Long-Term Prediction of Bikes Availability on Bike-Sharing Stations

Daniele Cenni, Enrico Collini, Paolo Nesi, Gianni Pantaleo, Irene Paoli
Distributed Systems and Internet Technologies Lab, Department of Information Engineering, University of Florence,
Florence, Italy, Https://www.disit.org, Https://www.snap4city.org (<name>.<surname>@unifi.it)

Abstract— Bike-sharing systems have been adopted in many cities as a valid alternative to traditional public transports since they are eco-friendly, prevent traffic congestions, reduce the probability of social contacts which are probable in public means. On the other hand, they also bring some problems which include the irregular distribution of bikes on the stations/racks/areas and 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 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 service. This paper presents a technique to predict the number of available bikes and free bike-slots in bike-sharing stations (which is still the best solution for e-bikes). To this end, a set features and predictive models have been compared to identify the best models and predictors for long-term predictions. 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 Gradient Boosting Machine (GBM) offers a robust approach for the implementation of reliable and fast predictions of available bikes in terms of flexibility and robustness with respect to critical cases, producing long-terms predictions in critical conditions (when available bikes are few).

Keywords- available bikes prediction, bike-sharing, machine learning, prediction models, smart city.

I. Introduction

In recent decades, the city has become an increasingly large and complex body. The number of inhabitants living in urban areas is increasing. Today, about 55% of the world’s population lives in urban areas, and the figure is expected to rise to 68% in 2050, according to the "World Urbanization Prospects 2018", published by the United Nations Department of Economics and Social Affairs [1]. Today, transportation is one of the most important causes of certain gas emissions and thus of air pollution. In this context, bike-sharing systems may represent a part of the solution. Bike-sharing systems are widely used in many cities, offering a more sustainable alternative to public transport and reducing congestion. The station can be smart when they are capable to detect the presence of the bike, their status, and can release the bike. The alternative could be floating bike-sharing in which the bikes are more intelligent, and capable to communicate with the central management servers their position, etc., such as Mobike solution. Floating solution are still not very effecting in the case of e-bike since the recharge can be easier on racks.

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 discomfort to the users when the station is full, and the user has to move to next and 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 to find a at least a bike at a 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 free slots) per station over time can be useful for managing the demands for bikes per station and to perform the redistribution in advance [2].

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 locations 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 behaviour and dynamics of a Bike-Sharing system.* In [4] and [5], clustering and forecasting techniques have been used on the network of bike-sharing stations in Barcelona to obtain useful information to describe the city's mobility. In [6], the authors interpreted the system as a dynamic network by analysing how bicycle flows distribute spatially along the network. In [7], different bike-sharing services are analysed highlighting the differences in bike flows and routes.

The last area concerns the prediction of bikes availability [8]. In [4], four different predictive models to estimate the availability of bikes in stations have been compared. The authors used a Bayesian network to predict the status of a bike-station (full, almost empty or empty) using bike-station information and providing predictions at 2 hours, with an accuracy of 80%. In [5], ARMA (AutoRegressive Moving Average) models has been used to predict the number of vacancies one hour in advance, while in [1], the authors presented a model system for predicting bike traffic of a bike-sharing network in Lyon.

DOI Reference number: : 10.18293/DMSVIVA2021-001

A. *Article Overview*

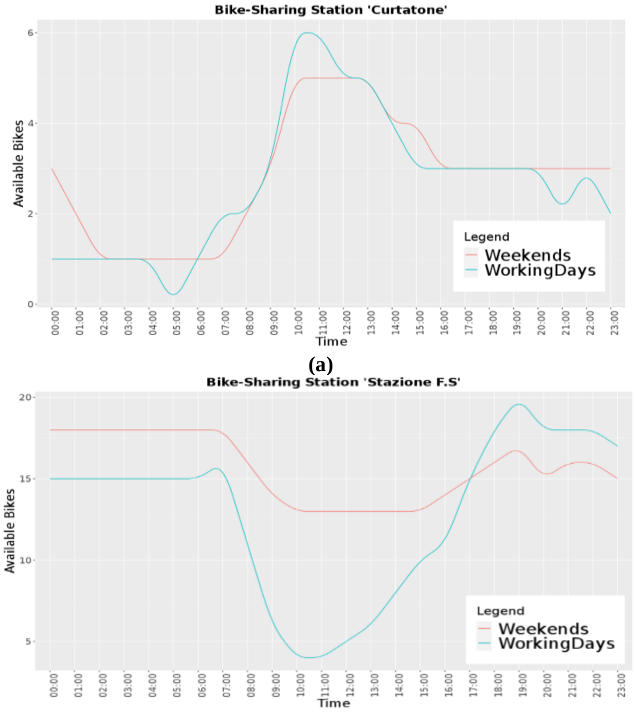
The **main contribution of this paper** consists in presenting a solution for real-time prediction of the available bikes on bike-sharing stations, and thus of the number of free slots by knowing the size of the station and the number of broken bikes. To this aim, a model has been identified to predict 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. 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 changes in 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 have been much better than those provided in the literature.

The solutions have been implemented in the context of Sii-Mobility project and infrastructure (national mobility and transport smart city project of Italian Ministry of Research for terrestrial mobility and transport, http://www.sii-mobility.org) solution based on Km4City model (https://www.km4city.org) and Snap4City tools [9], [10], [11]. Sii-Mobility project aimed at defining solutions for sustainable mobility, suggesting bikes availability status to users at least 15 minutes/1 hour in advance to allow them to take 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 II 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 III, the machine learning approaches adopted to identify and validate the predictive models and framework are presented. Conclusions are drawn in Section IV.

II. Data Description and Feature Identification

As mentioned in the introduction, the main goal was to find a solution to predict the bikes availability in each bike-station (and by knowing the size of the bike-station and the number of broken bikes on rack, we can derive the number of free slots). Typically, the status of each station is checked and registered by the server every 15 minutes. The data refers to 15 stations located in the municipality of Siena and 24 located in Pisa. In order to understand the typical time trend H24 (multiple seasonality may be present, daily, weekly and seasons over year) of bikes availability per station located in Siena and Pisa cities. Since the service is evolving quite rapidly over time, the seasonal trends taken into account are those daily and weekly. We taken into account data from June 2019 to January 2020 for Siena stations, and from December 2019 to March 2020 for Pisa stations. A clustering approach has been applied in order to classify Pisa and Siena stations based on their trend of bikes availability, 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 time trend H24, considering the normalized trend of bikes availability measures. 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 stations. The stations/racks belonging to **Cluster 1** are typically characterized by a decrement of bike availability at lunchtime, and are mainly located close to the railway stations, airport, etc. Bike racks belonging to **Cluster 2** are typically positioned in the central area of the cities and are characterized by an increment of the availability of bikes 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.



(a)

(b)

Figure 1. Working days/weekend trends of the (a) "Curtatone" bike-sharing stations in Siena and (b) "Stazione F.S" stations in Pisa municipality

For example, in Siena municipality "*Terminal Bus*" station that is a bike-sharing station positioned near by the train station in Siena, "*Ospedale*" station positioned near the Hospital, "*Due Ponti*" station positioned near the bus station/terminal, "*Curtatone*" station positioned near the stadium and "Napoli" positioned in residential areas. In Pisa municipality "*Comune Palazzo blu*" station is near the municipality building, "*Ospedale Cisanello*" station positioned near the Hospital, "*San Rossore F.S*" station positioned near the train station San Rossore, "*Stazione F.S*" station positioned near the central train station and "*Marchesi*" station near the educational buildings.

Moreover, we have also detected some changes in the typical trends from working days and weekends as shown in **Figure** 1. **Figure** 1**(a)** reports the comparison between the trendsd for working days and weekends for "Curtatone" station in Siena, while **Figure** 1**(b)** shows the trends of working days/weekends for "Stazione F.S" in Pisa.

A. *Feature Identification*

With the aim of developing a prediction model, a set of features have been identified and tested. We analysed a large number of features for selecting the best, with the aim of conquering a higher precision with respect to the state-of-the-art solutions mentioned above. So that the hypothesis has been verified in the results reported in this paper for the case of bike-station status predictions. 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, working day or not, number of bikes on racks, etc., belong to this category. 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 describing the **differences over time**. Usually, the trend of 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. Thus, two other features have been included in the model for capturing: (i) the difference between the number of bikes captured at the same time in the previous time slot of previous week (dPw); (ii) the difference between the number of bikes captured at the same time in the successive time slot of previous week (dSw). The value of the number of bikes related to the previous week respect to the observed one at the same time has been considered as additional feature (PwB).

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
	Min Temperature
	Temperature
	Humidity
	Rain
	Pressure
	Wind Speed
	Cloud Cover Percentage
	Sunrise
	Sunset

Table 1. Overview of the feature used in the prediction models

Features belonging to the **real time weather and weather forecast** *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 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 has described in Section II. For each time sample, the features of **Table** 1 are collected and when needed estimated and stored. When the long terms prediction is performed 24 hours in advance, the training/learning is performed once a day for each bike rack. To perform the training more often is not producing better results, and it is very computational expensive.

III. PREDICTION MODELS

In this section, number of machine learning techniques are considered and compared to identify the best solution to predict the bikes availability at bike-sharing stations/racks and to identify the features that could be the best predictors for the purpose. During our research study a number of techniques have been discharged since they did not produce satisfactory results -- e.g., Bayesian Regularized Neural Network that achieves an R-squared (https://en.wikipedia.org/wiki/Coefficient_of_determination) about 0.4 for each bike-sharing station. On the other hand, among the techniques we have presented here the comparison of the most effective solutions, which are **Random Forest** (RF) [14], **Gradient Boosting Machine** (GBM) [15] and the more traditional statistical approach such as **Auto-Regressive Integrated Moving Average** approach (e.g., ARIMA) [16]. The accuracy of each model has been evaluated in terms of R-squared, MASE (Mean Absolute Square Error), RMSE (Root Mean Square Error), and processing time considering the representative station per cluster. The RMSE is calculated as follows:

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

The MASE is calculated as follows:

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

and

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

where $obs_t = observation\ at\ time\ t$, $pred_t = prediction\ at\ time\ t$, n is the number of the values predicted over all test sets (96 daily observations per 7 days). The MAE (Mean Absolute Error) is estimated as follows:

$$MAE = \frac{\sum_{i=1}^{n}|obs_i - pred_i|^2}{n}$$

Note that, MASE is clearly independent on the scale of the data. When MASE is used to compare predictive models, the best model is the one presenting the smaller MASE.

A. Experimental Results

In the general framework, three different approaches were tested, i.e., RF, GBM, and ARIMA models applied on the features presented in **Table 1**. 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 dataset (see **Table** 2).

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

Table 2. Hyperparameter ranges and types for GBM model

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 ARIMA model has been executed as multi-step forward with updated iteration technique: the forecast was computed one hour in advance. Then, the training set is updated with the observations recorded in the predicted hour and a new forecast is executed for the next hour. 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 5.**

ARIMA Model Results		
Siena Bike-Sharing Stations	**MASE**	**RMSE**
Curtatone	1.23	1.58
Napoli	0.51	1.10
Terminal Bus	1.15	1.32
Due Ponti	0.52	1.15
Ospedale	0.23	1.39
Pisa Bike-Sharing Stations	**MASE**	**RMSE**
C.Marchesi	0.51	1.21
Comune Palazzo Blu	0.27	1.33
Ospedale Cisanello	0.86	1.13
San Rossore F.S.	1.01	1.22
Stazione F.S.	0.10	2.22

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

Table 3 shows the results for the ARIMA model for the main bike-sharing stations in Siena and Pisa. ARIMA model cannot be used for medium-long term forecasts due to the large 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, 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 (see **Table 5**), so that 24 times per day per 15/20 bike-sharing stations per city, which is computationally more expensive than the others. For this reason, the solution has been discharged, despite to the fact that for the ARIMA, the obtained accuracy in terms of MASE is better than those obtained by machine learning techniques presented in **Tables 4** and **6**. Please remind that, the goal was to find a computationally viable solution to make satisfactory predictions in terms of precision for several different cases. As a further step, the comparison has been focused by considering RF and GBM on the whole set of bike-sharing stations in Siena (**Table 4**), exploiting all features presented in **Table 1**. The comparison of the predictive models has been estimated on a training period of 7 months. **Figure 4** reports the GBM predicted values vs real in a 96 time slots (24 hours) for "Curtatone" station in Siena city, which is a typical result.

"Curtatone" Station	**RF**	**GBM**
R2	**0.86**	0.78
MAE	2.42	**2.41**
MASE	0.82	**0.79**
RMSE	**2.90**	3.16
"Napoli" Station	**RF**	**GBM**
R2	**0.92**	0.81
MAE	2.22	**1.35**
MASE	1.10	**0.87**
RMSE	1.50	**1.45**
"Terminal Bus" Station	**RF**	**GBM**
R2	**0.91**	0.89
MAE	3.51	**3.37**
MASE	2.62	**2.52**
RMSE	2.2	**2**
"Due Ponti" Station	**RF**	**GBM**
R2	**0.96**	0.95
MAE	2.22	**1.85**
MASE	1.10	**0.92**
RMSE	2.60	**2.35**
"Ospedale" Station	**RF**	**GBM**
R2	**0.87**	0.79

MAE	**2.23**	2.35
MASE	**0.88**	0.92
RMSE	2.59	**2.35**

Table 4. Machine Learning Models results and comparison for different Siena stations

MASE and RMSE error measures have been estimated on a testing period of 1 week after the 7th January 2020. This comparison has highlighted that in Siena stations GBM approach achieved better results in terms of MASE, MAE and RMSE even if RF turned out to be the better ranked in terms of R-squared.

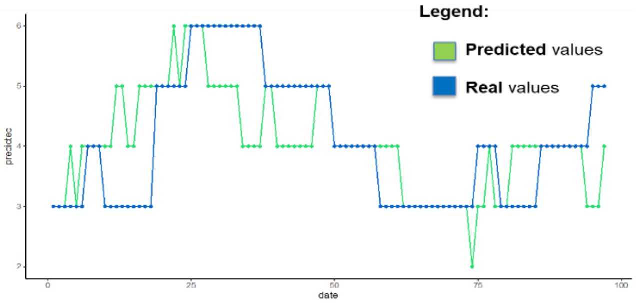


Figure 2. GBM predicted values vs real in a 96 time slots (24 hours) for "Curtatone" station in Siena

Table 5 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 respect the RF. GBM model can be considered the best solution for a real-time application.

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

Table 5. Forecasting Models comparison in terms of processing time

Figure 3 shows the GBM model feature relevance [15] for "*Curtatone*" bike-sharing station in Siena (a similar figure could be presented for Pisa while it has been omitted for the lack of space). The most important features are those related to *Time*, *Day of the Week* and Weather category as *Air pressure*, *humidity and temperature.* The same features relative influence has been obtained for the other stations in Siena municipality. This result shows that exists a strong dependence between bicycle use and weather conditions. For this reason, the only use of data related to bike-sharing station are not enough and cannot produce satisfactory and flexible results. In addition, to confirm the strong dependence with weather features, a new model exploiting baseline feature only has been trained for "*Curtatone*" station in Siena. Results show that the R2 decreases to 0.48 while error measure RMSE increase to 2.90.

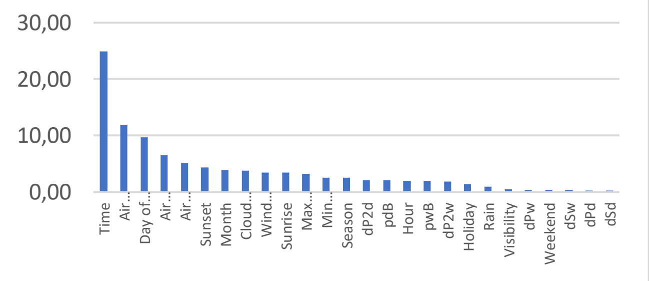


Figure 3. GBM model features relative influence for "*Curtatone*" station considering all features presented in Table 2.

The experiment above has been performed without applying feature selection. As additional analysis, a recursive feature elimination approach (RFE) based on RF was applied as a dimensionality reduction measure. The RFE technique implements a backward selection of the features by ranking their importance to an initial model using all predictors. The RFE selection method [17] is a recursive process that grades feature according to a certain degree of importance in order to filter unnecessary features and achieve a better performance of GBM model.

The RF-RFE optimization procedure has been applied to find the best performing subset of features, before applying the GBM model. The RF-RFE method identified a subset of 18 features (see **Figure** 4), in particular: *Time*, *Air Pressure*, *Day of the Week*, *Wind Speed*, *Cloud Cover Percentage*, *Hour*, *Air Temperature*, *Air Humidity*, *Sunset*, *Max Temperature*, *Min Temperature*, *Sunrise*, *Visibility*, *dP2d*, *dP2w*, *pwB*, *pdB.*

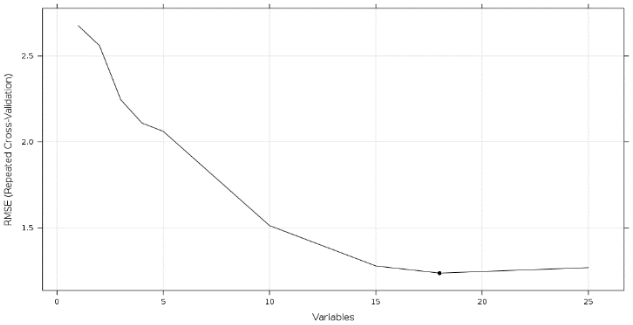


Figure 4. RF-RFE performance profile across different features subset sizes in terms of RMSE (the black point represents the best subset size of features, that is equal to 18)

Results from the GBM model trained on the identified subset of features have not shown a better accuracy with respect to those presented in **Table** 4. In conclusion, the features presented in **Table 1**, are those strictly necessary to obtain the best GBM performance in terms of R-square MASE and RMSE. In order to produce predictions, two GBM models have been trained to test the capabilities in predicting bike rack status in the next 15 and 30 minutes, respectively.

Predictions	MASE	RMSE
15 minutes	1.02	2.8
30 minutes	1.27	2.98

Table 6. GBM model predictions 15 and 30 minutes for "*Curtatone*" station, where the MASE and RMSE have been computed with respect to the true values by using 10 consecutive predictions.

Table 6 reports GBM models results for predictions showing that the RMSE does not improve much (passing from 2.9 to 2.8 for the 15 minutes), and the MASE seem to be worst with respect to the results presented in **Table 4** (the precision decreases with the distance from the last actual value, such as for ARIMA solutions). In these cases, the online learning has been performed at every time slot of 15 minutes, which is very expensive. It means that, in a perspective of online prediction and similar cost of a traditional method as the ARIMA, which can be preferable since the results in terms of error measures (in particular in terms of MASE) are better (see **Table** 3).

The same machine learning models presented above have been applied and compared on the bike-sharing stations of Pisa, considering all the features presented in **Table** 1. Note that, the amount of data available for the city of Pisa is lower than for Siena municipality. The comparison of the predictive models has been estimated on a training period of 3 months (from 1st December 2020 to 1st March 2020). **Table 7** reports the results of the comparisons of RF and GBM models for five representative stations presented in Section II. Contrary to the results achieved for stations in Siena, RF approach achieved slightly better results in terms of R-squared, MASE, MAE and RMSE in all five representative stations in Pisa, while the training period for the RF model remains significantly longer than for GBM. MASE and RMSE error measures have been estimated on a testing period of 1 week after the 1st March 2020.

"C. Marchesi" Station	**RF**	**GBM**
R2	**0.90**	0.84
MAE	**2.43**	2.77
MASE	**0.78**	0.89
RMSE	**2.91**	3.36
"Comune Palazzo Blu" Station	**RF**	**GBM**
R2	**0.94**	0.91
MAE	**1.91**	2.25
MASE	**0.95**	1.16
RMSE	**2.32**	3.01
"Ospedale Cisanello" Station	**RF**	**GBM**
R2	**0.92**	0.91
MAE	**1.18**	2.02
MASE	**0.90**	1.01
RMSE	**2.32**	2.60
"San Rossore" Station	**RF**	**GBM**
R2	**0.91**	0.86
MAE	**3.59**	4.01
MASE	**1.16**	1.25
RMSE	**4.08**	4.94
"Stazione F.S" Station	**RF**	**GBM**
R2	**0.94**	0.92
MAE	**5.11**	5.65
MASE	**0.73**	0.92
RMSE	**5.21**	6.15

Table 7. Machine Learning Models results and comparison for different Pisa stations

Also, the feature relative influence of RF model for Pisa shown that weather category variables are no longer the most influential in forecasting available bikes. The most important features are those related to the previous days/week information: in order of importance, are *PwB*, *PdB*, *Time*, *Day of the Week*, *Hour*, *DP2d* followed by *Air Pressure*, *Air Humidity* and *Air Temperature*. In the case of "Stazione F.S", the error trend was also evaluated based on the different times of the day. The results are shown in the **Table 8**.

"Stazione F.S" Station	*MAE*	*MASE*	*RMSE*
Night	5.54	1.94	6.58
Morning	**4.83**	**0.56**	**5.98**
Afternoon	4.93	0.98	6.05
Evening	5.52	0.85	6.31

Table 8. RF results for "Stazione F.S" station in Pisa per times of the day

As additional result, the same model has been trained for "Stazione F.S" bike-sharing station exploiting baseline feature only. The trained model seems not to be much worse in terms of prediction errors than the RF model results presented in **Table 8.** The RF predicted values vs real value in 24 hours for "Stazione F.S" station in Pisa, number of free bikes is reported in **Figure 5**. Results in terms of R2, MASE, MAE and RMSE are respectively 0.88, 0.76, 4.67 and 5.54.

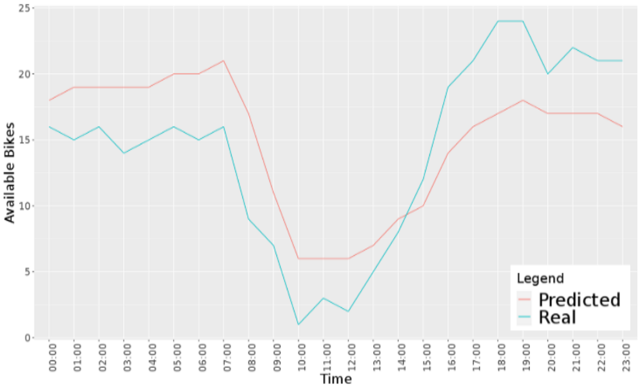


Figure 5. RF predicted values vs real in 24 hours for "Stazione F.S" station in Pisa, number of free bikes.

IV. Conclusions

In this paper, we proposed machine learning methods to predict bike availability for each station in bike-sharing systems. The proposed methods use a model which takes high dimensional time-series data from each station and uses real-time and forecast weather information as input to perform long term prediction the next 24 hours bikes availability for each bike-sharing station. The proposed solution demonstrated that in case of prediction (1 hour in advance), the ARIMA models may outperform in short time the predictions obtained using the RF and GBM algorithms. However, ARIMA model cannot be used for medium-long term forecasts 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 medium-long term predictions that produce satisfactory results in terms of precision and able to suit for several cases.

In the models, we have considered several features, such as the *historical data, difference in days and weeks, and the weather conditions and forecast.* In almost all predictive models, the baseline/historical data and weather information have demonstrated high predictive capabilities in explaining the number of available bikes. The weather features have improved the accuracy of forecasting available bikes. Please note that, despite the changes in Siena and Pisa in both cases the identified features and model have been the same, thus demonstrating the validity of the derived results. 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, while they are provided on mobile applications, 30 minutes, 1 hour in advance directly, and if requested also a day in advance as possible general trend. The solution has been deployed as an additional feature on Smart City 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] Kuhn I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene selection for cancer classification using support vector machines”, Machine Learning, vol. 46, no. 1, pp. 389-422, 2002.