National College of Ireland

A Machine Learning Framework to identify the Optimal location for Electric Vehicle Charging Stations

MSc Research Project
Data Analytics

Nayan Sharma
Student ID: X21167818

School of Computing
National College of Ireland

Supervisor: Prof. Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nayan Sharma
Student ID:	X21167818
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Prof. Paul Stynes
Submission Due Date:	15/12/2022
Project Title:	A Machine Learning Framework to identify the Optimal location for Electric Vehicle Charging Stations
Word Count:	4378
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Framework to identify the Optimal location for Electric Vehicle Charging Stations

Nayan Sharma
x21167818

December 2022

Abstract

Electric Vehicles are a next-generation mode of transportation. An Electric Vehicle charging station (EVCS) is a place that supplies electrical power to the electric vehicle. It is also known as a charge point or electric vehicle supply equipment (EVSE). Electric Vehicle charging stations (EVCS) are required to meet future demand and to solve vehicle charging availability and compatibility issues. However, the challenge is to identify the optimal location for the electric vehicle charging station. This research proposes a framework to predict the optimal location to install Electric Vehicle charging stations. The proposed framework combines machine learning clustering and the regression model with the spatial model. Data sets containing location postcodes with energy utilization, footfall and passerby, car ownership, and traffic details are used to train the machine learning clustering and the regression model. The Spatial model is built to plot the geographical map and point to the optimal location. An experiment is conducted to identify the optimal location. Results demonstrate Linear regression is the best Machine learning regression model to predict the results based on accuracy and loss. All of the models presented in this paper are evaluated based on their accuracy and loss. The findings from this research highlight a generalized model that enables Distribution Network Operators (DNOs) to identify optimal locations and to plan infrastructure for EVs as excessive electrical power requirements caused by EV integration may have a negative impact on the distribution network.

1 Introduction

The market for electric vehicles (EV) is expected to grow by 70% by Gov.uk (2022), which is faster than the market for traditional cars. Distribution Network Operators(DNOs) are responsible for the distribution of electricity from the transmission grid to most end users. The DNOs need to plan better to fill the demand-supply requirements for transitioning to Electric vehicles Brinkel et al. (2020). An in-depth understanding of Electric Vehicle charging station networks and the current pattern of electricity usage is required which may lead to the optimal placement of charging stations. Higher energy utilization of the location indicates a higher possibility of EVCS setup. The transition to EVs will increase the demand for publicly accessible charging stations, and DNOs must plan a well-structured EV charging network across the United Kingdom. The machine learning model framework can help DNO to plan the network efficiently.

The aim of this research is to investigate to what extent a Machine learning framework can predict the optimal location to install Electric Vehicle charging stations. To address the research question, the following specific sets of research objectives were derived:

The first objective is to investigate the state of the art broadly around various models' implementation to find the optimal location. Design a framework that combines Machine learning clustering and a regression model with the Spatial model to answer the research question. Implement the framework which includes machine learning clustering and regression model with the spatial model. The evaluation of this framework is based on the accuracy and loss of machine learning models. The major contribution of this research is a novel framework that combines Machine learning clustering and a regression model with the Spatial model that assists operators in finding the optimal installation location.

This paper discusses the state of the art around machine learning clustering and the regression model with the Spatial model implementation to find optimal location in section 2 related work. In the following section, the research methodology is explained. Discussion on the design components for the machine learning framework and its implementation is explained comprehensively in sections 4 and 5, respectively. Section 6 presents and discusses the evaluation results. The conclusion and future work are given in the last section of the paper.

2 Related Work

Finding an optimal location for an Electric Vehicle charging station plays a significant role in the charging process as actual costs related to the construction of stations and its impact on the electricity load distribution system. *Sustainability* (Tai-Ran Hsu - Nov 2013) discussed the sustainability of the electric power required to drive Electric Vehicles. This research represents a comprehensive assessment showing that rapid adoption of EVs will result in a tremendous increase in demand for electric power generation, far exceeding what the United States electric power generating industry is now capable of providing. Furthermore, if we keep using the same technology to generate most of our electricity from fossil fuels, this demand will have severe negative effects on the environment. The significant amount of electricity needed to generate batteries that power EVs and the negative repercussions related to recycling used batteries are two further realities about the program's impacts that are rarely accounted for. Research about the best size and placement of EV charging infrastructures has been developing in parallel with the expanding EV market Csonka & Csiszár (2017), Jia et al. (2012). Ahmad et al. (2022) discussed the economics of charging infrastructure, with a particular emphasis on the actual costs associated with the building of stations.

Catalbas et al. (2017) discussed optimization algorithms to find the optimum locations in Ankara, the capital of Turkey. The average number of EVs on the road and the average range are examined to design the model. The spectral clustering method is implemented to find optimal EV charge locations. The result of this research suggests that the optimal location for the fast-charging station is high-density areas of residential housing. Traffic information plays a significant role in finding the optimal location. However, the accuracy of finding locations can be improved by considering other datasets related to energy utilization de- tail per postcode, footfall, passersby information, and EV car ownership. Traffic count only considers the moving vehicle information around the locations, but these datasets help to determine the consumption of electricity which may impactthe output to find the optimal location.

The best places for public EV charging stations have been identified by Chen et al. (2013). From the Washington, Puget Sound, and Regional Council's Household Activity Survey, data on more than 30,000 trips were used to decide where parking lots should be put. More the demand for parking lots more the charging point requirement. Regression equations can be used to estimate parking demand variables based on how easy it is to get to a site, how many jobs are in the area, how many people live there, and what the trip is like (i.e. total

vehicle hours per zone or neighborhood and parked time per vehicle trip). The Regression algorithm is set up in a way that minimizes the costs for EV users to get to a station. The research presented here gives readers a starting point for planning for parking demands and optimizing the placement of electric vehicle charging infrastructure in unfamiliar environments or under varying constraints (on access costs and stationavailability).

Phonrattanasak & Leeprechanon (2014) look at two primary goals for locating fast charging stations along residential electric power distribution networks: 1. fast charging station construction. 2. Distribution grid transmission line loss. The Ant colony algorithm is implemented to determine the residential neighborhood with traffic and power system security constraints a fast charging station should be installed. The result of this paper suggests that the first fast charging station should be located nearest the first bus. The quick charging stations are all situated in highly populated regions. The transmission line loss of the power distribution grid is significantly impacted by the power usage of rapid charging stations.

Roy & Law (2022) discussed the complex interactions between social, economic, and demographic factors that may be causing these new differences in EVCS placements. various complex interactions between social, economic, and demographic factors are likely contributing to the growing chasm in the number of electric vehicle charging stations across the United States. The difficulty in locating and installing charging infrastructure for electric vehicles (EVs) is likely exacerbated by several factors, including Social resistance to EV ownership, especially in rural and suburban areas. - Limited awareness of and interest in EVs among consumers. Limited charging infrastructure (both public and private) in many parts of the United States. Lack of incentive programs or regulatory supportfor the development of EV charging infrastructure.

Hamada & Naizabayeva (2020) discuss how the KMeans Clustering Algorithm can be used to find the best store location based on social network events. The aim is to develop a decision-support system to help the store manager to find the best store location. Data is extracted from social media platforms using the "Octoparse API" as a web data extraction tool and the K-means algorithm is used to cluster the same location with higher business profit and help users make better and more accurate decisions, which reduces the chances of making bad business decisions and cuts down on business losses. A similar idea has been implemented in this research to cluster the same location with energy consumption and help tofind the optimal location for EVCS.

In conclusion, the state of the art is broadly around the implementation of clustering and regression models with the density of traffic. This art can be improved by considering other features like footfall, passersby, car ownership, and energy utilization per postcode. The idea is to predict the energy utilization of postcodes using traffic count, footfall, passerby, and car ownership datasets. Higher energy utilization, higher the chances of Electric vehicle charging station

utilization. To find the best regression model Logistic regression, K Neighbours, Decision Tree, Random Forest, Gradient Boosting, and Ada-boost regression Machine learning models have been built and compared to predict the energy utilization per postcode. There is a need for a dynamic model that can input latitude and longitude information and plot the geographical location on the map with different colors for each location. This can help Distribution Network Operators (DNOs) to identify optimal locations by looking at the image provided by the Spatial model so that they can facilitate better infrastructure for EVs and improve EV adoption. The primary goal of DNOs is to support the United Kingdom government's goal towards net zero with the end of the sale of new petrol and diesel cars by 2030. To achieve this goal government needs better infrastructure for Electric Vehicle charging stations and promote Electric Vehicle adoption.

3 Methodology

The five steps of the research methodology are data collection, data pre-processing, data transformation, data modeling and conversion, evaluation, and results shown in Figure 1.

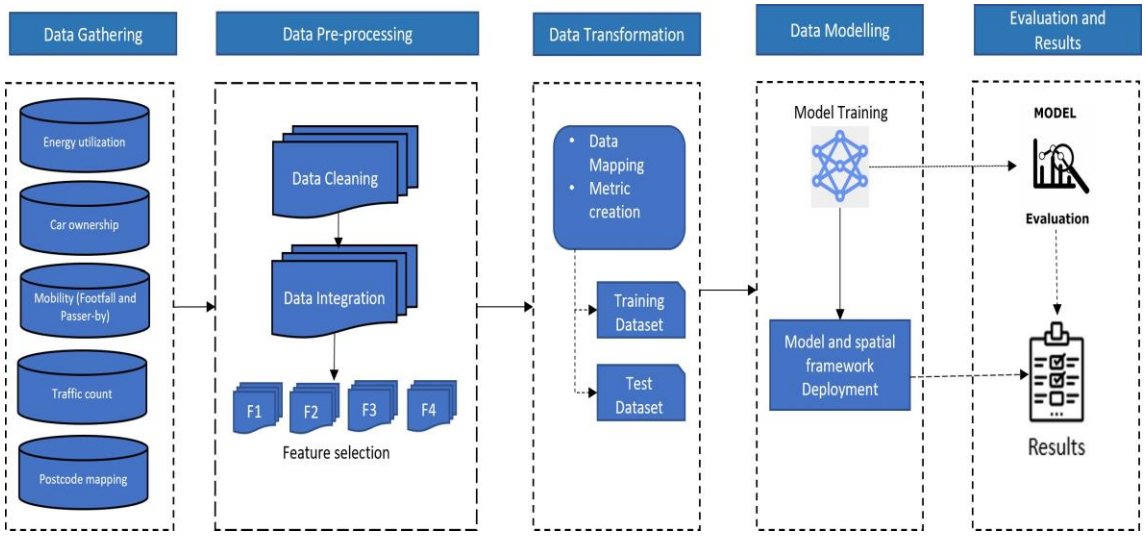

Figure 1: Research Methodology

The first step, Data gathering involves collecting seven datasets.

1. UK_postcode_dataset.csv(*uk-postcode-electricity-consumption* (2022))
2. UK_Car_ownership_dataset.csv (Transport (2022))
3. Mobility_Footfall_dataset.csv(*Uk Mobility Footfall* (2022))
4. Mobility_passerby_dataset.csv(*Uk Mobility passerby* (2022))
5. UK_traffic_count.csv(*Traffic Count data.gov.uk* (2022))
6. UK_Postcode_mapping.csv (*uk postcode-NSPL* (2022))
7. UK_Outputareas_with_population.geojson (*uk postcodegeojson* (2022))

UK_postcode_dataset.csv dataset is from the official UK government website and contains 0.8 million records with location postcode, number of meters, utilization(kWh), Mean utilization (kWh), and Median utilization (kWh). UK_Car_ownership_dataset.csv dataset represents the number of the total car owned in the United Kingdom per quarter per year per postcode in 2020 and 2021. Anonymized data is generated for UK mobility data which contains two datasets: Mobility_Footfall_dataset.csv and Mobility_passerby_dataset.csv as per the postcode for the year 2020-2021 with columns year, month, date, age_band, gender, dwell_hrs, distance_home_km_range, dwell_hrs_range, footfall, passerby. UK_traffic_count.csv contains the yearly traffic count in the United Kingdom postcode for the year 2020-2021. UK_Postcode_mapping.csv has postcode mapping to latitude and longitude which will help in the Spatial model to plot the geographical map. The UK_Outputareas_with_population.geojson file contains the geometry to plot the United Kingdom geography for the Spatial Model.

The second step, Data pre-processing involves data cleaning, and the removal of null and outliers to insure clean datasets. An outlier is a piece of data that is far away from the rest of the data in the set. Outliers are identified using the box plot for the energy utilization, car ownership count, traffic count, total footfall, and total passersby data as shown in figure 2 which may impact the model training. Outliers can be removed using the Interquartile Ranges method. All the datasets were analyzed individually to check for bad data which contains null values and junk characters in the postcode. All the irrelevant columns like id, latitude, and longitude are removed to select the features from individual datasets. After data cleansing, data integration has been performed to create a uniform dataset for feature selection.

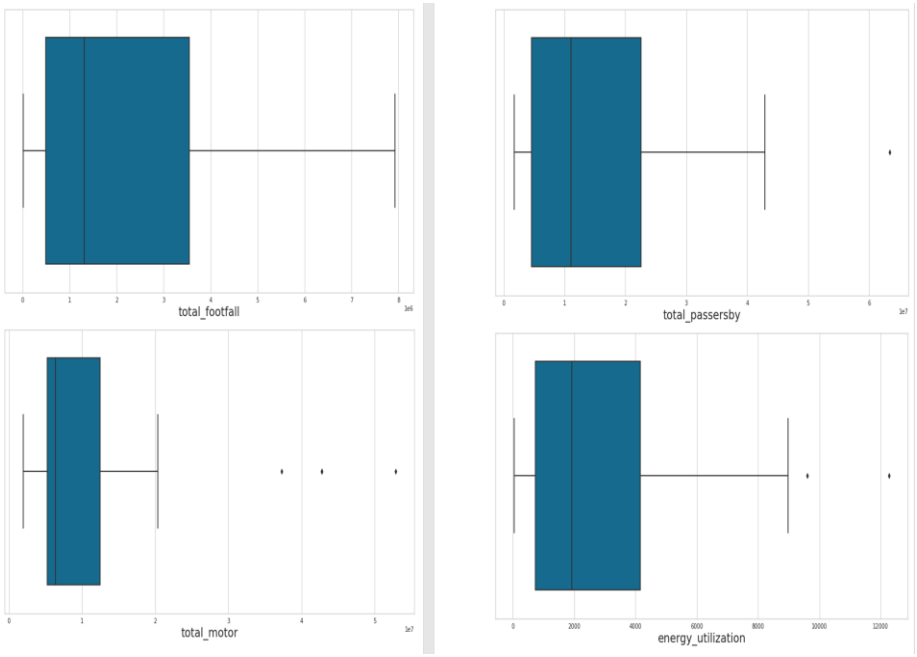

Figure 2: Box Plot

The third step, Data transformation involves data mapping, one hot encoding of categorical features, and metric creation. All the datasets are combined based on the postcode monthly level and one hot encoding is performed on categorical features. All the numerical columns have been scaled so that the feature columns look like a normal distribution. This makes it easier to learn the weights for the machine-learning algorithm. The further dataset was split into a ratio of 80:20 for training and testing.

The fourth step, Data modeling involves model training and spatial model deployment. The cleaned dataset is split into train and test datasets. The training dataset is used to train seven machine learning models: KMean clustering, Logistic, K Neighbours, Decision Tree, Random Forest, Gradient Boosting, and Ada-boost regression. All these models are regression and supervised machine learning models except Kmean clustering which is an unsupervised learning algorithm where dataset values are assigned to the group of clusters based on the Euclidian distance and then the number of iterations tries to partition the data into distinct clusters. All features are reassigned to their nearest cluster centroid as shown in Figure 3 using t-SNE and each data point belongs to one cluster. The K-means clustering algorithm in data mining uses a set of initially randomly selected centers as initialization points for each cluster as it learns how to process the input. The optimal locations for the centers are then determined by performing numerous computations. After training all seven models are fit to test data.

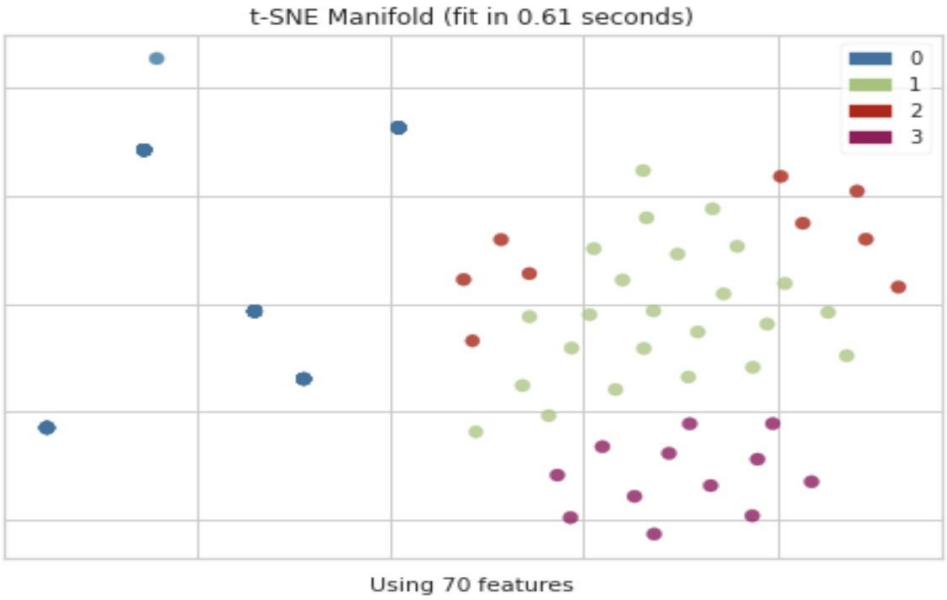

Figure 3: T-SNE Manifold

Output from the trained model is used as an input for the spatial model. The spatial model is deployed using a coordinate reference system(CRS), and UK_Outputareas_with_population.csv file.
Longitude and Latitude need to convert into points so that geometry works out correctly. Now the coordinates, act as a single data point in one convenient place thejob of a Point. The CRS value used in this research is epsg:32630.

The fifth step, Evaluation, and Results involve evaluating the performance of each of the clustering and regression models using accuracy and loss. The optimal model was selected from the experimentation and integrated with the spatialcomponent.

4 Design Specification

The optimal location machine learning framework combines Machine learning clustering and a regression model with the Spatial model as shown in Figure 4. The framework includes captured dataset, machine learning clustering and regression model in section 4.1, and the spatial model discussed in section 4.2.

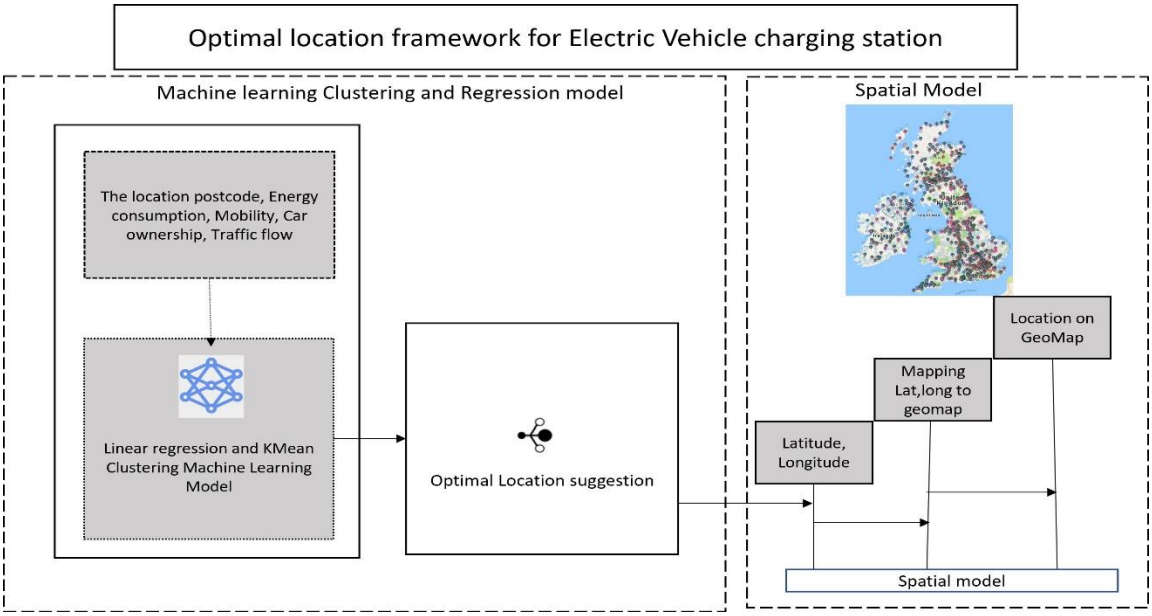

Figure 4: Machine Learning Framework

4.1 Machine learning Clustering and Regression model

In this research KMean clustering and Logistic Regression model have been implemented. The model starts when a data pipeline load seven datasets: UK_postcode_dataset.csv,UK_Car_ownership_dataset.csv , Mobility_Footfall_dataset.csv, Mobility_passerby_dataset.csv,UK_traffic_count.csv, UK_Postcode_mapping.csv, UK_Outputareas_with_population.geojson as input, it will predict the energy utilization for each postcode. The KMean clustering and Logistic Regression model will predict the energy utilization for each postcode and give a rank to each postcode from high to low on the basis of energyutilization. Based on the ranking DNOs can prioritize the postcode to install the Electric Vehicle charging station.

4.2 Spatial Model

The Spatial model is created using the UK_Outputareas_with_population.geojson file. It includes a geographical map showing the high-ranked location with different colors. With the help of a Map and postcode, Distribution Network Operators can decide the optimal geographical location to place the Electrical Vehicle charging station.

5 Implementation

The framework is implemented using Python Notebooks on Databricks Azure cluster with 9.1 LTS (includes Apache Spark 3.1.2, Scala 2.12) configuration. Seaborn-0.12.1 (*seaborn documentation* (2022)) library is used to perform charts and graphs used for analysis. scikit-learn-1.1.3 yellowbrick-1.5 learn library (*scikit-learn documentation* (2022)) is used to implement the Machine learning framework. All the datasets are present in CSV formatted and uploaded to Azure Data Lake Storage Gen2 storage and then, imported into the data frame using PYSPARK and Pandas libraries. After importing data sets into the data frame exploratory data analysis has been performed shown in Figures 5 and 6 to understand the relationship of the target variable with all the independent variables. Figure 5 is the

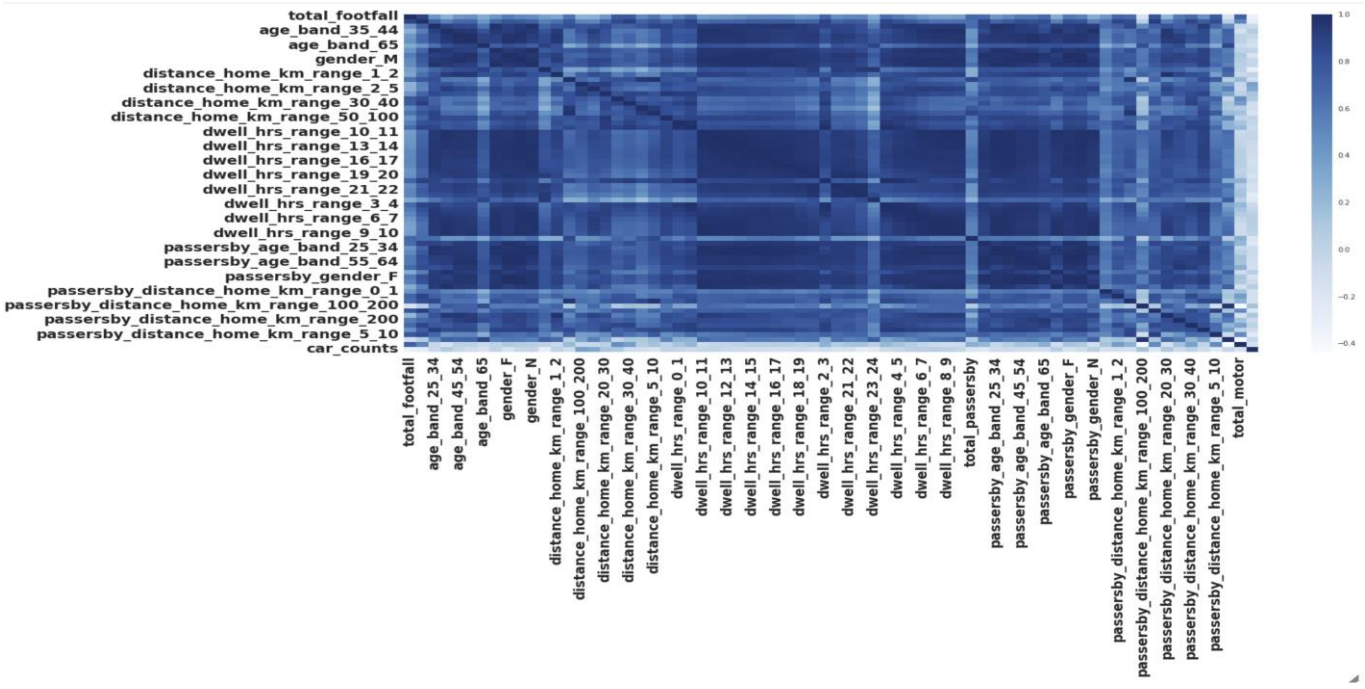

Figure 5: Correlation Matrix

correlation matrix between all the variables present in the dataset. The correlation matrix makes it easy to see if there's a link between two variables. Total footfall is correlated to dwell hours range 10_11, 12_13, and 16_17. Similarly, distance home km range 1_2 and 2_5 is correlated with age band 25_34 and dwell hour

range 2_3. Figure 6 shows the correlation bar between the target variable and all the other independent variables in the dataset. If the correlated value of the independent variable is greater than 0.5 then, the variable has a strong correlation with

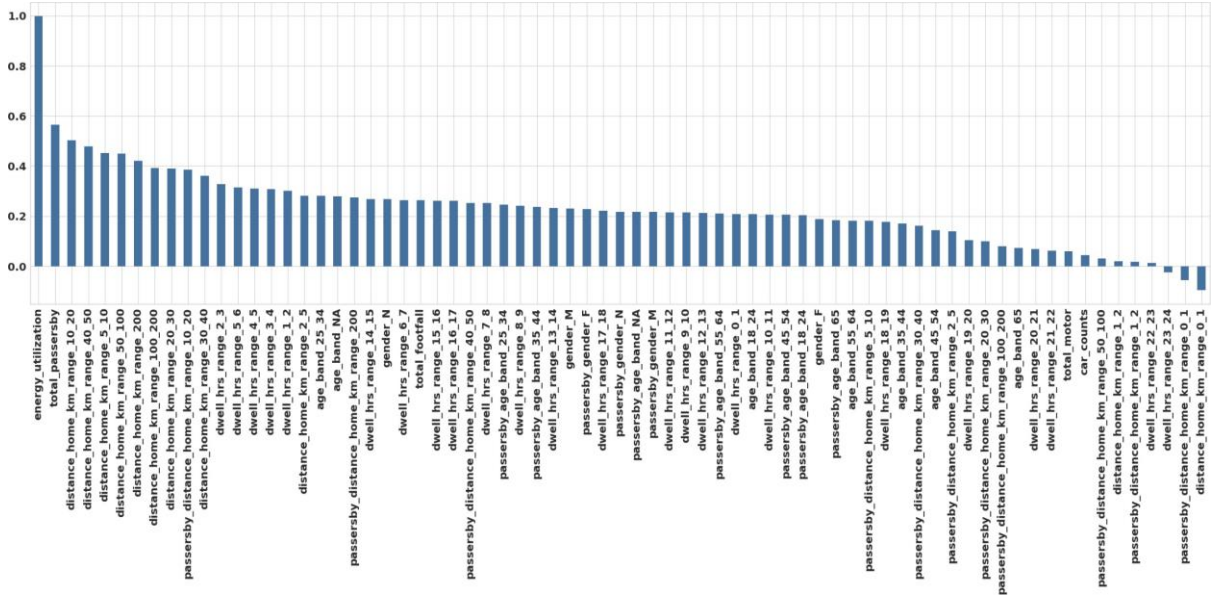

Figure 6: Correlation Bar with Target

the target variable. Total passerby, Distance km range 10_20, Distance km range 40_50, and dwell hours show a promising relationship with the energy utilization target variable. Total missing data and total missing data per column have been identified using isna() method of pyspark.

The spatial model is implemented in a different notebook. The first step is to find out the correct UK Outputareas with population geojson file which is open source and available at Geojson (2022). The matplotlib library is used to make maps, and geopandas is a high-level interface to it. The next thing to do is to acquire the data in a suitable format. To accomplish this, it is necessary to convert a standard Pandas DataFrame into a geo-DataFrame. The output postcodes with rank as per energy utilization received from the Machine learning clustering and a regression model act as input to the Spatial model. The Spatial model plots all the postcodes on the geographical map and give different color to each rank.

6 Result and Discussion

The aim of this experiment is to find the optimal Machine learning clustering and a regression model by comparing the accuracy and loss using the metric R-Squared and Root mean squared error (RMSE). As a part of the first experiment, state of art has been replicated using the total motor count in a postcode with the UK_traffic_count.csv. The R squared from the Linear regression model is 0.3147 and the root mean square error is 0.042 which is further improved by adding otherfeatures to the Linear regression model.

Table: 1 shows the comparison of the Six models: Logistic, K Neighbours, Decision Tree, Random Forest, Gradient Boosting, and Ada-boost regression based on the root mean square error and R square values. This result indicates that the Linear regression model has an R squared of 0.433 which is the highest among all the models. Similarly, Kmean clustering with, K equal to three gives the best results and it will predict the outputwith greater efficiency.

Table 1: Model Comparison

Model	RMSE	R Squared
Linear Regressor	0.17322	0.433643
K Neighbours Regression	0.196496	0.424461
Decision Tree Regressor	0.221557	0.123884
Random Forest Regressor	0.181933	0.234373
Gradient Boosting Regressor	0.189542	0.241024
Adaboost Regressor	0.178321	0.064220

The Root Mean Squared Error (RMSE) is a statistic that measures the accuracy with which a regression model predicts the absolute value of the response variable. The RMSE is a good indicator of the accuracy of a regression model and can be used to identify any problematic aspects of the regression model. R-squared indicates how well a model can predict the percentage value of the response variable. Higher values of R-squared indicate that the model is better at predicting outcome values than random chance would suggest. Contrary to the supervised learning model, KMean model performance evaluation is done in a different way. It does not learn from the data as it requires K as input. This research uses two metrics: 1. within-cluster sum squared 2. Silhouette Method to find the value of K. Elbow method is based on the sum of the squared distance between data points and cluster centroid. Look for the value of K where the elbow formed

as shown in Figure 7. Silhouette uses a degree of separation between clusters. The optimal number of clusters is identified as three.

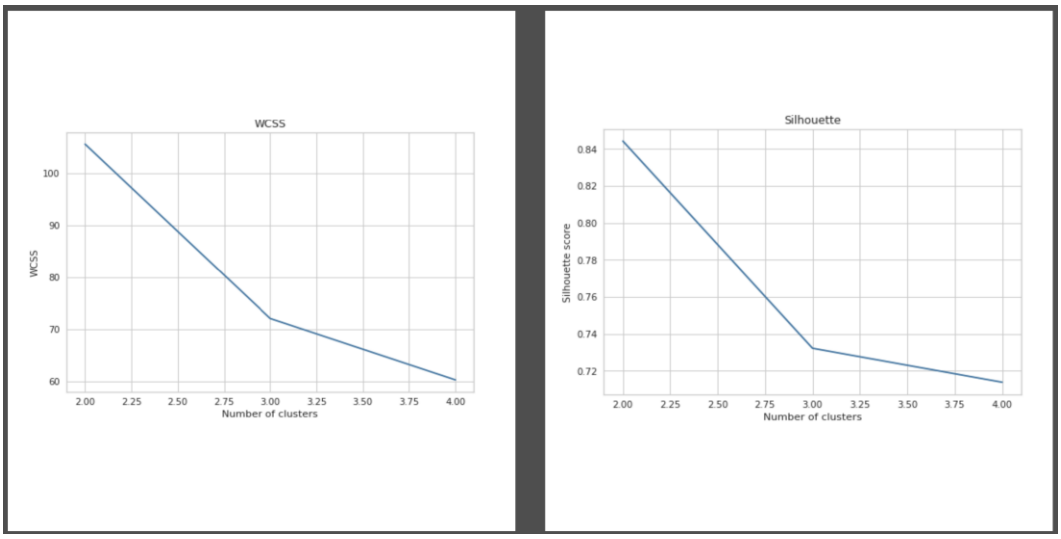

Figure 7: KMean Cluster Size Metrics

The result from the Spatial model is shown in Figure 8. Output is geographical map that shows all the postcodes plotted at different locations as per the latitude and longitude information provided in the dataset. Different Color is assigned to each venue as per the rank provided by the Machine learning clustering and a regression model.

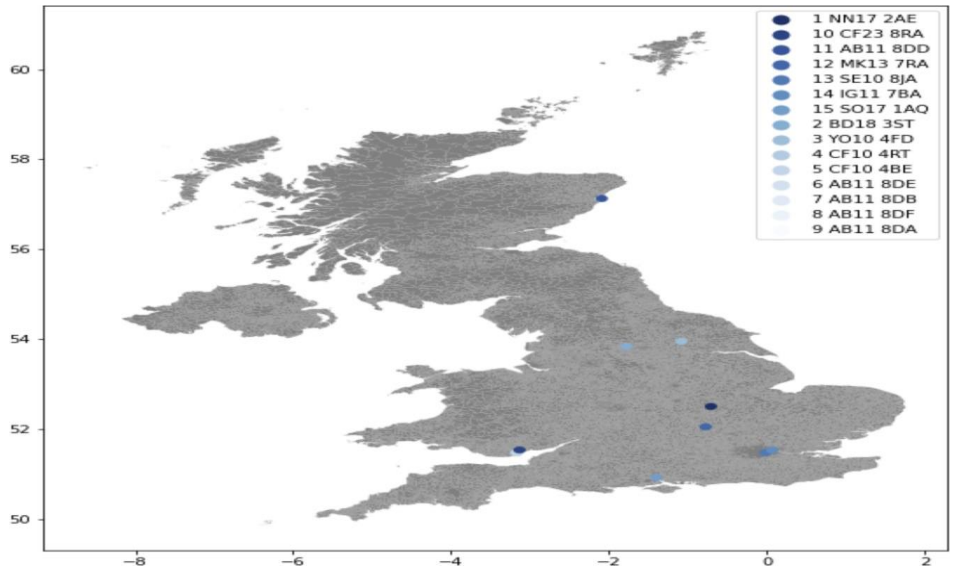

Figure 8: Spatial Model output

7 Conclusion and Future Work

The aim of this research was to is to investigate to what extent a Machine learning framework can predict the optimal location to install Electric Vehicle charging stations. This research proposes a framework that combines Machine learning clustering and a regression model with the Spatial model. Results with R squared 0.433643 and RMSE 0.17322 demonstrate that Linear regression shows promise if the motivation is for accuracy and KMean clustering with the number of clusters three. Results indicate that the state of the art can be improved by including other features like footfall, passersby, and car ownership. This study is limited by the size of the dataset and the low accuracy of the algorithm which can be improvedby training the model with more volume of datasets.

This research can potentially help Distribution Network Operators(DNOs) to identify optimal locations and facilitate infrastructure for EVs. This work can be enhanced by optimizing the models and extensive research can be conducted on this work by adding more features, such as demographic factors and the nearest popular locations, as inputs to the model.

References

Ahmad, F., Iqbal, A., Ashraf, I., Marzband, M. & khan, I. (2022), 'Optimal location of electric vehicle charging station and its impact on distribution network: A review', *Energy Reports* **8**, 2314–2333.
URL: *https://www.sciencedirect.com/science/article/pii/S2352484722001809*

Brinkel, N., AlSkaif, T. & van Sark, W. (2020), The impact of transitioning to shared electric vehicles on grid congestion and management, *in* '2020 International Conference on Smart Energy Systems and Technologies (SEST)', pp. 1–6.

Catalbas, M. C., Yildirim, M., Gulten, A. & Kurum, H. (2017), Estimation of optimal locations for electric vehicle charging stations, *in* '2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE

Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe)', pp. 1–4.

Chen, T. D., Kockelman, K. & Khan, M. (2013), 'Locating electric vehicle charging stations', *Transportation Research Record: Journal of the Transportation Research Board* **2385**, 28–36.

Csonka, B. & Csiszár, C. (2017), 'Determination of charging infrastructure location for electric vehicles', *Transportation Research Procedia* **27**, 768–775. 20th EURO Working Group on Transportation Meeting, EWGT 2017, 4-6 September 2017, Budapest, Hungary.
URL: *https://www.sciencedirect.com/science/article/pii/S2352146517310128*

Geojson, U. (2022), 'Detailed statistics about vehicle licensing and registered vehicles in the united kingdom.'.
URL: *https://statistics.ukdataservice.ac.uk/dataset/2011-census-geography-boundaries-middle-layer-super-output-areas-and-intermediate-zones*

Gov.uk (2022), 'Sales of electric vehicles reach an all-time high while uk boasts one of the most extensive networks of rapid chargers in europe.'.

Hamada, M. A. & Naizabayeva, L. (2020), Decision support system with k-means clustering algorithm for detecting the optimal store location based on social network events, *in* '2020 IEEE European Technology and Engineering Management Summit (E-TEMS)', pp. 1–4.

Jia, L., Hu, Z., Song, Y. & Luo, Z. (2012), 'Optimal siting and sizing of electric vehicle charging stations', *IEEE International Electric Vehicle Conference* .

Phonrattanasak, P. & Leeprechanon, N. (2014), Optimal placement of ev fast charging stations considering the impact on electrical distribution and traffic condition, *in* '2014 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE)', pp. 1–6.

Roy, A. & Law, M. (2022), 'Examining spatial disparities in electric vehicle charging station placements using machine learning', *Sustainable Cities and Society* **83**, 103978.
URL: *https://www.sciencedirect.com/science/article/pii/S2210670722002980*

scikit-learn documentation (2022).
URL: *https://scikit-learn.org/stable/*

seaborn documentation (2022).
URL: *https://seaborn.pydata.org/*

Sustainability (Tai-Ran Hsu - Nov 2013), https://www.researchgate.net/profile/Tai-Ran-Hsu/publication/258849529$_{O}n_{t}he_{S}ustainability_{o}f_{E}lect$ $the - Sustainability - of - Electrical - V ehicles.pdf$?$origin = publication_{d}etail.$ [$Accessed$24 − Nov − $2022].

$Traffic_{c}ount−data.gov.uk$(2022), $https$: $//www.data.gov.uk/dataset/208c0e7b$ − 353f − 4e2d − 8b Nov − 2022].

Transport, D. f. (2022), 'Vehicles statistics'.
URL: *https://www.gov.uk/government/collections/vehicles-statistics*

Uk Mobility Footfall (2022).
URL: $https://my.api.mockaroo.com/ev_{f}ootfall.json$?$key$ = 9eb8e7e0

Uk Mobility passerby (2022).
URL: *https://api.mockaroo.com/api/67042b30?count=1000key=9eb8e7e0*

uk-postcode-electricity-consumption (2022), https://www.data.gov.uk/dataset/e7d4c1cf-45a0-4070-878f-24ad9641f655/domestic-electricity-and-gas-estimates-by-postcode-in-great-britain. [Accessed 24-Nov-2022].

$uk_{p}ostcode_{g}eojson$(2022), $https$: $//geoportal.statistics.gov.uk/$.[$Accessed$24− Nov − 2022].

$uk_{p}ostcode−NSPL$(2022), $https$: $//geoportal.statistics.gov.uk/datasets/national - statistics -$ Nov − 2022].