

Parking Availability Prediction and Recommendation Using Random Forest

Juveriya Khan¹ and Dr. Kiran D. Tikar²

¹Department of Master of Computer Applications

²Assistant Professor

Sadhu Vaswani Institute of Management Studies for Girls, Pune, Maharashtra, India

¹juveriyakhan212@gmail.com and ²kirantikar@svims-pune.edu.in

Abstract: *Finding the right parking facility can be challenging as there may be a chance that the parking lot may get full till the time driver reaches the destination, as the parking facility highly depends on time, day and place. This study introduces a system which uses machine learning to predict future parking occupancies and recommend the most suitable parking based on historical data. For this study, Aachen parking data from 2021 was used, which consisted of 447,074 observations and 12 parking facilities. As a result of the analysis, Linear Regression, Decision Tree, Random forest Regressor and Gradient Boosting were tested, and the best performer was the Random Forest Regressor. It had MAE of 11.35, RMSE of 19.21 and R^2 of 0.9870. Predicted occupancies were than used in recommendation system which considered facilities, within a 2 km radius. Facilities were prioritized based on their predicted occupancy and distances. If the preferred facility is expected to be full or highly occupied than the model will recommend another nearby facility. This study concludes that the use of occupancy prediction and recommendation system can help users in finding the best parking facilities*

Keywords: Parking Availability Prediction, Parking Occupancy, Random Forest Regression, Machine Learning, parking Recommendation

I. INTRODUCTION

Nowadays finding a suitable parking facility in a busy area might be challenging as there is a chance that the parking spot that was previously available might be full by the time driver reaches the destination. Hence, predicting the future availability will be more helpful for the parking decision rather than the current availability.

Machine learning can be used to predict the future parking occupancy based on the past occupancy. While previous studies have investigated various approaches to forecasting the parking occupancy and availability, relatively little focus has been made on applying these forecast to choose an appropriate parking location.

This research introduces a clear and understandable framework that merges forecast of future parking availability with parking recommendations. A Random Forest Regressor predicts the percentage of future occupancy based on the historical percentage of occupancy, time of day, the past availability, previous capacity, and the parking location. The predicted occupancy is then converted to the number of available spots at the given time, and the parking options are evaluated at the user's planned arrival time. For example, if parking A is likely to become full till the time driver arrives the system will suggest parking B which have more vacant occupancy.

This study make the use of Aachen parking dataset covering the year 2021. The prediction model is evaluated using the MAE, RMSE and R^2 . The recommendation component takes into account the geographic distance of the target location and parking facilities within 2km. The proposed framework shows how the historical parking data can be used to predict the future parking occupancy and recommend the suitable parking locations without any requirement of any special equipment such as sensors, GPS, or IoT devices.



II. LITERATURE REVIEW

Several research works have been conducted to address the problem of parking occupancy and availability prediction using machine learning approaches. Bagosher et al. [1] compared the performance of Linear Regression, SVR, Random Forest, and LSTM algorithms for parking availability prediction task using the data of a campus. As a result Random Forest performed better than other algorithms, however, the data was collected within a short period of 3 days and the evaluation did not consider multiple parking facilities. Kaltenbrunner et al. [2] proposed an explainable prediction model based on the data collected from 8 different locations in Barcelona. The model achieved low prediction errors. However, the study concentrated on prediction of parking occupancies rather than the recommendation of facilities and future parking availability. Vakrinou et al. [3] conducted a comparative study on the performance of the XGBoost, LSTM with Attention mechanism, and Transformer model, by using parking, traffic, metro and weather data. Although the algorithms yielded an excellent prediction performance, the use of various external data made the system complicated, and the study did not consider the provision of parking recommendations. Bassetti et al. [4] used the machine learning models, including Random Forest for the purpose of classifying the availability of parking spaces based on the simulated data of parking search. The result indicated that Random Forest achieved 97% accuracy for one of the cases. However, the used synthetic data and did not proposed alternative parking facilities.

Huang et al. [5] have suggested an innovative deep learning system, which uses a significant amount of multi-sourced data from Chengdu. Although the accuracy of their prediction is impressive, the approach is complicated and depends on various factors such as spatio-temporal modeling, which reduces its overall practicality.

Overall, the research presented shows that machine learning can be used for predicting parking occupancy and availability. However, there are still some opportunities to combine such predictions with the actual recommendations. Hence, the current study extends the reviewed articles by finding a simple and explainable solution that considers both the prediction availability and distance to the parking.

Research Gap

The majority of previous research has focused on the utilization on machine learning, statistical, and deep learning approaches for parking prediction. Bagosher et al. [1] and Bassetti et al. [4] demonstrated that the use of Random Forest can be effective. Kaltenbrunner et al. [2] presented a simple and explainable occupancy model. Vakrinou et al. [3] and Huang et al. [5] attained promising forecasting results by applying mobility, traffic, weather, and spatio-temporal features. However, the majority of them utilize external, sensor-based, synthetic, or complex data, and they mainly focus on the prediction task, rather than recommending a suitable parking facility based on the predicted availability.

This study contributes to parking prediction by developing a Random Forest-based approach for approximately 30 minutes ahead parking occupancy prediction and distance-based parking recommendation using the historical dataset of 12 parking facilities in Aachen. Using the predicted parking occupancy, the parking availability and status are predicted, while the distance is used to indicate closest parking facilities for recommendation purpose. In other words, this paper expands the parking prediction into a simple, interpretable, and applicable parking recommendation framework using historical parking data

III. METHODOLOGY

Dataset Description

This study follows quantitative and experimental research approach. For this research, Aachen Parking Dataset from parking-data repository was used. The dataset consisted of historical parking data for Aachen, Germany, for 2021. Historical data for the year 2021 was collected for 14 Aachen facilities, which were listed in metadata. After filtering 12 valid facilities were selected for further analysis. The dataset had facility ID, name, capacity, address, latitude and longitude. The dataset recorded available/free spaces rather than occupied ones. Therefore, the occupancy was



calculated based on the capacity of each facility. The latitude and longitude of each parking facility was kept for further parking recommendation according to distance task.

Data Preprocessing

The historical files were put together into a single time-series dataset, which contained approximately 82,355 timestamp based records from all 12 facilities. The timestamps were converted from UTC to the Europe/Berlin timezone, and duplicate values were also removed. As the parking observations were not recorded regularly, a forward fill was performed on the resulting time series for each facility to fill any gaps, unless the gap between the previous valid data point and next the valid data point was longer than 1 hour. The availability of parking was checked against the capacity of the facility. The values that did not meet the condition of $\text{Availability} < 0$ or $\text{Availability} > \text{Capacity}$ were marked invalid and replaced with missing. Invalid values were found for Parkplatz Luisenhospital.

Feature Engineering and Target Variable

The source dataset provides available parking spaces, occupancy was derived from:

$\text{Occupancy} = \text{Capacity} - \text{Availability}$

For Example, if the capacity is 605 and availability is 542: $\text{Occupancy} = 605 - 542 = 63$

This process resulted in 988,260 observations at facility level, with 5488,810 valid occupancy observations.

To compare the utilization of parking facilities with different capacities, the occupancy was expressed as a percentage:

$$\text{Occupancy Percentage} = \frac{\text{Occupancy}}{\text{Capacity}} \times 100$$

For simplicity, the occupancy percentage was divided into five groups:

Occupancy	Statue
<40%	Low
40% - < 75%	Average
75% - < 90%	High
90% - <100%	Nearly Full
100%	Full

Several time related features were then extracted from the local Aachen timestamp. They include Hour, Day, Day of Week, and Weekend. The Day of Week was encoded, the day as a number from 0 for Monday to 6 for Sunday. The Weekend feature was set to 1 for Saturday and Sunday and 0 for other days. Two lag-based features, previous availability and previous occupancy was also created using the previous valid observation of each facility

The prediction target was set as the Future Occupancy max 30 minutes ahead. The target time was defined by the following formula:

$$\text{Target Time} = \text{Current Time} + 30 \text{ minutes}$$

Only those observations that met the deadline of the next record within 8 minutes were considered. The records that did not have a valid occupancy value were removed, leaving about 447074 observations for the final modelling set.

Final Model Features and Train-Test Split

The final model used eight features Hour, Day, Day of Week, Weekend, Facility, Capacity, Previous Occupancy, and Previous Availability. The target variable was Future Occupancy. Since parking occupancy follows a time-series pattern the data was sorted by date and time. Then it was split into training and testing sets using 80/20 ratio. This meant 357,659 observations were used for training and 89,415 were used for testing. The training set covered the beginning of 2021. The test set covered October to December of 2021 letting use see how the parking occupancy model performs on data.



Encoding

The categorical facility variable was converted into numerical using one-hot encoding. The model input therefore included 19 features, including the numerical variables and 12 facility-indicator variables.

Machine Learning Model

There were 4 different regression models tested using the same set of feature and chronological train-test split, which included Random Forest, Gradient Boosting, Decision Tree, and Linear Regression methods.

The models were evaluated using three standard regression metrics:

Model Evaluation

Mean Absolute Error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}|$$

Root Mean Squared Error

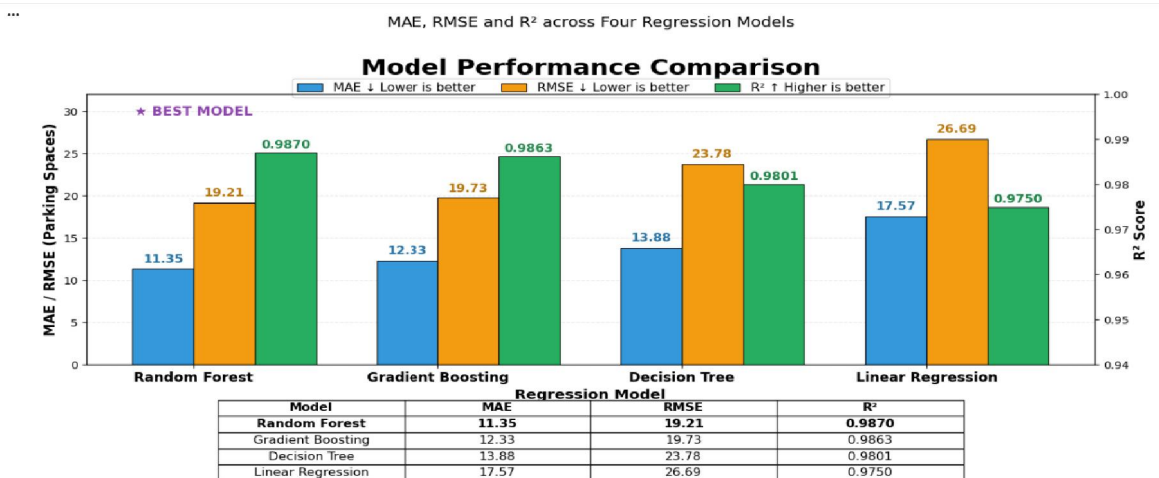
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2}$$

Coefficient of Determination

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Model Comparison and Model Selection

The results were as follow:

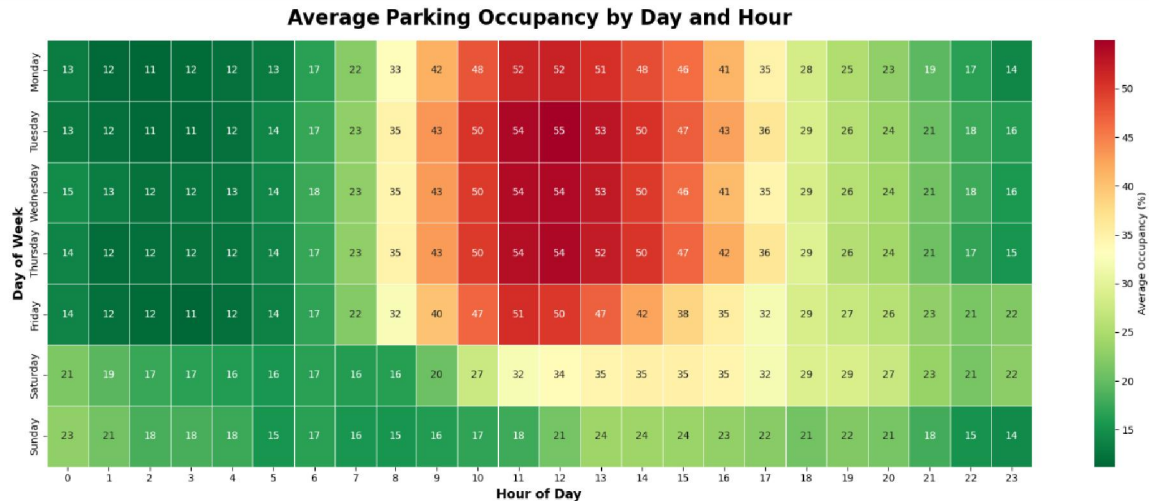


As the random forest had the lowest MAE and RMSE and the highest R² among the tested algorithms. As a result, it was selected as main algorithm for proposed parking recommendation system.



Historical Parking Pattern Analysis

The historical parking data was analyzed by each day, hour, and facility to determine the parking patterns.



Tuesday: highest average occupancy – 37.34%, Sunday: lowest – 20.54%, 12:00 PM: highest overall occupancy – 46.61%, 11:00 AM – 1:00 PM: busiest period.

Geographic Distance – Based Recommendation

The recommendation module calculates the distance for each parking facility using latitude and longitude to determine its distance from the user's destination using Haversine formula:

The location information is not incorporated into prediction model. Only facilities within a specified distance (eg., 2km) are recommended to the user.

$$d = 2R \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta\phi}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\Delta\lambda}{2} \right)} \right)$$

Proposed Parking Recommendation System

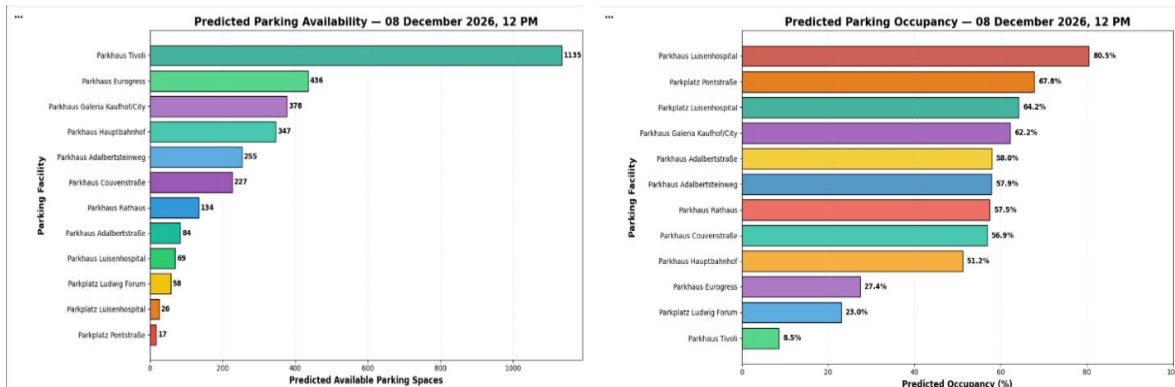
The system takes the user's destination, date, and arrival time as input. It will filter out the farthest parking facilities based on the distance and apply the random forest model to predict the occupancy of the filtered parking facilities 30 minutes after the user's arrival. Predicted availability is calculated as:

Predicted Availability = Capacity – Predicted Occupancy

Predicted Occupancy % = (Predicted Occupancy / Capacity) x 100

Predicted Availability % = (Predicted Availability / Capacity) x 100





Based on the availability percentage, the parkinging facilities with the nearest distance is assigned. The recommendation will be made based on 60% availability and 40% distance. In case of high predicted occupancy at preferred parking facility, an alternative parking facility will be recommended.

Major Findings

The present research created a machine learning framework for predicting future parking occupancy. The results were obtained using historical data of 12 facilities in Aachen, Germany. There were 82,355 timestamps refined into 447,074 final observations. The Random Forest Regression resulted in the best performing algorithm for the chronological test set. It had an MAE of 11.35, RMSE of 19.21, and an R^2 of 0.9870. The previous occupancy was proved to be the most significant feature comprising 96.87% of the total importance. It was followed by Hour, which comprised 2.09% of total importance. According to the historical data analysis, Tuesday was the busiest day, with an average of 37.34% occupancy, while Sunday was the calmest, with an average of 20.54% occupancy. The busiest time per Hour was at 12 PM, with an average of 46.61% occupancy, while the number of almost full observations peaked at 11 AM. The parking facilities also differed based on their average occupancy, busiest day, and busiest hour. Finally, the suggested framework for recommending parking lots used the calculated future occupancy and the calculated distances. The latter was determined using the Haversine formula to calculate the distance between the user and the parking lot.

Limitations

The study has certain limitations. The study uses historical data, which might not present the patterns specific for the other years, in particular for Aachen in 2021. Further, the data employed in the current research had missing and irregular entries that required cleaning and exclusion of several observations. The model was developed to predict the occupancy rate 30 minutes ahead instead of predicting the exact time of 30 minute intervals for each entry. The final algorithm does not rely on any particular sensors, cameras, GPS, or traffic feed, the current analysis did not consider these records. Only 12 parking facilities in Aachen were found to have sufficient historical data for the regression task. The current algorithm considers distance but not real time traffic conditions. In the future, it will be possible to use real-time parking and traffic data to improve the accuracy and relevance of predictions and recommendations.

IV. CONCLUSION

The present research provides a machine-learning based framework for predicting future parking occupancies and recommending suitable parking facilities, based on historical findings. Four regression models were trained and tested on Aachen's data set, which illustrated that the random forest regression achieved the best result, with an MAE of 11.35, RMSE of 19.21, and R^2 of 0.9870 on the chronological test set. The results proved that current parking occupancy is the signal for short-term prediction while different days of week and time intervals show their own patterns. By introducing the estimated occupancy with the distance to the user, the recommendation system will be



capable of suggesting available parking facilities nearby based on the predicted arrival time. In general, the research has proved the possibility of creating an explainable parking prediction and recommendation system based on historical data, utilizing machine learning methods.

REFERENCES

1. Madyan Bagosher, Tala Mustafa, Mohammad Alsmirat, Amal Al-Ali and I. M. Al Jawarneh (2025) : A Cost-Effective Framework for Predicting Parking Availability Using Geospatial Data and Machine Learning
2. Andreas Kaltenbrunner, Josep Ferrer, David Moreno and Vicenc Gomez (2025) : A Simple and Explainable Model for Park-and-Ride Car Park Occupancy Prediction
3. Katerina Vakrinou, Eleni G. Mantouka and Eleni I. Vlahogianni (2025) : Leveraging multisource data for off-street parking occupancy prediction: a twofold analysis
4. Enrico Bassetti, Andrea Berti, Alba Bisante, Andrea Magnante and Emanuele Panizzi (2022): Exploiting User Behavior to Predict Parking Availability through Machine Learning
5. Yin Huang, Yongqi Dong, Youhua Tang and Li Li (2025) : Parking Availability Prediction via Fusing Multi-Source Data with A Self-Supervised Learning Enhanced Spatio-Temporal Inverted Transformer

