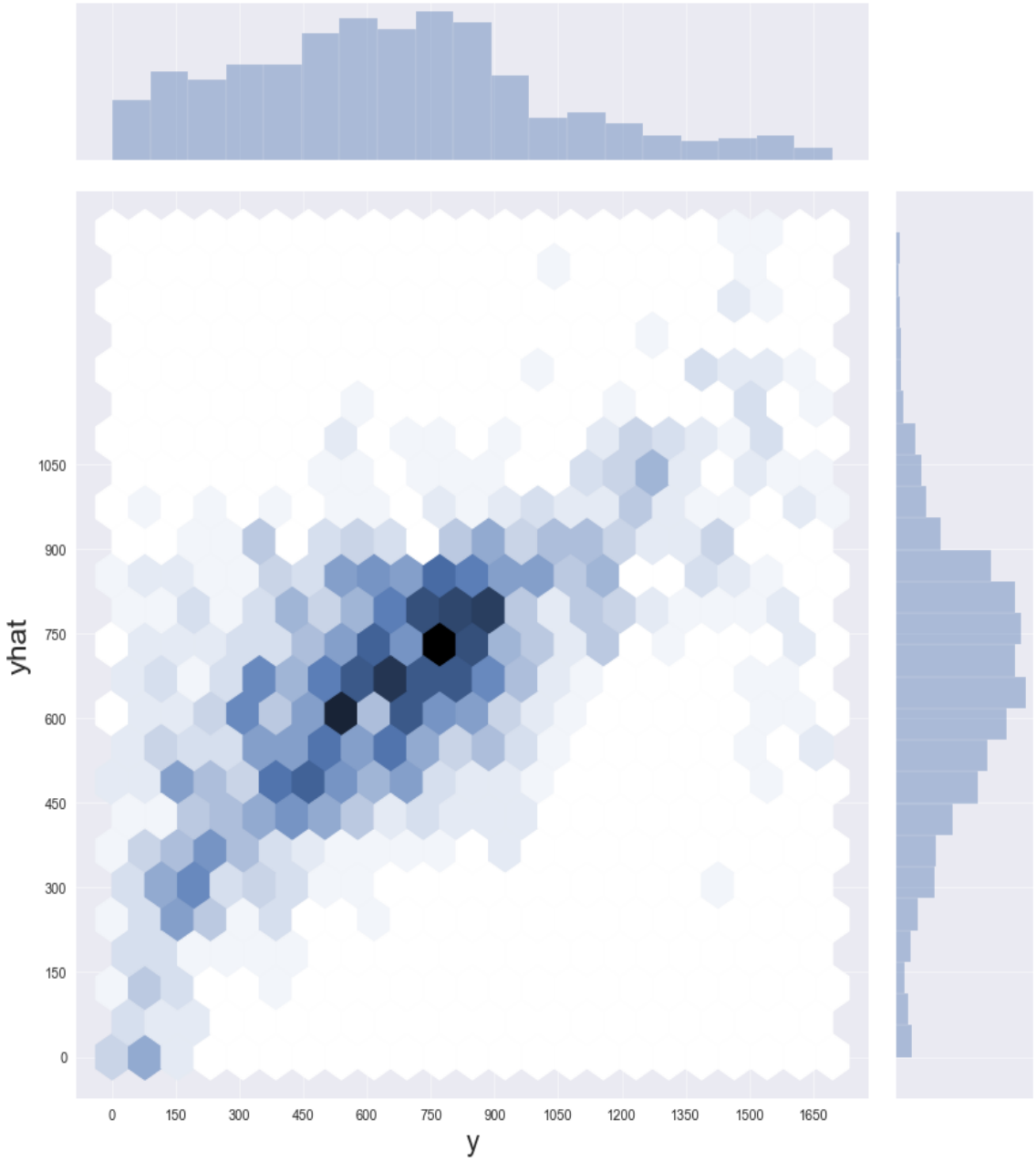

Figure 11: *Correlation plot of observed values vs. predicted values.*

V. Discussion

Boulder, Colorado, has a great cycling community that is ever growing and the data can be used to further improve the infrastructure in place. Forecasting bicycle traffic has provided an insight into when and where people bike given a small amount of information on the environment. Originally, the goal of this research was to provide a broad analysis of every intersection provided by the data. This could provide a more specific analysis of when and where people ride including northbound and southbound traffic during morning and evening work commutes or if the quality of the bike lane affects traffic. The shift in focus to forecasting came with the hopes of predicting traffic by understanding how the weather affects cyclists. In the future there may be a practical use for real-time bicycle traffic monitoring and predicting to improve the commuting experience for riders and to collect useful rider data to improve the infrastructure. While the daily forecasting returned decent results, there is plenty of room for improvement in the future. With more time we would like to explore more features such as the length of the day, or hourly weather data to better predict sub-daily traffic. The focus of this research was daily forecasting, however, sub-daily forecasting is possible in the future given more time and could be more useful. This research was the first step towards better understanding the cyclist community in Boulder, Colorado, and we are excited to continue this work into the future.

References

[1] [AC Shilton and the Bicycling Magazine Editors, 2018]
The Best Bike Cities in America,
https://www.bicycling.com/culture/a23676188/best-bike-cities-2018/

[2] [People For Bikes, 2019]
2019 City Ratings: Top Overall Cities,
https://peopleforbikes.org/blog/2019-city-ratings-top-5-overall-cities/

[3] [City of Boulder, 2019]
City of Boulder Bicycle Traffic Counts,
https://bouldercolorado.gov/open-data/bicycle-traffic-counts/

[4] [NOAA, 2019]
NOAA Boulder Daily Data,
https://www.esrl.noaa.gov/psd/boulder/getdata.html

[5] [Scikit Learn, 2019]
sklearn preprocessing StandardScaler,
https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[6] [Scikit Learn, 2019]
sklearn LinearRegressor,
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegres

[7] [Hyndman, R.J., Athanasopoulos, G, 2018]
Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia,
OTexts.com/fpp2

[8] [Prophet, 2019]
Facebook Prophet,
https://facebook.github.io/prophet/

[9] [StatsModels, 2019]
StatsModels p-values,
https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLSResults.pvalues.html

[10] [Pandas, 2019]
Pandas Correlation,
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html