Analytics and Application - Team 6 Assignment 2023/24

Seminar Paper

Author: Esther Mitterlehner (7419216), Emma Lux (7419032), Nico Guldin (7424632), Jasper Fülle (7419157), Ferdinand Popp (7424516)

Supervisor: Univ.-Prof. Dr. Wolfgang Ketter

Co-Supervisor: Janik Muires

Department of Information Systems for Sustainable Society
Faculty of Management, Economics and Social Sciences
University of Cologne

January 28, 2024

Executive summary

EVs as increasing mean of transport to reduce mobility emissions depend on reliable charging hubs which are able to meet the amount of energy requested by connected EVs at any time. Understanding underlying usage patterns of the customers charging their EVs, the management of the charging hubs can be supported by investment and transformation decision.

The given data contains charging sessions from to different charging sites between 25.04.2018 and 14.09.202 with multiple features recorded. During data preparation, duplicates, unnecessary features and outliers were deleted. Extreme data was adapted and some features were prepared further to enable correct conclusions during analysis.

The analysis was started with some description, where first of all, charging site 1 was defined as the private site whereas site 2 is treated as public. Charging mostly happened between 7 am and 5 pm on weekdays and charging sites on weekends are significantly less used. The current hub operation can be supervised by the charging speed of each spaceID, an overview of connection counts between registered and non-registered users, the insight in the difference between doneChargingTime and disconnectTime and parallel between the delivered and requested kWh.

In order to know the different types of users for marketing campaigns or other individual offerings, a cluster analysis was applied and with different means the number of three clusters was found within the data set. Cluster 1 are therefore students or shoppers mostly charging in the afternoon. Cluster 2 are workers frequently driving to the office with regular connection patterns and low energy requests. Cluster 3 are in contrast workers which now and then drive to the office with higher energy requests.

For future business transformation and profit maximization, utilization prediction is needed. After testing various predictions models, a decision tree was recommend since its explainability helps the management to select specific features which can be influence hub usage. By using calculated usage predictions, pre-buying and -charging of EVs with low price energy and later transferring within the charging system and connected EVs promises higher profits.

For further data analysis, data sets with higher quality and better contextual reference needs to be recorded. And give users the incentive to publish more user data can lead to better charging hub performance and transitions.

Github repository: https://github.com/emitterl/AA-project-team-6

Contents

List of Figures

1 Problem description

As the second-largest share of total EU emissions is made up of greenhouse gases related to transportation, there needs to be a change in our approach to mobility in order to meet decarbonization targets. At present, traditional urban mobility is heavily reliant on internal combustion engine (ICE) vehicles. A shift to electric vehicles (EV) is therefore a key factor in reducing personal mobility emissions. However, the charging infrastructure and underlying power grid must be capable of providing adequate capacity at the right time.

To combat the fear of not being able to complete a trip due to uncertain battery capacity, users tend to plug EVs in during parking to ensure full battery load. This creates a lot of pressure on operators of such charging hubs, since the charging stations may not have the capacity to accommodate all users who want to charge their vehicles at the same time. By understanding the utilization of charging hubs now and in the future, EV charging hub operation management can decide about necessary transformation and optimization within their value creation.

Analyzing charging data according to performance patterns and indicators can help achieve this goal, as well as finding archetypal charging sessions to differentiate between different types of users or customers. In addition, usage forecasts enable discussions about future charging patterns and the resulting necessary adjustments.

2 Data Description and Preparation

ToDo: Was haben wir mit missing values gemacht?

The given data set contains 66,450 data records for EV charging sessions that took place between 25.04.2018 and 14.09.2021 at two different charging sites. Of these, 35,042 tuples originate from charging site 1 and 31,408 from charging site 2. The table contains 12 columns, one of them is the column userInputs which contains seven more individual columns. We merged them together and thus created one table with 19 columns. Afterwards we examined every column and its value for our analysis. Table 1 contains a description and the changes we executed on each column. Afterwards we took a further look at the values of the remaining columns and prepared the data as follows:

In data preparation, a fundamental distinction must be made between outliers (values that do not contribute to correct conclusions in the target population) and extreme values (values that are extremely large or small). Implausible outliers are

always removed from the basic population of the data set, while extreme values require a more differentiated approach (e.g. adjustment of the value).(Höfler, 2019) When analyzing the data set, we noticed the following outliers, which we removed:

1 here were 0-values in the kWhRequested and milesRequested columns. Since it is not reasonable/ possible to charge a car with 0 units, we have excluded these values.
2 In the minutesAvailable column, there was a single value that was more than 5 times as large as all the others. As this is most likely an incorrect entry, this value has been removed.

The website https://ev-database.org was used to identify extreme values. This contains current information on electric cars such as range, consumption, battery performance, etc. The extreme values identified were adjusted as follows:

1 All values from WhPerMile above 474.8 were set to 474.8.
2 All values from WhPerMile below 223.8 were set to 223.7.
3 All values from kWhRequested above 123 were set to 123.
4 All values from milesRequested above 425.6 were set to 425.6.
5 Sometimes the doneChaegingTime was after the disconnectTime. This could be due to measurement inaccuracy, so in this cases we set the value of doneChaegingTime to the value of disconnectTime
6 Some valzes from doneChargingTime were earlier then connectionTime. In these cases doneChargingTime was set to NaT.

Another extreme point that was identified is the value 108.8 from the kWhDelivered column. Although this value is significantly higher than the other values in this column, it is a realistic value. Accordingly, it was neither changed nor removed. We also checked the dataset for duplicates and eliminated them. Missing data...

3 Data Analytics

3.1 Descriptive Analytics

3.1.1 Temporal Patterns and Seasonality

A look at the use of the charging sites over the course of the day, week or year reveals various patterns 1.

In terms of the day, there is a clearly recognizable peak between 7 and 9 am for connecting and between 4 and 5 pm for disconnecting the EV. So most of the people are charging their cars during work or study courses. Very few are leaving the EV over night at the charging station.

There is also a structure of charging events over the week. As the majority of work and study takes place during the week, Saturdays and Sundays are not nearly as busy as the other days of the week.

Looking at the level of the seasons, there is a difference between the use of charging stations in winter and spring versus summer and fall. This can be explained by a lower range in colder temperatures due to higher energy consumption for heating and battery warming (Rudschies, ADAC, 2022). Consumers may therefore be more likely to switch to the more flexible public transport.

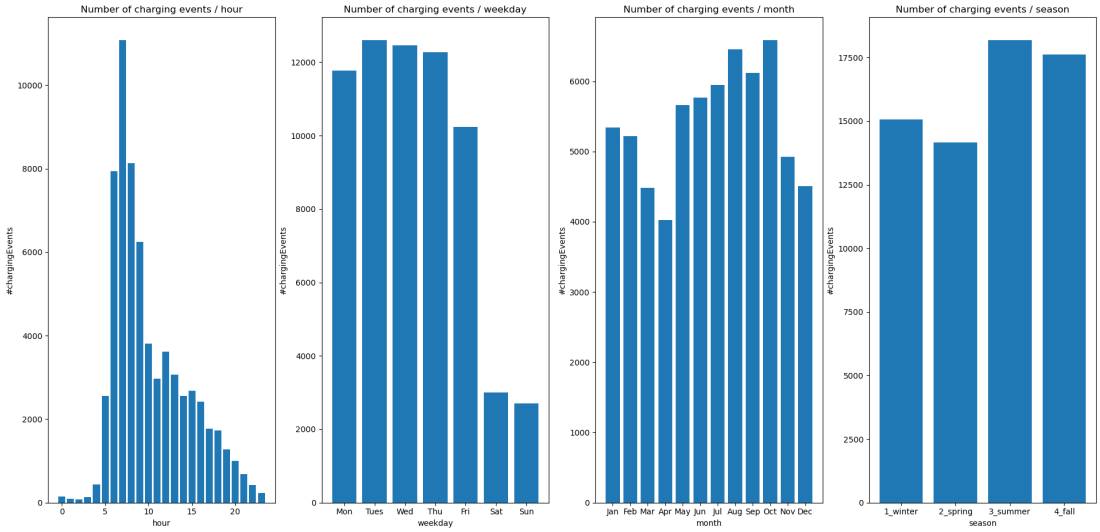

Figure 1: Total Connection Counts per hour, day, month, season

3.1.2 Site Characteristics

After investigating the temporal patterns over all charging events, a distinction between the charging sites is made. As already mentioned, it is known that the two charging sites are once public at a university and once private for the

employees of a company. Despite the lack of allocation, the following hypothesis is made and proven with the analysis below:

While site 2 is public, site 1 is a private charging station.

As can be seen in the illustration 2, site 1 is used very regularly every month whereas site 2 shows a high variance due to variable semester times.

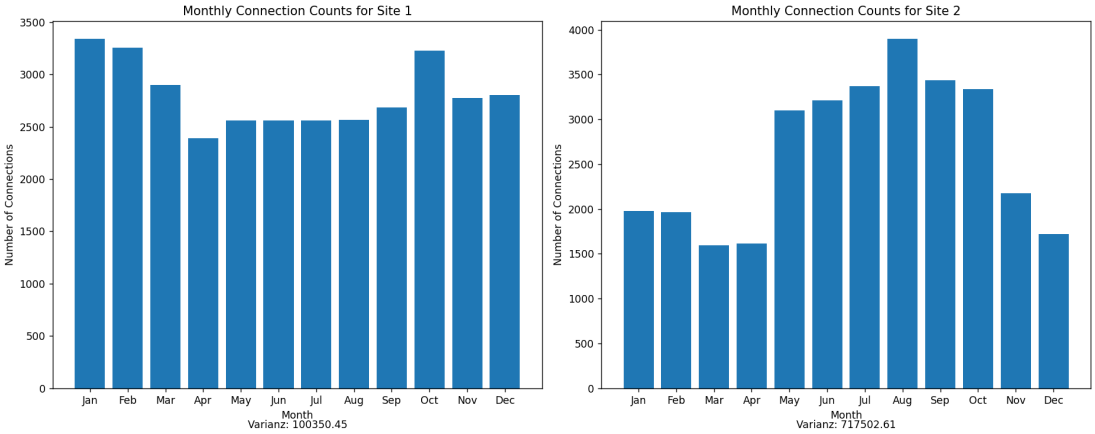

Figure 2: Monthly Connection Counts per Site

The same pattern reflects within the daily connection and disconnection times shown in 3. The average values of site 1 correspond approximately to a normal working day. However, the high variance of site 2 in particular can be explained by the very different timetables at the university, which means that students connect and disconnect more frequently at times other than at the beginning and end of the working day.

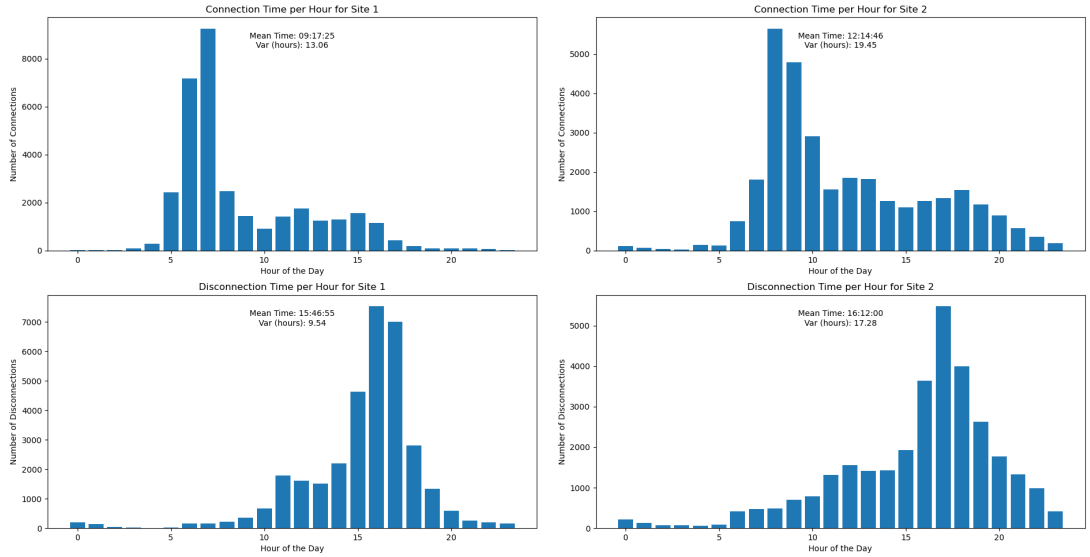

Figure 3: Daily Connection and Disconnection Counts per Site

Another indication is the low usage of Site 1 at weekends (4). Very few

employees work on Saturdays and Sundays, so the charging site could be closed or only accessible to the management level or those responsible for the building. On the other hand, students with partially open libraries have more reason to use Site 2 on weekends. In addition, it can be assumed that other people who are out and about in the city at the weekend can also park and connect on the public charging site.

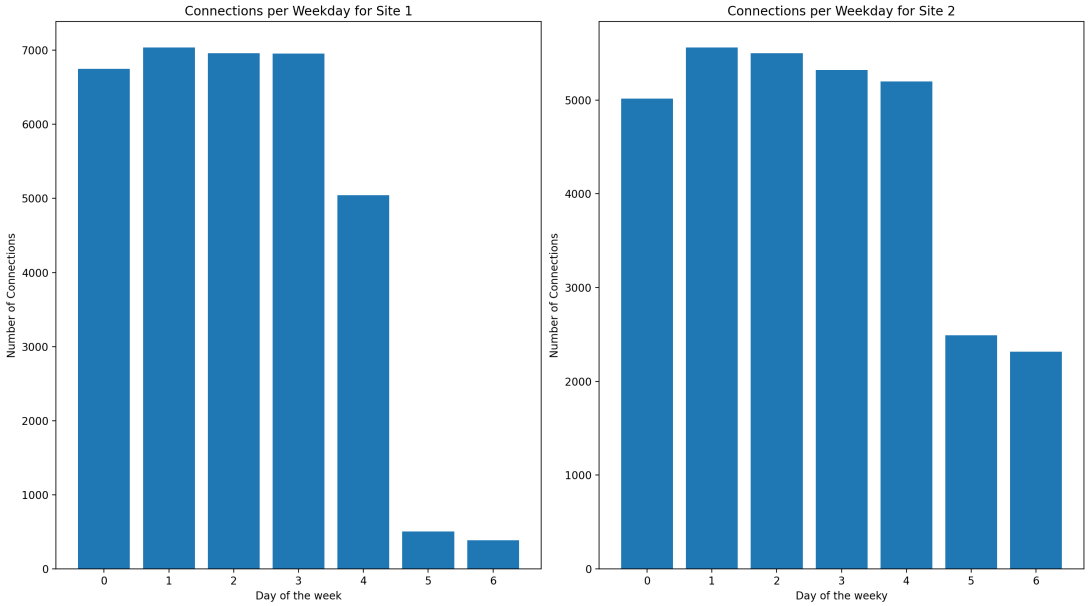

Figure 4: Weekly Connection Counts per Site

In summary, all analysis indicates for site 1 being the private and site 2 being the public charging site.

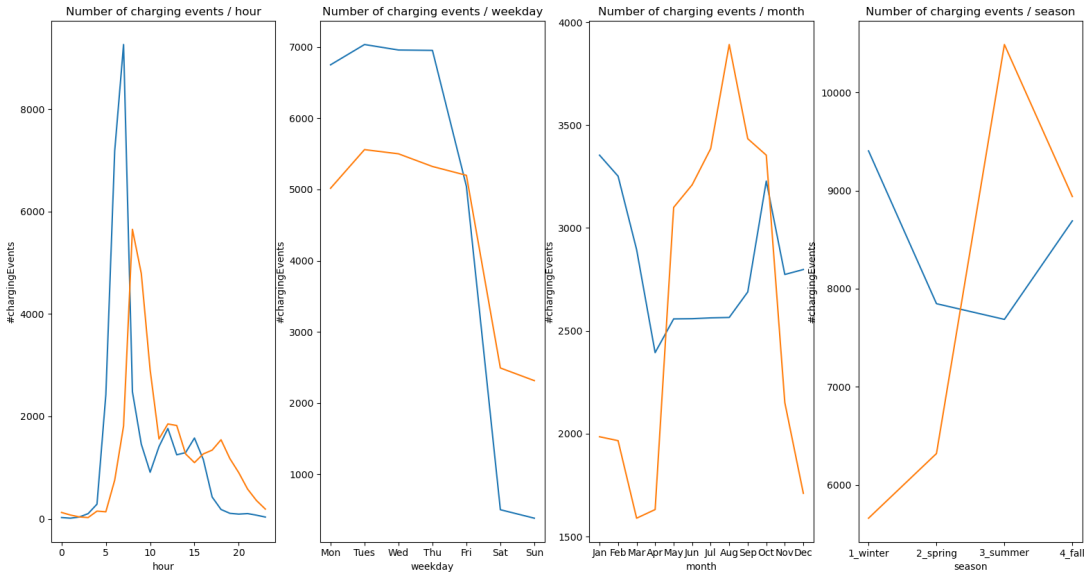

Figure 5: Weekly Connection Counts per Site

3.1.3 Key Performance Indicators (KPIs)

In order to have an overview over the current hub operation, various KPIs were defined, starting with the connection counts per hour between registered and non-registered users (6). Besides the fact, that the user base consists of more registered users, that KPI indicates the different usage patterns between registered and non-registered users and could get extended with the disconnection counts per hour. Showing ads at displays at the charging station could be a use case, where knowing if the hour for connecting is rather used by already registered users or not, can be helpful to chose the right content.

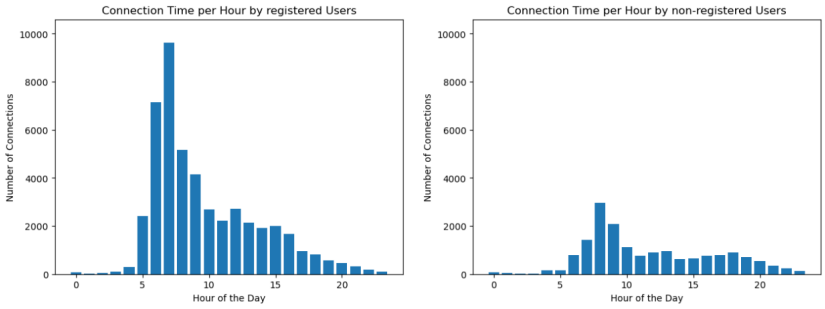

Figure 6: Connection Counts between registered and non-registered Users

The second KPI deals with the discrepancy between the doneChargingTime and disconnectTime (7). As EVs occupying a charging space without actual charging is unused revenue potential, knowing how much potential is lost and if measures against that have an impact is severe for higher profits. Besides outliers with parking times over days, most of the charging sessions are ended within one until six hours since the full charge was accomplished. Additional fees for the parking time without charging could provide incentives for customers to minimize that discrepancy.

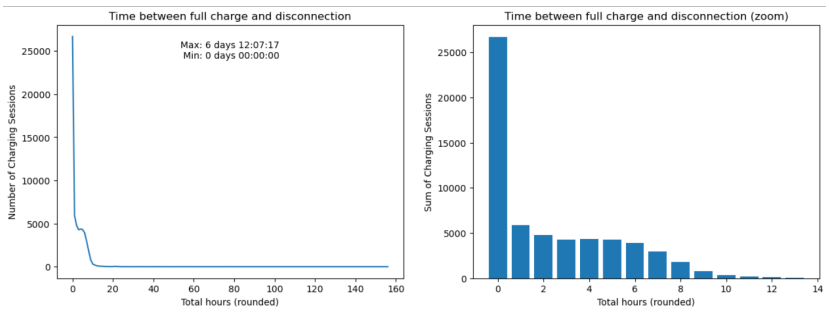

Figure 7: Difference between doneChargingTime and disconnectTime

The next KPI is only possible with charging sessions of registered users, as it compares the requested amount of kWh and the actual loaded amount (8). Only regarding the charging sessions (red), where the EVs are still connected after

doneChargingTime, this difference is either the result of misjudging the capacity of the EV by the user or the amount of not delivered energy maybe due to a supply deficit which means missed revenue. Thinking further, by providing the requested amount of kWh and the parking duration by the user, the charging of multiple EVs during a supply deficit could be scheduled, so EVs with longer parking duration could get charged later.

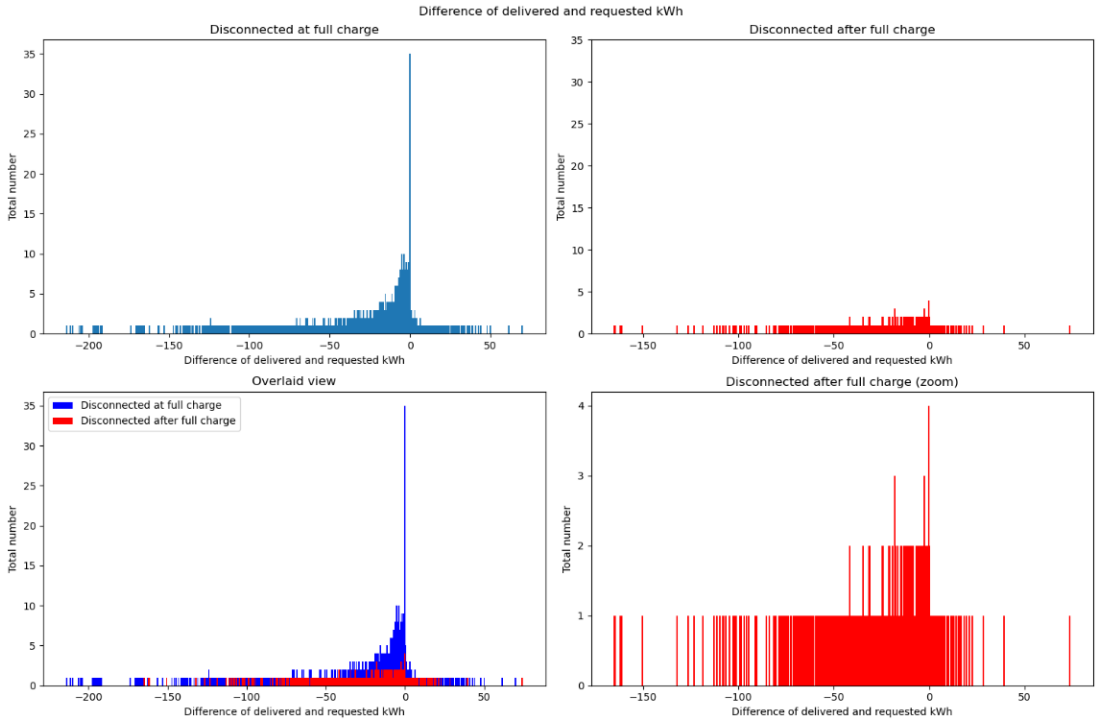

Figure 8: Difference between delivered and requested kWh

In case of improving the technical capacities of charging sites, it is always helpful to know where to start. And the last KPI, showing the mean charging speed per site (9) allows to prioritize. The generally better charging speed at site 1 can be related to the site being private and therefore maybe getting more investments for better performance.

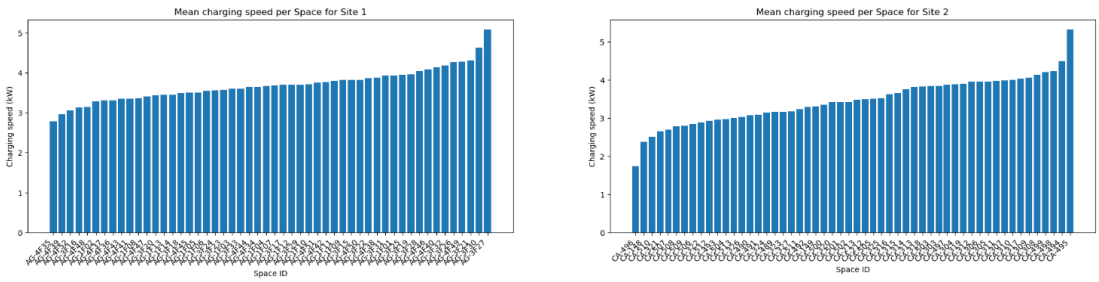

Figure 9: Mean charging speed of spaceID

3.2 Cluster Analysis

Before talking about the archetypical charging events, the decision for three clusters in explained. As base for the cluster analysis, the following features were selected. All others were neglected in other to reduce complexity: connectionTime, disconnectTime, doneChargingTime, kWhDelivered, kWhRequested, milesRequested, minutesAvailable

the residual loss dependent on the number of clusters was calculated (10). Only up to three does the loss decrease significantly with a higher number of clusters. This corresponds to the result from hierarchical clustering (11), where there is a large distance between the first two separations of the data set and the next cluster formation.

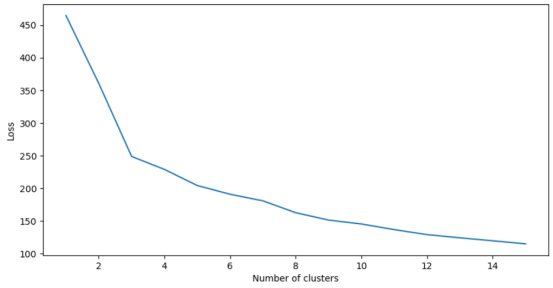

Figure 10: Residual loss of dependent on number of clusters

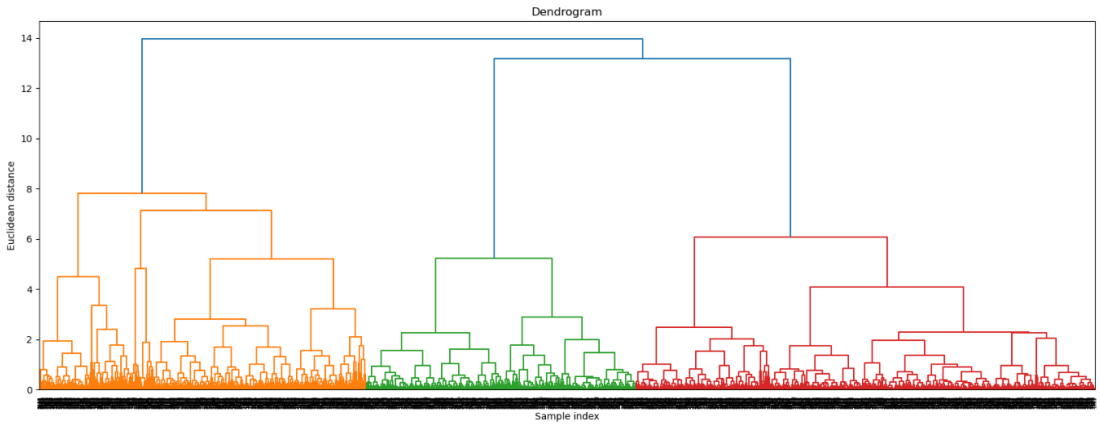

Figure 11: Hierarchical clustering

The formed clusters are described in the following:

Clusters 2 and 3 are very similar in their connection patterns. Both types of charging sessions usually connect at around 7:00 a.m. cluster 1 with connection times mainly scattered in the afternoon is very different. However, all clusters usually have a disconnecting hour around 6 pm. This difference is also evident in the user inputs, where users from cluster 1 state that they have significantly less time to load than in clusters 2 and 3, since the two latter are mostly connected

over the whole day. Another significant distinction can be made on the basis of the kWh charged. Here, cluster 2 has significantly higher amounts than clusters 1 and 3. This also fits with the longer period of time that EVs from cluster 2 need to charge, in contrast to cluster 1, although both were connected at around the same time. This differentiation is also evident in the user entries, as cluster 2 also requested higher amounts of kWh and miles.

Cluster 1 (blue): Students or shoppers who are mainly out and about in the city in the afternoon and charge the EV during lectures or shopping.

Cluster 2 (green): Workers who drive to work from time to time. The connection and disconnection time corresponds to a normal working day. However, there is a greater distance between their office days and thus their charging sessions, which is why more kWh must be charged.

Cluster 3 (orange): Workers who drive to the office every day. The connection and disconnection time corresponds to a normal working day. And since they connect their EVs every day, they do not need as much energy.

The value of identifying different types of charging sessions is not only the possibility to identify market trends, target advertising efforts, and develop new services or products that cater to specific user clusters. Understanding usage patterns also allows for better placement and optimization of charging stations and enables energy management to prevent a lack of energy supply.

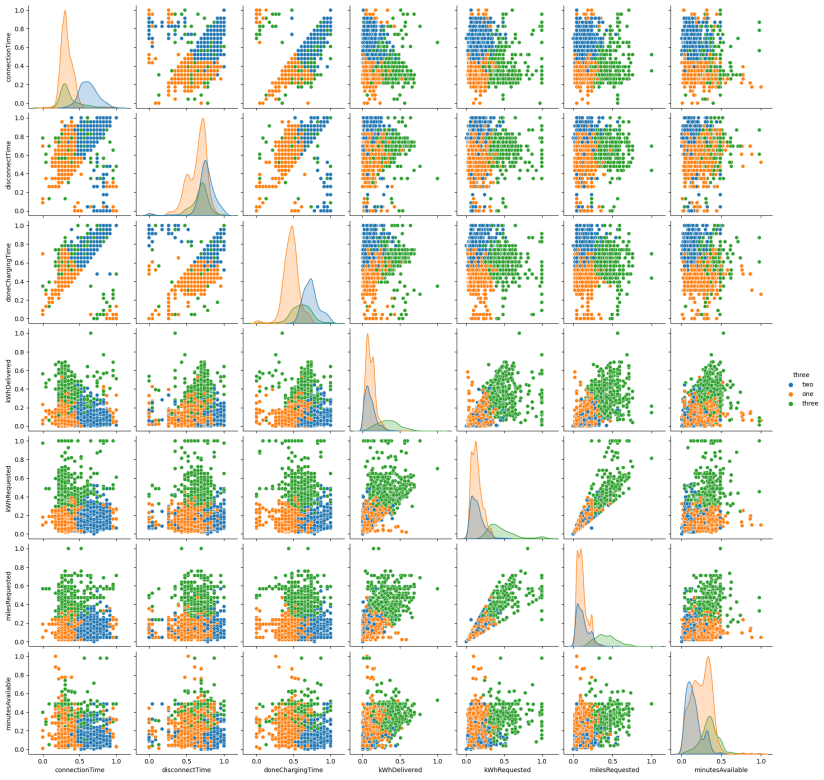

Figure 12: Pairplot with 3 Clusters.

3.3 Utilization Prediction

After analysing the data on EV charging sessions, a model for predicting hourly utilization of the two sites is developed. For training this model, a new data frame had to be created. That contains the siteID, how many charging spaces are occupied, the weekday, the temperature, a weather description and if it is a holiday for each hour of the day over multiple years (time, siteID, occupiedCount, isHoliday, temperature, weatherDescription, Weekday). The dependent value is occupiedCount, all other feature can be used for predictions. The data frame is then split into 70/30 for training and test purposes.

To find the model with the best performance, multiple model types were developed. After the comparison between a linear regression, a polynomial regression and a decision trees with varying hyperparameters, the decision tree provides the best performance with the following hyperparameters, found by cross-validation on the training data: max depth is 10 with each leaf containing at least 50 samples. To avoid overfitting and reduce the complexity of the tree, a pruning parameter of 0.1 was added. As a restriction in the number of features used leads to worse results during cross-validation, maxFeatures was set to NONE. That results in a Mean Squared Error of 17.98 on the test data.

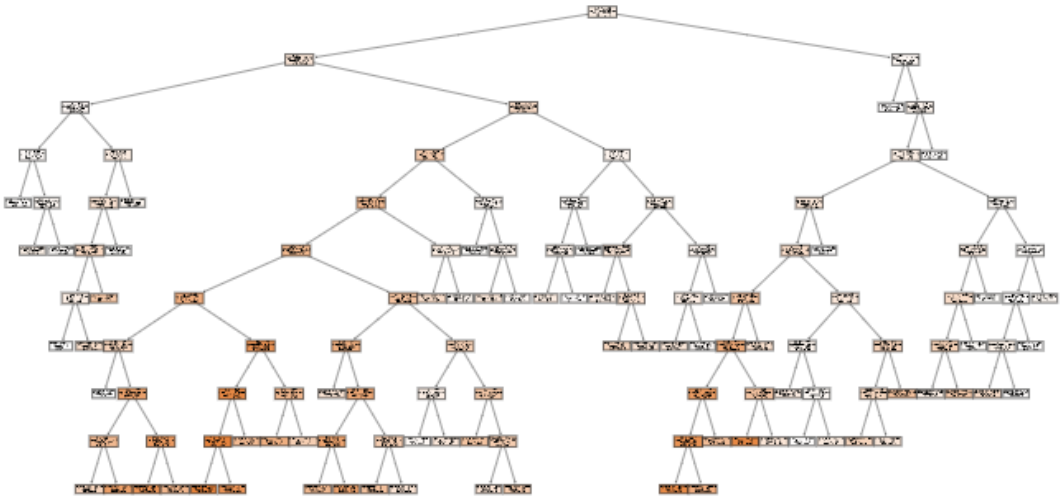

Figure 13: Decision Tree (Original in repository under decisiontree.svg)

This model was compared to the performance of a neural network constructed for each charging session which consists of one input layer with 24 features, five hidden layers, two of which are dropout layers with an ration of 0.1/0.3 to reduce overfitting, and one output layer. ReLU was chosen as activation function and each hidden layers consists of 25 nodes. The model was trained with 30 epochs and has an equal performance as the decision tree with a Mean Squared Error of 16.05 on the test data.

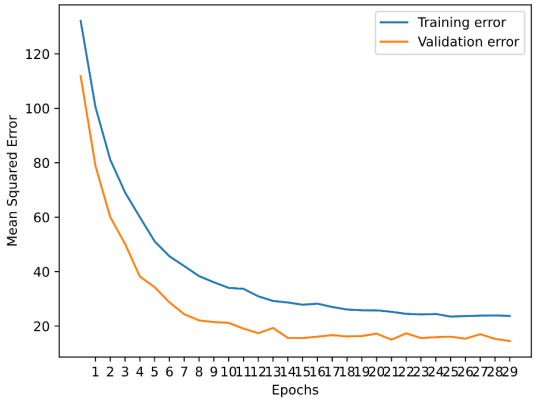

Figure 14: MSE dependent on amount of Epochs

As a proposition, the further deployment of the decision tree is suggested, since it has an equal performance to the neural network but is by far more explainable and you can investigate which features have an impact on the predictions.

Use Case: These predicting models could enable the following method to reduce the costs of buying energy for the charging hubs. As described on the following website (https://1komma5grad.com/de/magazin/strommarkt/intraday-day-ahead-handel), the price for energy is varying over the hours of the day in the day-ahead-market as well as in the intraday-market. In the optimal case, the hub operator only buys energy when the price is the lowest. But the peak connection time is often during a higher price level. Predicting the amount of vehicles connecting to the charging stations in the next few hours could enable pre-buying energy at a lower cost and charging already connected EVs to 100 percent capacity, maybe even if it was not requested. This surplus of charging can then be transferred to EVs that connect according to the prediction in the next hours with peak energy prices. Using EVs as storage unit is already applied in other networks, so it could also be used to ensure lower energy costs. This leads either to higher profit or lower prices for the customers, which could help to undercut competitors. This can be combined with a scheduling algorithm, so that EVs don´t necessarily start to charge once connected but are scheduled for a time with lower energy prices. For all this to work, the first requirement is the availability of user input data. Without knowing the requested amount of energy and the available minutes, this transfer of energy from one EV to another and the scheduling is not possible. Another prerequisite is the low bias of utilization prediction and that the user data is valid. As this algorithm would rely on probability and predictions, certain fallicies could happen, which the operator would need to include into the price setting system. Since users could disconnect their cars earlier than stated in their user inputs, the EVs could either be more charged than requested

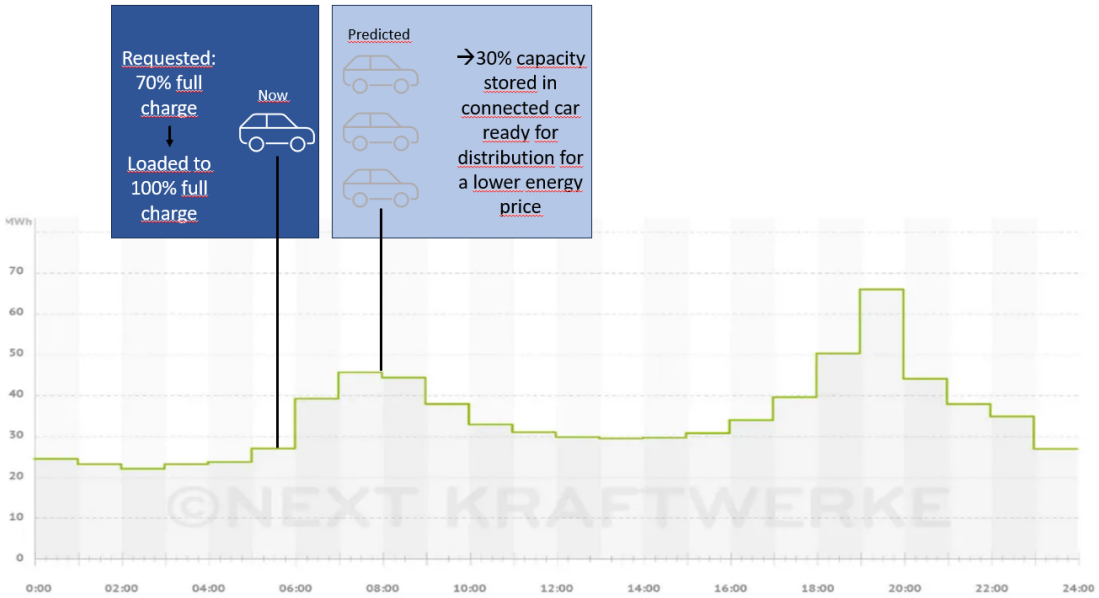

Figure 15: Pre-charging of EVs based on utilization prediction

due to pre-charging for other EVs, or not charged up to their requested load for example due to scheduling. But with good algorithms and predictions, this could lead to significant energy cost reductions and therefore higher profits.

4 Conclusions

After the analysis of charging data from different EV charging sites, different clusters could be exposed and multiple models were tested to find the best performance of predicting the utilization of the sites. Here, the decision tree was recommended. In general, one must say, that the data quality for these kind of analysis was insufficient and for future investigations, more contextual data and information is needed to draw the right conclusions. Recommendation for the charging site operator is the application of the described pre-charging and energy storing and transforming method in order to use predictions for profit maximisation. And for other analysis, creating incentives for users to provide their data enables the distinctions of more usage patterns and better adaptation of the business.

5 Appendix

5.1 Table 1: Data Description and Preparation

Column Name	Description	Change
NaN	Unnamed column with non-unique numbers between 1 and 9947.	This column offers no added value and has been removed.
id	Unique assignment of the charging session using a combination of numbers and digits.	No change made.
connectionTime	Timestamp of date plus time indicates when the EV was connected to the charging station.	The time zone has been converted from UTC to America/Los_Angeles.
disconnectTime	Timestamp when the EV was disconnected from the charging station.	The time zone has been converted from UTC to America/Los_Angeles.
doneChargingTime	Timestamp which shows until when the EV was actively charging.	The time zone has been converted from UTC to America/Los_Angeles. 4.088 records are missing, no adaptations needed.
kWhDelivered	The amount of electricity (kWh) that the EV has charged during the charging session.	No change made.
sessionID	Unique identifier for every record. Consists of sationID and connectionTime.	As two unique identification numbers (see id) are not required, this column has been removed.
siteID	The dataset contains data from two charging sites, 1 and 2. One of these sites is public and the other is private. See more in chapter 5.1.2	No change made.
spaceID	Subdivision of the sites (see siteID), represented in 107 different parking lots, of data type string.	The "spaceID" 11900388 is missing from the schema. However, this has no effect on the analysis, so no change was made.
stationID	Unique identifier for each spaceID.	spaceID is unique, so this column offers no further value and has been removed.

Column Name	Description	Change
timezone	Contains only one value, indicating that both charging sites are located within the America/Los_Angeles time zone.	After all timezones were changed from UTC to America/Los_Angeles this columns adds no further value and has been removed.
userID	Unique identifier for the users from column userInputs.	No change made.
userInputs	Contains a list of information that are shown in the next seven rows. As this entry was voluntary for users, some data is missing.	The content of this column (if available) has been resolved into its own columns. Then this column has been deleted.
WhPerMile	Efficiency of the vehicle (Wh per mile).	No change made.
kWhRequested	Requested Energy (kWh)	No change made.
milesRequested	Requested Row (miles)	No change made.
minutesAvailable	Estimated departure time according to the user.	No change made.
modifiedAt	Time of the user input.	No change made.
paymentRequired	Boolean whether the user had to pay for the charging session.	The content of this column is always true, so it adds no further value and has been removed.
requestedDeparture	Estimated departure time according to the user.	No change made.

References

Höfler, M. (2019). Umgang mit extremwerten und ausreißern. *TU Dresden*.