# Bicycle-Sharing System Trip Duration and Checkout Prediction

**Apurvaa Kaamesh, Bhavya Navuluri, Vidhya Lakshmi Sankaranarayanan**

## Abstract

Bicycle-sharing is a low-carbon and environment friendly mode of public transport. It consists of many bikes placed in the docking stations. These bikes can be rented and returned to any of the docking stations after usage. Therefore, they are normally used as a short distance trip supplement for private vehicles as well as regular public transportation. As stations are located all over the city, the bike usages are very imbalanced. Some stations have too many incoming bikes, thus lack docking space for new bikes and some stations have too many outgoing bikes that lack enough bikes for people to checkout. Therefore, it becomes important for service providers to manually re-dispatch bikes. To do this, they need to have proper understanding of the bikes demand at a particular station and also the arriving time of each individual trip beforehand. The objective of this project is to provide a trip prediction model that analyses both the checkout demand of bikes at stations and the trip duration.

## 1. Introduction

Bike sharing system is one of the transportation services which is getting popular in the current health conscious trend in the society. Usually, bikes are preferred for short duration and are environment friendly. Chicago bike sharing system from Divvy bike sharing system is used for analysis in this paper. Bike sharing system in Chicago has different users. Customers are those who use the bike with one time pass e.g. tourists. Subscribers have monthly pass which is relatively cheaper e.g. office commuters. Different quarters of the year had different demand in different stations. Quarter 2 and quarter 3 had more trips compared to other quarters. As bike usage is different for each station, either stations run out of bikes or filled up with bike with no empty docks for other returning bikes. Hence, station level demand specific to the month and time is necessary in order to re-dispatch the bikes wherever necessary. Jia Jiang [3] studied about predicting cyclist destination using neural network in three steps such as learning user behavior through Long short-term memory (LSTM), relationship between origin and destination stations learnt through convolution neural network and external features through fully connected neural network. Jiawei Zhang in his paper [1] predicted trip duration which aids in estimating the time when the bikes will be available. Yexin Li [7] predicted the bike usage depending on weekday/weekends and weather. Station level clustering is performed. Various studies focused on predicting the trip duration, usage patterns, destination station and so on. Markov chain-based Demand prediction [11] did comparison between demand forecast for all stations and station level demand and showed that station level demand provides good prediction.

Our paper is mainly focused on time series analysis of predicting station level checkout and non-time series analysis on predicting trip duration. Station level checkout is the number of trips from any station on a day. Dataset has start time feature which tells date and time of the checkout from a station. Checkout needs to be predicted for station level as the demand for bikes varies for different stations. Before applying any time-series models, data needs to be stationary. Data is tested for stationarity using Dickey fuller test and later stationary data is applied on the model. To predict station level checkout, Auto Regressive Integrated Moving Average (ARIMA) is used. In order to understand the nature of the demand, model auto regression, moving average are applied individually and compared. The best model is selected using Akaike Information criterion (AIC) and Bayesian Information criterion (BIC).

To predict trip duration, various features are identified from the features given. Features like age from birth year, month, day, hour, Manhattan distance from latitude and longitude are used. Model used for trip duration prediction are Lasso Regression, Ridge Regression, Decision Tree and Random Forest.

The remaining of the paper is organized as follows. Section 2 elaborates about the dataset and exploration of data. Section 3 describes about testing the data stationarity. Prediction models are discussed on section 4. Section 5 provides details on experiments and results. Section 6 discusses the conclusion and Future work.

## 2. Dataset Exploration

### 2.1 Data Description

In this project we performed our trip prediction analysis on the Chicago’s Divvy Bike Sharing Dataset for the year 2015 for months April, May and June. The dataset consists of two parts, first being Trip description and second Station description.

The Trip description has attributes such as "trip ID", "trip start time", "trip end time", "bike ID", "from station ID", "from station name", "to station ID", "to station name", "user type", "gender" and "birth year". The "trip start time" and "trip end time" is in the form of date and time, "from station ID" and "to station ID" has unique ID's of departure and arrival station of the trip respectively. Similarly, "from station name" and "to station name" has the departure and arrival station names corresponding to their IDs respectively. "User type" has two types of values Customer and Subscriber. The Gender column has values Male or Female and missing values for Customer. Also, birth year has the year of birth for Subscriber and missing values for Customer. There are 893890 trips.

The Station description has attributes such as "Station ID", "Station name", "Latitude", "Longitude", "docking capacity". "Station ID" is a unique value for each station, "Station name" has the name of each station corresponding to its ID, "Latitude" and "Longitude" have the respective latitude and longitude values for that station. "docking capacity" states the number of docks available at that station. This is a constant value for each station. There are 475 stations here.

The correlation of the attributes in the Trip description dataset is as shown in figure 1.

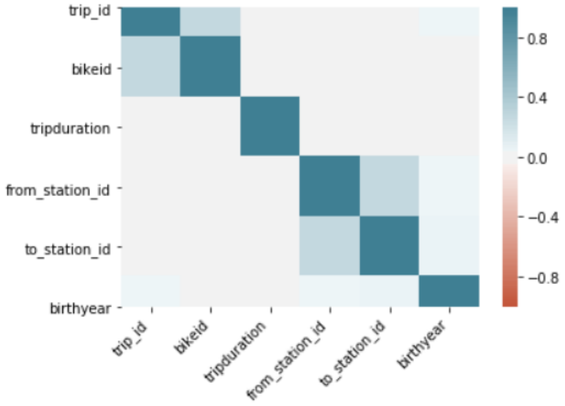


*Figure 1 Initial Trips dataset*

On performing explorational analysis, we found that on weekdays i.e. Monday to Friday Subscriber takes more trips (figure 2) compared to Customer. On the contrary, customer takes more trips on the weekend i.e. Saturday and Sunday.

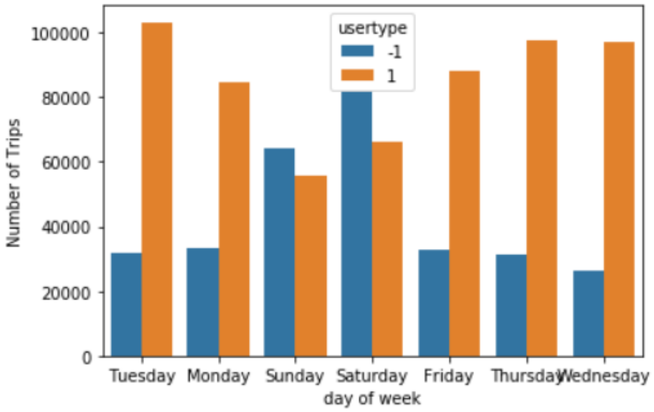


*Figure 2 Weekday Checkout for different users -1 means Customer 1 means Subscriber*

One of the objectives of this paper is to find the checkout for all stations but the demand or the checkout was different for different stations. Some stations had more checkouts while others have only check-ins. In quarter 2, popular station with more checkouts (around 20,000 trips) was Streeter Dr and Illinois St (figure 3).

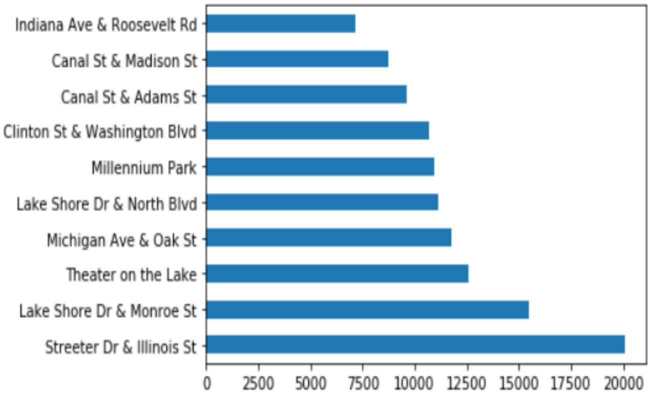


*Figure 3 Popular station with maximum checkout in Quarter 2*

Figure 4 shows the hourly demand changes for each quarter. In quarter 2, more bikes were checkout starting at morning 8am and similar demand through the afternoon till reaching the peak at around 5pm and slowly comes down at night.

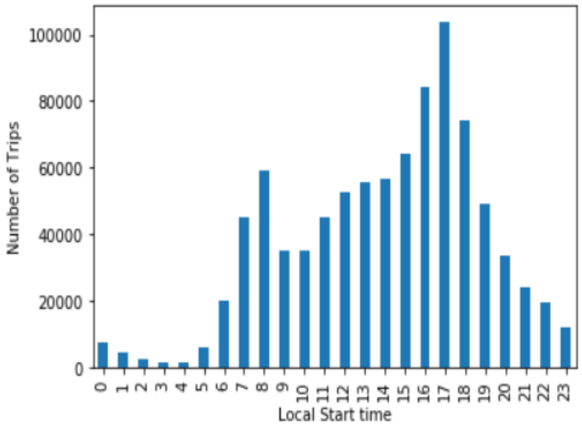


*Figure 4 Hourly checkout Peaked around 5pm*

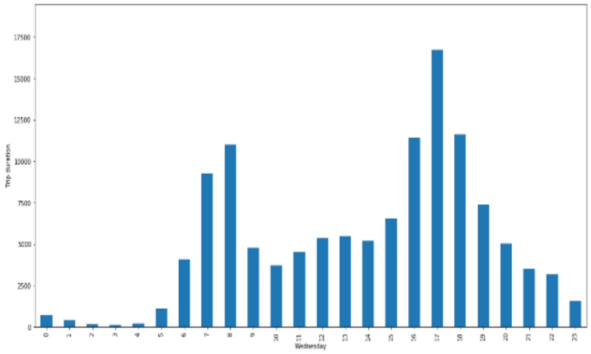


*Figure 5 Trip duration on Wednesday*

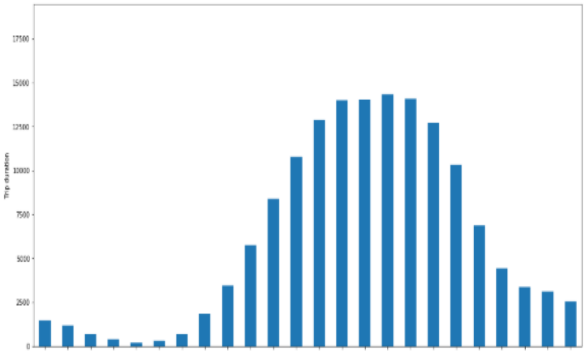


*Figure 6 Trip duration on Saturday*

Also, the trip duration varied depending on weekday and weekends (figure 5) with weekends having more trip duration compared to weekdays. Weekends had more trip duration throughout the noon (figure 6).

### 2.2 Data Stationarity

To check stationarity, we need to analyze the data in order to check whether the data has any trend, or it is stationary.

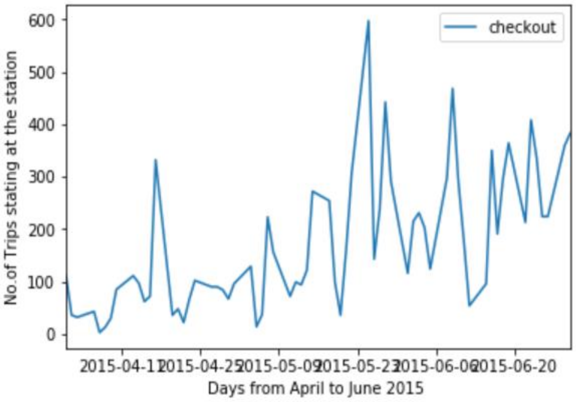


*Figure 7 Trip checkout for April, May, June*

The above graph shows that there is an upward trend for our data. In order to achieve good predictions, we must make our data stationary. A log transform can be used to flatten out exponential change back to a linear relationship.

To perform time series analysis, data should be stationary. Data is stationary if it has constant mean, constant standard deviation throughout the time period. To make the dataset stationary, detrending and shift by 12 lag differencing is used. Also, Dickey fuller test is carried out. In both detrending and 12 lag differencing, data has become stationary and hence the data after detrending is used for time series analysis. Trend means upward or downward change in checkout for a time period. Detrending means subtracting from the mean and dividing by standard deviation. 12 lag differencing means shifting the data by previous 12th value. (in figure 8). P-value in dickey fuller test came down to 0 after making the data stationary (in table 1).

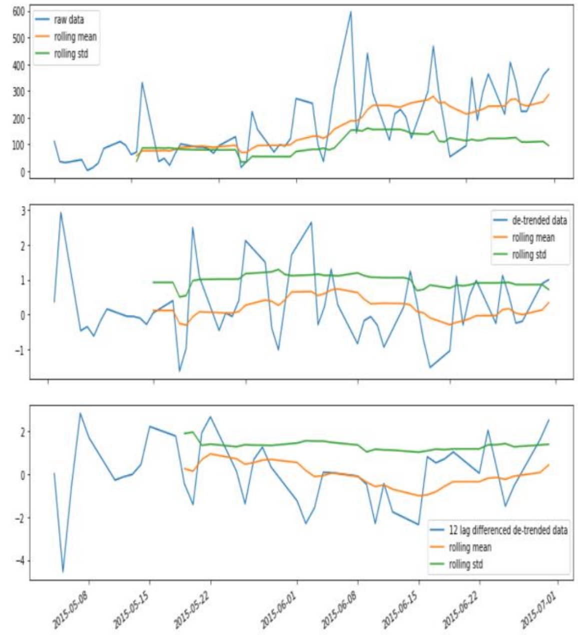


*Figure 8 Test for Data Stationarity Sub Graph 1(Top) Actual data, Sub Graph 2 Detrended data, Sub Graph 3 (Bottom) 12 Lag differenced data*

| **Data** | **Test Statistic** | **P-value** |
|---|---|---|
| Raw Data | -2.338 | 0.160 |
| De-trend | -7.198 | 0.000 |
| 12-lag difference | -6.464 | 0.000 |

*Table 1 Dickey Fuller Test. Data is stationary for detrended data and 12 lag differencing.*

From the Table 1 , it is clearly shown that the P-value and Test-Static Values are best for 'de-trended data' (the more the negative value of Test-static will be the better one and the less the P-value is the better one ). So, we consider de-trended data for further predictions.

## 3.Model Formulation

The trip prediction model analyzing both trip duration and checkout demand at a station has been provided as two separate problem statements. The trip duration prediction model has been looked at as a Regression Model and the checkout level demand at a station is looked as a Regression and Time-Series problem both.

### 3.1 Trip Duration Prediction:

For the trip duration prediction, the features have been divided into three different categories based on the User, Departure time and the Stations.

**A. User**: The previous data exploration shows that the type of the User, their Gender and Age have different roles in terms of trip duration. So, we have three features based on this category. The first feature is the User type, if the user is a Customer, we have the value set as -1 and in case of Subscriber we have it as +1. The second feature is the User Gender. For Female subscribers we have -1, male subscribers as +1 and as customers have missing values for gender, we take it as 0. The third feature is User Age, birth year information is available only for subscribers, so we calculate their age as of year 2015, since the dataset is for year 2015. And for customers as year of birth is missing, we have their age as 0.

**B. Departure Time**: The trip duration has a big influence based on the time of the day or the day of the week, the user checks out the bike. So, we have three features based on Month of the Trip, Weekday or weekend and hour of the trip. For the month, since this dataset is based only on quarter 2, we have the values 4, 5 and 6 for April, May and June respectively. In case of weekday i.e. from Monday to Saturday we have value set to 1 else in case of Sunday, we have set to 0. The hour of the trip has values 0 to 23 based on the hour the trip was checkout out. For example, in case the bike was checked out between 7AM and 8AM, the hour value would be 7.

**C. Stations:** Depending upon the user, the trip may either start from a tourist attraction, shopping spot, office or school and thus have different trip durations as customer tends to travel for short time duration compared to subscriber. Thus, we consider the distance between the source and destination station here. For distance, we use the latitude and longitude values provided for each station. We use Manhattan distance as the distance measure.

Then, we finally have the target as trip duration. The trip duration is taken as a logarithmic value as when used in seconds the range is too large. Figure 9 shows the correlation of all the features in our model:

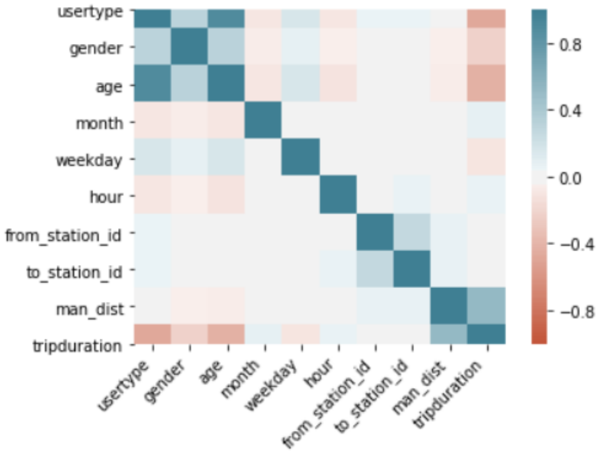


*Figure 9 Preprocessed dataset*

Different regression models have been applied which is discussed in the Experimental Results section.

### 3.2 Station level checkout demand prediction:

**Feature Selection:** We have selected the data for popular station 'Streeter Dr & Illinois St' for the year 2015 _Quarter_2 (includes April, May, June data only). The data with 'starttime', 'from_station_name', 'trip_id' are selected as features and from these a new column is created with the name 'checkout'. The columns 'starttime' and 'trip_id' are used for model analysis. The 'checkout' column is considered as the target variable.

- **Step 1: Check stationarity**: If a time series has a trend or seasonality component, it must be made stationary before we can use ARIMA to forecast.
- **Step 2: Difference:** If the time series is not stationary, it needs to be made stationarized through differencing. Take the first difference, then check for stationarity. Take as many differences as it takes. Make sure you check seasonal differencing as well.
- **Step 3: Filter out a validation sample:** We have used three different test sample like 10, 20 and 30% as test data.
- **Step 4: Build the model:** Build the model and set the number of periods to forecast to N (depends on your needs).
- **Step 5: Validate model:** Compare the predicted values to the actuals in the validation sample.

## 4 Experimental Results

Again, since we have two different problem statements, the results are analyzed differently as below:

### 4.1 Trip Duration Prediction:

We divided the entire dataset into 3:1 ratio of training and testing samples. The split was done randomly. The models have been compared based on two methods, one considering all the samples and other by removing the abnormal trips.

### 4.1.1 Removing Abnormal Trips:

As we can see in the frequency plot, we removed all the trips that had duration above 2 hours as we consider them as outliers. Also, in case of trips with same origin and destination station, we considered trips only above 5 mins as this could be a case where a user recognizes some issue with the bike and returns it. Similarly, we consider only the trips having different origin and destination station with duration greater than 5 minutes as some stations are so close that user might have discovered some issue with the bike and returned immediately.

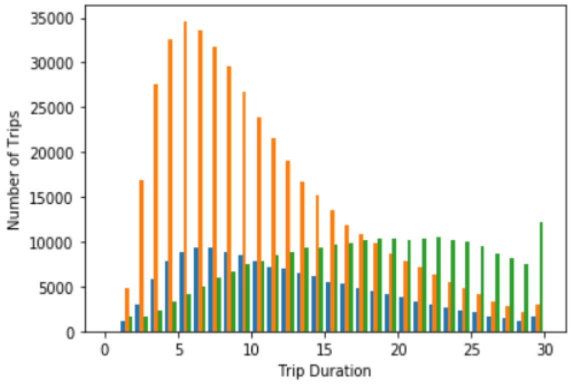
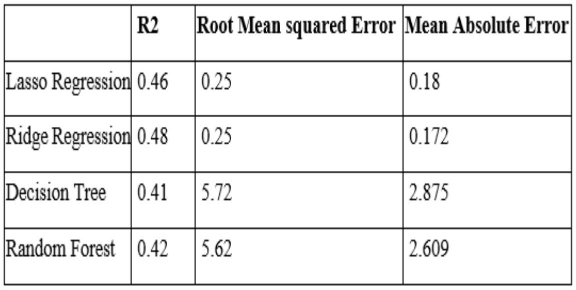


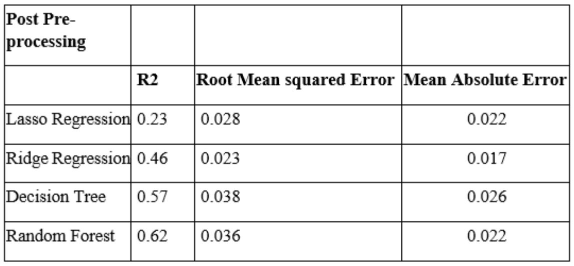
*Figure 10 Trip duration Green - Female, Orange - Male, Blue-Customers.*

Three Regression Models have been applied Linear Regression with Lasso and Ridge Regularization, Decision Tree regression and Random Forest Regression. The results are shown in table 2 and 3. It can be observed that removing abnormal values and outliers has improved the performance of the model. The R squared values of Decision tree regression and Random Forest regression is 57% and 62% respectively. Also, the Root mean squared error value and mean absolute error value has improved from 0.25 to 0.023 in case of Linear regression and 5.72 to 0.038 in case of Decision tree regression. Post removal of abnormal values and outliers, even though R square values have improved for Decision tree regression and Random Forest Regression to 57% and 62% respectively, compared to Linear Regression at 46%, error seems to be close to similar. We know that Random Forest regression model tends to sometimes overfit the data. For trip duration prediction, none of the models did an excellent job. Also, on checking the time difference between actual and predicted values for Linear regression we found that for 3% of the test data, difference exceeds 15 minutes, which is not bad.



| | R2 | Root Mean squared Error | Mean Absolute Error |
|---|---|---|---|
| Lasso Regression | 0.46 | 0.25 | 0.18 |
| Ridge Regression | 0.48 | 0.25 | 0.172 |
| Decision Tree | 0.41 | 5.72 | 2.875 |
| Random Forest | 0.42 | 5.62 | 2.609 |

*Table 2 Initial Trip duration prediction results*



| Post Pre-processing | | | |
|---|---|---|---|
| | R2 | Root Mean squared Error | Mean Absolute Error |
| Lasso Regression | 0.23 | 0.028 | 0.022 |
| Ridge Regression | 0.46 | 0.023 | 0.017 |
| Decision Tree | 0.57 | 0.038 | 0.026 |
| Random Forest | 0.62 | 0.036 | 0.022 |

*Table 3 Trip duration prediction results after preprocessing*

## 4.2 Station level checkout demand prediction:

**4.2.1 Model identification:** Since our data is time-series we have to use time-series models like Auto Regressive(AR(*p*)), Moving Average (MA(*q*)) ,ARMA(*p* ,*q*) it is a combination of AR and MA , ARIMA(*p*, *d*, *q*)is a combination of ARMA and integration and many other time-series models . The p, d, q values refers to Periods to lag, no. of differencing transformations are needed by the timeseries in order to make the data stationary, denotes the lag of the error component. We have applied AR, MA, ARMA and ARI MA models for our data to predict the feature trends. The Akaike Information Criteria AIC and Bayesian information criterion BIC values are used as metrics in order to identify the best model for our data. The lower the AIC and BIC values (figure 11 and 12) the best is the model and generally it is considered as closer to the real data. The results are as below:

| Model | AIC | BIC |
|---|---|---|
| AR (1) | 217.29743710979014 | 224.44351701381177 |
| AR (2) | 215.66969308750814 | 225.19779962620368 |
| AR (3) | 216.71137801168436 | 228.62151118505378 |
| MA (1) | 214.9907236236894 | 222.13680352771104 |
| MA (2) | 216.98513295809755 | 226.51323949679306 |
| MA (3) | 217.2986153586771 | 231.5907751667204 |
| ARMA (1,1) | 216.9873423952704 | 226.5154489339659 |
| ARMA (2,1) | 211.62850661743863 | 223.53863979080805 |
| ARIMA (2,1,1) | 133.72631717422132 | 143.08232222876077 |

*Table 4 AIC and BIC values for different models*

From the table 4, it is clearly shown that AIC and BIC values are less for ARIMA (2,1,1) model.

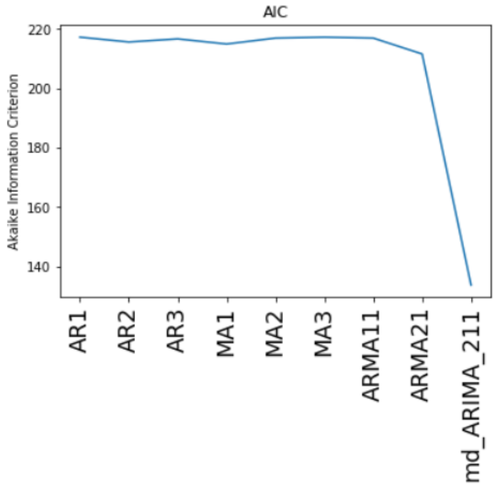


*Figure 11 Akaike Information Criterion (AIC) for various models*

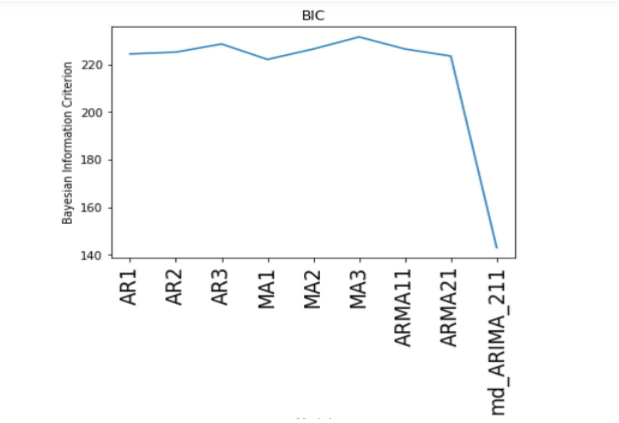


*Figure 12 Bayesian Information Criterion (BIC) for various models*

**4.2.2 Parameter Estimation:** The co-efficients and intercept for the best ARIMA model we obtained are as follows:

| Metric | AR (1) | AR (2) | MA (1) | Intercept |
|---|---|---|---|---|
| **Co-efficients** | 0.3962 | -0.2050 | -1.0000 | 0.0029 |
| **Standard Error** | 0.111 | 0.110 | 0.034 | 0.005 |

*Table 5 Evaluation Metrics for ARIMA*

After choosing the best model, we have tested the model on test data (last week of June). The below is the graph obtained. In figure 13, the Blue line indicates the forecast values (predicted values) and the yellow line indicates the Z_data (Actual data). From the figure 13 and 14, reversal of the trend was observed for predictions of few days. Predictions were good for the last week of June (figure 15).

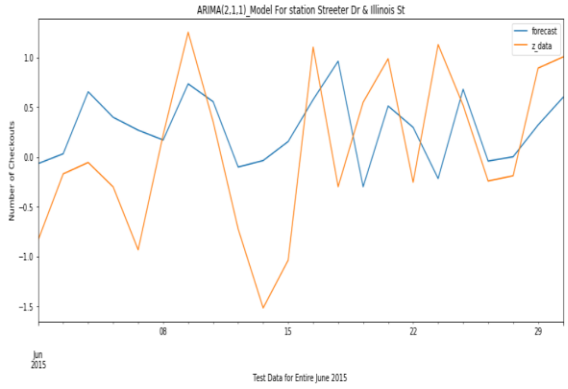


*Figure 13 ARIMA model Prediction of entire June month*

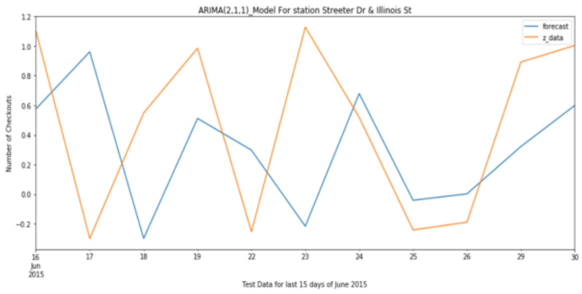


*Figure 14 ARIMA model Prediction for June month last 15 days*

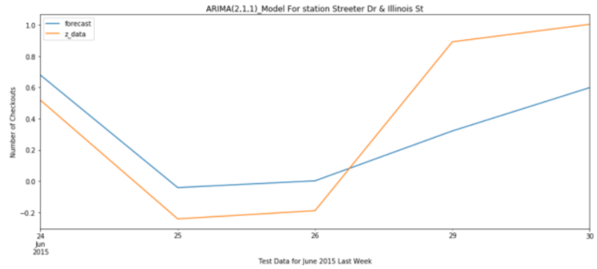


*Figure 15 ARIMA model prediction for last week of June*

| Metric | 1 week (Last week Of June) 10% | 15 days (15'th June to 30'th June)20% | 1 Month (Entire June Month) 30% |
|---|---|---|---|
| **MSE** | 0.9045 | 1.30142 | 1.33966 |
| **RMSE** | 0.9510 | 1.1408 | 1.1574 |
| **Standard Error** | 1.5243 | 4.49320 | 4.6059 |

*Table 6 Evaluation Metric for different test samples*

We considered Mean Squared Error for comparing the better model. Out of three sample test data (10,20 30 %) we got better results for 10 % (Figure 15) test data.

## 5. Conclusion and Future Work

In this paper, we focused on two main objectives of bike sharing systems which are station level checkout and trip duration prediction. Both Time series and non-time series analysis aids in better understanding of the dataset. As the AIC and BIC values are very low for ARIMA model, it is the best model for predicting station level checkout for this dataset. Predicting checkout will be very useful for the service providers to provide bikes depending on the demand. Station level Check-in demand can also be performed similar way so that overloaded stations can be taken care. In future research, we need to incorporate weather and understand the impact of it on the trips. As the data is seasonal, that is checkout pattern varying for each quarter, seasonal ARIMA for all quarters should be applied for the better analysis.

## 6. References

1..Jiawei Zhang, Xian Pan, Moyin Li, Phlip S. Yu, Bicycle sharing system analysis and Trip Prediction. 2016 17th IEEE International Conference on Mobile Data Management.

2.Divya Singhvi, Somya Singhvi, Predicting Bike Usage for New York City's Bike Sharing System. 2015 AAAI Workshop

3.Jian Jiang, Fei Lin, Jin Fan, A Destination Prediction Network Based on Spatiotemporal Data for Bike-sharing. Hindawi Volume 2019.

4.Xiaomei Xu, Zhirui Ye, Jin Li and Mingtao Xu, Understanding the Usage Patterns of Bicycle-Sharing Systems to Predict Users' Demand: A Case Study in Wenzhou, China. Volume 2018, Article ID 9892134, 21 pages

5.Loh, Wei-Yin. (2011). Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 1. 14 - 23. 10.1002/widm.8

6.Leo Breiman, Random Forest. Machine Learning, 45, 5–32, 2001.

7.Yexin Li and Yu Zheng, Citywide Bike Usage Prediction in a Bike-Sharing System. 12 February 2019.

8.Romain Giot, Predicting Bikeshare System Usage Up to One Day Ahead. IEEE Symposium Series in Computational Intelligence 2014 (SSCI 2014).

9.P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," Journal of Public Transportation, vol. 12, no. 4, pp. 41–56, 2009.

10.Ratnadip Adhikari, R. K. Agrawal, An Introductory Study on Time Series Modeling and Forecasting.

11. Yajun Zhou, Lilei Wang, Rong Zhong, and Yulong Tan, A Markov Chain Based Demand Prediction Model for Stations in Bike Sharing Systems. Hindawi, Mathematical Problem in Engineering, Volume 2018.

12.Lasse Drevland, Patrick Finseth, Evaluating Machine Learning Methods for Ci Bike Demand Prediction in Oslo, June 2018.