Journal of Visual Language and Computing

journal homepage: www.ksiresearch.org/jvlc/

Long-Term Predictions of Bike-Sharing Stations' Bikes Availability

Enrico Collini[a], Paolo Nesi[a,*] and Gianni Pantaleo[a]

[a]Distributed Systems and Internet Technologies Lab, Department of Information Engineering, University of Florence, Florence, Italy, Https://www.disit.org, Https://www.snap4city.org

ARTICLE INFO

Article History:
Submitted 3.1.2021
Revised 6.1.2021
Second Revision 8.1.2021
Accepted 8.15.2021

Keywords:
Available bikes prediction
Bike-sharing
Machine learning
Prediction models
Smart city

ABSTRACT

Bike-sharing systems are present in many cities as a valid alternative to fuel-based public transports since they are eco-friendly, prevent traffic congestions, reduce the probability of social contacts. On the other hand, bike-sharing present some problems such as the irregular distribution of bikes on the stations/racks/areas (still very used for e-bikes) and for the final users the difficulty of knowing in advance their status with a certain degree of confidence, whether there will be available bikes at a specific bike-station at a certain time of the day, or a free slot for leaving the rented bike. Therefore, providing predictions can be useful for improving the quality of e-bike services. This paper presents a technique to predict the number of available bikes and free bike slots in bike-sharing stations (the best solution for e-bikes). To this end, a set of features and predictive models have been developed and compared to identify the best prediction model for long-term predictions (24 hours in advance). The solution and its validation have been performed by using data collected in bike stations in the cities of Siena and Pisa, in the context of Sii-Mobility National Research Project on Mobility and Transport and Snap4City Smart City IoT infrastructure. The Random Forest (RF) and Gradient Boosting Machine (GBM) offer a robust approach for the implementation of reliable and fast predictions of available bikes in terms of flexibility and robustness to critical cases, producing long-terms predictions in critical conditions (i.e., when there are only few remaining available bikes on the rack).

1. Introduction

Today, about 55% of the world's population lives in urban areas, and this figure is expected to reach 68% in 2050, according to the "World Urbanization Prospects 2018", published by the United Nations Department of Economics and Social Affairs [17]. Fuel-based transportations are one of the most important causes of certain gas emissions and thus of air pollution. Bike-sharing systems may represent a part of the solution. Therefore, their use is increasing in many cities, being a more sustainable alternative to public transportation reducing congestion and pollution. The bike-sharing solution adopting bike rack stations are capable to detect the presence of the bike, to assess their status, to recharge e-bikes, and release/manage the bike-sharing system. In this case, the bikes can be very simple even when they are e-bikes. The alternative solution could be floating bike-sharing systems in which the users can take the bikes from the road and leave them in any place, in some cases with specific rules and areas. The bikes have to be more intelligent, and capable to communicate with the central management servers their position, etc., such as Mobike solution. Floating solutions are still not very effective in the case of e-bike since the recharging can be easier on racks. The recharging of floating bikes has to be performed by collecting bikes and/or by recovering them to put a charged battery. All these activities are very expensive to be performed for a large number of bikes.

In the context of this article, the solution with simple bikes (even e-bike) and smarter stations is addressed. The bikes can be typically released at any station providing that a free slot is available, this may create

*Corresponding author
enrico.collini@unifi.it *(E. Collini)*; paolo.nesi@unifi.it *(P. Nesi)*;
gianni.pantaleo@unifi.it *(G. Pantaleo)*
ORCID(s): *0000-0002-1304-5545 (E. Collini)*; 0000-0003-1044-3107 *(P. Nesi)*; 0000-0002-9235-437X *(G. Pantaleo)*

DOI reference number: 10.18293/JVLC2021-N1-001

discomfort to the users when the station is full, and the user has to search for an empty slot in near bike racks to leave the bike, and then return by walk. One of the problems of bike-sharing is related to the irregular distribution of bikes among the various stations and the impossibility to know with a certain confidence where to find at least a bike at the desired station in a precise time slot of the day, or just few minutes in advance. The same for the possibility to find a free slot to leave the bike. Therefore, predicting the availability of bikes (as well as to predict the presence of free slots) per station over time can be very useful for managing the demands for bikes per station and to plan/schedule a bike redistribution [2].

1.1 Related Works

In recent years, many researchers have studied urban bike-sharing systems, mainly on four main areas of interest.

The first area is the *design of Bike-Sharing Systems*. In [3], a mathematical model has been proposed to determine the number of docking stations needed, their location and the possible structure of the cycle path network, as well as models to make predictions about possible routes taken by users between stations of origin and destination.

The second area is related to the *analysis of the behavior and dynamics of a Bike-Sharing system*. In [4] and [5], clustering and forecasting techniques are used on the network of Bike-Sharing stations in Barcelona to obtain useful information to describe the city's mobility. In [6], the authors studied the Vélo' system. They interpreted the system as a dynamic network by analysing how bicycle flows distribute spatially along the network. In [8], clustering techniques are used to analyse the Vienna docking station network. In [7], different Bike-Sharing services are analysed highlighting the differences in bike flows and routes.

The third area is referring to the *redistribution of bicycles among stations of the city* that is necessary to compensate for the imbalance created during their use. For example, in [18], [19], [20], the authors studied the optimization of the routes taken by vehicles with the aim of balancing the number of bicycles in each station.

The last area concerns the *prediction of bikes availability*.
In [4], four different predictive models for estimating the availability of bikes in stations have been compared. The authors use a Bayesian network to predict the status of a bike station (full, almost empty or empty) using bicycle parking information only 2 hours in advance. They achieve a forecast accuracy of about 80%. In [5], ARMA models have been used to predict the number of vacancies one hour in advance, while in [1], the authors present a system for predicting bike traffic of a bike-sharing network in Lyon. In [21], data mining and cluster techniques based on historical data series are used to estimate pickup and return activity patterns at bike stations in Vienna; while in [22], the authors presented an ARIMA model that takes into account both spatial and temporal factors to predict the number of available seats in each docker station.

In most cases, the prediction algorithm aimed at understanding the total number of used bikes in the whole network over time, which is a topic of interest for the operator. This is also much simpler than the prediction of the bike or slot on single rack.

In [23], the authors presented a predictive model of the state of the public bike-sharing stations in Barcelona 2 days in advance. Thus, the Random Forest has been applied to predict the status of a station (i.e., when a station is full, almost full, if there are slots and bikes available, almost empty or empty, two days in advance) with a maximum accuracy of about 75%. The authors also consider in the model some external factors as holidays, and weather information observing that the inclusion of these external factors was not relevant. In [24] and [25], a probabilistic approach based on dynamics modelling of a single bicycle parking using Markov chains in continuous time has been proposed. In [24], the authors predict the number of available bikes per bike station in Paris with an error measure of about 3.5 in terms of RMSE (which is very high for small size racks), for a prediction horizon of one hour in slots of 10 minutes. In [25], The authors use statistical methods to model the spatio-temporal shifts of bikes between stations, and then estimate bike check-in results based on the model and online check-out records. A random forest-based prediction mechanism is further proposed to model and forecast the users' check-out behaviours. In [26], an approach based on Graph Convolutional Neural Network has been used to analyse the dynamics between the different bike-sharing stations in the city. In [27], a deep learning model for short-term prediction of the number of available bikes is presented. In [27], the authors adopt LSTM and GRU models to predict the number of bikes per docker station 1, 5 and 10 minutes in advance with one-month historical data, and they apply a Random Forest as a benchmark. In [28], the authors present an approach based on the application of machine learning models as Random Forest and LSBoost algorithms to create univariate models to predict the number of available bikes at each of the 70 stations of the Bay Area Bike Share network. RF with a MAE of 0.37 outperformed LSBoost with a MAE of 0.58 bikes/station with a prediction horizon of 15 minutes. The authors also apply a Partial Least-Squares Regression to model available bikes at the spatially correlated stations of each region obtained from the trip's adjacency matrix. Results show that the MAE was approximately 0.6 bikes. Finally, in [29], the authors propose a framework based on recurrent neural networks to predict bike demand for each station in a bike-sharing system one hour in advance. Table 1 shows a comparative overview of the most relevant related works.

Table 1: Related Work implementation overview

Paper	Models Type	Features	Pred. Horizon	Accuracy/ Error Measures
[4]	Bayesian network (station status: full; almost empty; empty)	Holiday, Weather and Historical data	2 hours	Accuracy about 80%
[23]	Random Forest (RF) (station status: full; almost full; bikes available; almost empty; empty)	Weather and Historical data	2 days	Accuracy about 70%
[24]	Markov chains (#available bikes)	Historical data	1 hour	RMSE about 3.5
[27]	- GRU -LSTM - RF (#available bikes)	Historical data	3-time intervals: 1, 5, 10 min	-
[28]	- RF - LSBoost - PLSR (#available bikes)	Historical data	From 15 min to 120 min	MAE (RF) about 0.37 for 15 min pred. horizon
[29]	RNNs (#check-in/check-out)	Weather and Historical data	1 hour	MAE about 1.2

1.2 Article Overview

The **main contribution of this paper** consists in presenting a solution for long-term prediction of available bikes on bike-sharing stations, and thus of the number of free slots by knowing the size of the station and the number of broken bikes. To this aim, a model has been identified to predict the availability of bikes 24 hours in advance (long-term predictions) with a resolution of 15 minutes, and thus also the free slots in the stations. The prediction of available bikes is a non-linear process whose dynamic changes involve multiple kinds of factors, coming from the context. To this end, the solution has been obtained by taking into account different cities and locations, and despite the differences characterizing the two cities (namely Siena and Pisa), in both cases the identified features and model have been the same, thus demonstrating the validity of the derived results. The precision obtained for long terms prediction has been much better than those provided in the literature.

The solutions have been implemented in the context of Sii-Mobility project (national mobility and transport smart city project of Italian Ministry of Research for terrestrial mobility and transport, http://www.sii-mobility.org) and Snap4City infrastructure (https://www.km4city.org) [9], [10], [11], which in turn is based on Km4City model. Sii-Mobility aimed at defining solutions for sustainable mobility, engaging city users, providing predictions on parking, suggesting bikes availability status to users at least 15 minutes/1 hour in advance to allow them to make a conscious decision, and maybe change their own plan. As a result, the solution has been capable to produce reliable prediction even 24 hours in advance.

The paper is structured as follows. Section 2 provides a description of the bike-sharing data and their characterization in terms of clustering in groups. In addition, the identification of several features at the basis of the predictive models is reported. In Section 3, the machine learning approaches adopted to identify and validate the predictive models and framework are presented. Section 3.1 presents the metrics for the assessment, Section 3.2 the ARIMA model. In Section 3.3, the machine learning approaches are presented. A computational cost analysis on the proposed solutions is presented in Section 3.4 . The feature relevance of the predictive model is discussed in Section 3.5 and Section 3.6 reports the results based on a feature reduction analysis. Conclusions are drawn in Section 4.

2. Data Description And Feature Identification

As mentioned in the introduction, the main goal was to find a solution to predict the number of bikes available in each bike station 24 hours in advance. Thus, by knowing the size of the bike station and the number of broken bikes on the rack, we can derive the number of free slots to leave the rented bike. Typically, the status of each bike station is checked and registered on the central server every 15 minutes. The data we adopted refer to 15 stations located in the municipality of Siena and 24 stations located in Pisa. In order to understand the typical time trend H24 (multiple seasonality may be present, i.e.: daily, weekly and seasons over the year) of bikes availability per station. Since the service acceptance is evolving quite rapidly over time, the seasonal trends taken into account are the daily and weekly ones. This means that the learning and predictions have to be continuously updated. We took into account data from June 2019 to March 2020 for Siena and Pisa stations. A clustering approach has been applied in order to classify together Pisa and Siena's stations based on their time trend of bikes availability over the day, which is also correlated to the typical services in the neighbourhoods. In detail, the K-means clustering method has been applied to identify clusters. In K-means clustering, there is an ideal center point that represents a cluster. The clustering has been performed on the basis of the H24 time trend, considering the normalized trend of bikes availability measure. The optimal number of clusters resulted to be equal to 3, and it has been identified by using the Elbow criteria [12]. In particular, each cluster represents a group of bike-sharing stations. For each cluster, we selected the representative bike rack as the one closer to the center of the considered cluster. Figure 1 reports the typical

trends during the day of the representative bike rack for each cluster.

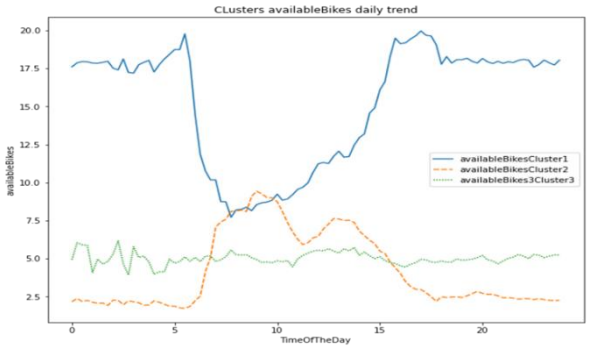
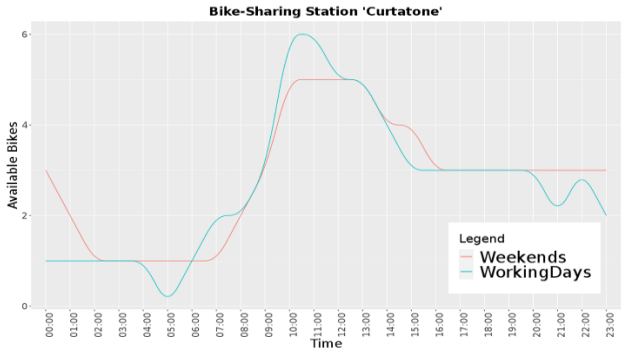


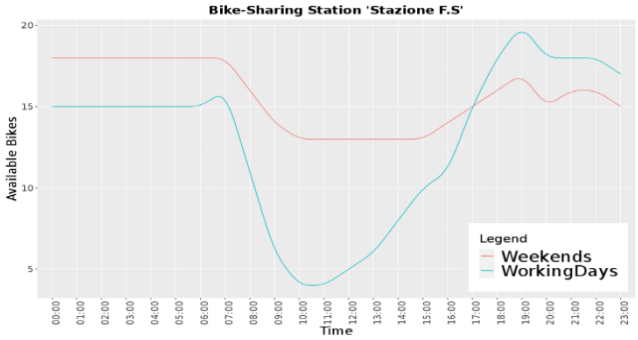
Figure 1: Typical day trend of available bikes of bike rack clusters. Cluster 1 is represented by "Stazione FS." in Pisa, Cluster 2 by "PoloMarzotto" in Pisa, and Cluster 3 by "Due Ponti" rack in Siena.

The bike stations/racks belonging to **Cluster 1** are typically characterized by a decrement of bike availability at lunchtime, and they are mainly located close to the railway stations, airport, mobility hubs, etc. Bike racks belonging to **Cluster 2** are typically positioned in the central area of the cities, and they are characterized by an increment of the bikes availability in the central part of the day (lunch hours, since most of the people are parking their bikes to get lunch). **Cluster 3** presents an almost uniform trend in the bike availability and bike racks are mainly positioned in the peripheral areas of the city.

Moreover, we have also detected some changes in the typical time trends from working days and weekends as shown in Figure 2. Figure 2a reports the comparison between the trend for working days and weekends for "Curtatone" station in Siena, while Figure 2b shows the trends of working days/weekends for the bike rack called "Stazione F.S" in Pisa.



(a)



(b)

Figure 2: Working days/weekend trends of the (a) "Curtatone" bike-sharing stations in Siena and (b) "Stazione F.S" stations in Pisa municipality.

2.1 Feature Identification

With the aim of designing a prediction model, a set of features have been proposed, identified, and tested. Typically, the values are recorded every 15 minutes. Please note that the temporal window for the training is not based only on 15 minutes, but the measures over months are taken every 15 minutes.

Features belonging to the ***Baseline*** **(time series)** category refer to aspects related to the direct observation of bike status over time as in [13]. Date and time when measures are taken, the number of bikes on racks, information on weather the observation day was a weekend etc., belong to this category.

We considered also features describing the **differences over time**. Usually, the trend of the number of bikes is similar from one week to another for the same day (e.g., Monday to prev/next Monday), in the same month for example. Therefore, the following features have been included and refer to the number of available bikes at the observation time *t* in the day *d*, with respect to the previous week *(d-7)* (PwB) and the previous day *(d-1)* (PdB), as:

$$\text{PwB} = availableBikes_{d-7,t}$$
$$\text{PdB} = availableBikes_{d-1,t}$$

And thus, other features have been included in the model for capturing the difference between the number of bikes captured at the observation time (time slot t and day d) and the available bikes in the:

- previous time slot (t-1) of previous week (d-7) dPw:
$$dPw = availableBikes_{d,t} - availableBikes_{d-7,t-1}$$
- successive time slot (t+1) of previous week (d-7) dSw:
$$dSw = availableBikes_{d,t} - availableBikes_{d-7,t+1}$$
- previous time slot (t-1) of the previous day (d-1) dPd:
$$dPd = availableBikes_{d,t} - availableBikes_{d-1,t-1}$$
- successive time slot (t+1) of the previous day (d-1) dSd:
$$dSd = availableBikes_{d,t} - availableBikes_{d-1,t+1}$$

Other features have been included in the model for capturing the difference between the number of bikes captured at the observation time:

- week (d-7) and the one of two weeks prior (d-14) dP2w:

$$\text{dP2w} = availableBikes_{d-7,t} - availableBikes_{d-14,t}$$

- day (d-1) and the one of two days prior (d-2) dP2d:

$$\text{dP2d} = availableBikes_{d-1,t} - availableBikes_{d-2,t}$$

Features belonging to the **real-time weather and weather forecast** are also collected every 15 minutes (i.e., temperature, humidity and rainfall). Please note that, according to our analysis, the significant values for the weather are those related to the current time and the hour just before the measured bike availability time. For example, in order to predict the number of available bikes at the rack at 3 pm, the weather features at 2 pm and at the current time are relevant. Thus, the weather conditions influence the decisions on using the bike or other transportation means. Similarly, the weather forecast influences the plan to get the bike.

The data collected from historical values of each bike rack are in practice all the data in the learning window (several weeks or months) of the past, as described in Section 2. For each time sample, the features of Table 2 are collected and when needed estimated and stored.

Table 2: Overview of the feature used in the prediction models

Category	Feature
Baseline-Historical	Available Bikes in the past
	Time, month, day
	Day of the week
	Weekend, Holiday
	Previous week (PwB)
	Previous day (PdB)
Diff. from actual values and prev. observations	Previous observation's difference of the previous week (dPw)
	Subsequent observation's diff. of the previous week (dSw)
	Previous observation's difference of the previous day (dPd)
	Subsequent observation's difference of the previous day (dSd)
	Previous observation's difference between the previous week and two weeks earlier (dP2w)
	Previous observation's difference between the previous day and two days earlier (dP2d)
Real-time weather and weather forecast	Max Temperature Forecasted
	Min Temperature Forecasted
	Temperature
	Humidity
	Pressure
	Wind Speed
	Cloud Cover Percentage

When the long-term prediction is performed 24 hours in advance, the training/learning is performed once a day for each bike rack. Please note that performing the training more often may not produce significantly better results, and it is very computational expensive since the prediction should be performed for each bike rack.

3. Prediction Models

In the study of the model, we have tested several machine learning solutions to predict the number of available bikes at bike-sharing stations/racks. Several techniques have been discharged since they did not produce satisfactory results for long-term prediction, among them: Bayesian Regularized Neural Network that achieves an R2 (defined in the sequel) of about 0.4 for each bike-sharing station.

In this section, the results of the two best solutions are considered and compared to predict the number of available bikes at bike racks and to identify the features that could be the best predictors for the purpose. Thus, the techniques compared and reported in this paper are those that resulted to be the most effective. And in particular: **Random Forest** (RF) [14], **Gradient Boosting Machine** (GBM) [15] and the **Auto-Regressive Integrated Moving Average** (e.g., ARIMA) as a representative of the traditional statistical approaches [16]. Those solutions have been applied on the features presented in Table 1.

3.1 Assessment metrics

The accuracy of the resulting models has been evaluated against different metrics. Thus, before presenting the results, the assessment metrics are presented in this subsection.
The R-squared which is defined as:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n} (obs_i - pred_i)^2}{\sum_{i=1}^{n} (obs_i - \overline{y})^2} \right)$$

Where

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} obs_i$$

The MASE (Mean Absolute Scaled Error) which is calculated as:

$$MASE = mean\,(|q_t|), \qquad t = 1, \dots, n$$

where

$$q_t = \frac{obs_t - pred_t}{\frac{1}{n-1} \sum_{i=2}^{n} |obs_i - obs_{i-1}|}$$

And $obs_t = observation\ at\ time\ t$, $pred_t = prediction\ at\ time\ t, n$ is the number of the values predicted over all test sets (96 daily observations per 7 days). The RMSE (Root Mean Square Error) calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (obs_i - pred_i)^2}{n}}$$

The MAE (Mean Absolute Error):

$$MAE = \frac{\sum_{i=1}^{n} |obs_i - pred_i|}{n}$$

Among them, the MASE is clearly independent from the scale of the data. When MASE is used to compare predictive models, the best model is the one presenting the smaller MASE.

3.2 ARIMA model

The ARIMA model has been executed as multi-step forward with updated iteration technique: the forecast was computed one hour in advance. The best ARIMA model has been identical for all the clusters and resulted to be a so called (1,1,2), respectively for p, d, q order in AutoArima. ARIMA model cannot be used for medium-long term forecasts due to the relevant errors produced. An approach to cope with this problem could be to apply the forecasting ARIMA technique as a multi-step forward to make 24-hour predictions (96 time slots). In other words, to compute 24 forecasts (i.e., 1 hour in advance per 24 times): the real observations recorded in that hour (four slots of 15 minutes) are inserted into the training set, and the prediction for the next hour is computed with the new information. Therefore, the model needs to be trained every hour, so that 24 times per day per 15/20 bike-sharing stations per city, which is computationally more expensive than the others. Moreover, this approach cannot be claimed as long-term prediction. Then, the training set is updated with the observations recorded in the predicted hour and a new forecast is executed for the next hour. Table 3 shows the results for the ARIMA model for the main bike-sharing stations in the different clusters for short-term prediction.

Table 3: ARIMA multi-step forward (short term online predictions) with updated iteration results in terms of MASE and RMSE per station in Siena.

ARIMA Model Results			
MASE	**RMSE**	**Cluster**	**city**
0.10	2.22	1	Pisa
1.23	1.58	2	Siena
0.52	1.15	3	Siena

For this reason, the solution has been discharged, despite the fact that for the ARIMA, the obtained accuracy in terms of MASE on the short-term is better than those obtained by machine learning techniques for long terms, as presented in Table 5. Please remind that, the goal was to find a computationally viable solution to make satisfactory long-term predictions in terms of precision for several different cases.
The comparison of the needed processing time per each bike-sharing station, among the models considered above, is also relevant and it is reported in Table 6.

3.3 Experimental Results via machine learning

In detail, for **GBM** a regression tree with a maximum depth of 9 was used as a basic learner and the total number of trees was increased to 500 while the minimum number of observations in each leaf was increased to 5. The learning rate has been set to 0.1. Note that, determining the optimal (hyperparameter) settings for the model is crucial for the bias-reduced assessment of a model's predictive power. The choice of GBM parameters has been obtained by a hyperparameter tuning implementation. Different combinations of parameter values have been tried on the dataset (see Table 4).

Table 4: Hyperparameter ranges and types for GBM model

Hyperparameter	Type	Start	End	Default
n.tree	Integer	100	10000	100
shrinkage	Numeric	0.01	0.3	0.1
interaction.depth	Integer	3	10	1
bag.fraction	Numeric	0.1	1	0.5

The **RF** has been set with number of trees composing the forest equal to 500 and the candidate feature set equal to 1/3 of the number of the data set variables.

The result of RF and GBM machine learning solutions are compared in Table 5 with respect to the clusters, exploiting all the features presented in Table 2. The predictive models have been estimated on a training period of 7 months. MAE, MASE, RMSE and R2 measures have been estimated on a testing period of 1 week after the 7th January 2020. This comparison has highlighted that both the approaches produce similar results. On the other hand, RF is more precise in most cases obtaining a better R2. The GBM approach achieved better results only in cluster 3, which presents almost stable trends (see Figure 1) and thus less critical cases since the risk to find the rack empty is low. Moreover, the values are not very far from those obtained by RF in the same cluster.

Table 5: Machine Learning Models results and comparison for different clusters. In bold the best results for the comparison

"Stazione FS" (cluster 1)	**RF**	**GBM**
MAE	**3.467**	3.481
MASE	**0,600**	0,603
RMSE	**4.136**	4.296
R2	**0.989**	0.820
"Polo Marozotto" (cluster 2)	**RF**	**GBM**
MAE	**3.108**	3.214
MASE	**1.209**	1.250
RMSE	**3.605**	3.764
R2	**0.985**	0.763
"due Ponti" (cluster 3)	**RF**	**GBM**
MAE	1.632	**1.529**
MASE	0,999	**0,936**
RMSE	2.148	**1,991**
R2	**0,966**	0,655

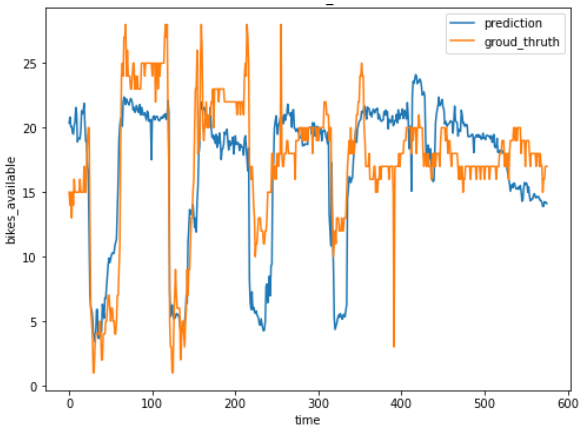


Figure 3: RF predicted values vs real in testing period for Cluster 1 reference bike rack.

3.4 Computational Costs

Table 6 shows that almost all the approaches may produce predictions every hour for the next hour in a reasonable estimation time. On one hand, in order to produce satisfactory predictions, the ARIMA approach needs to re-compute the training every hour (even if the online training can be seen as an alternative it is also a computational cost). This is a quite expensive cost of about 30s for each bike-sharing station, due to the fact that the charging stations can be hundreds. On the other hand, machine learning models (i.e., GBM and RF) provide predictive models with 96 values in advance with quite satisfactory results, they produce better results with less effort with respect to ARIMA. GBM processing time is quite low and results in terms of error measure are better with respect to the RF. GBM model can be considered the best solution for a real-time application.

Table 6: Forecasting Models comparison in terms of processing time

Processing Time	**ARIMA**	**RF**	**GBM**
Average training time	30.9 sec	410.3 sec	**21.8 sec**
Training frequency	1 time per hour	1 time per day	1 time per day
Training period	1 months	7 months	7 months
Forecast window	1 hour	1 day	1 day

3.5 Feature Relevance

In Figure 4, the feature's relevance [15] for the three clusters has been reported by considering RF and GBM. From the comparison it should be noted that both techniques present almost the same features in the first 5 most relevant features.

The most important features are those related to the past values of the time series (available bikes), to *Time*, *Day of the Week*, *weekend (yes or no)*, *Day*. The information regarding weather such as Air pressure, humidity and temperature are less relevant.

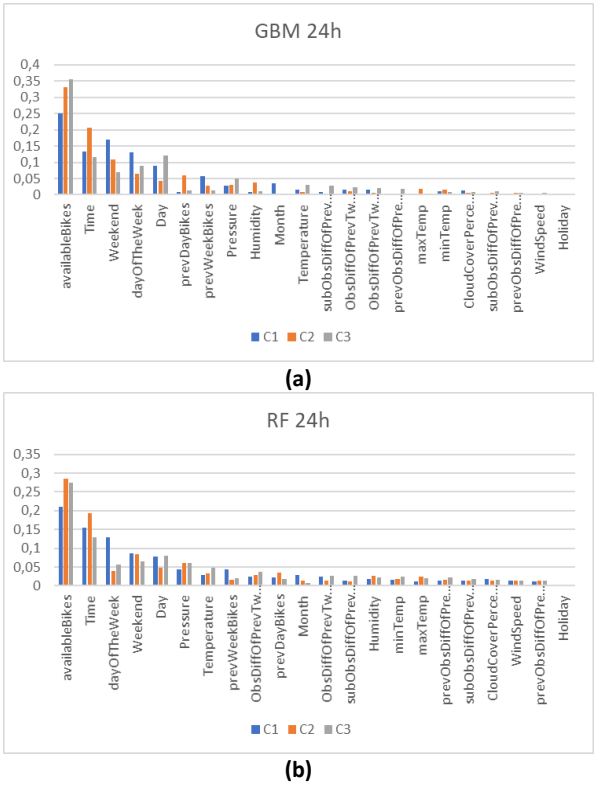


Figure 4: Feature relevance for the RF and GBM with respect to the clusters of bike racks.

3.6 Feature Reduction

According to Table 2, the features are classified into three main groups: temporal, weather, and differential. In addition, it can be observed that the top 5 features are those belonging to the temporal category. In Table 7, the impact of reducing the feature space is reported, the case in which all features are considered has been already reported in Table 5.

Table 7: Impact of feature reduction to precision of predictions in the different clusters: C1, C2 and C3.

		RF			**GBM**		
		c1	**c2**	**c3**	**c1**	**c2**	**c3**
Temporal	**MAE**	4.36	3.85	4.22	3.93	4.27	1.85
	MASE	0.75	1.33	0.73	0.68	1.47	1.13
	RMSE	5.71	4.61	5.03	4.89	4.88	2.41
	R2	0.98	0.98	0.98	0.78	0.72	0.63
Temporal + Weather	**MAE**	4.22	3.12	1.68	3.69	3.26	1.52
	MASE	0.73	1.18	1.03	0.64	1.22	0.93
	RMSE	5.03	3.6	2.19	4.38	3.84	1.96
	R2	0.98	0.98	0.96	0.81	0.76	0.65
Temporal + differential	**MAE**	3.19	3.89	1.72	3.32	4.21	1.58
	MASE	0.55	1.35	1.05	0.57	1.47	0.97
	RMSE	3.87	4.50	2.31	4.19	4.88	2.08
	R2	0.98	0.98	0.96	0.79	0.73	0.65

It can be noted that the RF results to be the best ranked in terms of R2 with respect to GBM in all cases.

In addition, it can be observed a general improvement of performance with the increment of features, as usual in RF and GBM. The weather, as well as the differential features, may lead to gain about 1% in terms of MAE (the average MAE for RF is about 4.14 for Temporal only, and 3.01 for Temporal + Weather, and 2.93 for Temporal + Differential). This analysis is providing some evidence that to compute all the features may increase the precision of a small amount at the expense of much higher computational costs.

4. Conclusions

In this paper, we proposed machine learning methods to predict the number of available bikes 24 hours in advance in any station of bike sharing systems. The proposed methods use a model which takes high dimensional time-series data from the smart bike station and uses real-time and forecast weather information as input to perform the long-term prediction. ARIMA model cannot be used for long term forecasts (24 hours in advance) because the iterative forecasting model should be trained at least 24 times per day per several bike-sharing stations per city. To this aim, RF and GBM algorithms have been considered as alternative finding a satisfactory computationally viable solutions to make long-term predictions that produce satisfactory results in terms of precision.

In the models, we have considered several features, such as the *Baseline-Historical data, the difference among actual values and previous observations, the Real-time weather and weather forecast*. In almost all predictive models, the top 5 features are those belonging to the Baseline-Historical category according to the feature relevance analysis performed. Please note that, despite the different trends of the clusters, in all cases the identified features and model have been the same, thus demonstrating the validity of the derived results. Using all the features may increase the precision of the models of a small amount compared to reducing the feature space to the top 5 or including also the weather or the differential metrics.

The entire approach resulted to be very flexible and robust with respect of the sporadic lack of data samples. The predictive models can produce predictions 24 hours in advance via mobile Apps. The solution has been deployed as a feature of Smart City Mobile Apps in the Tuscany area to encourage sustainable mobility. https://play.google.com/store/apps/details?id=org.disit.toscana

Acknowledgment

The authors would like to thank the MIUR, the University of Florence and companies involved for co-founding Sii-Mobility national project on smart city mobility and transport. Km4City and Snap4City (https://www.snap4city.org) are open technologies and research of DISIT Lab. Sii-Mobility is grounded and has contributed to Km4City open solution. In addition, the authors would like to thank to Gabriele Bruni for his support in the early experiments of this research.

References

[1] Flandrin P. Robardet C. Rouquier J. Borgnat P., Abry P. and Fleury E. “Shared Bicycles in a City: a Signal Processing and Data Analysis Perspective,” *Advances in Complex Systems*, vol.14, n.3, 2011, pp.415-438.

[2] Hulot, Pierre, Daniel Aloise, and Sanjay Dominik Jena. "Towards station-level demand prediction for effective rebalancing in bike-sharing systems." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018

[3] Lin, Jenn-Rong, and Ta-Hui Yang. “Strategic design of public bicycle sharing systems with service level constraints,” *Transportation research part E: logistics and transportation review*, vol.47, n.2, 2011, pp.284-294.

[4] Froehlich, Jon Edward, Joachim Neumann, and Nuria Oliver. “Sensing and predicting the pulse of the city through shared bicycling,” *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[5] Kaltenbrunner, Andreas, et al. “Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system,” *Pervasive and Mobile Computing*, vol.6, n.4, 2010, pp.455-466.

[6] Flandrin P. Robardet C. Rouquier J. Borgnat P., Abry P. and Fleury E. “A dynamical network view of Lyon’s Velo’v shared bicycle system,” *Dynamics On and Of Complex Networks*, Volume 2. Birkhäuser, New York, NY, 2013, pp.267-284.

[7] Gupta S. Ma D. Bargar A., Gupta A. “Interactive visual analytics for multicity bikeshare data analysis,” *The 3rd International Workshop on Urban Computing*, New York, USA, Vol. 45, 2014.

[8] Colace, Francesco, et al. "A multilevel graph approach for predicting bicycle usage in London area." Fourth International Congress on Information and Communication Technology. Springer, Singapore, 2020.

[9] C. Badii, P. Nesi, I. Paoli. “Predicting available parking slots on critical and regular services exploiting a range of open data, *IEEE Access*, 2018, https://ieeexplore.ieee.org/abstract/document/8430514/

[10] C. Badii, E. G. Belay, P. Bellini, D. Cenni, M. Marazzini, M. Mesiti, P. Nesi, G. Pantaleo, M. Paolucci, S. Valtolina, M. Soderi, I. Zaza. “Snap4City: A Scalable IOT/IOE Platform for Developing Smart City Applications,” *Int. Conf. IEEE Smart City Innovation*, China 2018, IEEE Press. DOI: https://ieeexplore.ieee.org/document/8560331/

[11] C. Badii, P. Bellini, A. Difino, P. Nesi. “Smart City IoT Platform Respecting GDPR Privacy and Security Aspects,” IEEE Access, 8 (2020): pp.23601-23623.

[12] Kodinariya, T. M., & Makwana, P. R. “Review on determining number of Cluster in K-Means Clustering,” *International Journal*, vol.1, n.6, pp.90-95, 2013.

[13] Kim, Kyoungok. "Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations." Journal of transport geography 66 (2018): 309-320.

[14] Breiman, Leo. “Random forests,” Machine learning, vol.45, n.1, 2001, pp.5-32.

[15] J. H. Friedman. “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol.29, n.5, pp.1189–1232, 2001.

[16] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. “Time series analysis: forecasting and control,” John Wiley & Sons, 2015.

[17] United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition

[18] Chappert B. Taille A. D. L. Laroche F. Meunier F. Benchimol M., Benchimol P. and Robinet L. “Balancing the stations of a self- service “bike hire” system,” RAIRO-Operations Research, vol.45, n.1, 2011, pp.37-61.

[19] Meunier F. Chemla D. and Wolfler-Calvo R. “Balancing a bike-sharing system with multiple vehicles”, Proceedings of Congress annual de la société Française de recherche opérationelle et d’aidea la décision, ROADEF2011, Saint-Etienne, France, 2011.

[20] Morency C. Contardo C. and Rousseau L. “Balancing a dynamic public bike-sharing system,” Vol. 4. Montreal, Canada: Cirrelt, 2012.

[21] Mattefeld D. C. Vogel P., Greiser T. Understanding bike-sharing systems using data mining: “Exploring activity patterns,” Procedia-Social and Behavioral Sciences, vol.20, pp.514-523, 2011.

[22] Calabrese F. YoonJ. W., Pinelli F. “Cityride: a predictive bike sharing journey advisor”, 13th International Conference on Mobile Data Management, IEEE, 2012.

[23] Boris Bellalta Gabriel Martins Dias and Simon Oechsner. “Predicting occupancy trends in barcelona’s bicycle service stations using open data,” sai intelligent systems conference (intellisys), IEEE, 2015.

[24] Daniël Reijsbergen Mirco Tribastone n Nicolas Gast, Guillaume Massonnet. “Probabilistic forecasts of bike-sharing system for journey planning,” Proceedings of the 24th ACM international on conference on information and knowledge management, 2015.

[25] Yuanchao Shu Peng Cheng Jiming Chen Thomas Moscibroda Zidong Yang, Ji Hu. “Mobility modeling and prediction in bike-sharing systems,” In Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, pp.165-178.

[26] Lin, Lei, Zhengbing He, and Srinivas Peeta. “Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach,” Transportation Research Part C: Emerging Technologies, vol.97, 2018, pp.258-276.

[27] Wang, Bo, and Inhi Kim. “Short-term prediction for bike-sharing service using machine learning,” Transportation research procedia, vol.34, 2018, pp.171-178.

[28] Ashqar, Huthaifa I., et al. “Modeling bike availability in a bike-sharing system using machine learning,”, 5th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), IEEE, 2017.

[29] Chen, Po-Chuan, et al. “Predicting station level demand in a bike-sharing system using recurrent neural networks,” IET Intelligent Transport Systems, 2020.