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. 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. 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 differencing 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.

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 weak 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.

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. 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.

The figure **??** depicts the registered, casual and total rides per station. Only the top ten stations are chosen in terms of the number of rides.

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.

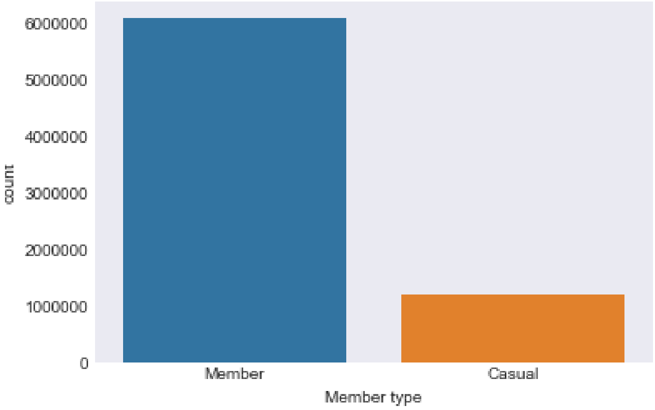

Fig. 1.

3) Duration Related: Below are the duration statistics and plot describe the probability density function of duration. Higher values are lower and the plot has a long tail.

IV. Implementation

A. Analysis

The figure **??**, 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

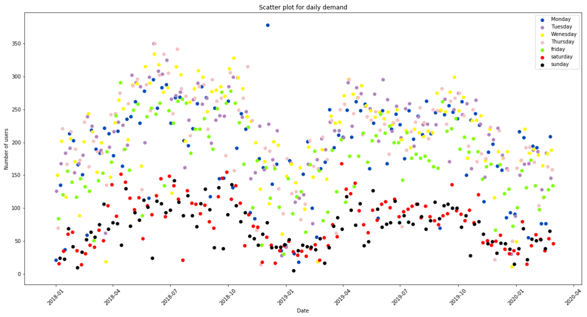

Fig. 2.

The figure 2, 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 3 shows how forecast values are align with the original test series. It is clear that the model is able to produce

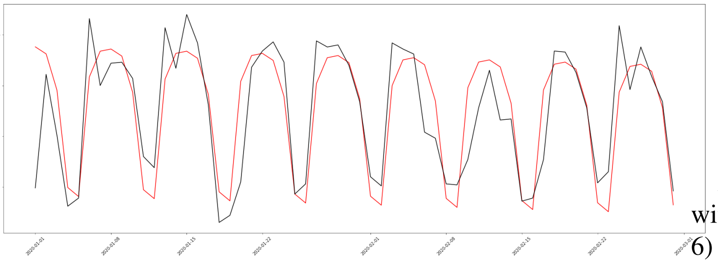

Fig. 3.

cyclic, trend, pattern for the test data and able to align with spikes and dips.

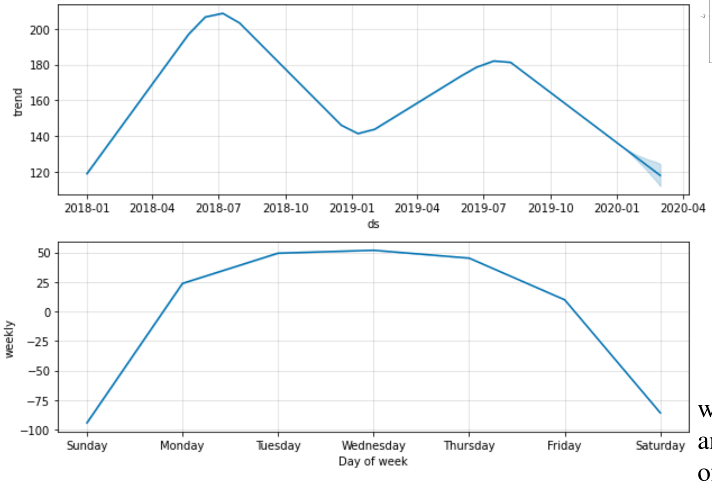

Fig. 4.

By figure 4 diagram, it is clear that the model learns the 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 calculation is calculated on the initial training data and checked for the stationarity.

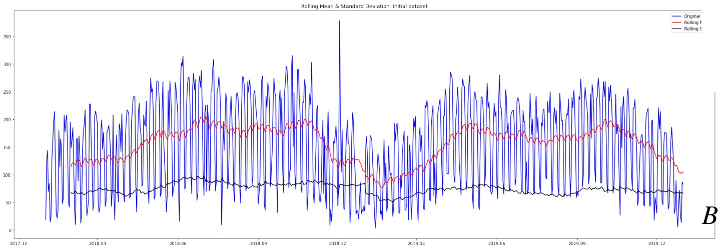

Fig. 5.

ADF output of the initial training data: figure 5

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

As it does not show the stationarity then the same is done with the converted log values. The ADF with log values (figure 6) gave the output for the stationarity as follows.

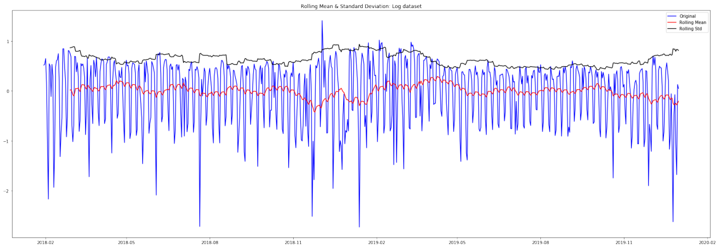

Fig. 6.

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

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.

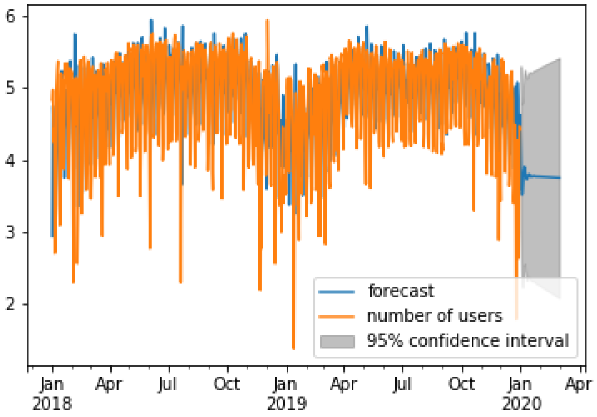

Fig. 7.

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.

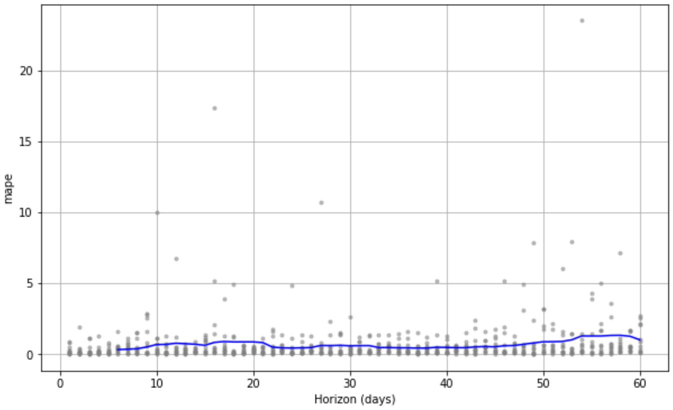

Fig. 8.

This cross-validation procedure can be done automatically for a range of historical cutoffs.

In the figure 8, we can see that MAPE value for the Cross-validation is almost always less than 10 percent, 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.

For the MAPE calculation for the demand forecasting by the member type returned the value around 4.5 percent.

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

VI. Conclusion

To forecast we were able to achieve less than 10 percent 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 percent error with ARIMA . In addition to point forecasting, the Prophet provided a confidence interval which we used to make decisions under uncertainty.

VII. References

References

[1] Data Source: https://s3.amazonaws.com/capitalbikeshare-data/index.html
[2] License: License: https://www.capitalbikeshare.com/data-license-agreement