Car Parking Prediction

1. Problem Statement

Traffic volume in urban areas is continuously growing due to massive urbanization. This makes urban life more congested and polluted, and consequently, increasing energy consumption, airborne diseases, and global. According to NASA analysis, 2020 has been recorded as the hottest year.

In this scenario, ensuring efficient traffic movement in smart cities is a major goal, minimizing the negative impact on the environment and public health. Nowadays, a common problem of huge cities is based on searching for a free parking space, which generates an increase in traffic congestion, fuel consumption, and greenhouse gas pollution in urban areas.

A potential solution to mitigating parking searches consists of the accurate prediction of parking availability. Thus, reducing drivers waiting time and traffic congestion while looking for a free parking slot.

2. Scope of the work

2.1. Objective

In this work a predictive model of parking availability is developed to improve the vehicle parking experience in the city of Freiburg in Baden-Württemberg. Germany. In the posterior stage, the model will be integrated into a new app, which will inform the customer of the slot availability at arrival time.

2.2. Specific Objectives

In this project the following tasks are performed:

- Data analysis
- Parking garages comparison
- External data sources incorporation: Public holidays and weather forecasts
- Data processing and Feature engineering
- Building a predictive model of parking availability.

2.3. Study Assumptions

The following assumption were assumed in the development of the predictive model:

- The parking capacity of both garages remains constant in the all-time data range.
- Specific social events on historic data are negligible.

2.4. Questions to be addressed

1. Which of the two parking garages has a higher utilization rate?

The City center station garage has a higher utilization rate than the Train Station garage. This could be attributed to the location and the train transportation service. The utilization rate was calculated by dividing the total number of cars observed over the

study period by the total capacity of the parking garage (both garages have the same number of observations). The occupation ratio was defined as follows:

$$Occupation\ Ratio\ =\ \frac{City\ Center\ Garage\ Occupancy}{Train\ Station\ Garage\ Occupancy} \quad (1)$$

The occupation ratio was higher than one in all train station garage occupancy range, indicating that city center has a higher ratio occupation. However, this response is biased, because the ratio was calculated by removing the occupancy of the train station with a higher value than the total capacity assumed (272). The total instances with this condition in raw data were 252, representing 0.0459% of the total population. Thus, in this case, the uncertainty of the results is negligible.

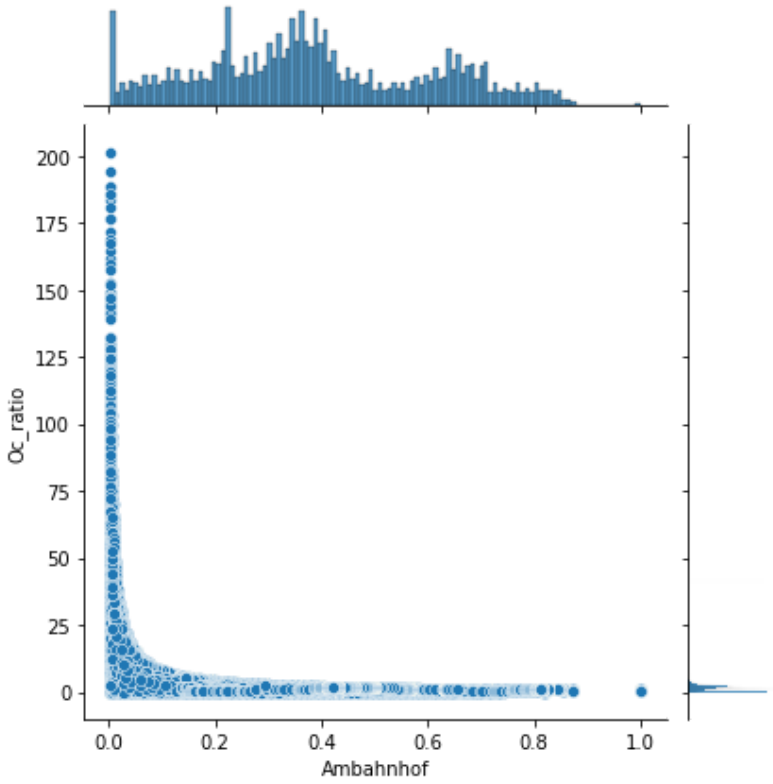

Fig 1. Occupation Ratio vs train station garage occupancy.

2. Will the occupancy of the parking at the train station be less than 100

slots on February 1st, 2022, and on April 15th, 2022, between 12:00 and

13:00?

The prediction for the indicated dates was carried out using Model B (see Supplementary section). The predictions were performed with a certain interval with of 90%. The parking occupancy predicted for 2022/02/01 and 2022/04/15 between 12:00 and 13:00 was 64 ± 15 and 141 ± 32, respectively. Thus, the occupancy of the parking at the train station be less than 100 for 2022/02/01 and higher for 2022/04/15 between the indicated timeslot. This is in concordance with the observation that the occupancy of the train station garage is higher in April than in February (especially the 0.75 quartiles), as can be observed in the corresponding boxplot (Supplementary Information Section)

3. Please quantify your certainty about the answers to questions (1) and (2).

The certainty is expressed in questions (1) and (2).

4. Based on the data which factors increase the usage of parking garages?

According to the data considered in this study, the features with an appreciable correlation coefficient are the following (more information of correlation matrix coefficient can be found in Supplementary Section):

a. **Train station garage:** Year (0.57), Weekend (0.33), Weekday (0.27), and Hour (0.11).

b. **City Center garage:** Hour (-0.15), Tmax (-0.14), Tmin(-0.14), Tavg(-0.13), Month(-0.1) and Year (0.1).

Train station garage was no sensitive to weather forecast, showing strong variation with time variables, especially with year. This point out that the of train station garage will be increased over time, particularly during weekend at rush hours. From an overview, the seasonal behavior of train station garage occupancy will be more pronounced over time.

In contrast, city center garage occupation is mainly affected by hour and temperature. It would be expected that that occupation rate growth in the following years with a lower rate than train station garage. The coefficient correlation between city center garage occupancy and different temperatures was similar, due to these variables are strongly correlated with each other, showing a negative effect in the occupancy. On the other hand, the daily timeslot with lower occupancy is between 8-16 h, which corresponds with working hours. This could indicate a repetitive customer cycle.

5. Suggest requirements for future data collection

The basic requirements for future data collection consists of IoT gadget to register the entry and the exit of a new car into the garage. Determining the initial amount of car in the parking, the occupancy over time can be obtained from the difference between the flow-in/flow out of cars. Iot sensors could be located in gateway (entry and exit) and a sensor in each parking slot, providing detailed information of occupied slot. This could be useful for reserved parking spaces. Raw data can be storage in a SQL database on cloud (e.g. S3). An alternative could be integrating a computer vision model to cameras located in each parking garage, quantifying the live occupation rate of each parking garage.

In this work, holiday and historic data was analyzed. However, other data type can be considered, for instance pedestrians or traffic conditions. This information can be obtained from API services and be analyzed only in those cases where it is a relevant parameter for the model prediction on selected parking garage.

6. Suggest high-level implementation pipelines of the parking app

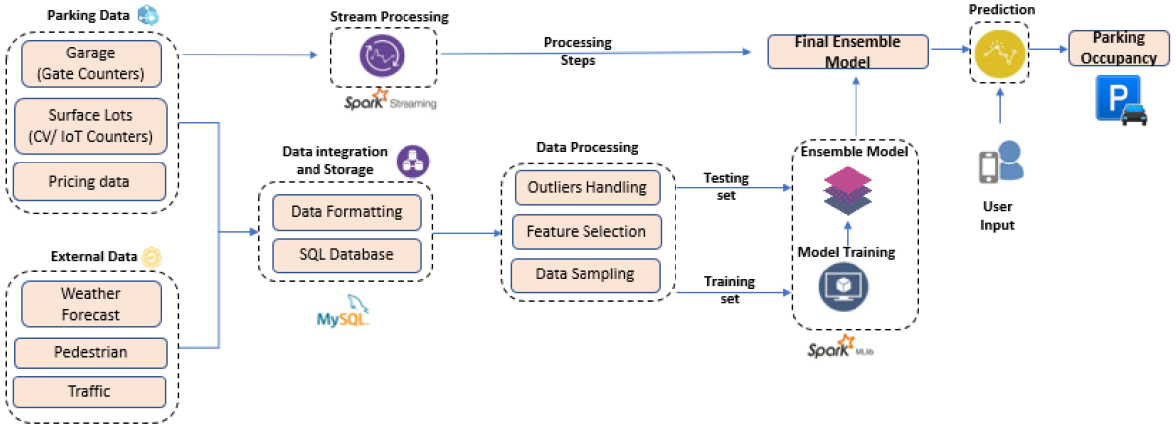

Fig 2. The pipeline of the parking app.

Supplementary information

Exploratory Data Analysis (EDA)

'parking_data' was obtained by merging different files of raw data from two parking garages.

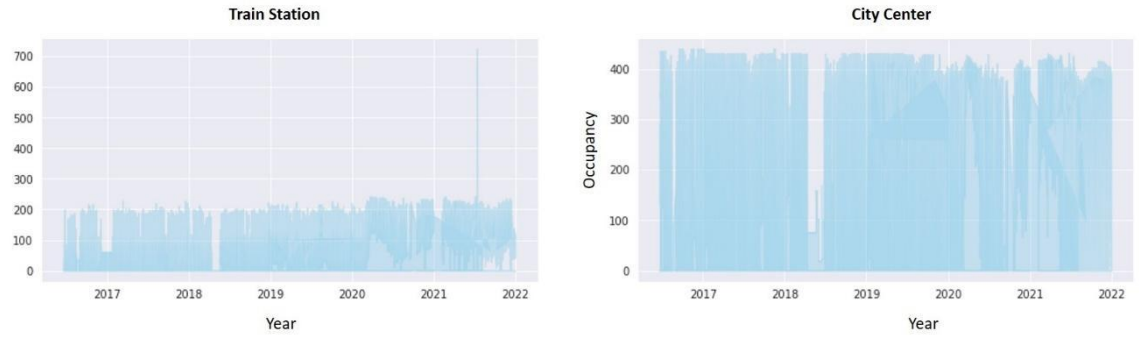

Fig 3. Historic parking data for train station and city center garage

The total capacity of train station and city center garage is 272 and 516, respectively [1]. The historic data of train station exhibited higher values than the maximum capacity, in consequence, the data was cleaned (It was assumed that the total capacity of both garages remains constant over time).

In order to obtain a better understanding of timeseries and outliers' distribution, the date was classified as 'Weekend' variable (yes/no) and 'Weekday'. Both parking garage shows outliers during weekend, especially on Sunday. This observation is evidenced in Fig 4.

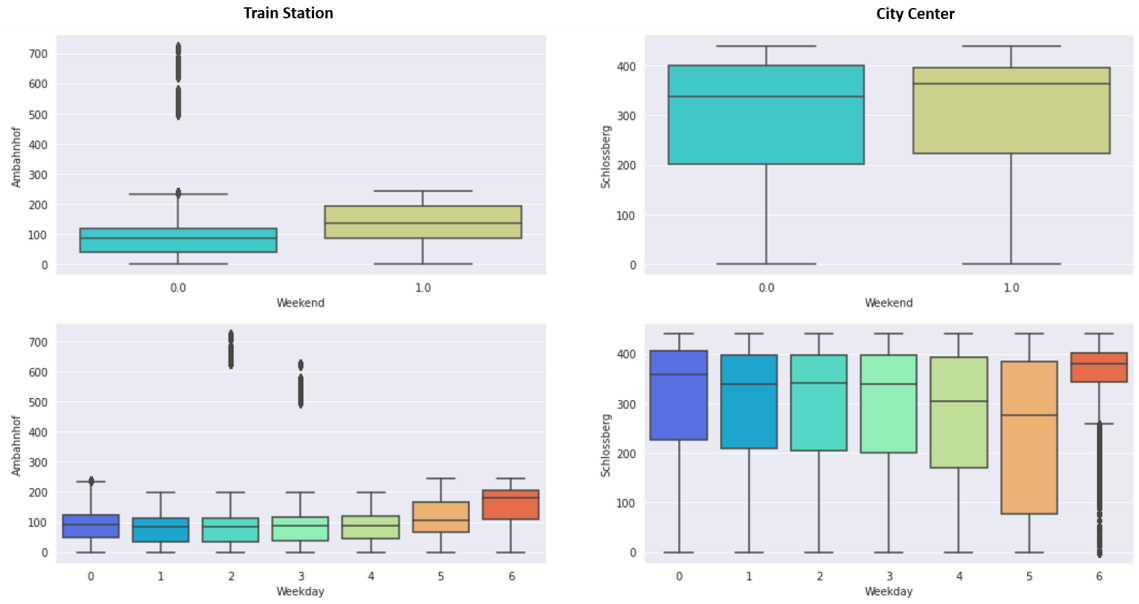

Fig 4. Boxplot of parking garage according to the weekday. Ref.: M: Monday, Tu: Tuesday, W: Wednesday, Th: Thursday, F: Friday, Sat: Saturday, Su: Sunday.

The time-series also was classified according to year, month and hour. The corresponding boxplots and presented in Fig 5.

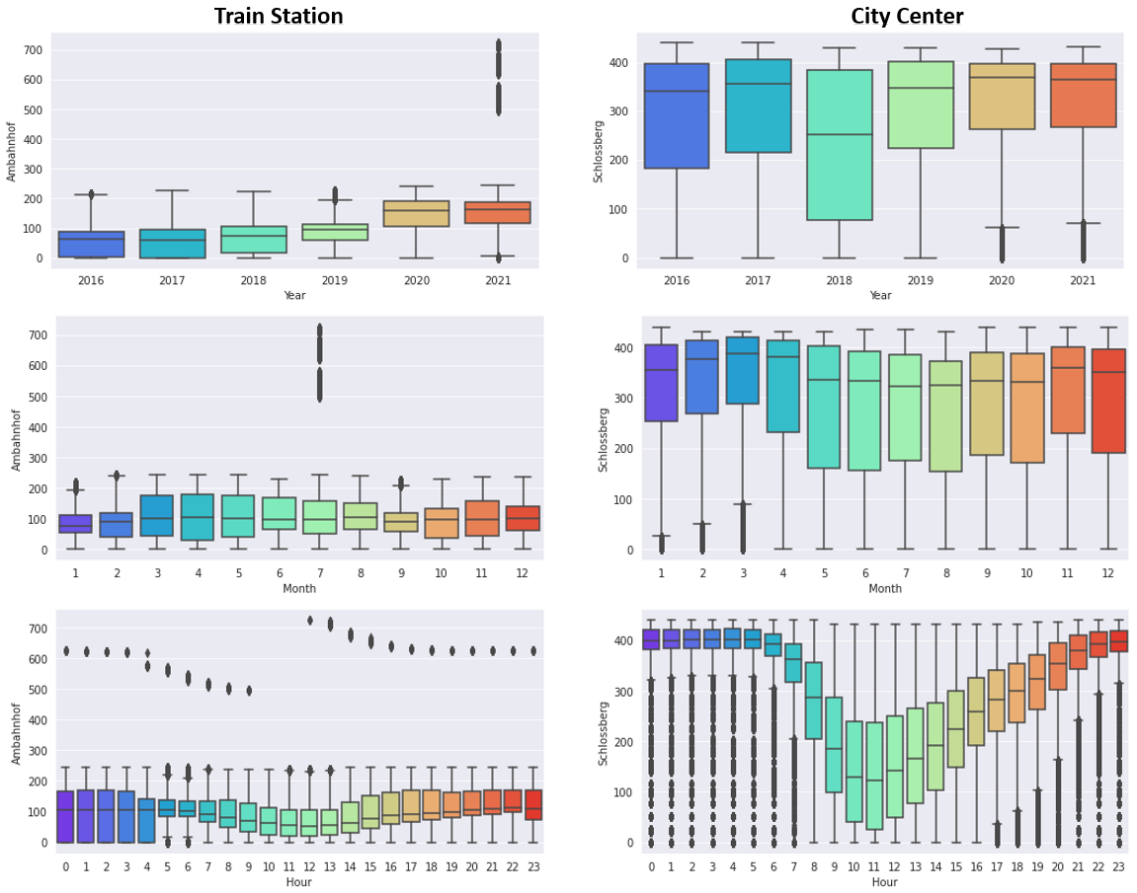

Fig 5. Boxplot of parking garage according to year, month and hour.

A huge number of outliers occurs in 2019 and 2021 for the train station, and in 2020 and 2021, for the city center. The major number of outliers are observed in July for the train station, and between 0-9 and 12-23 h. In contrast, city center parking shows the opposite behavior, where the majority of outliers take place in cold months (January, February, and March). Also, outliers are observed outside the time range of 8-16 h.

The reduction of occupancy in working hours may be explained by the hypothesis that the majority of parking spots are used by workers near the site. Thus, releasing the spot during working hours.

Feature engineering

The feature with a higher effect on parking garage occupancy are selected. The heatmap correlation matrix was employed to visualize the correlation coefficient.

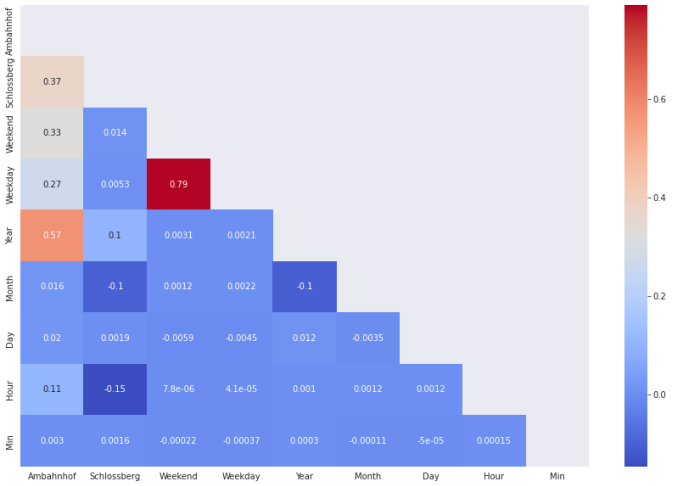

Fig 6. Matrix correlation heatmap of raw data with five minutes shift.

From the correlation matrix heatmap it can be observed that the features with more predictive power in decreasing order are:

a. **Train station garage:** Year (0.57), Weekend (0.33), Weekday (0.27), and Hour (0.11).

b. **City Center garage:** Hour (-0.15), Month(-0.1) and Year (0.1).

The minute time base exhibited the lowest correlation coefficient for both parking garage with five minutes rounding.

Extra data considering weather features and holidays was also analyzed. Weather historic data was obtained from Meteostat [2]. The features considered were: Daily maximum temperature (Tmax, °C), daily minimum temperature (Tmin,°C), daily average temperature (Tavg, °C), precipitation (Prcp, y/n), freezing temperature or snow (Freezing, y/n), average wind speed (Wspd, km/h), and wind peak gust (Wpgt, km/h). The maximum total precipitation in the historic data was 48.7 mm (in a day), with an average of 2.09. Considering this variation, precipitation and Freezing features were adopted as binary variables.

The Baden-Württemberg Public Holidays Register was obtained from Public Holidays [3]. Holiday was adopted as a binary variable for the specific day (y/n).

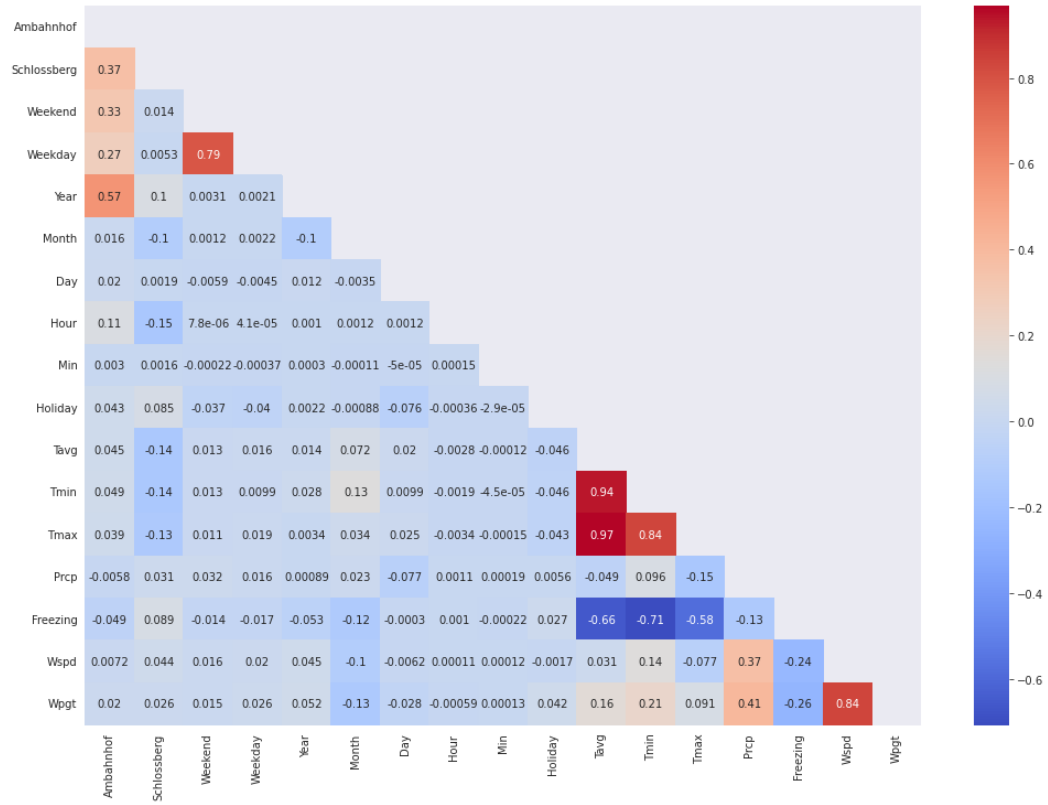

Fig 7. Matrix correlation heatmap of parking occupancy with the weather forecast and holiday data.

As can be observed, the train station garage occupancy is mainly correlated to the year, weekend, weekday, and hour. On the other hand, no strong correlation with weather features was observed. In contrast, city center garage occupancy is correlated with time basis variables (daily hour, month, and year), and weather features, especially with Tmax, Tmin, and Tavg. It is noteworthy that both parking garages are covered.

Regarding the Holiday variable, the city center garage exhibited a double correlation coefficient compared to train station garage occupancy. However, this feature does not present a relevant correlation with the label outcome under the conditions covered in this work.

From the correlation matrix heatmap can be observed that the features with more predictive power in decreasing order are:

a. **Train station garage:** Year (0.57), Weekend (0.33), Weekday (0.27), and Hour (0.11).
b. **City Center garage:** Hour (-0.15), Tmax (-0.14), Tmin(-0.14), Tavg(-0.13), Month(-0.1) and Year (0.1).

Outliers Handling

Outliers are unusual values in the dataset, which can distort statistical analyses. Removing outliers can cause data loss. In this case, it is required used various scaling and transformation techniques. Quantile Transformer is a robust preprocessing method able to transform the features following a uniform or a normal distribution. Thus, this

transformation tends to spread out the most frequent values, reducing the effect of marginal outliers. The variation of correlation coefficients using Quantile Transformer was determined. Table 1 shows the results obtained.

Table 1. Variation of correlation coefficient of selected features employing Quantile Transformer methodology.

Feature	Train Station			City Center		
	Before	After	Abs. Dif [%]	Before	After	Abs. Dif [%]
Weekend	0.33	0.3	9.1	0	0	0
Weekday	0.27	0.26	3.7	0	0	0
Year	0.57	0.57	0	0.1	0.061	39.0
Month	0	0	0	0.1	0.12	20.0
Hour	0.11	0.12	9.1	0.15	0.22	46.7
Tavg	0	0	0	0.13	0.16	23.1
Tmin	0	0	0	0.14	0.17	21.4
Tmax	0	0	0	0.14	0.14	0

Higher variation in city center garage correlation coefficients was observed compared to train station garage. A reduction of 39% in the correlation coefficient for Schlossberg-Year was obtained by employing Quantile Processing. The selection feature criterion employed in this work is indicated by a correlation coefficient higher than 0.1. In this way, there is no strong correlation between city center occupancy and the year variable. Seasonalities were decomposed using stats models seasonal_decompose function.

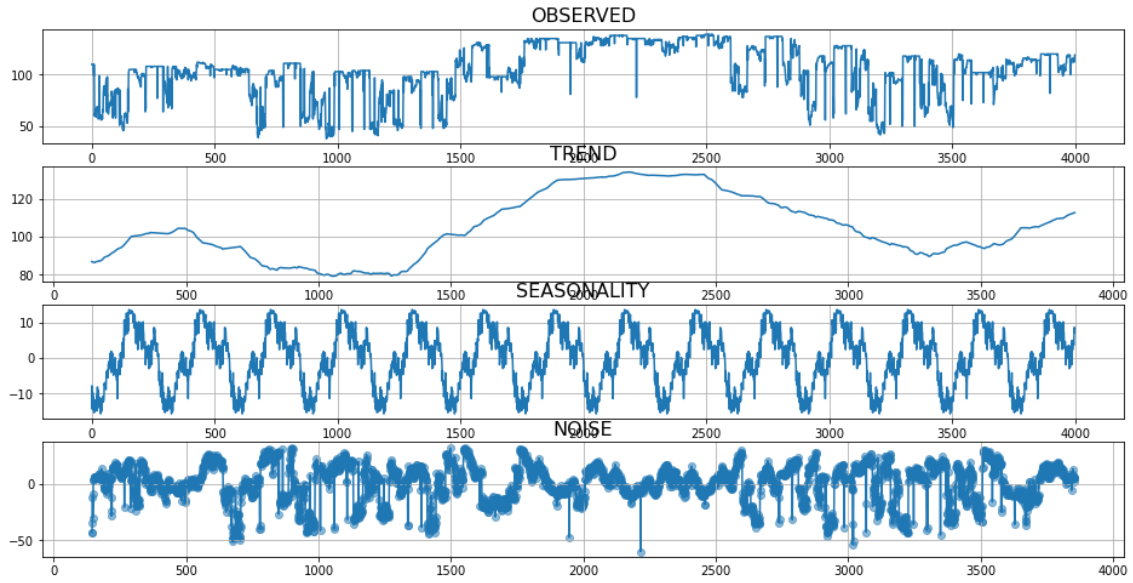

Fig 8. Seasonality of train station garage (period: 48 h).

As can be observed in Fig 6, raw data, even with Quantile processing, present a considerable level of noise. On the other hand, a defined seasonality is appreciated for a 24 h period, where three peaks are followed by a valley (parking occupation reduction).

Model Development

Three predictive models (ARIMA, Prophet Univariate, Prophet Multivariate) in combination with different data processing were developed. These machine-learning models were selected considering the daily variation of raw data. Prophet, or "Facebook Prophet," is an open-source library time series forecasting developed by Facebook. Prophet is especially useful for datasets that:

a. Contain an extended time period (months or years) of detailed historical observations (hourly, daily, or weekly)
b. Have multiple strong seasonalities
c. Include previously known important, but irregular, events
d. Have missing data points or large outliers
e. Have non-linear growth trends that are approaching a limit

Prophet is an additive regression model with a piecewise linear or logistic growth curve trend. It includes a yearly seasonal component modeled using the Fourier series and a weekly seasonal component modeled using dummy variables. Prophet can be considered a nonlinear regression model of the form:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t \quad (2)$$

where g(t) describes a piecewise-linear trend (or "growth term"), s(t) describes the various seasonal patterns, h(t) captures the holiday effects, and ε_t is a white noise error term.

a. The knots (or changepoints) for the piecewise-linear trend are automatically selected if not explicitly specified. Optionally, a logistic function can be used to set an upper bound on the trend.
b. The seasonal component consists of Fourier terms of the relevant periods. By default, order 10 is used for annual seasonality and order 3 is used for weekly seasonality.
c. Holiday effects are added as simple dummy variables.
d. The model is estimated using a Bayesian approach to allow for automatic selection of the changepoints and other model characteristics.

The knots (or changepoints) for the piecewise-linear trend are automatically selected if not explicitly specified. Optionally, a logistic function can be used to set an upper bound on the trend.

The seasonal component consists of Fourier terms of the relevant periods. By default, order 10 is used for annual seasonality and order 3 is used for weekly seasonality. Holiday effects are added as simple dummy variables. The model is estimated using a Bayesian approach to allow for the automatic selection of the changepoints and other model characteristics.

In general terms, Prophet exhibited a better fit than the ARIMA model. Thus, the hyperparameters of this model were tuned. The metrics or the models were Mean Absolute Error (MAE) and RSME (root-mean-square error). The train/validation split was 0.85.

The optimal hyperparameters of the Prophet model were:

a. changepoint_prior_scale=0.001,
b. seasonality_prior_scale=10,
c. yearly_seasonality=True,
d. weekly_seasonality= True,
e. daily_seasonality=True,
f. interval_width=0.90

“changepoint_prior_scale” parameter shows a major effect on model performance, determining the flexibility of the trend, followed by yearly_seasonality and daily_seasonality. This could be in concordance with the findings obtained from the correlation matrix heatmap, where the higher correlation coefficient of the train station corresponds to the year variable. Table 2 shows the attributes of the most promising Prophet model.

Table 2. Attributes of different Prophet models for Train Station garage (All models with outlier remotion through Quantile processing).

Model	Model variate	Data[c]	Time shift [min]	MAE	RSME
A	Univariate	2021	60	0.12	0.15
B	Univariate	2016-2021	60	0.18	0.21
C	Univariate	2016-2021	5	0.22	0.26
D	Multivariate	2021	60	0.13	0.16
E	Multivariate	2016-2021	60	0.16	0.20
F	Multivariate	2016-2021	5	0.21	0.24

[c]Data range for training and validation.

According to the metrics obtained, the selection of data training and validation exhibited a paramount effect on metric models. Another relevant attribute consists of time shifts. Both metrics were favored when representative train data was selected and daily occupancy variation was reduced. On the other hand, no significant improvements were observed using Multivariate Model. The features for this model were selected according to Feature Engineering: ‘Weekend’, ‘Weekday’, ‘Year’, ‘Month’, ‘Day’, and ‘Hour’. These variables had not a pronounced effect on model performance. However, Multivariate Prophet involves major computational costs and latency, being a paramount parameter in stream data science products.

Overall conclusions

- There are anomalies and a lack of data in the raw data corresponding to the train station garage: 2016/08/06 – 2016/08/31, 2016/12/06 – 2017/25/06, 2018/04/11 – 2018/05/18, 2020/09/03 – 2020/09/11, 2020/09/20 – 2020/10/16 and 2021/06/18 - 2021/07/21. These anomalies in data need to be explored in order to select the appropriate value for model training and validation.
- Different data range selections had a paramount effect on model metrics, surpassing even the effect of the hyperparameters tuning.
- A different seasonal behavior was identified from the second half of 2020 to the first quarter of 2021, leading to a significant increment in daily minimum occupation. This behavior was followed by a decrease in the occupation rate of the train garage. The monthly and seasonal variations observed in 2020 and 2021 influenced the model performance on prediction, especially to April 15th,

2022, between 12:00 and 13:00, considering that the available data cover until 2021/12/31.

- The appropriate selection of training and validation data represents the most important parameter to enhance the model performance, followed by time shift and hyperparameter tuning. For this reason, the ensemble method could be an affordable solution to increase prediction accuracy.

Future works

According to the findings obtained in this work, it is suggested to collect more information associated with problem framing. In this way, it can be established appropriate criteria for data selection. In this scenario, the development of the ensemble method is encouraged, considering the yearly variations, especially those observed during 2020 and 2021. This would contribute to increasing the model prediction in future months or weeks.

References

[1] Freiburg, Parken. https://www.parken-in-freiburg.de/bahnhofsgarage (accessed October 22, 2022).

[2] Meteostat. https://meteostat.net/en/station/10803?t=2017-05-03/2022-05-25 (accessed October 22, 2022).

[3] Public Holidays. https://publicholidays.de/baden-wurttemberg/ accessed October 22, 2022).