Machine Learning-Based Parking Occupancy Prediction Using OpenStreetMap Data

Shanqi Zhan *

Posted Date: 13 June 2025

doi: 10.20944/preprints202506.1149.v1

Keywords: Parking occupancy prediction; OpenStreetMap; machine learning; XGBoost; smart parking; urban mobility; spatial analysis; intelligent transportation system

 doi:10.20944/preprints202506.1149.v1

Article

Machine Learning-Based Parking Occupancy Prediction Using OpenStreetMap Data

Shanqi Zhan

Municipal Parking Services, Inc. (MPS), Golden Valley, USA; Zhans.scholar@gmail.com

Abstract: Accurate parking occupancy prediction is essential for reducing traffic congestion and optimizing urban mobility. Traditional monitoring methods are costly and difficult to scale, making machine learning a viable alternative. This study employs XGBoost (eXtreme Gradient Boosting) to predict parking occupancy using OpenStreetMap (OSM) data, with simulated occupancy rates based on proximity to the central business district (CBD). The model achieves a Mean Squared Error (MSE) of 0.0022 and an R^2value of 0.8922, demonstrating strong predictive accuracy. Results confirm the significance of spatial factors in parking demand. Future work will integrate real-time data and explore deep learning models to further enhance prediction accuracy.

Keywords: Parking occupancy prediction; OpenStreetMap; machine learning; XGBoost; smart parking; urban mobility; spatial analysis; intelligent transportation systems

I. Introduction

As urban populations continue to grow, cities face increasing challenges in managing transportation infrastructure, particularly parking availability. Limited parking spaces and inefficient allocation often lead to traffic congestion, increased fuel consumption, and environmental pollution. Accurate parking occupancy prediction is essential for optimizing urban mobility, improving traffic flow, and reducing the time spent searching for parking spaces. Traditional parking management approaches, such as manual monitoring and sensor-based detection, can be costly and difficult to scale. In contrast, machine learning models provide a data-driven approach to forecasting parking occupancy, enabling more efficient allocation of parking resources.

Machine learning (ML) techniques have emerged as effective tools for predicting parking occupancy, leveraging diverse data sources, including OpenStreetMap and real-time sensor data. Various models, such as Random Forest (RF), Adaptive Graph Convolutional Networks (AGCN) combined with Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM) networks, have been evaluated for their predictive capabilities. For instance, RF has shown superior performance in IoT-enabled environments, enhancing accuracy and reducing congestion [1]. The AGCRU model, which integrates spatial and temporal dynamics, achieved a mean absolute error (MAE) as low as 0.0156, indicating its effectiveness in urban settings [2]. Additionally, advanced methods like time series decomposition with LSTM and adaptive neuro-fuzzy inference systems (ANFIS) have demonstrated significant improvements in prediction accuracy, with ANFIS-LSTM achieving a 34.04% enhancement in mean squared error (MSE) over standalone LSTM models [3,4]. These findings underscore the potential of hybrid ML approaches in optimizing parking management systems and enhancing urban mobility strategies [5].

In this study, we develop a machine learning-based parking occupancy prediction model using OpenStreetMap (OSM) data. OSM is a publicly available, open-source geographic information system that provides detailed spatial data on road networks, parking facilities, and other urban infrastructure. Although OSM does not contain real-time occupancy data, we simulate parking demand by assigning higher occupancy rates to locations closer to the central business district (CBD), reflecting well-established urban mobility patterns.

 doi:10.20944/preprints202506.1149.v1

We employ XGBoost (eXtreme Gradient Boosting) as the predictive model due to its efficiency, scalability, and ability to handle structured data. Model performance is assessed using key metrics, including Mean Squared Error (MSE) and R- Squared (R^2), to determine its accuracy in forecasting parking demand. The key contributions of this research include: (1) Utilizing OSM data as a cost-effective and scalable alternative for parking prediction. (2) Implementing XGBoost for accurate parking occupancy forecasting. (3) Analyzing the influence of spatial factors, such as proximity to the CBD, on parking demand.(4) Evaluating model performance and identifying potential improvements for future research.

The remainder of this paper is organized as follows. Section II describes the dataset preparation process, including the extraction of parking-related features from OpenStreetMap (OSM) and the simulation of occupancy rates based on spatial proximity to the central business district (CBD). It also introduces the XGBoost model and outlines the hyperparameter tuning strategy. Section III presents the experimental results, including statistical analysis of the dataset, model evaluation metrics, and discussions on the significance of spatial features in parking occupancy prediction. And, limitations of the current research and future improvement plans are discussed to guide subsequent studies. Section IV concludes the study by summarizing the key findings, highlighting practical implications, and proposing directions for future work.

II. Methodology

A. Dataset

OpenStreetMap (OSM) is a free, open-source geospatial database that is collaboratively updated and maintained by a global community of contributors [6]. Data is collected through surveys, aerial imagery tracing, satellite images, and imports

from publicly available geospatial datasets. OSM operates under the Open Database License (ODbL) and is widely used for electronic mapping, navigation, humanitarian aid, and geospatial analysis.

For this study, we extracted OSM parking-related data for a selected urban area—San Francisco, California, USA. Since OSM does not provide real-time parking occupancy data, we assigned simulated occupancy rates based on spatial characteristics. Specifically, parking locations closer to the central business district (CBD) were assumed to have higher occupancy rates, reflecting real-world demand patterns. Additionally, distance from the CBD was incorporated as a feature to model demand trends.

The dataset consists of four primary attributes, as detailed in Table 1. To ensure data quality, missing values were handled using the dropna() function, effectively removing incomplete records. The dataset was then divided into training and testing subsets in an 80:20 ratio. The training set was used to develop and fine-tune the predictive model, while the test set served as an independent evaluation benchmark. This approach ensures that the model is assessed on unseen data, providing a realistic estimate of its generalization performance.

Table 1. Attributes in the dataset.

Feature Name	Data Type	Description
lon	float64	Longitude coordinate of the parking location
lat	float64	Latitude coordinate of the parking location

 doi:10.20944/preprints202506.1149.v1

distance_to_cbd	float64	Distance of the parking location from the city center (CBD), used to simulate demand trends
occupancy_rate	float64	Simulated parking occupancy rate (proportion of occupied spaces)

B. XGBoost

XGBoost (eXtreme Gradient Boosting) is an optimized machine learning algorithm based on gradient-boosted decision trees [7]. It is designed for efficiency, scalability, and high predictive accuracy, making it particularly well-suited for structured data analysis. As an ensemble method, XGBoost sequentially builds decision trees, where each new tree focuses on correcting the residual errors of the previous ones, thereby improving overall model performance.

XGBoost was selected for this study due to its robustness in handling tabular data and its integration of advanced optimization techniques. The algorithm minimizes a differentiable loss function iteratively, refining predictions at each step. To prevent overfitting, XGBoost employs both L1 (Lasso) and L2 (Ridge) regularization, making it effective for datasets with complex, high-dimensional feature spaces. Additionally, it can handle missing data natively by learning optimal split directions, ensuring that incomplete records do

not compromise model performance. Another key advantage of XGBoost is its ability to efficiently utilize computational resources by performing parallelized training, significantly reducing training time while maintaining scalability. Given these characteristics, XGBoost is well-suited for predicting parking occupancy using OpenStreetMap data, which consists of numerical and spatial features.

C. Model Hyperparameters

To enhance the predictive performance of the XGBoost model, key hyperparameters were fine-tuned using Randomized Search [8,9]. Hyperparameter tuning was conducted using RandomizedSearchCV [10] with 5-fold cross- validation, ensuring that the model was trained and validated on multiple data splits to improve generalization. The following hyperparameters were optimized:

(1) n_estimators: Defines the number of decision trees in the ensemble. An optimal value of 250 was chosen, balancing model complexity and computational efficiency. While increasing the number of trees generally improves performance, excessive values may lead to diminishing returns and overfitting.

(2) max_depth: Controls the maximum depth of each tree. A depth of 7 was selected to allow sufficient model complexity while mitigating overfitting.

(3) learning_rate: Determines the step size for updating model weights. A value of 0.1 was used, ensuring incremental learning and preventing drastic model adjustments that could lead to instability.

A total of 10 hyperparameter configurations were evaluated across five cross-validation folds, resulting in 50 training iterations. The final optimized model configuration— n_estimators = 250, max_depth = 7, and learning_rate = 0.1— demonstrated strong predictive performance while maintaining generalizability to unseen data. Hyperparameter tuning was crucial in refining the model's ability to balance complexity and avoid overfitting, ensuring its applicability in real-world parking occupancy prediction.

 doi:10.20944/preprints202506.1149.v1

III. Experimental Results

A. Statistics of the Dataset

To gain a deeper understanding of the dataset and its underlying trends, three visualizations were generated to illustrate key statistical and spatial characteristics of parking occupancy. Figure 1 presents the spatial distribution of parking occupancy, highlighting how demand varies geographically across different locations. The visualization indicates that parking occupancy tends to be higher in areas closer to the central business district (CBD), aligning with the expectation that commercial hubs experience greater demand for parking. Conversely, locations farther from the city center generally exhibit lower occupancy rates, reflecting reduced parking pressure in less congested areas.

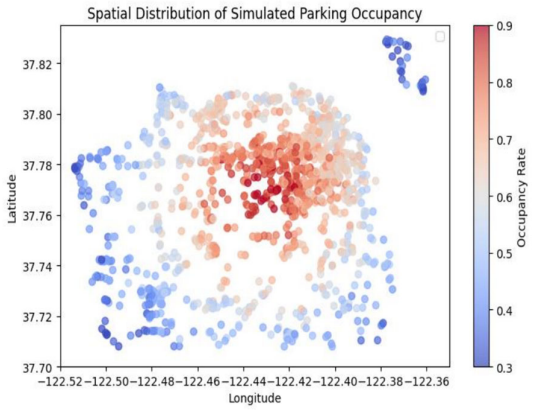

Figure 1. Spatial distribution of parking occupancy.

Figure 2 displays the histogram of occupancy rates, which shows the frequency distribution of parking utilization across all sampled locations. This histogram provides insights into the overall pattern of parking space usage, identifying commonly observed occupancy levels and potential outliers. The distribution helps in understanding whether parking occupancy follows a uniform, normal, or skewed trend, which is essential for developing an accurate predictive model.

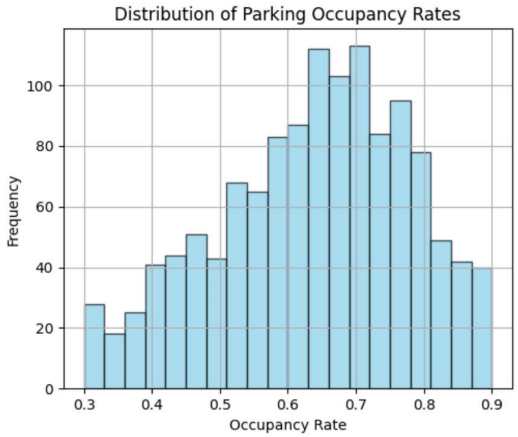

Figure 2. Spatial distribution of parking occupancy.

Figure 3 illustrates the scatter plot of distance to the CBD versus occupancy rate, which examines the relationship between proximity to the city center and parking demand. The plot reveals whether there is a strong inverse correlation, indicating that parking occupancy tends to

 doi:10.20944/preprints202506.1149.v1

decrease as distance from the CBD increases. Identifying this trend is crucial for validating assumptions about urban parking behavior and incorporating spatial dependencies into the predictive model.

These visualizations collectively provide a comprehensive overview of the dataset, offering valuable insights into spatial patterns and occupancy trends that inform the subsequent modeling process.

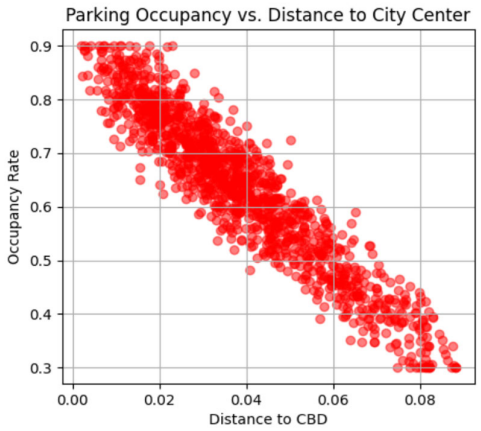

Figure 3. Scatter plot of distance to the CBD versus occupancy rate.

B. Results of the Machine Learning Model

The predictive performance of the XGBoost model was assessed using key evaluation metrics to quantify its accuracy in forecasting parking occupancy [11,12]. The Mean Squared Error (MSE) for the model was 0.0022, indicating a low average squared difference between actual and predicted occupancy rates. A smaller MSE value suggests that the model is capable of making precise predictions with minimal deviation from observed values. The R-Squared (R^2) value was calculated as 0.8922, signifying that approximately 89.22% of the variance in parking occupancy is explained by the model. This high R^2 value suggests a strong fit between the model's predictions and actual occupancy trends, demonstrating its effectiveness in capturing the underlying patterns in the dataset, as shown in Figure 4. These results confirm that the XGBoost model is well-suited for parking occupancy prediction, effectively leveraging spatial and contextual features to generate accurate forecasts.

To further evaluate model performance, a scatter plot of actual versus predicted occupancy rates was generated. The plot reveals that most points are closely clustered around the 45-degree line, indicating a high level of agreement between the observed and predicted values. Only a small number of predictions exhibit noticeable deviations, suggesting that the model maintains robust performance across the majority of the dataset.

Additionally, feature importance analysis was conducted to better understand the model's decision-making process. Among all input features, distance to the central business district (CBD) was identified as the most influential predictor of parking occupancy, followed by longitude and latitude. This finding reinforces the hypothesis that proximity to urban centers plays a dominant role in determining parking demand, aligning with real-world urban mobility patterns.

The residual analysis also indicated no significant systematic bias across the prediction range. Residuals were randomly distributed around zero, implying that the model did not consistently overestimate or underestimate occupancy rates across different locations.

Overall, the XGBoost model demonstrates excellent predictive capability with strong generalization to unseen data. Its ability to automatically handle missing values, incorporate regularization techniques, and model complex non-linear relationships contributed significantly to

 doi:10.20944/preprints202506.1149.v1

its high accuracy and stability. These findings validate the feasibility of using machine learning approaches combined with open-source spatial data for scalable, cost-effective parking occupancy prediction in urban environments.

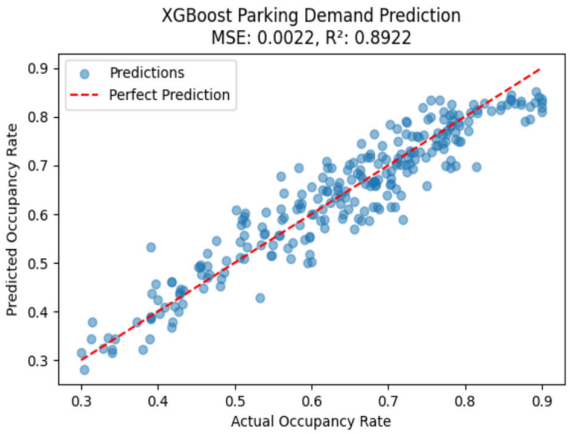

Figure 4. Parking demand predictions.

C. Discussion

The results demonstrate that the XGBoost model performs well in predicting parking occupancy, as evidenced by a low Mean Squared Error (MSE) of 0.0022 and a high R^2value of 0.8922. These metrics indicate that the model effectively captures patterns in the dataset and generalizes well to unseen data. The high R^2suggests that key features, such as proximity to the central business district (CBD) and spatial distribution, contribute significantly to parking occupancy trends.

One of the primary advantages of using XGBoost is its ability to handle complex, structured data while preventing overfitting through built-in regularization techniques. The strong predictive performance can be attributed to the model's ability to capture non-linear relationships between input features and occupancy rates. Additionally, the feature importance analysis reveals that distance to the CBD plays a crucial role in predicting occupancy, confirming urban planning insights that demand tends to be higher in central areas.

However, despite the strong performance, some limitations should be considered. First, the dataset relies on simulated occupancy rates rather than real-time sensor or historical data, which may introduce biases or inaccuracies. Additionally, external factors such as weather conditions, time-of-day variations, and special events were not included in the model, which could enhance predictive accuracy if incorporated in future iterations.

To further improve performance, future work could explore deep learning approaches, such as recurrent neural networks (RNNs) or transformer-based models, to capture temporal

dependencies in parking trends. Moreover, integrating real- time traffic data and incorporating additional socioeconomic indicators may enhance the robustness of the predictions. Overall, the results highlight the feasibility of machine learning-based parking occupancy prediction and provide a foundation for further enhancements in urban mobility optimization.

D. Limitations and Future Plan

While the current study demonstrates strong predictive performance using XGBoost and simulated spatial data, several limitations should be acknowledged. First, the occupancy rates used for model training were synthetically generated based on the proximity to the central business district (CBD) rather than real-time or historical parking usage data. Although this approach aligns with established urban mobility patterns, it may not capture the full complexity of real-world

 doi:10.20944/preprints202506.1149.v1

parking dynamics influenced by factors such as weather, special events, pricing policies, or seasonal variations.

Second, the feature set in this study was limited to spatial attributes derived from OpenStreetMap (OSM) data, primarily longitude, latitude, and distance to CBD. Additional contextual features, such as traffic volume, public transportation accessibility, land use patterns, socioeconomic indicators, and temporal factors (e.g., time of day, day of the week), were not incorporated. Including these factors could significantly improve model accuracy and generalizability.

Third, while XGBoost proved effective for tabular spatial data, the model does not explicitly capture temporal dependencies or sequential parking behavior. Future research could explore the use of time-series modeling techniques, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, or Transformer-based architectures, to incorporate temporal patterns in parking occupancy.

Moving forward, future work will focus on several enhancements. First, real-world parking occupancy datasets, such as sensor-based parking data or transaction records from municipal parking services, will be integrated to validate and further improve the predictive model. Second, a broader range of features encompassing environmental, social, and economic factors will be incorporated to enrich the input space. Third, more advanced machine learning models, including hybrid models that combine spatial and temporal learning, will be evaluated to achieve even higher prediction accuracy. Finally, the deployment of the developed models into practical smart parking management systems will be explored, aiming to provide actionable insights for city planners, transportation authorities, and mobility service providers.

IV. Conclusions

This study explored the use of machine learning, specifically the XGBoost model, to predict parking occupancy using OpenStreetMap (OSM) data. The results demonstrate that the model effectively captures spatial and contextual factors influencing parking demand, achieving a high R^2value of 0.8922 and a low Mean Squared Error (MSE) of 0.0022. These metrics indicate strong predictive performance, confirming the feasibility of using machine learning for parking

occupancy estimation. The study highlights the significance of distance to the central business district (CBD) as a key predictor of parking occupancy, reinforcing established urban mobility trends.

References

1. Dahiya, Anchal, Pooja Mittal, Yogesh Kumar Sharma, Umesh Kumar Lilhore, Sarita Simaiya, Ehab Ghith, and Mehdi Tlija. "Machine Learning-Based Prediction of Parking Space Availability in IoT-Enabled Smart Parking Management Systems." *Journal of Advanced Transportation* 2024, no. 1 (2024): 8474973.
2. Zhao, Xiaohang, and Mingyuan Zhang. "Enhancing predictive models for on-street parking occupancy: Integrating adaptive GCN and GRU with household categories and POI factors." *Mathematics* 12, no. 18 (2024): 2823.
3. Ye, Wei, Haoxuan Kuang, Jun Li, Xinjun Lai, and Haohao Qu. "A parking occupancy prediction method incorporating time series decomposition and temporal pattern attention mechanism." *IET Intelligent Transport Systems* 18, no. 1 (2024): 58-71.
4. Elomiya, Akram, Jiří Křupka, Stefan Jovčić, and Vladimir Simic. "Enhanced prediction of parking occupancy through fusion of adaptive neuro-fuzzy inference system and deep learning models." *Engineering Applications of Artificial Intelligence* 129 (2024): 107670.
5. Lyu, Mengqi, Yanjie Ji, Chenchen Kuai, and Shuichao Zhang. "Short- term prediction of on-street parking occupancy using multivariate variable based on deep learning." *Journal of Traffic and Transportation Engineering (English Edition)* 11, no. 1 (2024): 28-40.

 doi:10.20944/preprints202506.1149.v1

6. Mooney, Peter, and Marco Minghini. "A review of OpenStreetMap data." *Mapping and the citizen sensor* (2017): 37-59.
7. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016.
8. Bergstra, James, and Yoshua Bengio. "Random search for hyper- parameter optimization." *The journal of machine learning research* 13, no. 1 (2012): 281-305..
9. Probst, Philipp, Marvin N. Wright, and Anne-Laure Boulesteix. "Hyperparameters and tuning strategies for random forest." *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9, no. 3 (2019): e1301.
10. Vishnu, M. K., VR Vishal Rupak, S. Vedhapriyaa, M. Sangeetha, R. Manjuladevi, and C. Sagana. "Recurrent gastric cancer prediction using randomized search cv optimizer." In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-5. IEEE, 2023.
11. Tatachar, Abhishek V. "Comparative assessment of regression models based on model evaluation metrics." *International Research Journal of Engineering and Technology (IRJET)* 8, no. 09 (2021): 2395-0056.