Using Machine Learning to Assess the Relevance of Weather Conditions for Short-Term Demand Predictions in Bike-Sharing Systems

Marzieh Afsari*[1], Mousaalreza Dastmard[2], Guido Gentile[3]

[1] PhD student, Department of Infrastructure and Transport, Sapienza University, Italy

[2] MSc Graduated, Department of Data Science, Sapienza University, Italy

[3] Professor, Department of Infrastructure and Transport, Sapienza University, Italy

SHORT SUMMARY

Bike-sharing systems (BSS) are being studied for their potential to enhance urban accessibility and sustainable mobility. Despite its popularity, effectively managing BSS encounters challenges due to demand-supply imbalances. The significance of BSS lies in accurately predicting bike demand at various stations, a task we approach using regression methods such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Regularized Linear Regression (Ridge), and Least Absolute Shrinkage and Selection Operator (LASSO), for short-term predictions. The study utilizes data from Los Angeles on bicycle-sharing and weather conditions, alongside the p-median method for station clustering. The results demonstrate that combining both datasets yields accurate predictions at the city level, with an error rate of 0.1%, and at the station level at 18%. Notably, RF emerges as the most accurate method among the regression models examined.

Keywords: Bike-Sharing, Machine Learning, P-median.

1. INTRODUCTION

BSS plays a crucial role in improving urban accessibility, fostering diverse transportation options, and contributing significantly to sustainable mobility endeavors. Cities worldwide are progressively adopting these shared mobility solutions to confront the expanding challenges associated with urban transportation, surging air pollution, and dynamic shifts in mobility behaviors. There are three different systems of BSS: station-based, dockless, and hybrid. Each type comes with advantages and challenges, impacting user experience and system operations in different ways. Station capacities limit bike allocations, but forced rerouting (returning bikes to nearby stations when target stations are full) indirectly re-balances station-based systems. On the one hand, this advantage is diminished in dockless systems, partially offsetting the flexibility in bike allocation. On the other hand, dockless users can save 10%–15% of their trip time with bike access/return (Kou and Cai, 2021).

Nevertheless, the effective administration of bike-sharing resources encounters complexities owing to the persistent imbalance between demand and available supply. To address these challenges, two main approaches are employed: demand re-balancing and demand forecasting. Demand re-balancing adopts a redistribution strategy to determine the number of bike station and their location. This approach can be operator-based, i.e., using vehicle carriers, or user-based, i.e., offering incentives for BSS re-balancing (Harikrishnakumar and Nannapaneni, 2023). Bike demand prediction methods fall into city-level, cluster-level, and station-level categories. Models at

the city and cluster levels facilitate grouping stations for more effective planning, while the individual station-level model faces challenges in anticipating rapid changes in bike demand patterns. Bike-sharing is a hot topic in academics, and one key area of interest is predicting bike usage using algorithms. Accurate predictions help service providers plan their activities better. Ensemble learning techniques like Random Forest, Gradient Boosting Machines, and Deep Neural Network are emphasized. For example, one research in London worked on the challenge of predicting the hourly, daily, and monthly bike-sharing numbers, used machine learning regression techniques. The algorithms employed include RF, Bagging Regressor (BGR), XGBoost, and Ada Boosting (AB) regressor. Notably, RF, BGR, and XGBoost demonstrated superior performance based on metrics such as R2, MAE, Mean Squared Error (MSE), and Root Mean Squared Log Error (RMSLE) (Abdellaoui Alaoui and Koumetio Tekouabou, 2021) .

Given that bike usage data is typically characterized as time-series data, numerous research endeavors employ data mining techniques predicting demand for bike sharing in metropolitan areas. Predicting the aggregate demand for all stations becomes notably more straightforward with city-level forecasting (Wang, 2016) (Giot & Cherrier, 2014).

In addition, there are several studies work on prediction in station level; however, a limitation of these approaches is their inapplicability to free-floating BSS. Wu et al., used three common machine learning models—RF, Gradient Boosting Regression Tree (GBRT), and Neural Network (NN) to accurately predict hourly changes in the number of bikes at the station level. They tested two training methods: one predicting check-ins, another predicting check-outs, and also trained the model directly with processed bike number change data. The findings indicate that training the model on bike number change data is more effective, with the GBRT model outperforming the other models (X. Wu et al., 2019).

The bike-sharing networks display a small-world property, featuring a brief average path length and a substantial clustering coefficient. Geographical correlation is observed among clusters of stations, influencing the demand prediction model's calculation of trip rates for each cluster. This assumes that the demand within a cluster will self-equilibrate, anticipating users to locate an available bike within the same cluster (Cantelmo et al., 2020). Wu and Kim utilized complex network theory and spatial autocorrelation analysis methods to scrutinize the structural properties of bike-sharing systems. Their research not only quantified the significance of bike stations within the network but also evaluated spatial clustering patterns (C. Wu & Kim, 2020). It's essential to note that the choice of clustering methods can significantly impact prediction outcomes, adding an additional layer of consideration to the analysis.

Paper Contributions: This paper presents several key contributions, with a primary emphasis on the exploration of the bike-sharing dataset through a multilayered approach. Firstly, we initiated the process by cleaning the dataset. Subsequent to this, an extensive analysis and preprocessing phase were conducted to enhance our understanding of the underlying dynamics. To augment the dataset, external weather data were incorporated, with a noteworthy alignment of each bike-sharing station with its nearest weather station.

Additionally, our methodology involved the application of four distinct machine learning methods to discern the optimal approach. Specifically, a p-median aggregation approach was employed to consolidate a subset of stations. The degree of aggregation directly influenced the consolidation of stations into representative stations, with higher aggregation levels resulting in the integration of more stations. This systematic approach allowed us to systematically evaluate the performance of different machine learning methods and determine which one yielded the most effective outcomes.

Paper organization: In Section 2, we delve into the data regarding BSS and weather, outlining the methodology and framework for the proposed BSS forecasting. Section 3 provides an in-depth discussion of the numerical results obtained from the proposed approach. Lastly, Section 4 offers a summary of the conclusions drawn from the study and outlines avenues for future research.

2. METHODOLOGY

In this section, we outline our approach to forecasting the aggregate demand for bike shares, commencing with a description of the data employed. We initiate the process by utilizing a station clustering method to categorize bike-sharing stations. Following this categorization, machine learning algorithms are deployed to predict the quantity of shared bikes.

Data

The Bike-sharing dataset employed in this research is derived from the publicly accessible repository, originating from the Metro Bike Share initiative of the city of Los Angeles (Https://Bikeshare.Metro.Net/about/Data/, n.d.). This dataset stands offers an extensive array of information. It comprises a diverse range of attributes, such as start and end time, and location of trips, duration, and trip categories.
In Figure 1, there's a histogram showing the distribution of trip durations, and alongside it, there are bar charts displaying the distribution of trip categories. The histogram analysis reveals that a majority of trips were completed within a time frame of less than 12 minutes, with instances of trips exceeding 100 minutes being infrequent. A significant observation is that the majority of bike-sharing system users utilized the service for one-way trips, underscoring its convenience and practicality in this context.

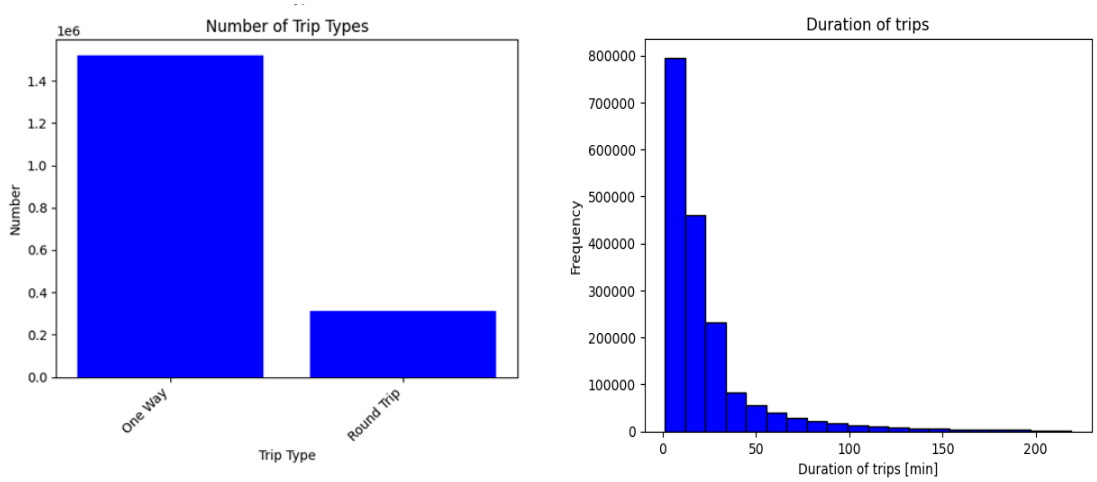

Figure 1: Number of trips type, Duration of trips [min]

The temporal utilization pattern of share-bikes is represented in Figure 2. The analysis reveals a prominent trend wherein bicycle usage experienced a significant increase from 2016 to 2019, followed by a decline until 2020 during the Covid-19 restrictions. Subsequently, a recovery in usage is observed. When examining the monthly distribution, it becomes evident that the months of June and September emerge as the most favored periods for utilizing share bikes, whereas December and February experience the lowest levels of usage. Further examination of daily distributions indicates a prevalent trend of bike usage during weekends. Additionally, an hourly distribution analysis demonstrates two distinct peaks in usage: during lunchtime (almost 12 o'clock) and during the period of returning home from work (almost 17 o'clock).

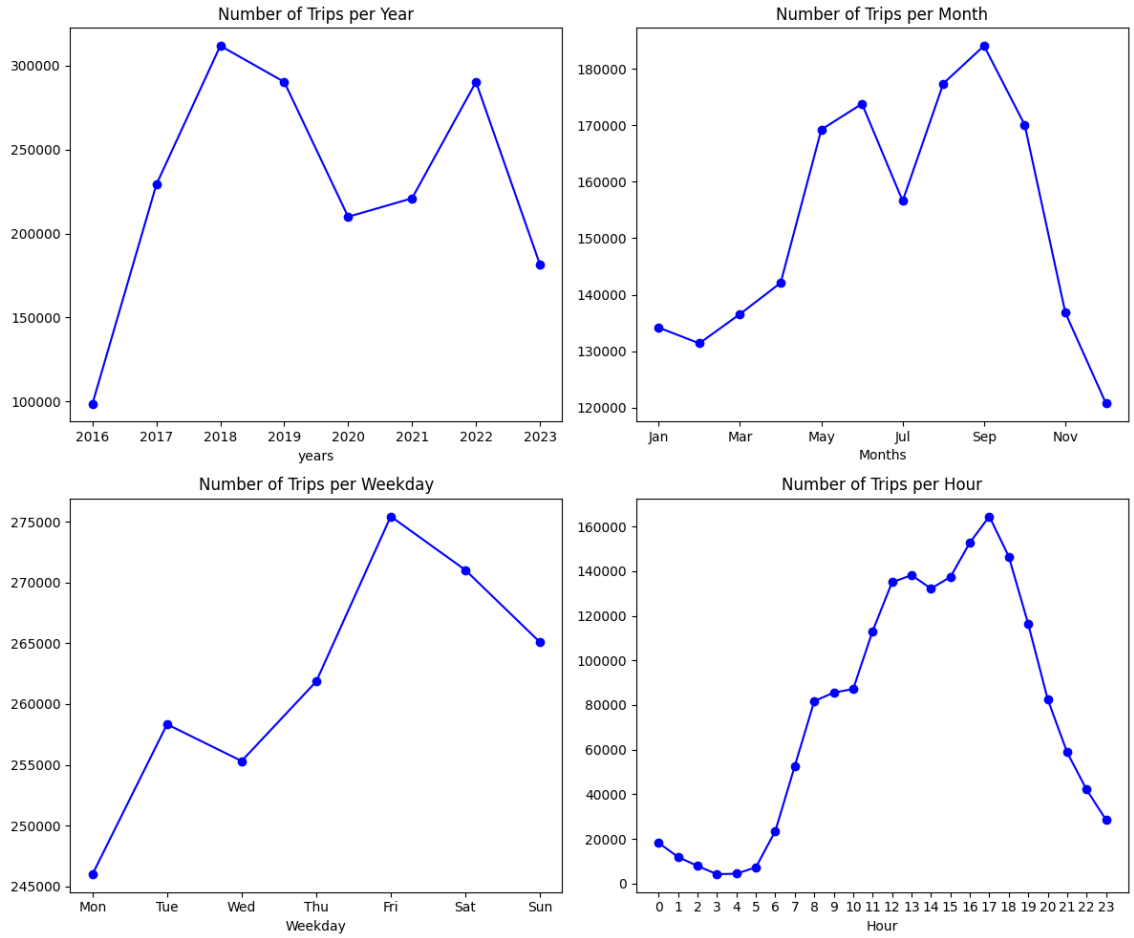

Figure 2: Temporal distribution of share-bikes using

Weather data for Los Angeles was sourced from (*Https://Www.Visualcrossing.Com/*, n.d.). This comprehensive dataset incorporates multiple weather stations distributed across the city. To ensure accurate weather information for each bike station, we adopted a strategy of associating each bike station with its nearest weather station, as depicted in Figure 3**Error! Reference source not found.**. The dataset encompasses a diverse range of weather attributes including temperature, humidity, snowfall, visibility, and windspeed.

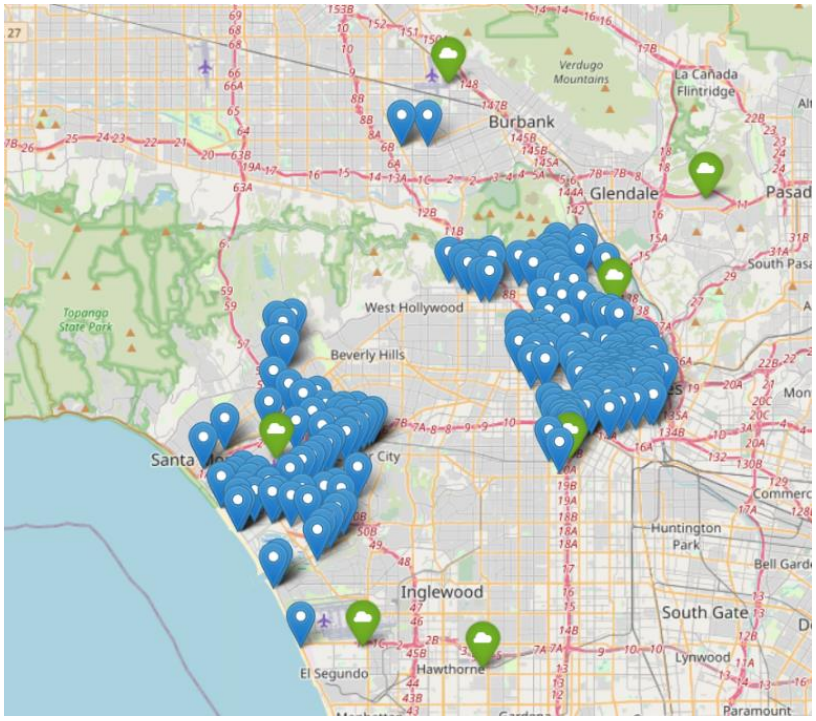

Figure 3: Bike-sharing stations (blue) and weather stations (green)

P-Median

P-Median is a mathematical optimization problem used in operations research and facility location analysis. The goal of the P-Median problem is to determine the optimal placement of a predetermined number (denoted as 'P') of facilities within a given geographic area, so as to minimize the overall cost or distance between the facilities and the demand points, they serve. Demand points are typically represented by a set of locations where customers or clients require services.
The P-Median method, also can be adapted as a robust clustering tool for demand prediction. The method identifies clusters of demand points with shared characteristics. By considering factors like transportation costs, it efficiently configures a network to serve these clusters. This adaptation is valuable for businesses aiming to predict demand within each identified cluster, providing insights for tailored strategies and enhanced operational efficiency.
The variable 'P' determines how many representative stations, treated as medians, are included in this aggregation approach Its values can span ranging from 1 (city-level) to 235 (station-level), permitting a systematic investigation into its impact on the outcomes. Any number between 1 to 235 will be considered cluster-level.

Regression methods:

In this investigation, we've incorporated four well-regarded machine learning algorithms known for their proven success in handling timeseries tasks. Here's an overview of the algorithms utilized in this study: RF (is an ensemble method combining multiple decision trees for time series analysis capturing complex patterns by averaging or voting among trees), XGBoost (is a sequential ensemble learning algorithm effective in forecasting time series, correcting errors of previous trees, known for speed and performance.), LASSO (is a regression technique with absolute value-

based penalty, useful in time series for feature selection and parameter shrinkage, tackling high-dimensional data), RIDGE (is a linear regression with regularization term penalizing squared coefficients, effective in time series to prevent over fitting and enhance stability, especially in multi collinear scenarios).

It is worth mentioning that the performance of the model was assessed using Mean Absolute Percentage Error (MAPE) which is calculated as the Mean Absolute Error (MAE) divided by the Mean Grand Truth (MGT). MAPE represents the average absolute differences between predicted and actual values. We use MGT to rescale errors because it ensures that models are comparable on the same scale. Without rescaling, the grand truth for lower P values is significantly higher. As a result, comparing MAE when P is one with MAE when P is 235 wouldn't be fair.

Concluding our methodology, we started creating time series features and conducted an exhaustive evaluation of our predictive models. Central to our analysis were the definition and utilization of the modeling parameters, which played a pivotal role in shaping the trajectory of our investigations:

- Window length: It's the short-term forecast window length, and can be Δ in 1, 2, 4, and 8 hours.
- Level of aggregation P: The number of medians which can range from 1 to 235 (1, 10, 20, 50, 100, 235).
- Number of trips: Forecast the average number of trips originated from a median in $(t, t + \Delta]$.

3. RESULTS AND DISCUSSION

In this section, we conduct an evaluation and comparison of the ML models utilized in our study. Initially, we compare the overall performance of different models configured with their default architecture. The violin graph of Figure 4 compares four different ML methods based on MAPE. We see RF exhibits a lower average MAPE, which provide more accurate predictions compared to the other three methods. This could be due to its ability to capture complex relationships in the data, handle noisy features, and effectively generalize patterns. The lower maximum MAPE indicates that RF is more robust and less susceptible to extreme outliers or instances where predictions deviate significantly from the actual values. The combination of lower average MAPE and lower maximum MAPE suggests that RF has a strong generalization capability.

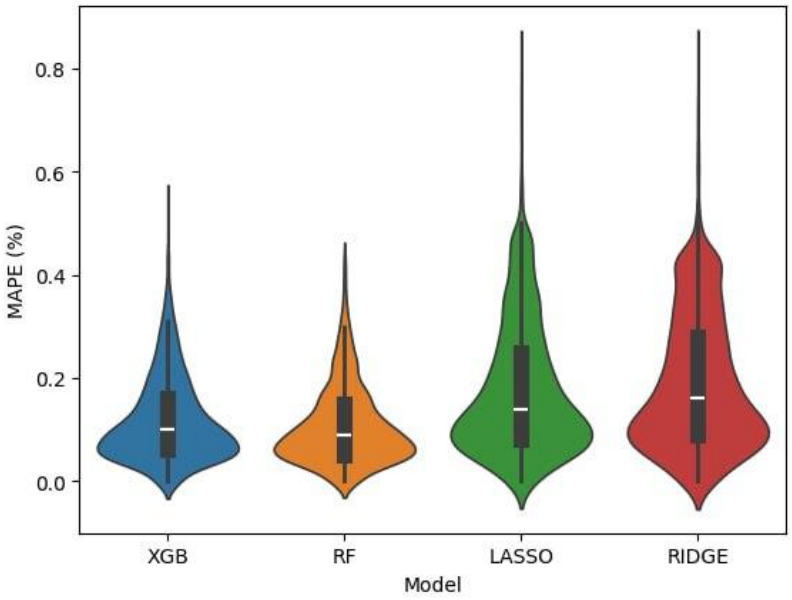

Figure 4: Comparing the ML methods

Figure 5 illustrates how the performance methods is influenced by the level of aggregation P. A notable trend is observed: as the number of medians increases, so does the error, which can be attributed to the inherent difficulty in predicting medians containing fewer stations. However, a compelling finding emerges as RF consistently outperforms the other methods in this context.
The connection between the level of aggregation and prediction error is evident, indicating that as the number of medians increases, prediction tasks become more challenging. This is logically expected because predicting medians that cover a smaller number of stations inherently poses greater difficulty. The challenge arises from the limited information available for modeling when dealing with fewer stations, making accurate predictions more demanding. So, there is a tradeoff between the number of P and error in prediction.
The connection between the aggregation level P and prediction error in BSS has practical implications for system optimization. Achieving a balance between granularity and minimizing errors at this level helps make cost-effective operational decisions. These findings offer valuable insights for BSS operators looking to optimize their networks efficiently.

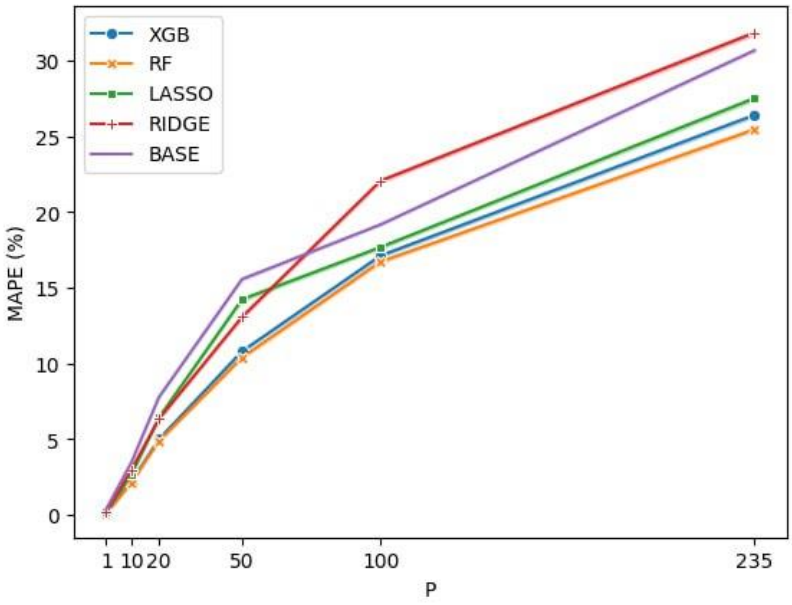

Figure 5: The effect of the level of aggregation P in the performance of ML methods

Concentrating on the RF model, Figure 6 illustrates variations in median versus MAPE across different delta values. Notably, the findings reveal a consistent trend wherein RF exhibits superior prediction accuracy when the delta, representing the future window size in hours, is lower. Specifically, the model demonstrates more accurate predictions when forecasting demand for the next hour compared to an 8-hour forecast. This observation holds true across various levels of P values, indicating that the model's predictive performance is influenced by the temporal horizon of the forecast.

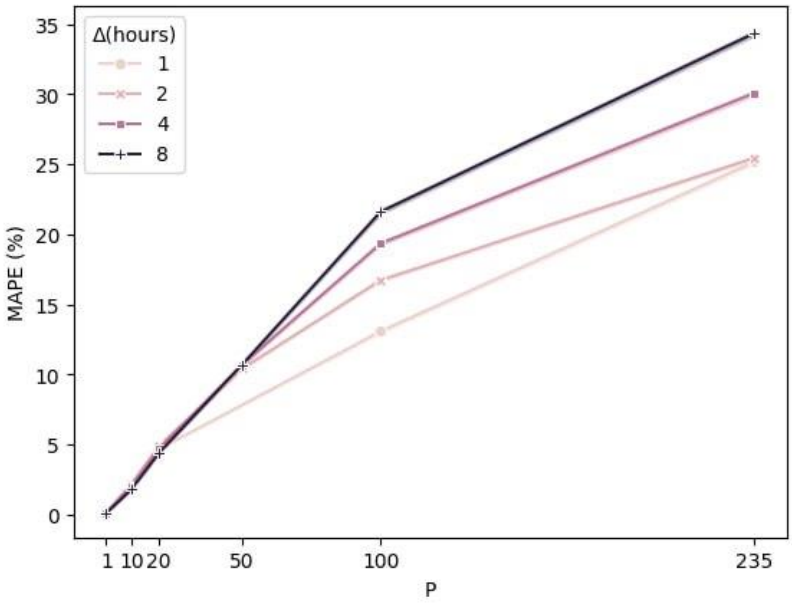

Figure 6: Variations in median versus MAPE across different delta values

Figure 7 compares the model's performance when incorporating weather information against its performance without such data, aiming to illuminate the impact on predicting demand. Notably,

the exclusion of weather-related information results in a diminished performance of the RF model. This decline is particularly pronounced when the aggregation level P is higher. The MAPE escalates to nearly 26% at the station level when weather data is disregarded, compared to 18% a markedly lower percentage when weather information is considered.

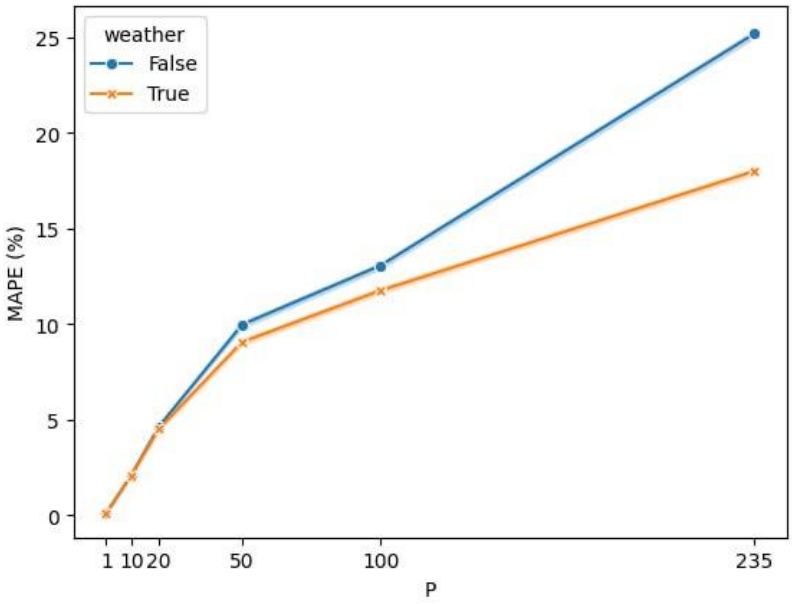

Figure 7: The model's performance by considering against not considering the weather information (one hour prediction)

Figure 8 depicted the feature importance analysis of the RF method for predicting one hour ahead at the city level. It explains that the demand from the last 2 hour and the preceding 1 hour exerts the most significant influence on the prediction. The third position in terms of importance is occupied by hourly sine seasonality. Additionally, the weather-related features, specifically temperature and humidity, exhibit nearly equal effects on the predictive model.

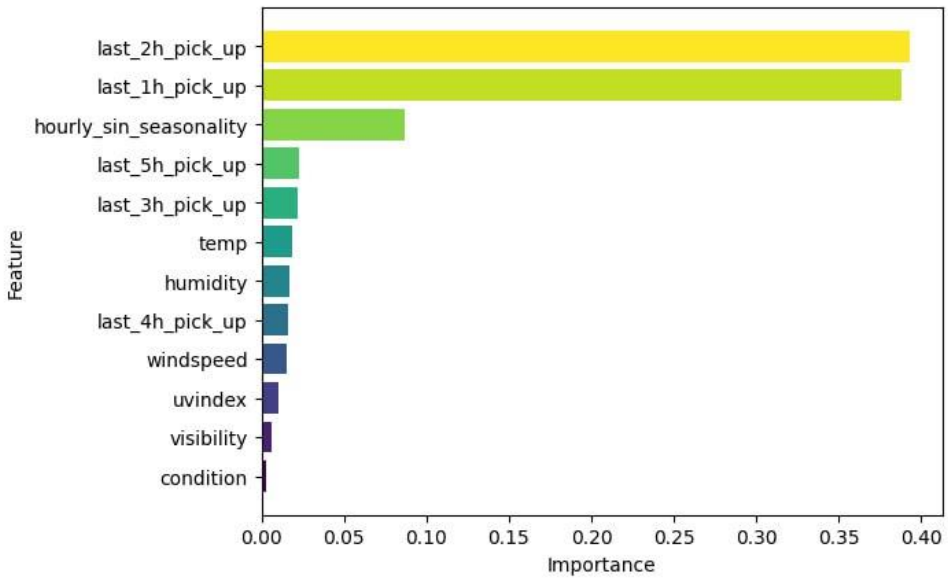

Figure 8: Feature importance analysis for RF method

4. CONCLUSIONS

In conclusion, our study addresses the challenges of demand prediction of BSS by integrating P-Median clustering and machine learning algorithms for short-term. Using Los Angeles as a case study, we applied regression methods such as RF, XGBoost, Ridge, and LASSO. Combining BSS and weather data with P-Median clustering yielded acceptable predictions at both station and city levels, with RF consistently outperforming other models. Emphasizing the importance of balancing aggregation (P) in P-Median clustering, we highlighted a tradeoff between medians and prediction error, contributing to improved user experience and cost-effective decisions. Our investigation also underscores the influence of the forecast window size on RF's predictive performance, emphasizing the need for tailored forecasting models. This research provides valuable insights for system operators, offering a promising approach for short-term demand prediction in BSS, bridging academic discourse with practical guidance. Future research directions may explore real-time data incorporation and further model refinements to address the dynamic nature of demand patterns in evolving urban mobility scenarios.

REFERENCES

Abdellaoui Alaoui, E. A., & Koumetio Tekouabou, S. C. 2021. Intelligent management of bike sharing in smart cities using machine learning and Internet of Things. *Sustainable Cities and Society*, *67*, 102702. https://doi.org/10.1016/j.scs.2020.102702

Cantelmo, G., Kucharski, R., & Antoniou, C. 2020. Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data. *Transportation Research Record*, *2674*(8), 132–144. https://doi.org/10.1177/0361198120932160

Giot, R., & Cherrier, R. 2014. Predicting bikeshare system usage up to one day ahead. *2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS)*, 22–29. https://doi.org/10.1109/CIVTS.2014.7009473

Harikrishnakumar, R., & Nannapaneni, S. 2023. Forecasting Bike Sharing Demand Using Quantum Bayesian Network. *Expert Systems with Applications*, *221*, 119749. https://doi.org/10.1016/j.eswa.2023.119749

Https://bikeshare.metro.net/about/data/. (n.d.).

Https://www.visualcrossing.com/. (n.d.).

Kou, Z., & Cai, H. 2021. Comparing the performance of different types of bike share systems. *Transportation Research Part D: Transport and Environment*, *94*, 102823. https://doi.org/10.1016/j.trd.2021.102823

Wang, W. 2016. *Forecasting Bike Rental Demand Using New York Citi Bike Data*. https://api.semanticscholar.org/CorpusID:167378491

Wu, C., & Kim, I. 2020. Analyzing the structural properties of bike-sharing networks: Evidence from the United States, Canada, and China. *Transportation Research Part A: Policy and Practice*, *140*, 52–71. https://doi.org/10.1016/j.tra.2020.07.018

Wu, X., Lyu, C., Wang, Z., & Liu, Z. 2019. Station-Level Hourly Bike Demand Prediction for Dynamic Repositioning in Bike Sharing Systems. *Smart Innovation, Systems and Technologies*. https://api.semanticscholar.org/CorpusID:196179464

Xu, M., Di, Y., Yang, H., Chen, X., & Zhu, Z. 2023. Multi-task supply-demand prediction and reliability analysis for docked bike-sharing systems via transformer-encoder-based neural processes. *Transportation Research Part C: Emerging Technologies*, *147*, 104015. https://doi.org/10.1016/j.trc.2023.104015