Bike Sharing Analysis and Demand Forecasting

Purusanth Shanmukanathan
Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
purusanths.20@cse.mrt.ac.lk

Kanarupan Kularatnarajah
Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
kanarupan.20@cse.mrt.ac.lk

Jayani Hellarawa
Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
jayani.hellarawa.19@cse.mrt.ac.lk

***Abstract*—This paper discusses about the time series analysis that can be used for forecasting the demand of the bike rentals. Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. But, one of the major limitations of such an automated process is addressing the varying demand for tangible/limited resources across multiple operational points. Therefore the availability of information related to future business demand plays a big role while smoothing out the operation. The outline of the paper is as follows. Section one gives the introduction to the research, section two and three talks about the related work and the methodology. Section four to six explains the data set, implementation details and results and discussion. Then finally, the conclusion.**

***Index Terms*—time series analysis, forecasting, ARIMA, Prophet, bike-sharing**

I. Introduction

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Capital bikeshare is a metro DC's bike sharing service with 4,500 bikes and 500+ stations across 7 jurisdictions. The system is designed for quick trips with convenience in mind to make it fun and an affordable way of getting around. Through these systems, the user is able to easily rent a bike from a particular position and return back to another position. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

With all those real-world importance of these bike-sharing systems, the characteristics of data being generated by these systems make them attractive for the research and have the ability to be used as a way of optimization and even a way to address some of the business-critical issues. The old data with features that are recorded through these systems. As these systems give the user the flexibility to easily rent a bike from a particular position and return back at another, rental stations face the difficulty of resource distribution due to the varied demand with limited numbers of bikes. Therefore the bike rental stations have to address the rotation of bikes in order to keep up with the demand. If an organization has the capacity to better forecast the sales quantities of a product, it will be in a more favourable position to optimize inventory levels. Therefore the purpose of this research is to perform both exploratory data analysis and predictive analysis of this bike sharing data set to give proper insight of what is currently happening with the business as well as to forecast what the future demand will be in several dimensions.

II. Related work

A series of data points that are indexed (or graphed) in time order is known as a time series. A time series can be broken down into 3 components [3]. That are

a) Trends:: Upward downward movement of the data with time over a large period of time

b) seasonality: : Seasonal variance

c) noise: : Spikes troughs at random intervals

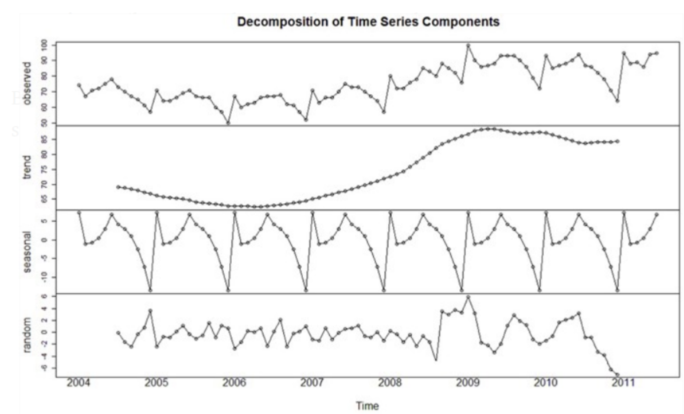

Fig. 1. [3]

If a time series is stationary and has a particular behaviour over a given time interval, then it is safe to assume that it will have same behaviour at some later point in time. In order to perform time series analysis, most of the statistical modeling methods assumes or requires the time series to stationary. Which means

- The mean of the series should not be a function of time
- The variance of the series should not be a function of time
- The co-variance of the ***i***th and the ***(i+m)***th should not be a function of time.If the spread becomes as the time increases, then the co-variance is not constant over the time.

Therefore, before applying any statistical model on a time series, we have to ensure it's stationary. There are two primary way to determine whether a given time series is stationary.

- **Rolling Statistics:** Plot the rolling mean and rolling standard deviation. The time series is stationary if they remain constant with time (with the naked eye look to see if the lines are straight and parallel to the x-axis)
- **Augmented Dickey-Fuller (ADF) Test:** The time series is considered stationary if the p-value is low (according to the null hypothesis) and the critical values at 1%, 5%, 10% confidence intervals are as close as possible to the ADF Statistics

In this research we have performed both methods to make sure that the data is stationary.

A. Auto Regressive Integrated Moving Average (ARIMA) Model

ARIMA stands for Auto Regressive Integrated Moving Average. There are seasonal as well as non-seasonal models that can be used for time series forecasting.The non-seasonal ARIMA model is obtained with a combination of the differences with autoregression and a moving average model. In addition to the non-seasonal ARIMA models, a seasonal ARIMA model is formed by including additional seasonal terms to give the time series attribute to the model.

*1) **AutoRegressive Model (AR)**:* In an AR model, it operates on the assumption that past values have an effect on current values. And these models can be commonly used to analyse nature, economics, and other time varying processes. This model attempts to predict the current values given the previous day values.

*2) **Moving Average Model (MA)**:* In this model, it assumes that the value of the dependent variable on the current day depends on the previous days error terms.

ARIMA model is a combination of both AR and MA models with the order of differentiation added.

B. Prophet Model

The Prophet is an open-source forecasting model implemented and published by Facebook. It has been the key piece to improve a large number of trustworthy forecasts on Facebook. It completely automates the forecasting process with an analyst-in-loop approach to incorporate useful assumptions or heuristics. There are a limited number of people who can do a high-quality forecast because forecasting is a specialized data science skill. The traditional statistical models are tediously hard to fine-tune for analysts who don't have forecasting skills but strong domain knowledge about the problem at hand. The Prophet has intuitive parameters that can be fine-tuned with domain knowledge about the problem. Since not all forecasting can be solved by the same algorithm Prophet is optimized for the following characteristics

- Observations are daily, hourly, the monthly granularity with at least a few months or a few years of data.
- Strong seasonality(eg. Day of the week , the month of the year).
- Support a reasonable number of missing values or outliers.
- Support historical events like holidays.

The Model can be decomposed into three main components and an error term.

$$y(t) = g(t) + s(t) + h(t) + e(t) \tag{1}$$

Here, g(t) is the trend function that models the non-periodic changes in the values of the time series, s(t) represent periodic changes in the time series such as week seasonality or yearly seasonality, and h(t) represent holidays which may occur in irregular schedules. By the domain knowledge that you have, if you believe holidays or other recurring events that have an impact on the demand you can add those in both training and test sets. If they will not repeat in the future, the Prophet will model them and will not include them in the forecast.

The error term e(t) represents the idiosyncratic changes that are not accommodated by the mode; later we make parameter assumption that e(t) is normally distributed.

C. Mean Absolute Percentage Error (MAPE):

Mean absolute error (MAE) is the difference between the measured/predicted value and true/actual value and the MAPE is the percentage equivalent of MAE. This is used to evaluate the model accuracy.

III. DATA SET

A. Data set Introduction

1) Source: Capital Bikeshare (abbreviated CaBi) is a bike-sharing system that operates in Washington D.C and some other cities in the United States of America. They have published their data [1] under this [2] license inviting any interested parties to perform analysis, development, and visualization. The data set consists of the below fields as per their description.

- Duration – Duration of trip
- Start Date – Includes start date and time
- End Date – Includes end date and time
- Start Station – Includes starting station name and number
- End Station – Includes ending station name and number
- Bike Number – Includes ID number of bikes used for the trip
- Member Type – Indicates whether the user was a "registered" member or a "casual" rider

During the descriptive and predictive analytic processes, Holiday, weekend/weekday information (the type of the day), hour bin, etc also are added to this data.

2) Preparation Steps: Data from 2018 January to 2020 February is chosen for the analysis. Source data are kept as monthly files hence all relevant files are downloaded, validated to have the same columns and order and merged to form the overall data set.

Date conversion is done so that hourly bin based descriptive analysis could be carried out. But the main analysis and analytic are done on daily data. The number of unique stations identified in the data set is 583. The approach is to choose the most prominent station and limit the analysis and analytic to that scope only.

B. Data Descriptive Analysis

1) Member Type Frequency: The membership types of the riders fall into two categories, registered and casual. The rides with membership registered are much higher than the rides of casual type. Below is the overall data categorization in terms of member type, for all stations.

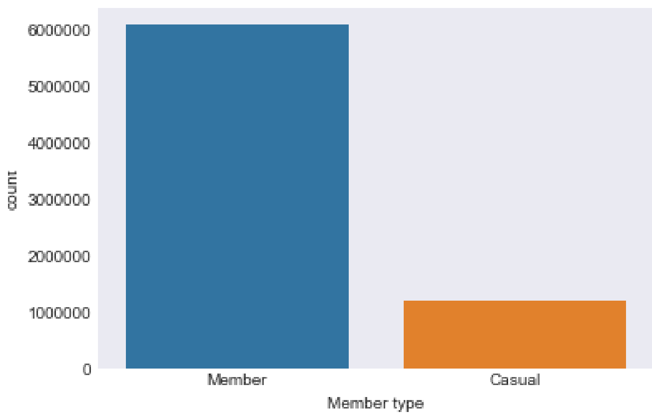

Fig. 2.

The figure 3 depicts the registered, casual and total rides per station. Only the top ten stations are chosen in terms of the number of rides. An interesting observation is that in certain stations the casual users are more than registered users. Some of these stations are are within the top ten as well (see below plot). Based on this trope the nature of the station could be pondered. For instance we could check whether the the particular stations are located near metro stations, residential area or remote tourist attractions.

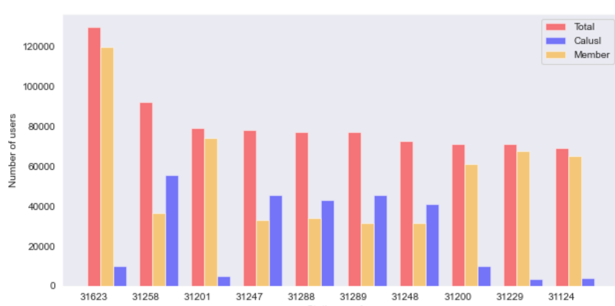

Fig. 3.

2) Select prominent station: As mentioned earlier the most prominent station is chosen based on having the highest number of rides throughout the chosen period. Columbus Circle / Union Station which is denoted with station identification number 31623 is chosen. It had 129514 total rides over the 26 months period.

3) Duration Related: Figure 4 and 5 are the duration statistics and plot describe the probability density function of duration. Higher values are lower and the plot has a long tail. The values are in seconds.

```
max:   84415
min:   60
count:   129514
median:   537.0
mean 744.35
std:   1506.28
var:   2268867.84
```

Fig. 4.

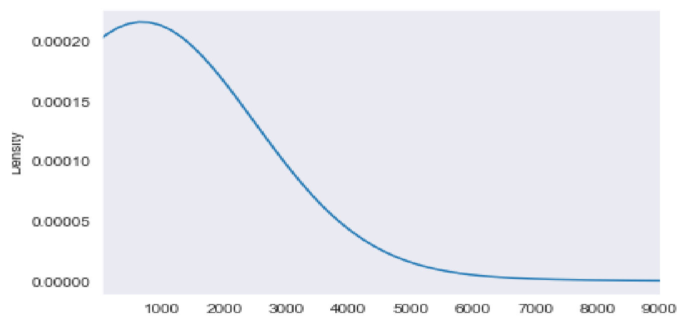

Fig. 5.

```
Considering duration more than 3600 an anomaly (normally rides are < 1800 secs)
Anomaly ratio:
Registered Member - 0.004626453609972392
Casual Member - 0.1287059294871795
Casual Member anomaly duration is 27.98 times more than Registered member
```

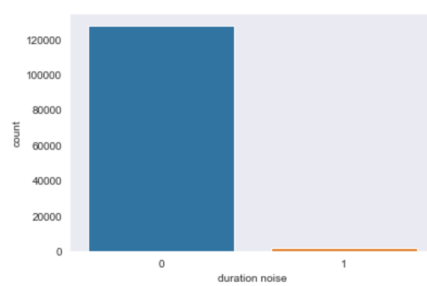

Fig. 6.

4) Membership-wise duration: Figure 6 and 7 reveal an interesting correlation between the membership type and the duration. Casual users in general are taking longer rides compared to registered users. And possibly anomaly values (above 3600 seconds; there are much higher values than that as one could see the max value is 84415) are higher in ratio perspectives among casual users. This could be resulted due

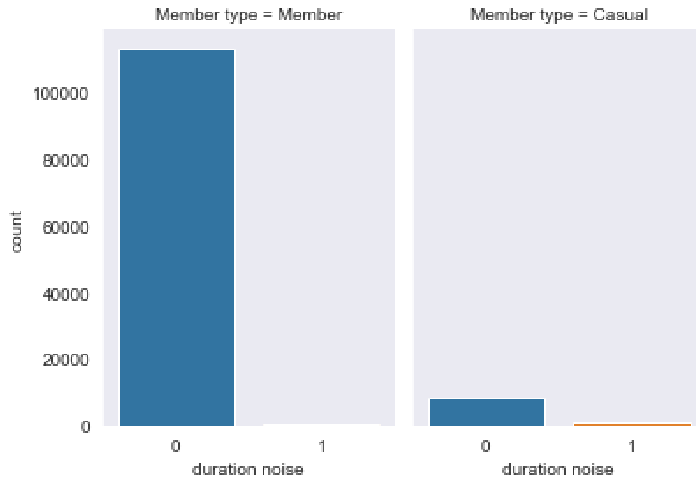

Fig. 7.

```
688.06
1084.95
1101.23
```

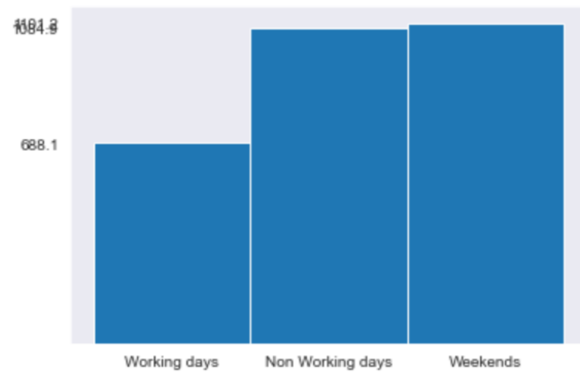

Fig. 9.

to technical faults, carelessness - leaving the bike unlocked, vandalizing etc.

5) Type of Day Related : Holiday: Related plot depicts the duration based on working days, non-working days and weekends. Here, working days: week days excluding national holidays Non working days: weekends and holidays

```
Mean duration seconds on working days: 605.62
Mean duration seconds on non-working days: 776.4
Mean duration seconds on weekends: 782.85
```

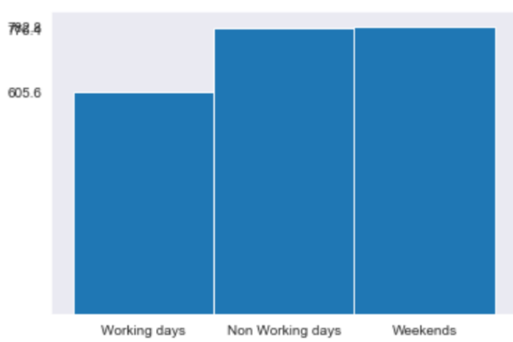

Fig. 8.

Similar values without eliminating the anomalies (above 3600) are given figure 9, working days, non-working days and weekends, respectively. On weekends the number of rides is less but the average duration is much higher than on weekdays.

But interestingly the weekend demand is lower compared that of weekdays. Hence, during the weekends despite the number of rides are lower, the average duration of those rides are considerably high compared to the former.

IV. Implementation

A. Analysis

The figure 10, stacked bar chart show number of users for prominent 10 station with member type discrimination. Stations 31258, 31247, 31288, 31289, are serving both registered and casual users equally likely and stations 31623, 31201, 31200, 31229, 31124 are dominated by registered users. The station 31623 is the prominent station which is dominated by registered users. For the rest of the analysis we are focus of the station 32632

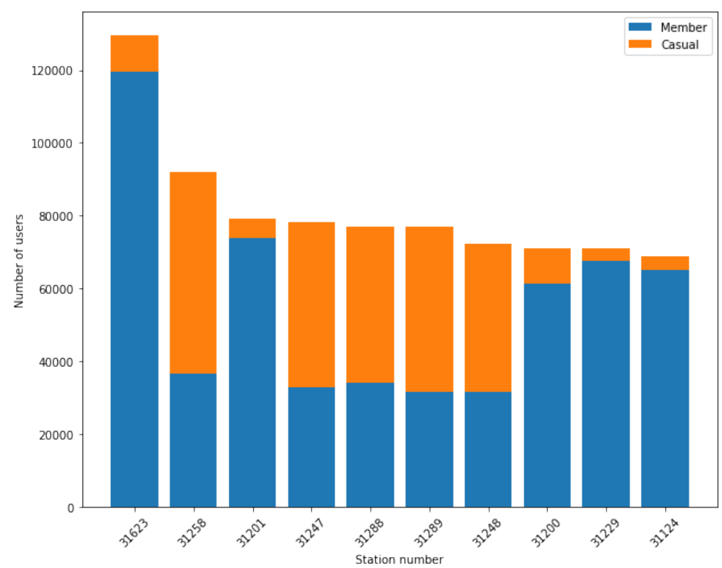

Fig. 10.

The figure 11, scatter plot there is a point for each day, and points are color-coded by day-of-week to show the weekly cycle. We can clearly visualize that the weekend demand is very low compared to other days and a higher number of people are using bikes on Wednesday. And a strong yearly seasonal pattern exists.

The figure 12 shows how forecast values are align with the original test series. It is clear that the model is able to produce cyclic, trend, pattern for the test data and able to align with spikes and dips.

By figure 13 diagram, it is clear that the model learns the

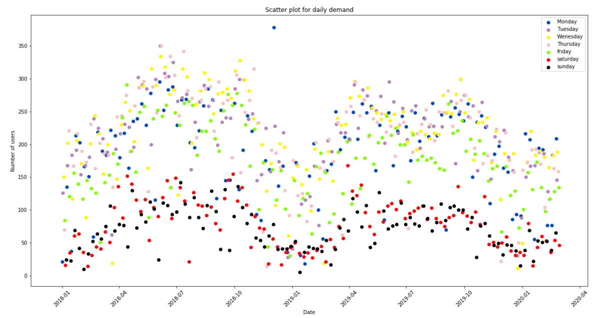

Fig. 11.

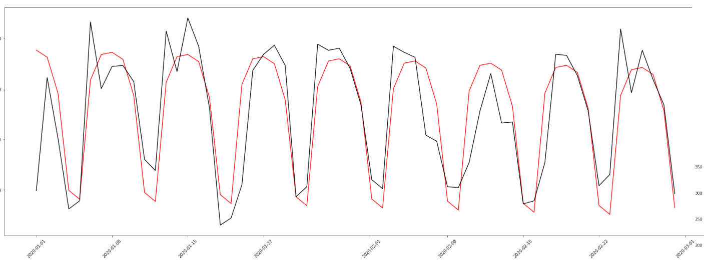

Fig. 12.

weekly seasonality that is weekend demand is very low and in the weekdays it is very high and relatively constant across the weekdays. There is an upward trend from January to June and a downward trend from June to December which is repeated for both 2018 and 2019.

In order to perform the predictive analysis of data related to member type, first the Augmented Dickey-Fuller (ADF) calculation and the rolling statistics are calculated on the initial training data to checked for the stationarity. Figure 14 shows the rolling statistics plots with rolling mean and the rolling standard deviation.

ADF output of the initial training data: figure 14

- ADF Statistic: -1.7881374833265393
- -value: 0.38636680443056337
- Critical Values:
- 1%: -3.439606888036868
- 5%: -2.865625121924057
- 10%: -2.5689454046801052

As the above ADF Statistic is far from the critical values and the p-value is greater than the threshold (0.05). Thus, we can conclude that the time series is not stationary. Therefore the same calculations are done with the converted log values of the data set. The ADF with log values (figure 15) gave the output for the stationarity as follows.

- ADF Statistic: -3.811579266087317
- p-value: 0.002791041144499671
- Critical Values:
- 1%: -3.4400031721739515
- 5%: -2.865799725091594
- 10%: -2.569038427768166

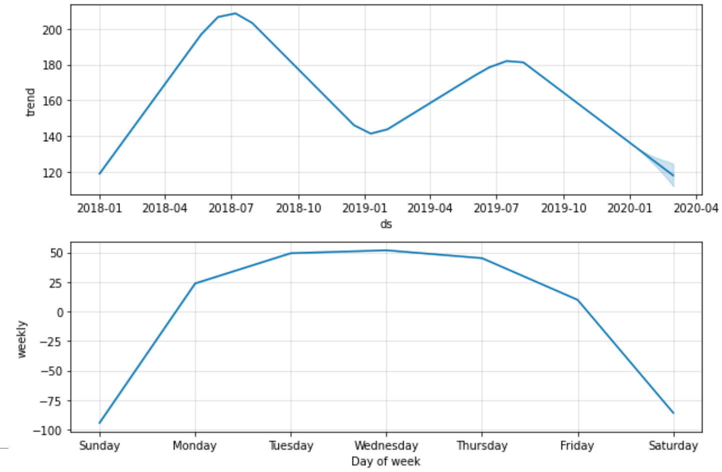

Fig. 13.

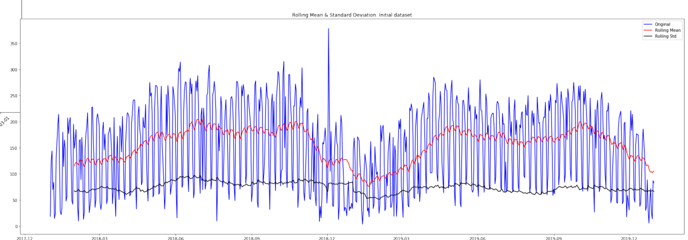

Fig. 14.

As the figure 15, the p-value is below the threshold of 0.05 and the ADF Statistic is close to the critical values. Therefore, the time series is stationary when using the log values for the data set.

Then the ARIMA model is build using the log data set with 'order=(2,1,2)'. The built model is then used to predict and validate the demand for the registered for a time period of 2020 January to 2020 February.

B. Cross-validation

Prophet includes functionality for time series cross-validation to measure forecast error using historical data. This is done by selecting cutoff points in history, and for each of them fitting the model using data only up to that cutoff point. We can then compare the forecast values to the actual values. This cross-validation procedure can be done automatically for a range of historical cutoffs.

In the figure 17, we can see that MAPE value for the Cross-validation is almost always less than 10%, which indicates that the model is able to learn the historical pattern very well and able to produce future patterns with high quality. The blue line is the moving average taken over the rolling window.

The MAPE calculation for the demand forecasting by the member type using the ARIMA model also returned the a value around 4.5%.

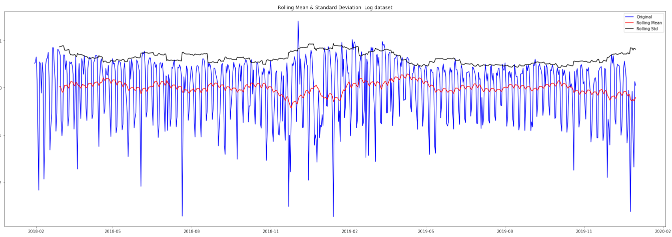

Fig. 15.

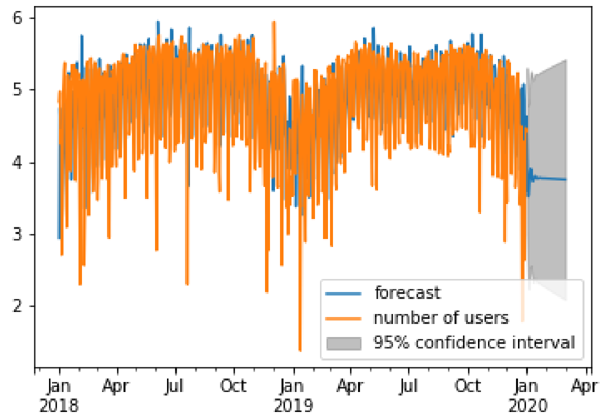

Fig. 16.

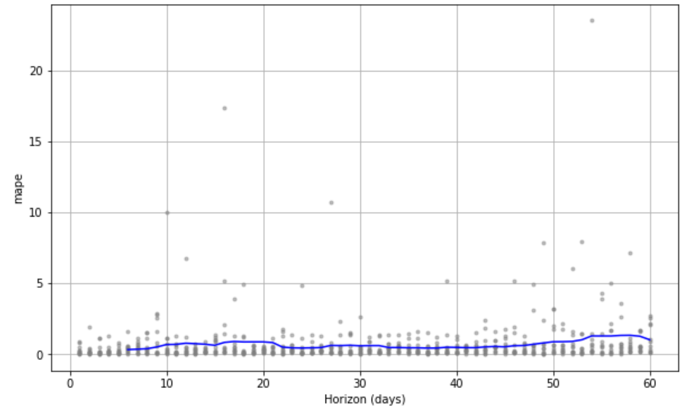

Fig. 17.

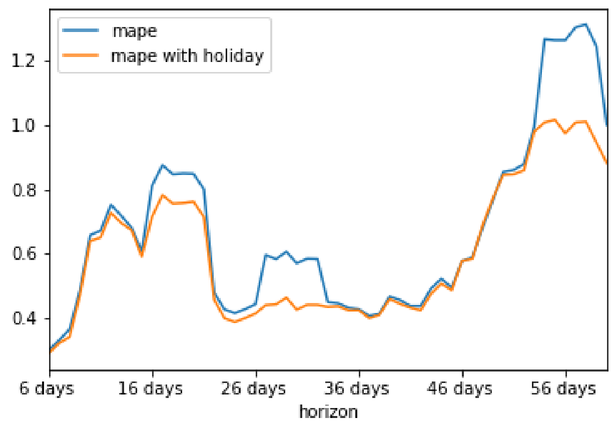

Fig. 18.

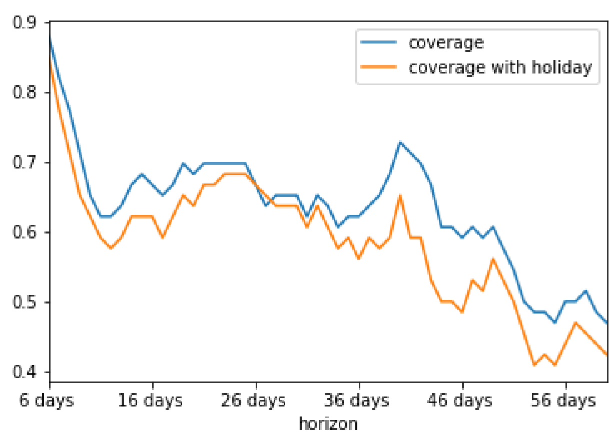

Fig. 19.

V. Results and Discussions

By the exploratory data analysis we confirmed our hypothesis that the most prominent stations are mainly serving for registered users.

The duration of users probability distribution as expected follows a chi-squared distribution. It is an indication that a user takes a bicycle from this station mainly for short trips and for long trips are unlikely, yet the number of longer trips that are possible anomalies requires attention.

An interesting observation in terms of addressing this is to check on casual users as the ratio of duration, as well anomaly duration is comparatively higher for casual users.

Even though the rides are less (least) on weekends the average duration is very high. Hence weekends requires special attention in terms of misuse and vandalizing. Ratio of casual and registered users in a particular station might reveal the location properties of the station.

The selected stations show a strong weekly and yearly seasonal pattern.

VI. Conclusion

To forecast we were able to achieve less than 10% error for the test data set with Prophet and when we add holiday information to the model the model performance remains the same. For the registered members, we were able to forecast with 4% error with ARIMA . In addition to point forecasting, the Prophet provided a confidence interval which we used to make decisions under uncertainty.

VII. ACKNOWLEDGEMENT

We would like to acknowledge Dr. Uthayasanker Thayasivam and Dr.Shehan Perera for supervising this research project. Also, we would like to acknowledge all the staff member of the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka for the support given to carry out this research.

VIII. REFERENCES

REFERENCES

[1] Data Source: https://s3.amazonaws.com/capitalbikeshare-data/index.html

[2] https://www.capitalbikeshare.com/data-license-agreement

[3] https://towardsdatascience.com/machine-learning-part-19-time-series-and-autoregressive-integrated-moving-average-model-arima-c1005347b0d7

[4] Prophet:https://facebook.github.io/prophet/docs/seasonality,$_holiday_effects,_and_regressors.htmlProphet : https : //research.fb.com/blog/2017/02/prophet - forecasting - at - scale/$