Predicting EV charging duration using machine learning and charging transactions at three sites

Zulkiflu Musa Sarkin Adar
Department of Engineering
King's College London
London, United Kingdom
k2257768@kcl.ac.uk

Abdulaziz Alhayd
Department of Engineering
King's College London
London, United Kingdom
K21169202@kcl.ac.uk

Grazia Todeschini
Department of Engineering
King's College London
London, United Kingdom
grazia.todeschini@kcl.ac.uk

***Abstract*—Predicting the duration of electric vehicle (EV) charging is relevant for the swift integration of EVs into the grid. It informs EV charging park operators on when to expect peak demand, and the time it takes to charge EVs. This paper describes an approach, based on three machine learning models, to predict the EV charging duration at three public charging sites in the city of Leeds (United Kingdom). The prediction is based on the use of a dataset comprising a total of 7271 charging sessions for the year 2019. In the first part of the paper, the characteristics of the considered dataset are described. The second part of the paper shows that accurate prediction performance can be achieved using features included in the charging transactions exclusively.**

***Index Terms*—Electric vehicle, charging duration, machine learning, feature engineering, prediction.**

I. Introduction

Electric vehicles (EVs) are growing in number worldwide due to a variety of reasons, including environmental concerns and fuel security. According to Ofgem, the government regulator for the electricity and downstream natural gas markets in Great Britain [1], the energy transition driven by electric mobility in the UK will be the most significant shift in the energy sector, and it projects that there will be up to 14 million electric vehicles on the road by 2030. Another projection [2] suggests that there will be 37.4 million EVs on the road in the UK by 2050. However, the adoption of EVs comes with various constraints. Some of these constraints include the impact and strain on the power distribution network caused by uncontrolled EV charging, and the planning for an extended EV charging infrastructure that is highly capital intensive. These factors contribute to reducing the pace of seamless integration of electric mobility into the power grid.

To improve the management of EV integration and identify potential constraints, various models are used. These models, which lead to the prediction of EV charging behaviour, can be characterized using numerous variables. These variables include charging duration, amount of power delivered to the EV, next-day charging prediction, estimating the times when EVs would arrive at a charging station, and charging speed.

For instance, the Gaussian mixture model is a probabilistic method used to characterize charging behavior due to its robustness in identifying similar charging behaviors and grouping them into clusters [3], [4].

Due to the number of variables and the complexity of problems related to forecasting EV charging, machine learning (ML) has gained more popularity in recent years. In [5], three machine learning techniques were adopted to predict charging duration in a two-year charging session for 500 EVs. Grey wolf optimiser, particle swarm optimiser, and genetic algorithm were used to enhance the robustness and accuracy of the prediction models. In [6], arrival time and departure time were predicted with mean absolute percentage error (MAPE) of 2.9 and 3.7 percent, respectively, using a support vector machine (SVM). In [7] the departure time was predicted through extreme gradient boost (XGB), achieving a mean absolute error (MAE) of 82 minutes.

In [8] machine learning (ML) and deep learning (DL) approaches were applied to the ACN-Data charging sessions to predict departure time and energy demand using a non-temporal regression model and a time series-based regression model. In the second case, ML and DL were applied to the same data set to predict arrival time, departure time, and energy absorbed, based on temporal regression modeling.

An emphasis on the data sparsity-to-data entropy ratio is considered in [9] while noting that prediction error increases with higher data entropy or lower data sparsity. Support vector regression (SVR) and random forest (RF) showed good prediction performance when data entropy to sparsity was low. On the contrary, the diffusion-based kernel density estimator (DKDE) showed better performance when the data entropy ratio was high.

In [10], an individual user's charging profile was considered to predict the following features, using long short-term memory (LSTM) networks: when an EV is expected to charge, the expected time slots when the EV will likely be charged, and the number of times the EV will charge during a day. In this paper, a 93.8 percent accuracy was achieved in predicting the charging duration. In a recent paper [11], RF, SVM, XGBoost and CatBoost were used to predict the energy consumption of EV and achieved an accuracy of 87.55, 89.33, 92.89 and 98.37 percent, respectively, using features such as EV charging station, location, session duration, weather, traffic and speed.

Table I summarizes the results of other research papers in

TABLE I
OVERVIEW OF RELATED WORKS.

Parameter	Model	Best result	Reference
Charging duration	ELM,FFNN,SVR	0.99	[5]
Departure time	SVR	0.011	[6]
Departure time	XGBoost	82mins	[7]
Arrival and departure times, energy	RF,SVM,XGB	1.529	[8]
Charging duration, other	SVR,RF, others		[9]
Charging duration,other	RF,SVM,others	50%	[10]
Energy consumption	RF,SVM,XGB	92.89	[11]

this area, indicating the parameter under consideration, the model used, and the best results. It is clear that RF, SVR and XGBoost are popular in the prediction of EV charging duration. These models were adopted in this paper to facilitate comparison with the results of other works in this domain. The difference between this work and the ones listed above, is that the focus is to determine whether a good prediction accuracy can be achieved by using only the characteristics recorded in the charging sessions, without adding external features.

In this paper, the following features are considered to predict EV charging duration (in hours) at three public parking sites: total power delivered to the EV (kWh), charging site, start date and time, end date and time. EV charging sessions are collected from three charging parks (sites), and each charging park contains multiple charging points, described by a unique identification label. Accurate prediction of EV charging duration enables charging park operators to optimize charging schedules in case charging stations become busy and it enables proper planning of charging infrastructure in the future.

The remainder of the paper is organized as follows. Section II describes the machine learning models adopted in this work, the data set used, and carries out an exploratory data analysis. Section III provides results for the three models and discussion, and Section IV gives the concluding remarks.

II. METHODOLOGY

This section includes a description of the models selected for the analysis, of the data sources, and then provides a short outline of the data characteristics. Finally, an overview of the metrics adopted in this work is provided.

A. Models

The Machine Learning (ML) models utilized in this paper are Random Forest (RF), Support Vector Regressor (SVR), and Extreme Gradient Boosting (XGB). This section provides a brief overview of the models, and references for further reading.

1) Random Forest (RF): The predictions of several decision trees are combined in the ensemble learning technique known as RF. Each tree is built separately, and the average of all predictions made by each tree, or the majority vote, is used to determine the final prediction. The RF model works well at managing the intricate relationships between the data, while

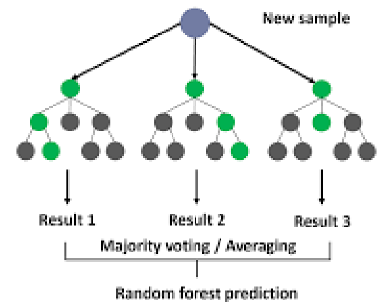

Fig. 1. Random Forest, [12].

preventing overfitting. Additionally, because of its flexibility, numerous applications exist [12]

The forecasting mechanism, $P(x)$, is given by:

$$P(x) = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

where $T_i(x)$ represents the prediction of the i-th tree for input x. [15].

The model decision-making process can be understood through graphical representations of decision trees and RF aggregation processes, as shown in Fig. 1.

2) Support Vector Regression (SVR): A regression method based on support vector machines is called SVR. With a margin of error, SVR seeks to identify the hyperplane that most accurately depicts the relationship between input features and output. This model is particularly helpful for capturing non-linear relationships in the data [13].

The SVR model involves solving a quadratic programming problem to find the optimal hyperplane parameters. The prediction function is given by:

$$P(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

where α_i and α_i^* are Lagrange multipliers, $K(x_i, x)$ is the kernel function and b is a bias term [16].

A visualization of the SVR hyperplane in feature space and the impact of support vectors is shown in Fig. 2.

3) Extreme Gradient Boost (XGB): The fast and effective gradient-boosting algorithm XGBoost is well known for its efficiency. XGB orderly constructs a sequence of decision trees, each of which fixes the mistakes of the one before it [14].

The XGBoost model involves optimizing an objective function, which combines the loss function and a regularization term. The total of all the trees' predictions yields the final forecast.

$$P(x) = \sum_{i=1}^{N} f_i(x) + b$$

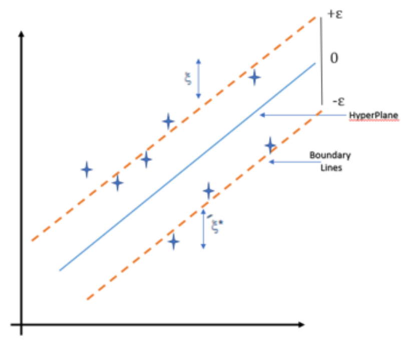

Fig. 2. Support Vector Regression (SVR) Model, [13].

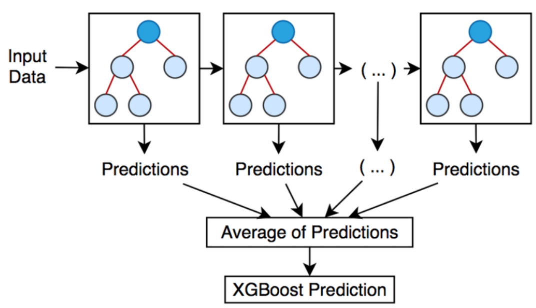

Fig. 3. XGBoost Decision Trees, [14].

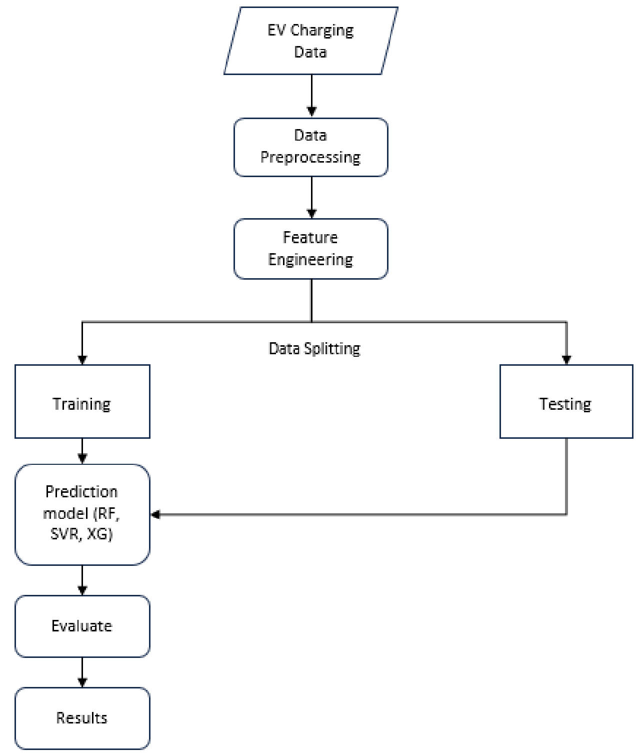

Fig. 4. ML Prediction workflow

where $f_i(x)$ represents the prediction of the i-th tree for input x, and b is a bias term. [17].

The decision trees are displayed graphically within the XGBoost framework and each tree's contribution to the final prediction is illustrated in Fig. 3.

4) Workflow adopted in this work: Fig. 4 shows the workflow adopted in this work, which follows a conventional procedure used in ML applications. EV charging data are selected and pre-processed, including cleaning, transformation and visualisation. Then, the data is fed into the three ML models described above, one at a time, with the aim of making predictions on different features. A comparative analysis of the RF, SVR, and XGB models, considering their strengths and weaknesses, will be presented. A detailed description of this workflow is the subject of the next two sections.

B. Data source and features

Several open EV data charging session data are available in the UK. These include EV charging transactions from Dundee, London Barnet, Leeds, Perth-Kinross and My Electric Avenue EV project.

The Leeds data was chosen because it provides a unique user identification, charging point identification, and names of charging sites. This allows an easier analysis of the charging behaviour [18]. For this data set, EV charging transactions were recorded in Leeds from 2 January 2019 to 31 December 2019 at four sites. One of the sites was discarded because it was a test site and not a regular EV charging site. The remaining sites analyzed in this work are: Wood Lane, Elland Road, and Temple Green.

The prediction of charging duration is modeled as a regression problem, starting with feature engineering. The prediction accuracy of the target feature (i.e., charging duration) is improved by adding more features to the model.

Table II shows all features of the Leeds data sessions. For this work, the following features were selected: total kWh, Site, Charging duration hours, Start and end date and time, and start and end day. Other features were discarded: Event ID, User ID, CP ID, connector and model. Columns were added for the days when EV users plug in their cars, and for the days when they unplug since some users charge overnight to the next day.

Overnight charging introduces an issue related to the cyclical nature of time-related features. In a standard data series, features are presented on a linear scale, and this means that cyclical features may appear distant in time even though they repeat cyclically. Sine and cosine transformations are well-established data processing techniques that allow capturing and taking into account the cyclical nature of the data [19].

C. Exploratory Data Analysis

A total of 7271 charging events are included in the data set, out of which 213 are unique users at Wood Lane, 205 at Elland Road, and 85 at Temple Green. In the context of this work, 'unique users' refers to those EV users who have

TABLE II
FEATURES IN THE LEEDS EV DATA SET.

Features	Description
Event ID	Unique charging session identifier
User ID	Unique EV user identifier
CP ID	Unique charging point identifier
Connector	Type of charging port
Model	Type of charger based on charging
Start date and time	Date and time when plugged in
End date and time	Date and time when plugged out
Total kWh	Amount of energy delivered to the car
Site	Name of charging site
Charging Duration hrs	Time taken to charge the EV in hours
Start day	Day when the car was plugged in for charging
End day	Day when the car was plugged out from charging

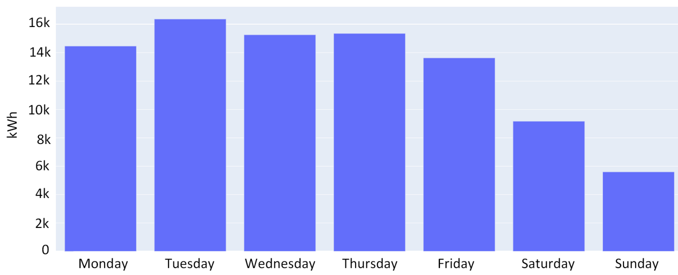

Fig. 5. Cumulative weekly usage for the year 2019.

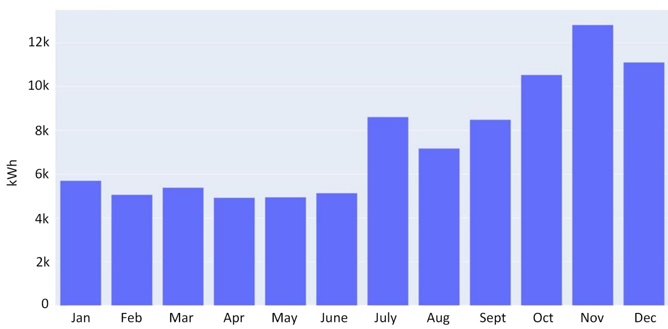

Fig. 6. Monthly usage for the year 2019.

plugged in their cars for charging in one of the charging points out within each park and ride site. These unique users include regular users and occasional users.

Table III shows some of the main site statistics. Woodhouse Lane is the busiest site, with the highest amount of users and charging. Elland Park has a similar number of users, but about half charging sessions. The average charging (kWh) per site is similar.

The power required to charge EVs is at higher between 12 pm to 6 pm, with a peak at 5 pm. There is a uniform charging trend from Monday to Friday, with a drop on the weekend, as shown in Fig. 5. Monthly trend shows a pick up in usage from July (Fig. 6 monthly. Seasonal trends are especially important when weather conditions and variables are selected as features to make predictions. However, this is beyond the scope of this paper.

TABLE III
SITE STATISTICS

Site name	Unique users	Total Charging (kWh)	Average Charging (kWh)
Woodhouse lane	213	55,000	12
Elland road	205	20,000	13
Temple green	85	10,000	10

D. Evaluation metrics

Three evaluation metrics were considered: mean absolute error (MAE), root mean square error (RMSE), and the co-efficient of determination (R-squared). These metrics will be briefly described below.

MAE measures the deviation between the actual value and the predicted value and is mathematically given by the formula:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|\text{predicted} - \text{actual}|$$

MAE is equivalent to the number of charging hours, making it one of the most straightforward metrics to interpret.

RMSE is the square root of the mean of the difference between the predicted and actual error value and is expressed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}|\text{predicted} - \text{actual}|^2}$$

The coefficient of determination, known as R squared (R^2), measures how well the model fits the data. R^2 ranges from 0 to 1, where 0 means that the model does not fit at all, and 1 means that the model entirely fits well with the data and is calculated as

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(\text{predicted} - \text{actual})^2}{\sum_{t=1}^{n}(\text{predicted} - \text{mean})^2}$$

III. RESULTS AND DISCUSSIONS

The data was split into 80 per cent (10 months of data) for training, and 20 percent (2 months of data) for testing. The models described in Section 2 were trained using the data, and the forecasted feature was charging duration for all sites. For each model and each site, the performance indicators described in the previous sections were calculated.

Table IV, Table V and Table VI show the results for XGB, SVR and RFR, respectively. In general, the models perform well and this demonstrates that they are adequate for the task of predicting EV charging time, for the selected dataset.

More in detail, the following comments apply:

- XGB (Table IV) is overall the best model, showing good performance at all sites. The MAE and RMSE are the

TABLE IV
RESULTS USING XGB.

Site	MAE	RMSE	R^2
Woodhouse lane	0.440	1.270	0.938
Elland road	0.486	1.860	0.837
Temple green	0.310	0.380	0.971

TABLE V
RESULTS USING SVR.

Site	MAE	RMSE	R^2
Woodhouse lane	3.21	4.32	0.290
Elland road	0.344	2.41	0.726
Temple green	0.0351	0.0417	0.999

TABLE VI
RESULTS USING RFR.

Site	MAE	RMSE	R^2
Woodhouse lane	0.358	1.19	0.946
Elland road	0.542	2.22	0.768
Temple green	0.189	0.334	0.977

lowest, while R^2 is the highest. As explained in the previous session, the closer R^2 is to unity, the better the model performs.

- SVR (Table V) has the best performance for the Temple Green charging site. The prediction performance at Woodhouse Lane is not very good with an MAE of 3.21, an R^2 of 0.290 which indicates that the model does not fit well with the data, but still an improvement compared to some results shown in the literature [20].
- RFR (Table VI) performs best for the Temple Green Park and Ride charging area with an MAE of 0.189 per cent, RMSE of 0.334, and R^2 of 0.977.
- Overall, the predictive performance at the Temple Green site is the best because its data has fewer outliers and has many identical users frequenting the site, and fewer random users than the other two charging sites. For example, a user with a unique user ID of 279 frequented the Temple Green charging site about 158 times from 1 May 2019 to 10 December 2019 and charged on average from around 10 am to 6 pm.

IV. CONCLUSION

This paper presented an approach to predict EV charging during using a UK dataset and three ML models. For this research, 2019 charging data from three public charging sites in Leeds were considered.

An overview of the ML models has been presented, and the main features of the dataset have been described. The performance of the models was evaluated by using three metrics: MAE, RMSE and R^2.

In this work, EV charging behaviour trends and patterns were drawn from the charging transactions based on: charging point usage, unique user usage, and hourly, weekly, monthly, and seasonal usage. The result indicates that excellent performance can be obtained by using the features included in the dataset, without the need to add other features from external sources, such as weather data.

All three ML models used are comparable in terms of predictive performance, with XGB providing overall best results. SVR performed poorly at one of the sites. It was observed that the type of data used and the amount of regular user behaviour also determine how well a model performs.

From this study, the use of features within the EV charging data was sufficient in achieving a good predictive performance of the charging duration, because of the frequent and identical EV users. However, other datasets, where the charging data have a more random pattern of user behaviour, will require additional features. These features may include weather data, because weather affects when and how frequently EV users charge their vehicle.

With a growing amount of EVs deployed in the UK, the importance of accurately forecasting charging patterns is growing. This paper aims to provide guidelines on models to be used for this type of study, the main features to be considered and the availability of datasets.

In this work, only the charging duration was considered as the target feature, but other factors such as the charging start and charging end will be considered in future works. Additionally, residential charging will be addressed.

ACKNOWLEDGMENT

This work is supported by the Petroleum Technology Development Fund (PTDF), Nigeria.

REFERENCES

[1] "Enabling the transition to electric vehicles - Ofgem," Enabling the transition to electric vehicles, https://www.ofgem.gov.uk/sites/default/files/2021-09/Enabling
[2] "Electric vehicles," ESO, https://www.nationalgrideso.com/future-energy/our-progress-towards-net-zero/net-zero-explained/electric-vehicles (accessed Dec. 12, 2023).
[3] Zachary J. Lee, Tongxin Li, and Steven H. Low. 2019. ACN-Data: Analysis and Applications of an Open EV Charging Dataset. In Proceedings of the Tenth ACM International Conference on Future Energy Systems (e-Energy '19). Association for Computing Machinery, New York, NY, USA, 139–149. https://doi.org/10.1145/3307772.3328313
[4] M. Nazari, A. Hussain, and P. Musilek, "Applications of clustering methods for different aspects of electric vehicles," MDPI, https://www.mdpi.com/2079-9292/12/4/790 (accessed Dec. 12, 2023).
[5] Irfan Ullah, Kai Liu, Toshiyuki Yamamoto, Md Shafiullah & Arshad Jamal (2023) Grey wolf optimizer-based machine learning algorithm to predict electric vehicle charging duration time, Transportation Letters, 15:8, 889-906, DOI: 10.1080/19427867.2022.2111902
[6] Z. Xu, "Forecasting electric vehicle arrival and departure time on UCSD campus using Support Vector Machines," eScholarship, University of California, https://escholarship.org/uc/item/1z43n3km (accessed Dec. 12, 2023).
[7] O. Frendo, N. Gaertner and H. Stuckenschmidt, "Improving Smart Charging Prioritization by Predicting Electric Vehicle Departure Time," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 10, pp. 6646-6653, Oct. 2021, doi: 10.1109/TITS.2020.2988648.
[8] S. Khan, B. Brandherm and A. Swamy, "Electric Vehicle User Behavior Prediction using Learning-based Approaches," 2020 IEEE Electric Power and Energy Conference (EPEC), Edmonton, AB, Canada, 2020, pp. 1-5, doi: 10.1109/EPEC48502.2020.9320065.

[9] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, “Ensemble machine learning-based algorithm for electric vehicle user behaviour prediction,” Applied Energy, vol. 254, p. 113732, Nov. 2019, doi: https://doi.org/10.1016/j.apenergy.2019.113732.

[10] A. S. Khwaja, B. Venkatesh and A. Anpalagan, ”Short-term Individual Electric Vehicle Charging Behavior Prediction Using Long Short-term Memory Networks,” 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Pisa, Italy, 2020, pp. 1-7, doi: 10.1109/CAMAD50429.2020.9209296.

[11] S. Salam, G. K. Veni, C. Kamale, A. L. Deepika and A. Bruhathi, ”Forecasting the Behavior of EV Charging Using Machine Learning,” 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 1765-1768, doi: 10.1109/ICACCS57279.2023.10112731.

[12] IBM, “What is Random Forest? — IBM,” 2023. [Online]. Available: https://www.ibm.com/topics/random-forest.

[13] Great Learning Team, ’Support Vector Regression in Machine Learning. What is SVM?,’ Great Learning Blog: Free Resources what Matters to shape your Career!, Sep. 2020. [Online]. Available: https://www.mygreatlearning.com/blog/support-vector-regression/.

[14] “What is XGBoost?,” NVIDIA Data Science Glossary. [Online]. Available: https://www.nvidia.com/en-us/glossary/data-science/xgboost/:~:text=XGBoost%2C%20which%20stands%20for%20Extreme.

[15] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324.

[16] Y. Grandvalet, F. d’Alché-Buc, and C. Ambroise, “Boosting Mixture Models for Semi-supervised Learning,” *Lecture Notes in Computer Science*, Jan. 2001, pp. 41-48, doi: 10.1007/3-540-44668-07. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-44668-0_7.

[17] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Jun. 2016. [Online]. Available: https://arxiv.org/pdf/1603.02754.pdf.

[18] L. C. Council, Electric vehicle charge points operated by Leeds City Council., https://www.data.gov.uk/dataset/5bb5c097-0e2f-42a3-8aee-1c0189a39082/electric-vehicle-chargepoints-operated-by-leeds-city-council (accessed Dec. 12, 2023).

[19] “Luke-ZM - Overview,” GitHub, https://github.com/Luke-zm (accessed Dec. 13, 2023).

[20] Irfan Ullah, Kai Liu, Toshiyuki Yamamoto, Md Shafiullah & Arshad Jamal (2023) Grey wolf optimizer-based machine learning algorithm to predict electric vehicle charging duration time, Transportation Letters, 15:8, 889-906, DOI: 10.1080/19427867.2022.2111902