Analysis and Prediction of Bikesharing Traffic Flow – Citi Bike, New York

2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) | 978-1-7281-8995-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/MT-ITS49943.2021.9529290

Salma Y. Y. Hamad
Department of Civil, Geo, and Environmental Engineering
Technical University of Munich
Munich, Germany
salma.hamad@tum.de

Tao Ma*
Department of Civil, Geo, and Environmental Engineering
Technical University of Munich
Munich, Germany
tao.ma@tum.de

Constantinos Antoniou
Department of Civil, Geo, and Environmental Engineering
Technical University of Munich
Munich, Germany
c.antoniou@tum.de

Abstract—**Bikesharing systems have witnessed unprecedented growth and significant scholarly attention in recent years. Technological advancement, environmental awareness, and demand for socially equitable transport modes were the major contributors to this development. However, with the ongoing expansion of these systems, companies are faced with the constant need to rebalance them in order to meet the growing demand. Operating companies are continuously searching for more effective and efficient tools for bikesharing traffic flow prediction. This research explores four different techniques for the traffic flow prediction of bikesharing traffic systems including three machine learning algorithms and a statistical time series model. The techniques were evaluated based on prediction accuracy and the best performing algorithm was identified and proposed. In addition, the study analysed the relationship between bike sharing utilisation, weather, and characteristics of bike users, and addressed the neglected aspect of multiple seasonality in time series models. The comparative results confirm that neural networks deliver the best performance. The research evidence suggests that complex seasonalities should be taken into account in traditional time series models.**

Keywords—Bike-sharing, Machine Learning, MLP, Random Forest, Gradient Boosting, ARIMAX

I. INTRODUCTION

Climate change, air quality, fluctuating fuel prices have recently become of primary concern and heightened the need for a more sustainable and green means of transport. In the last two decades, there has been a surge of interest in bikesharing systems, which have witnessed an extensive growth within major cities. There is concrete evidence that bikesharing systems play a crucial role socially, economically, and environmentally as a sustainable transportation mode. It provides notable health benefits. Bikesharing systems are inexpensive, efficient, flexible, and accessible. According to Shaheen et al. [1], bike sharing systems have the potential to avoid the production of 37,000 kg of carbon dioxide per day compared to cars travelling the same distance. The flexibility of the bikesharing system significantly increases the accessibility to the city or the area where it is provided, with additional benefit of reaching and contributing to local businesses [2].

The majority of existing bikesharing systems are docked systems. They have docking stations distributed around the city where users can self-checkout providing the choice of picking up a bike from any station and dropping it off to any of the available stations with free docks. In addition, the reasonable pricing of the system makes it appealing to commuters and students alike. It is also a reliable solution for the 'last mile' problem [1], which makes it even more appealing to commuters who use public transport. However, the limited docks and bikes available in each station constrain the flexibility of the system due to the overwhelming demand contributed by commuters in residential and business areas during the morning and evening rush hours.

The high demand during the morning and evening peaks results in an uneven distribution of bikes. Forecasting the bike demand is of tremendous commercial value because it helps to address the rebalancing issues of the bikesharing system. To this end, this study focuses on investigating the performance of traditional time series models and three machine learning algorithms commonly used in predicting the station-based bike flows. It sets out to analyse and visualize the general characteristic of the bikesharing system and the general behaviours of the bike users under different weather circumstances. Then it assesses three machine learning algorithms including Random Forest (RF), Gradient Boosting Regression Tree (GBRT), and Multilayer Perceptron (MLP), and the time series model ARIMAX in term of their prediction capacity. There are limited studies that the complex seasonalities of a bikesharing system are considered in time series models. A key strength of the present study is that ARIMAX considers multiple seasonalities for a bikesharing demand forecasting. Also, the computational cost of model training and hyperparameters tuning is used as a performance metric in the study.

II. RELEVANT WORK

A. *Shared Economy, Shared Mobility*

The past two decades have seen rapid growth in shared economy driven by the prominence of resources scarcities and the attempt to merge both offline and online worlds [3]. However, only in the last decade, the term "shared economy" has propelled to the forefront of research. According to Muñoz et al. [4], shared economy is defined as "a socioeconomic system that enables an intermediated set of exchanges of goods and services between individuals and organizations which aim to increase efficiency and optimization of underutilised resources in society." Accommodation and transport are amongst the major underutilised sectors that commonly implement shared economy concepts [3][5].

Shared mobility is a component and a forerunner of shared economy and technological advancement. It is defined as "the shared use of a vehicle that enables users to have a short-term

* Corresponding author: tao.ma@tum.de

978-1-7281-8995-6/21/$31.00 ©2021 IEEE

access to transportation modes on an 'as-needed' basis" [6]. Shared mobility includes carsharing and micro-mobility such as bikesharing.

Bikesharing systems were developed primarily as a result of environmental concerns by non-profit organisations. Bikesharing went through a series of four generations. The fourth-generation systems are characterised as demand-responsive and multimodal systems. It is integrated with public transportation and other mobility sharing systems. It also incorporates an innovative bicycle redistribution and advanced IT infrastructure, such as GPS tracking. [1][7][8]. Bikesharing systems follow three service models: station-based bikesharing system (SBBS), dockless bikesharing systems (DBSS) and hybrid bikesharing systems. The majority of trips made using these systems are commuting trips from/to place of work or education on weekdays, and trips for leisure and sightseeing on the weekends.

B. Modelling and Forecasting

A large and growing body of literature focuses particularly on forecasting bike or dock availability in station-based bikesharing systems, as predicting the behaviour of users and bikes flow between stations is essential. Previous research had examined both time series forecasting methods and machine learning methods in a univariate and multivariate settings, respectively. Kaltenbrunner et al. [9] based their approach on time series forecasting methods for predicting hourly station-level bike availability. Yoon et al. [10] extended their ARIMA model by using different clustering methods to obtain the spatial correlation between stations. With K-Nearest Neighbour (KNN), and Linear Regression based method, Gallop et al. [11] explored the relationship between weather and bike users' behaviour. It has conclusively been shown that temperature is a significant factor [12][13].

Machine learning algorithms are suitable for multivariate modelling, which have shown outperform linear models [14][15]. Li et al. [16] conducted a cluster-based prediction and integrated the weather information into their data. They used GBRT to predict the global lending amount of bikes for the next hour. Their baseline methods included historical average (HA), ARMA model, a hierarchical prediction based on K Nearest Neighbour (HP-KNN), and a uniform geographical grid clustering (GC), where GBRT showed outperformance compared to baseline methods with plausible results for anomalous periods. Some researchers adopted the hierarchical demand prediction method, where they either predicted the demand from a global level, i.e., the system demand, to a cluster-level, or from a cluster-level to station-level [17][18]. Liu et al. [18] predicted the demand before and after the expansion of CitiBike, New York bikesharing system. Yang et al. [19] predicted the total lending amount of station clusters for the next hour and compared the performance of linear and non-linear prediction models. Yang et al. [20] proposed a spatio-temporal mobility model to predict the hourly check-ins and check-outs on a station-level. Ruffieux et al. [21] studied real-time short-term and long-term predictions of bikes and slots availability focusing on two machine learning algorithms, i.e. RF and Convolutional Neural Network (CNN). Wang et al. [22] carried out a station-based short-term prediction of the number of available bikes using a month-long data without the inclusion of meteorology or time factors.

Thus far, previous studies tended to focus on clustering optimisation in terms of clustering-based forecasting, in which some stations lose their individuality. Nevertheless, some of those studies either neglected the weather data despite of their significant effect on bikesharing demand or time attributes which have shown to influence the ridership immensely [23][24]. On the other hand, the majority of station-based forecasting were centred on machine learning algorithms with no regard to time series forecasting methods except for ARMA, which was mainly utilised as a baseline method and has shown many drawbacks. Furthermore, studies that employed the ARIMA model neglected the seasonality effect on demand.

This study examines and compares the ARIMA model considering the complex seasonality of bikesharing demand and three prominent machine learning algorithms including Random Forest, Gradient Boosting Regression Tree, and Neural Network, specifically Multilayer Perception (MLP).

III. METHODOLOGY

A quantitative case-study approach was adopted to assess and determine the model that has the ability to deliver the most accurate prediction results of station-based check-ins and check-outs.

A. Data Source – Citi Bike, NYC

The data collected for this study is the open-source trip data taken from Citi Bike bikesharing system in New York. It is a privately owned bikesharing system launched in May 2013 and named after its first sponsor Citi Bank and currently operated by Lyft. The system is operational in New York City boroughs as well as in Jersey City, New Jersey. Citi Bike is a docked system with more than 750 stations. The system is continuously adding more stations and expanding to other boroughs in New York City. Moreover, e-bikes are added recently to the operational fleet in New York City.

The data sets used are the trip data made within New York City, specifically the area of Bronx, Brooklyn, Manhattan, and Queens, and New York City hourly weather data for the period between November 2018 and November 2019. The information in the trip data sets include trip duration, start time and stop time, date, start station name, end station name, station id, station latitude and longitude, bike id, user type, gender, and year of birth. The hourly weather data were obtained from a secondary source [25] and include information such as location, temperature, precipitation, dew point, wind speed and direction, weather type, and sea level pressure.

B. Data Preprocessing

Data pre-processing includes exploring, combining, and cleaning the data for further analysis and modelling. The first step in this process is to pre-process the weather data and extract the useful features needed for this study. Then, the trip data were pre-processed and split into check-in and check-out trips. Finally, the trip data set was aggregated to an hourly basis and merged with the weather data to extract the top three stations in term of bike volumes used for modelling.

C. Characteristics of Bikesharing Trips

1) Users and Trips

The location of a station is one of the key factors that influences the demand immensely [26]. Fig. 1 shows the density of the stations and trips within the study area, where the majority of the trips are highly concentrated in Manhattan. Manhattan is the highest populated borough in New York

City, which explains the high volume of demand and stations. They are located around a lot of touristic attraction sites, offices, and universities.

Fig. 2 shows that approximately 90% of the users were subscribers. It can be deduced that most users opted to become members. Fig. 3 shows the general utilization pattern of different user types during the week. Subscribers used the service more during the weekday rather than on the weekends, whereas customers were more likely to use it on the weekends, implying that subscribers are more likely to be employees or students.

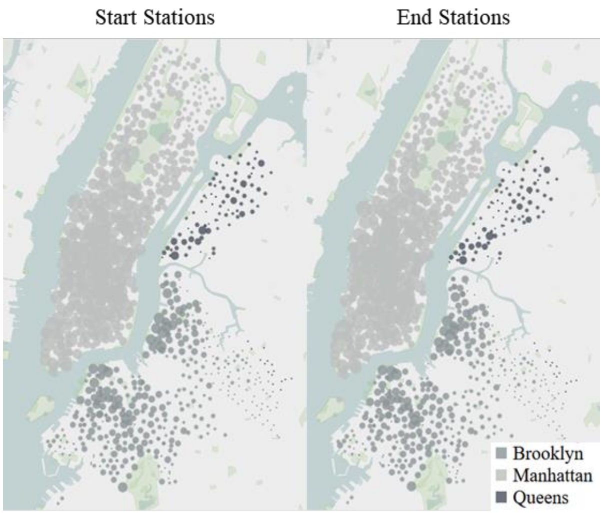

Fig. 1. Study Area

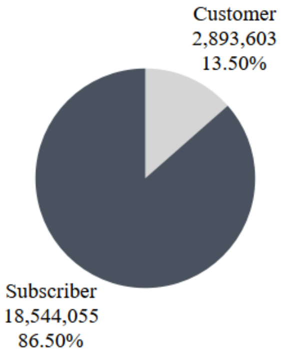

Fig. 2. User Type Distribution

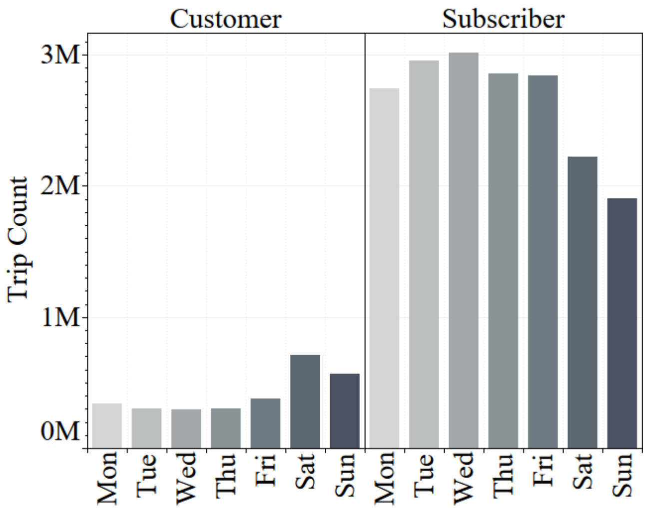

Fig. 3. User Type Distribution (Weekday)

As shown in Fig. 4, the trip duration distribution is skewed to the left with the mode being around 5 minutes. Trip durations spread from 2 minutes up to 120 minutes, which were considered in this study. The trips were considered anomalous if trip duration is more than 120 minutes. The highest density can be observed between 2 and 9 minutes.

Fig. 5 shows the trend of trips over time where there has been a steady decline from November 2018 to February 2019, then a gradual increase reaches its peak in September 2019, it drops again thereafter. The variability of trip volume is most likely attributed to weather conditions, as bike users tend to cycle more in warm conditions.

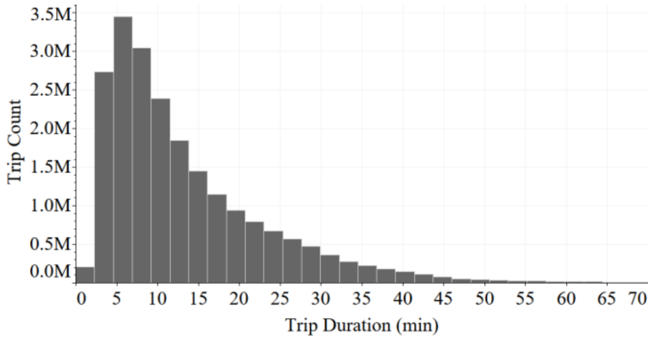

Fig. 4. Trip Duration Histogram

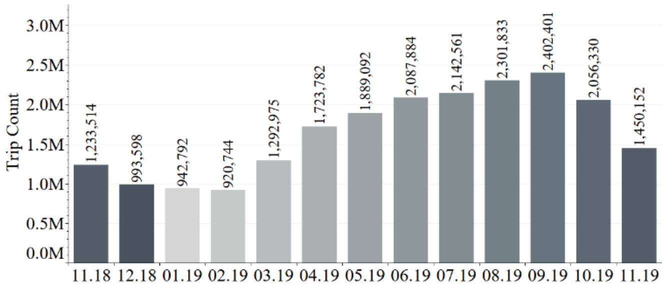

Fig. 5. Monthly Trip Count

Fig. 6 shows that there are two peak periods during the weekdays including 7:00 – 8:00 am in the morning and 5:00 – 6:00 pm in the afternoon, which are most likely influenced by the subscribers' proportion due to work and education trips.

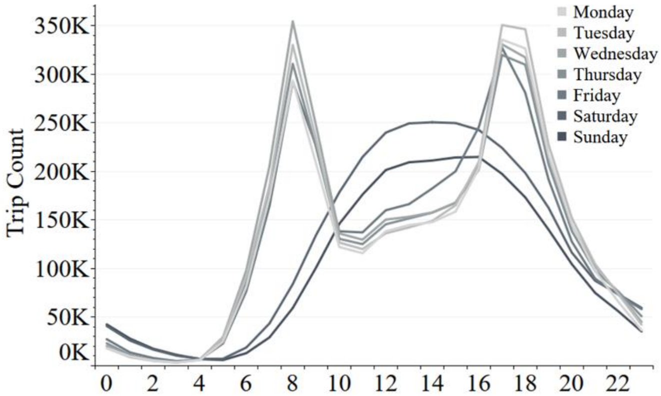

Fig. 6. Daily Trip Count – Start

2) Weather Impact

Bikesharing systems are highly affected by seasons. Fig. 7 shows that summer has the highest trip counts compared to other seasons, followed by autumn then spring. Customers were explicitly influenced by season when compared to subscribers, with a significantly smaller proportion in winter than subscribers. Fig. 8 further confirms that the weather influences the patterns of the bike usage. The amount of the trips made were remarkably high in clear weather condition and significantly decreased when the weather condition becomes worse. Fig. 9 reveals the fact that there has been a steady rise in the number of trips made when the temperature increased. It shows that bike users prefer warmer temperature than the colder temperature. However, the trip counts dramatically decline when the temperature exceeded a certain degree.

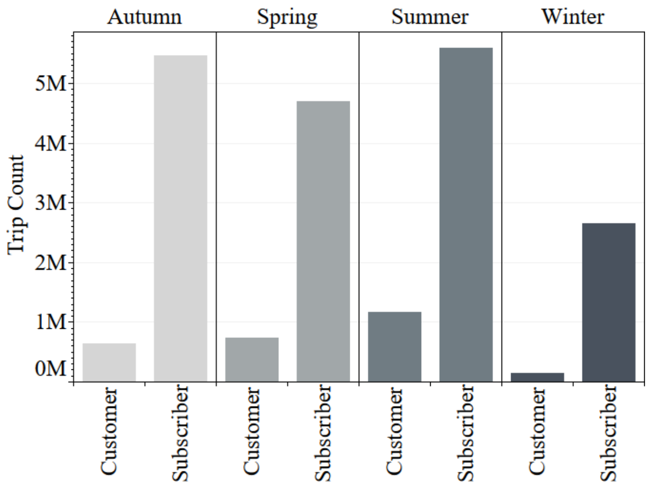

Fig. 7. Trip Count by Season

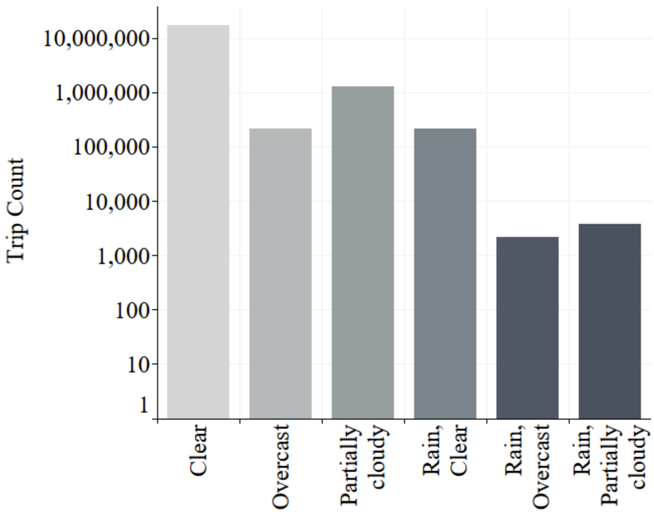

Fig. 8. Trip Count by Weather Condition

Other weather conditions such as relative humidity also greatly influence trip volume. Fig. 10 shows the relationship between trip volume and relative humidity. When the relative humidity passes a certain threshold outside the human comfort zone, the trip volume decreases.

D. Machine Learning Algorithms and Statistical Method

1) Random Forest & Gradient Boosting Regression Tree

Random Forest (RF) and Gradient Boosting Regression Tree (GBRT) are ensemble machine learning algorithms that can be used either as a classification or a regression model. Random Forest was introduced in 2001 by Leo Breiman [27], while Gradient Boosting Regression Tree was developed by Friedman [28]–[30] based on the work of Leo Breiman. Both algorithms can fit trees by selecting a random subset of the predictors from the original data and outputs a classification or a regression prediction. However, GBRT sequentially fits the trees. The final output depends on the original problem definition, i.e., whether it is a classification or a regression problem. The output is either the mode of the classes for classification problem or the mean prediction for regression problem. RF is less prone to overfitting due to the Law of Large Numbers [27], as a result of training multiple trees simultaneously. On the other hand, the trees in Gradient Boosting are considered weak learners individually. A robust model is built by minimizing the errors of the previous trees each time a tree is added, thus, attributing to a longer training time than RF.

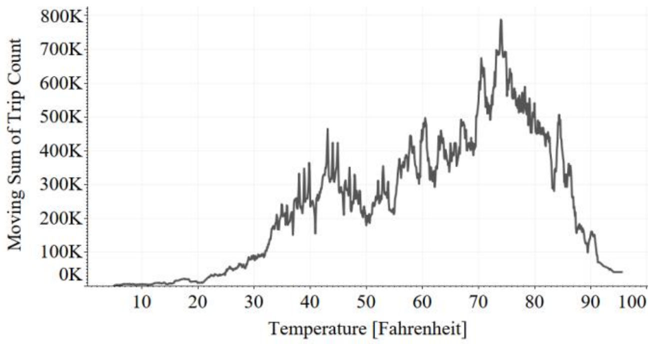

Fig. 9. Trip Count by Temperature

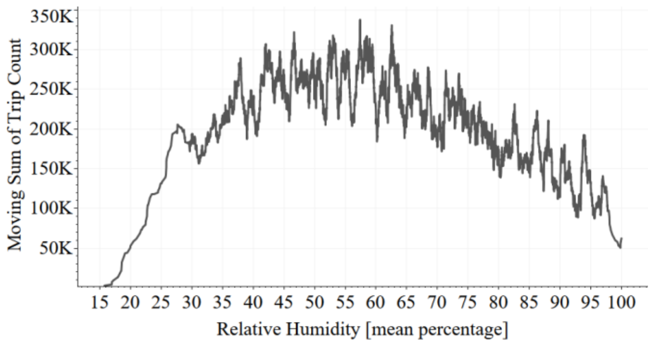

Fig. 10. Trip Count by Relative Humidity

2) Artificial Neural Network

Artificial Neural Networks (ANNs) are computational models that try to replicate the way how the human brain processes information. ANN consists of multiple layers of neurons. Each neuron is a processing unit, and all neurons are interconnected by links. A simple of ANN consists of a single neuron i (also called Perceptron) that receives the sum of multiplication of n inputs (x_1, x_2, …, x_n) by their weights (w_1, w_2, …, w_n), in which a bias is added (b), then an activation function is applied to get a single output, shown as equation (1).

$$f(x) = b + \sum_{i=1}^{n} x_i w_i \qquad (1)$$

ANNs have different topology according to the problem definition, so does the activation function. Multilayer Perceptron (MLP) with Back Propagation (BP) learning algorithm is a kind of feedforward artificial neural network. It is also called a multilayer feedforward neural network. MLP consists of three layers: an input layer, a hidden layer, and an output layer. Each layer may consist of multiple linear or non-linear activated neurons and a layer only affects the layer on its right. That is where the name feedforward come from [31]. An MLP with a BP learning algorithm tries to minimize the cost function by adjusting the weights between neurons via backpropagation. The cost function is to calculate the discrepancy between predicted value and actual value.

3) Autoregressive Integrated Moving Average

Time series is a sequence of data that is recorded in a timely manner over regular intervals. Time series data have different features that might not all be present, such as trend, seasonality, and cyclicality. A trend is an increase or decrease in values over time. Seasonality and cyclicality infer that the data have a pattern that either repeated with or without fixed intervals, respectively. Forecasting a time series falls into two categories: univariate time series forecasting and multivariate time series forecasting. The former uses only the previous values of the time series, while the latter uses additional predictors (i.e., exogenous variables) to forecast.

ARIMA is a univariate time series model. It is also written as ARIMA (p,d,q), where p is the order of the AR term, d is the differencing order, and q is the order of the MA term. The *AR* term describes the number of lags of the observations to use as predictors. The *I* term describes whether or not the data are differenced. The *MA* term describes the number of lags of the errors to use as predictors. [32]. The general formula of an ARIMA model can be written as equation (2):

$$\emptyset(B)(1-B)^d X_t = \theta(B) Z_t, \qquad (2)$$

where $\{Z_t\} \sim WN(0, \sigma^2)$, $\emptyset$(B) and θ(B) are polynomials of degrees p and q, respectively [33].

Data require differencing if it is non-stationary. Various tests are conducted to test stationarity, including the Augmented Dickey-Fuller test (ADH Test) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS test) (trend stationary). ADF test checks if the data has a unit root, thus indicating a non-stationary data. While KPSS test checks if the data is stationary around a deterministic trend. [34]

Bikesharing demand is a stochastic phenomenon, therefore, it was necessary to perform both tests. The results have shown that the demand is difference stationary, thus differencing the data yielded stationary stochastic data.

Traditionally, the ARIMA model is used for non-seasonal time series. In case the data exhibit defined seasonality, Seasonal ARIMA (SARIMA) is used instead. None the less, seasonal ARIMA does not perform well with long seasonality and cannot deal with multiple seasonality as well. Therefore, a modified version of ARIMA is used, where the seasonality is transformed into Fourier terms. The modified version is referred to as ARIMAX, where the X refers to exogenous variables. The Fourier term, in this case, are the exogenous variables. [35][36]

The data used in this study are hourly data. Hourly data exhibit three types of seasonality, i.e., yearly, weekly, and daily. The data used are only a year-long; therefore, only weekly, and daily seasonality are considered.

E. Hyperparameters Optimisation

Hyperparameter tuning is a fundamental part when training a model, as hyperparameters determine how the model is learning. For the machine learning models used in this study, these hyperparameters need to be initialised first before training. However, finding the optimal set of hyperparameters is a complicated task. Consequently, several time-saving optimisation algorithms are used to achieve better performance. An optimisation algorithm searches the hyperparameter space to find the best combination of hyperparameters to minimise the loss function of the validation set or cross-validation on the training set. The most common optimisation algorithms are: Grid Search, Random Search, and Bayesian Optimisation [37][38].

1) Random Forest

The initial hyperparameters were the default set by the library (sklearn), except for maximum depth (10), out-of-bag score (True), and evaluation function (Mean Absolute Error). The optimal set of hyperparameters was obtained using Bayesian Optimisation with 50 iterations, and they are: number of estimators, maximum depth, minimum samples split, minimum samples leaf, and maximum features.

2) Gradient Boosting

The initial hyperparameters were the default set by the library (sklearn), except for maximum depth (10), and evaluation function (Mean Absolute Error). Unlike RF, GBRT has fewer hyperparameters. The hyperparameter optimisation was done using Bayesian Optimisation with 50 iterations, and they are: number of estimators, learning rate, maximum depth, minimum samples split, minimum samples leaf, and maximum features.

3) Neural Network

In MLP, constructing a simple network helps to prevent overfitting problem, and thus the process of hyperparameter tuning could be easier. It is recommended using random search to obtain the optimal parameters, but the space of the given hyperparameter is considerably small. Therefore, grid search has been used. The MLP trained for each station consists of three layers, an input layer, a hidden layer, and an output layer. The hyperparameters were tuned separately in the same order, either separately or in pairs as stated: batch size and number of epochs, optimisation algorithms, learning rate, network weight initialisation, neuron activation function, output activation function, and the number of neurons.

The initial MLP was trained with 122 neurons in the input layer that equal to the number of variables. The weights were initialised using a uniform distribution and ReLu as an activation function. The Linear activation function was used in the output layer.

4) ARIMAX

ARIMAX do not require any hyperparameters tuning. However, a stepwise search similar to Grid Search was used to find the best order of Fourier terms for daily and weekly seasonality using AutoArima to include in the final model.

F. Model Training and Forecast

The hourly aggregated data set of each station was split into training and testing sets. For the purpose of comparison between the different models, the data dating from Nov. 01, 2018, till Oct. 31, 2019, was used as the training set and the first week of November (Nov. 01, 2019, till Nov. 07, 2019) was used as the testing set. The following station with the highest hourly demand in both check-ins and check-out were used: Allen St & Stanton St, W 41 St & 8 Ave, and 1 Ave & E 16 St for check-outs, and Allen St & Stanton St, 1 Ave & E 16 St, and Broadway & E 14 St for check-ins.

1) Data Transformation

The input variables used for machine learning models are shown in Table I.

TABLE I. DATASET INPUT VARIABLES

Variable	Description	Type
Month	1 – 12	Categorical
Day	Name of the weekday	Categorical
Hour	0 – 24	Categorical
Weekend	0 or 1, 1 indicating a weekend	Categorical
Season	Name of the season	Categorical
Holiday	0 or 1, 1 indicating a holiday	Categorical
Working Day	0 or 1, 1 indicating a working day	Categorical
Weather Type	Detailed description of the weather condition in a given hour, contains 52 unique values	Categorical
Conditions	A concise description of the weather in a given hour, containes only six unique values	Categorical
Temperature	In Fahrenheit	Numerical
Apparent Temperature	In Fahrenheit	Numerical
Relative Humidity	Mean percentage	Numerical
Preipitation	In inches	Numerical
Precipitation Cover	Percentage	Numerical
Dew Point	In Fahrenheit	Numerical
Visibility	In miles	Numerical
Wind Speed	Miles per hour	Numerical

RF and GBRT do not necessarily require the data to be transformed. Nonetheless, this option is not provided in the sklearn library. Consequently, they cannot handle categorical data. On the other hand, MLP requires the data to be normalised and transformed. As a result, the categorical variables were transformed using a one-hot encoder, and the numerical variables were normalised using standardisation. This process was repeated for three models.

2) Cross-Validation

Cross-validation is usually necessary for hyperparameter tuning. The data is split into k-fold where one-fold data is held back. The model is trained using the seen data and validated using held data. The process is repeated k times. The data used is a time-series data, hence, the time-series split by the sklearn library was used. It splits the data at a fixed interval without shuffling. In this study, 3-fold cross-validation was used.

3) Evaluation Metrics

The model performance is evaluated using three metrics: computational time, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

- Computational Time

The time it took each algorithm to complete a model using the training set is the main criterion to be considered. However, the time spent on hyperparameters optimisation is also considered.

- Mean Absolute Error

Mean absolute error (MAE) measures the average magnitude of the errors in a set of predictions without considering their direction. The MAE is calculated according to equation (3).

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \qquad (3)$$

- Root Mean Squared Logarithmic Error

Root mean squared logarithmic error (RMSLE) is a quadratic scoring that measures the average magnitude of the log of the error. The formula of RMSLE is shown as equation (4).

$$RMSLE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\log(y_j) - \log(\hat{y}_j))^2} \qquad (4)$$

IV. RESULTS & FINDING

GBRT and ARIMAX model tend to produce negative values. It is a common statistical practise to convert those negative values to zero in such case. Accordingly, the error terms were recalculated after obtaining the prediction dataset, which gave slightly different errors due to eliminating all the negative values.

Table II shows the average computational time to obtain the hyperparameters for each of the machine learning algorithms and the order of Fourier terms for daily and weekly seasonality as well as the order of the ARIMA model.

TABLE II. COMPUTATIONAL TIME (AVERAGE)

Model	Hyperparameter Optimisation	Model Training
RF	13 min 8 sec	1 min 4.3 sec
GBRT	16 min 35 sec	17.69 sec
MLP	3 min 9 sec	9.19 sec
ARIMAX	2 h 37 min 10 sec	1 min 8.7 sec

While the computational time of all the models is relatively small, getting the predication had the same computational cost except for ARIMAX, which had an average of 1 hour 24 minutes and 24 seconds. Table III shows the average errors of predictions performed with each model in term of MAE and RMSLE metrics. Fig. 11 shows an example of the comparison between predictions by four models versus observed bike flows.

TABLE III. PERFORMANCE METRICS (AVERAGE)

Model	MAE	RMSLE
RF	3.6109	0.4946
GBRT	3.4251	0.4678
MLP	3.1634	0.4462
ARIMAX	3.9571	0.5126

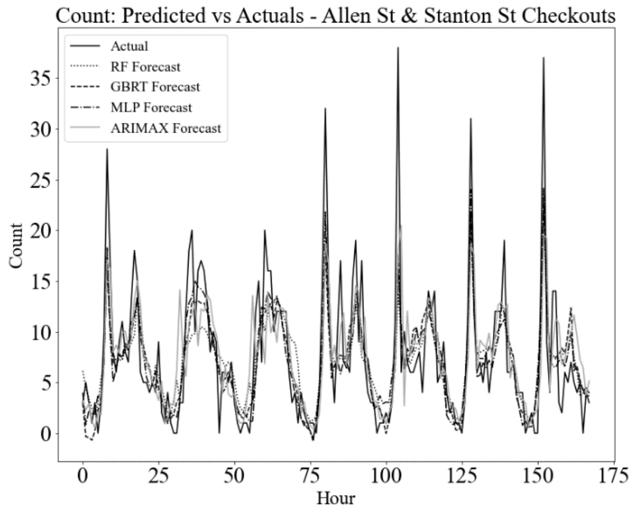

Fig. 11 Actual vs Predicted Bike Flow

V. CONCLUSION & DISCUSSION

This study was set to examine the bikesharing demand characteristics and four modelling techniques were adopted for the station-based prediction of bike flows. The techniques included three machine learning algorithms and the traditional time series model ARIMAX. Four models were developed and compared in term of their prediction performance. This study showed that neural networks have better predictive capacity than other three methods. The traditional time series model ARIMAX is usually treated as a benchmark for comparison in the literatures. An important finding of this study suggests that the traditional time-series model ARIMAX could also provide accurate station-based prediction of the bike demand if capture multiple seasonalities in the data. The finding formed a foundation for future research in the prediction of micro-mobility with multiple seasonality using traditional time series modelling. It would be a fruitful area for further work in investigating time series model that deals with multiple seasonality in the transport sector, as much of research can be seen in microeconomic and the stock market. A further study could assess how to expand these models to predict the bike flows of all stations simultaneously using the same micro-period or less, using such as bigQuery and other automation tools and techniques.

ACKNOWLEDGMENT

The research is partially sponsored by the 12.04 MO3 project, IGSSE TUM.

REFERENCES

[1] S. A. Shaheen, S. Guzman, and H. Zhang, ‘Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future’, Transportation Research Record: Journal of the Transportation Research Board, vol. 2143, no. 1, pp. 159–167, Jan. 2010, doi: 10.3141/2143-20.

[2] R. Buehler and A. Hamre, ‘Business and Bikeshare User Perceptions of the Economic Benefits of Capital Bikeshare’, Transportation Research Record: Journal of the Transportation Research Board, vol. 2520, no. 1, pp. 100–111, Jan. 2015, doi: 10.3141/2520-12.

[3] M. Cheng, ‘Sharing economy: A review and agenda for future research’, International Journal of Hospitality Management, vol. 57, pp. 60–70, Aug. 2016, doi: 10.1016/j.ijhm.2016.06.003.

[4] P. Muñoz and B. Cohen, ‘Mapping out the sharing economy: A configurational approach to sharing business modeling’, Technological Forecasting and Social Change, vol. 125, pp. 21–37, Dec. 2017, doi: 10.1016/j.techfore.2017.03.035.

[5] M. Hossain, ‘Sharing economy: A comprehensive literature review’, International Journal of Hospitality Management, vol. 87, p. 102470, May 2020, doi: 10.1016/j.ijhm.2020.102470.

[6] S. Shaheen, N. Chan, A. Bansal, and A. Cohen, ‘Shared Mobility: A Sustainability & Technologies Workshop: Definitions, Industry Developments, and Early Understanding’, Nov. 2015, Accessed: May 07, 2020. [Online]. Available: https://trid.trb.org/view/1375066.

[7] S. Shaheen, E. W. Martin, A. F. Cohen, and R. S. Finson, ‘Public Bikesharing in North America: Early Operator and User Understanding’, 2012. /paper/Public-Bikesharing-in-North-America%3A-Early-Operator-Shaheen-Martin/62f8a00c7590065d949aab5256165f0f935bec4b.

[8] S. Shaheen, S. Guzman, and H. Zhang, ‘In City Cycling’, in Bikesharing across the Globe, vol. 9, MIT Press. UC Berkeley: Transportation Sustainability Research Center, 2012.

[9] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, ‘Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system’, Pervasive and Mobile Computing, vol. 6, no. 4, pp. 455–466, Aug. 2010, doi: 10.1016/j.pmcj.2010.07.002.

[10] J. W. Yoon, F. Pinelli, and F. Calabrese, ‘Cityride: A Predictive Bike Sharing Journey Advisor’, in 2012 IEEE 13th International Conference on Mobile Data Management, Bengaluru, India, Jul. 2012, pp. 306–311, doi: 10.1109/MDM.2012.16.

[11] G. Cantelmo, R. Kucharski, and C. Antoniou, ‘Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data’, Transportation Research Record: Journal of the Transportation Research Board, vol. 2674, no. 8, pp. 132–144, Aug. 2020, doi: 10.1177/0361198120932160.

[12] C. Gallop, C. Tse, and J. Zhao, ‘A Seasonal Autoregressive Model Of Vancouver Bicycle Traffic Using Weather Variables’, i-manager’s Journal on Civil Engineering, vol. 1, no. 4, pp. 9–18, Nov. 2011, doi: 10.26634/jce.1.4.1694.

[13] K. Gebhart, ‘The Impact of Weather Conditions on Capital Bikeshare Trips’, p. 25, 2013.

[14] P.-C. Chen, H.-Y. Hsieh, X. K. Sigalingging, Y.-R. Chen, and J.-S. Leu, ‘Prediction of Station Level Demand in a Bike Sharing System Using Recurrent Neural Networks’, in 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, Jun. 2017, pp. 1–5, doi: 10.1109/VTCSpring.2017.8108575.

[15] D. Washimkar, ‘Forecast Use of a City Bikeshare System’, Dec. 15, 2014. http://darshanwashimkar.github.io/documents/Forecast-Use-Of-A-City-Bikeshare-System.pdf (accessed Apr. 27, 2020).

[16] Y. Li, Y. Zheng, H. Zhang, and L. Chen, ‘Traffic prediction in a bike-sharing system’, in Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS ’15, Bellevue, Washington, 2015, pp. 1–10, doi: 10.1145/2820783.2820837.

[17] S. Feng, H. Chen, C. Du, J. Li, and N. Jing, ‘A Hierarchical Demand Prediction Method with Station Clustering for Bike Sharing System’, in 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, Jun. 2018, pp. 829–836, doi: 10.1109/DSC.2018.00133.

[18] J. Liu, L. Sun, Q. Li, J. Ming, Y. Liu, and H. Xiong, ‘Functional zone based hierarchical demand prediction for bike system expansion’, 2017, vol. Part F129685, pp. 957–966, doi: 10.1145/3097983.3098180.

[19] H. Yang, X. Zhang, L. Zhong, S. Li, X. Zhang, and J. Hu, ‘Short-term demand forecasting for bike sharing system based on machine learning’, 2019, pp. 1295–1300, doi: 10.1109/ICTIS.2019.8883732.

[20] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, ‘Mobility Modeling and Prediction in Bike-Sharing Systems’, in Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys ’16, Singapore, Singapore, 2016, pp. 165–178, doi: 10.1145/2906388.2906408.

[21] S. Ruffieux, N. Spycher, E. Mugellini, and O. A. Khaled, 'Real-Time usage forecasting for bike-sharing systems: A study on random forest and convolutional neural network applicability', 2018, vol. 2018-January, pp. 622–631, doi: 10.1109/IntelliSys.2017.8324359.

[22] B. Wang and I. Kim, 'Short-term prediction for bike-sharing service using machine learning', 2018, vol. 34, pp. 171–178, doi: 10.1016/j.trpro.2018.11.029.

[23] D. Duran-Rodas, E. Chaniotakis, and C. Antoniou, 'Built Environment Factors Affecting Bike Sharing Ridership: Data-Driven Approach for Multiple Cities', Transportation Research Record: Journal of the Transportation Research Board, vol. 2673, no. 12, pp. 55–68, Dec. 2019, doi: 10.1177/0361198119849908.

[24] W. El-Assi, M. Salah Mahmoud, and K. Nurul Habib, 'Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto', Transportation, vol. 44, no. 3, pp. 589–613, May 2017, doi: 10.1007/s11116-015-9669-z.

[25] Weather Data & Mapping Visual Crossing. https://www.visualcrossing.com/ (accessed Jun. 03, 2020).

[26] E. Eren and V. E. Uz, 'A review on bike-sharing: The factors affecting bike-sharing demand', Sustainable Cities and Society, vol. 54, p. 101882, Mar. 2020, doi: 10.1016/j.scs.2019.101882.

[27] L. Breiman, 'Random Forest', Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[28] L. Breiman, Arcing The Edge. 1997.

[29] J. H. Friedman, 'Greedy Function Approximation: A Gradient Boosting Machine', The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, Accessed: Jul. 13, 2020. [Online]. Available: https://www.jstor.org/stable/2699986.

[30] J. H. Friedman, 'Stochastic gradient boosting', Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.

[31] Y.-S. Park and S. Lek, 'Artificial Neural Networks', in Developments in Environmental Modelling, vol. 28, Elsevier, 2016, pp. 123–140.

[32] H. G. Daellenbach and R. L. Flood, The Informed student guide to management science. London: Thomson, 2002.

[33] 'Statistical Learning and Data Analytics for TS', Technical University of Munich, Jul. 23, 2019, Accessed: Aug. 04, 2019. [Online].

[34] 'Stationarity and detrending (ADF/KPSS) — statsmodels'. https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html (accessed Aug. 04, 2019).

[35] R. J. Hyndman, 'Forecasting with daily data | Rob J Hyndman', Sep. 17, 2013. https://robjhyndman.com/hyndsight/dailydata/ (accessed Aug. 01, 2020).

[36] R. J. Hyndman, 'Forecasting with long seasonal periods | Rob J Hyndman', Sep. 29, 2010. https://robjhyndman.com/hyndsight/longseasonality/

[37] 'Hyperparameter tuning for machine learning models.', Jeremy Jordan, Nov. 02, 2017. https://www.jeremyjordan.me/hyperparameter-tuning/ (accessed Jul. 15, 2020).

[38] 'Common Problems in Hyperparameter Optimization', SigOpt, Mar. 29, 2017. https://sigopt.com/blog/common-problems-in-hyperparameter-optimization/ (accessed Jul. 14, 2020).