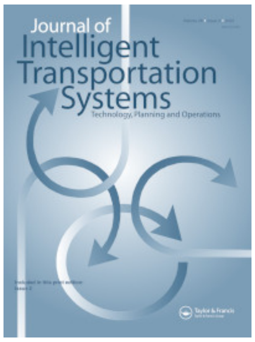

Taylor & Francis
Taylor & Francis Group

Journal of Intelligent Transportation Systems
Technology, Planning, and Operations

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gits20

Learning electric vehicle driver range anxiety with an initial state of charge-oriented gradient boosting approach

Yang Song & Xianbiao Hu

To cite this article: Yang Song & Xianbiao Hu (2021): Learning electric vehicle driver range anxiety with an initial state of charge-oriented gradient boosting approach, Journal of Intelligent Transportation Systems, DOI: 10.1080/15472450.2021.2010053

To link to this article: https://doi.org/10.1080/15472450.2021.2010053

Published online: 05 Dec 2021.

JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS
https://doi.org/10.1080/15472450.2021.2010053

Taylor & Francis
Taylor & Francis Group

Learning electric vehicle driver range anxiety with an initial state of charge-oriented gradient boosting approach

Yang Song and Xianbiao Hu

Department of Civil Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA

ABSTRACT

This manuscript focuses on the modeling of electric vehicle (EV) driver's range anxiety, a fear that a vehicle does not have sufficient range, or state of charge (SOC) of the battery pack, to reach its destination and would strand its occupants. Despite numerous research studies on the modeling of charging behaviors, modeling efforts to understand at what battery percentages do EV drivers charge their vehicles, and what are the associated contributing factors, are rather limited. To this end, an ensemble learning model based on gradient boosting is developed. The model sequentially fits new predictors to new residuals of the previous prediction and, then, minimizes the loss when adding the latest prediction. A total of 18 features are defined and extracted from the multisource data, which cover information on driver, vehicles, stations, traffic conditions, as well as spatial-temporal context information of the charging events. The analyzed dataset includes 4.5-year's charging event log data from 3,096 users and 468 public charging stations in Kansas City Missouri, and the macroscopic travel demand model maintained by the metropolitan planning organization. The result shows the proposed model achieved a satisfactory result with a R square value of 0.54 and root mean square error of 0.14, both better than multiple linear regression model and random forest model. To reduce range anxiety, it is suggested that the priorities of deploying new charging facilities should be given to the areas with higher daily traffic prediction, with more conservative EV users or that are further from residential areas.

ARTICLE HISTORY
Received 8 December 2020
Revised 8 November 2021
Accepted 19 November 2021

KEYWORDS
Charging behavior; electric vehicle; ensemble learning; mobility behavior; state of charge (SOC)

Nomenclature list

i, j, k	Index of the charging event, electric vehicle/driver, and charging station, respectively
CE_i	Charging event i
TC_i	Time spent on charging for event i
D_i	Duration of charging event i
$ISOC_i$ $ESOC_i$	Initial and End SOC of charging event i
EC_i	Energy that is charged in one charging event i
$FULC_i$	A binary variable indicating if an electric vehicle is fully charged in charging event i.
$THRH_a$	A threshold to determine if a vehicle is fully charged, based on the difference between charging duration time and charging time
$THRH_b$	A threshold to determine if a vehicle has sufficient amount of charging records
$ZPDR_j$	Zip code of vehicle j's registered address
$ZPCS_k$	Zip code of the charging station k
$ZPCSS$	Zip code space of all charging stations, $ZPCS_k \in \{ZPCSS\}$
$MAXE_j$	The maximum amount of energy that vehicle j can be charged
U_i	User ID of one charging event i
$CCAR_z$	Battery capacity of a popular vehicle model z.
$CSNUM_k$	The total number of charging stations in the zip code where station k is located at
$ARSZ_k$	The size of the zip code area in the unit of square miles
$ZPCD_i$	Zip code of the station where charging event i happened
$ZPDR_{ij}$	Zip code of charging event i's corresponding vehicle j's registered address, which usually is the driver's home zip code

Introduction

An electric vehicle (EV) uses one or more electric motors or traction motors for propulsion. A global transition to EVs offers a promising way to reduce greenhouse gas (GHG) emissions, as well as to mitigate fossil fuel dependency and promote a net-zero emissions economy (Ahmadi, 2019, Bonsu, 2020). The development and application of EVs, which have attracted significant attention from industry, researchers, consumers, and government agencies, are

CONTACT Xianbiao Hu ✉ xbhu@psu.edu Department of Civil Environmental Engineering, Pennsylvania State University, 221B Sackett Building, University Park, PA 16802-1408, USA

expected to significantly impact policy making, the urban environment, and the economy.

The increasing ownership of EVs comes with concerns about battery safety, vehicle quality, and more importantly, driving range, or "range anxiety" (Hao et al., 2020) – fear that a vehicle does not have sufficient range, or state of charge (SOC) of the battery pack, to reach its destination and would strand its occupants. Range anxiety is one of the most significant barriers to the large-scale adoption of EVs. In addition, the charging time for an EV is significantly longer than the refueling time required for a gasoline-powered vehicle. However, despite numerous researches on the modeling of charging behaviors, modeling efforts to understand range anxiety, and specifically, at what battery percentages do EV drivers charge their vehicles, and what are the associated contributing factors, are rather limited.

This manuscript aims to bridge this research gap by answering the question of "when do EV drivers charge their vehicles?" In other words, at what initial state of charge level (percentage of remaining electric battery capacity), and under what conditions is the EV charged. An analogy to this study is a smartphone user's charging behavior - some people with higher low-battery anxiety charge their phones whenever possible, no matter how high or how low their battery level is, while others may not charge until the very last minute. Oliver classified the former smartphone users as opportunistic chargers who were the most common cluster characterized by frequent and short charge durations; the latter were labeled as light-consumers who on average initiated a charge until their battery level had dropped to as low as 34%. The other smartphone user category was named as nighttime chargers who were distinguished from the other groups by charging predominantly during the night (Oliver, 2010). Despite the existing similar studies in the area of smartphone charging, such EV SOC oriented research, although important, has not been studied before, to the best of our knowledge. We argue that the learning of the initial SOC of the charging event is an important step toward gaining a fundamental understanding of EV range anxiety, and ultimately, for alternative energy promotion in transportation.

In this research, an ensemble learning model based on gradient boosting is developed. What we present in this paper is an efficient tree boosting-based regression model, and by the definition of ensemble learning, it combines multiple machine learning models to achieve better predictions. The model sequentially fits new predictors to new residuals of the previous prediction and, then, minimizes the loss when adding the latest prediction as long as the loss function is twice differential. The descent direction of objective function (summation of a training loss component and a regulation component) could be derived from last iteration, and combined with computation technics like parallel and distributed processing, the training process could be accelerated greatly compared with other ensemble learning methods. Besides, by minimizing the objective function, the model is designed to achieve good training results while controlling its complexity to avoid overfitting. Eighteen features are defined to feed into the learning model, covering information on the driver, station, vehicle, traffic, as well as spatial-temporal context information of the charging events. The proposed model is finally compared with two benchmark models to test its modeling accuracy and computational efficiency, including a multiple linear regression model (MLR) and a random forest model (RFM).

The reminder of this manuscript is organized as follows. Literature review is summarized in Section "Literature Review". Section "Learning Model" presents the proposed ensemble learning model to predict the initial SOC of an EV vehicle before charging, and the underlying features. Section "Data Description" gives a description of the real-world charging event log data used in this research. Numerical analysis is presented in Section "Modeling results". Finally, concluding remarks are given in Section "Conclusions".

Literature review

Various studies that focus on the charging behaviors of EV drivers can be found in the literature. For example, Sun et al. studied the charging timing decisions of EV drivers, in which a mixed logit model (with unobserved heterogeneity) was applied to panel data extracted from a 2-year field trial on EV battery usage in Japan. The results suggested that, the state of charge interval in days before the next travel day, and vehicle-kilometers to be traveled on the next travel day, were the main predictors of whether a user charged the vehicle, or not (Sun et al., 2015). Yang et al. proposed two multinomial logit-based and two nested logit-based models to explore charging and route choice behavior of battery electric vehicle (BEV) drivers. It was found that SOC at origin was the most important variable affecting charging decision and the SOC at destination became an important impact factor affecting BEV drivers' route choice behavior (Yang et al., 2016). Chen et al. used the EV's parking information from Puget Sound Regional Council's household activity survey to determine public parking locations and durations (Chen et al., 2013). Jabeen et al. used Western Australia EV trial data to study a driver's preference of a charging location. It was found that, with different charging costs, durations, and time of day, drivers preferred to charge an EV at home or work rather than at a public charging station (Jabeen et al., 2013). Morrissey analyzed EV drivers' preferences for different charging facilities, based on the Ireland data, and found that fast chargers recorded the highest usage frequencies, indicating that a public fast charging infrastructure was most likely to become commercially viable in a short- to medium-term (Morrissey et al., 2016). Khwaja et al. predicted charging behavior, including the charging start time slot, charging range of minutes and the number of charging times the next day, with long short-term memory networks (Khwaja et al., 2020). Kim et al. developed a hazard-based duration model of inter-charging times. It was revealed that most plug-in electric vehicles (PEV) users inclined to charge randomly at public charging stations with short charging durations and long charging intervals (Kim et al., 2017). Those studies on the charging behaviors emphasized on the influencing factors that triggered charging events, but did not quantitatively answer at which specific SOC level would a driver charge the vehicle.

When it comes to the methodological approaches, both traditional statistical methods and machine learning models were utilized to analyze the charging behavior and charging demands of EVs. Amini et al. presented an autoregressive integrated moving average method to forecast the charging demand with the integrated and auto-regressive order parameters optimized, which could facilitate the time series charging load data (Amini et al., 2016). Yi et al. adopted a deep learning approach-Sequence to Sequence (Seq2Seq) to predict monthly commercial EV charging demand, which produced satisfactory prediction performance in terms of both one-step and multi-step prediction (Yi et al., 2021). Shepero and Munkhammar used spatial Markov chain to forecast charging load of EVs in cities, which simulated the mobility of EVs considering various charging profiles (Shepero & Munkhammar, 2018). Yavasoglu et al. used decision tree to estimate the remaining driving range of BEVs, and improved the accuracy by 11.3% than rated ones (Yavasoglu et al., 2019). Chung et al. proposed an ensemble machine learning-based algorithm to predict the stay duration and energy consumption to optimize the EV charging schedule. The result showed that the synergy of support vector regression, random forest and diffusion-based kernel density estimator enhanced the prediction performance, although such self-defined ensemble learning model may be computationally expensive (Chung et al., 2019). Therefore, a highly efficient and accurate machine learning method is needed for charging decision study.

Aside from the existing discussions of range anxiety from the aspect of psychological phenomenon, there have been many studies focusing on its empirically-based understanding. For example, Hao et al. estimated electricity consumption and charging patterns of BEVs across different driving applications and seasons by fitting gamma distribution. It was simply stated that such accurate estimates of energy consumption and range can potentially reduce consumer anxiety, though not specifically justified. Also, the statistical analysis of charging frequency failed to reveal the influencing factors on charging decision and range anxiety (Hao et al., 2020). Neubauer and Wood took minimum range margin at the end of each trip as a proxy for range anxiety, i.e., a small minimum range margin represented low range anxiety and vice versa. It was found that driver range anxiety had a significant effect on the achieved utility. However, minimum range margin at the end of each trip was not directly related to range anxiety, as it reflected more BEV driver's risk-taking sensitivity of traveling with limited range. Meanwhile, the BEV trips in the research was simulated instead of directly collected, which may compromise its practicality (Neubauer & Wood, 2014). Further, SOC was considered by more existing studies as a acknowledged synonymous variable for range anxiety (Nilsson, 2011). Unlike minimum range margin at the end of each trip, Yang et al. took SOC at destination as an indicator of range anxiety and assumed that the higher SOC BEV has at destination, the less range anxiety the driver feels during travel. However, such discussion of SOC at destination was in the context of route choice without giving insight on how SOC value impacted charging decision with anxiety level as a bridge. Meanwhile, the data was collected by a Stated Preference survey instead of from real-world EV usage, which may question the fitness to true situation (Yang et al., 2016). Given that BEV drivers' utility would reduce much faster when the SOC approached zero, Xu et al. assumed a nonlinear relationship between initial SOC before charging and range anxiety, though not explicitly presented (Xu et al., 2017).

Table 1. Definitions of features.

Category	Feature notation	Definition	Data type
Driver Profile	DCF_j	Daily charging frequency of Driver *j*	Numerical
	$NHCS_j$	Existence of near-home charging station for driver *j*	Binary
	DCH_{ij}	Distance between charging event *i* and driver *j*'s home	Numerical
	$NHCG_i$	If event *i* is a near-home charging	Binary
Vehicle Profile	C_j	Battery capacity of vehicle *j*, in kWh	Numerical
Station Profile	DN_k	Density of charging stations in the zip code area where station *k* is located at	Numerical
Traffic Profile	$AADT_k$	Annual average daily traffic on station *k*'s nearby roads	Numerical
	TP_k	Trip production of the TAZ where station *k* is located at	Numerical
Temporal Context Information	WD_i	Workday feature: if event *i* happens in a workday	Binary
	TD_i	Time of day feature: if event *i* happens at night	Binary
	$SN1_i$	Season feature: if charging event *i* happens in Spring	Binary
	$SN2_i$	Season feature: if charging event *i* happens in Summer	Binary
	$SN3_i$	Season feature: if charging event *i* happens in Fall	Binary
	$SN4_i$	Season feature: if charging event *i* happens in Winter	Binary
Spatial Context Information	$LDU1_i$	Land use feature: if event *i* happens in residential area	Binary
	$LDU2_i$	Land use feature: if event *i* happens in industrial area	Binary
	$LDU3_i$	Land use feature: if event *i* happens in a business area	Binary
	$LDU4_i$	Land use feature: if event *i* happens in the airport area	Binary

Frank et al. and Guo et al. proposed that range anxiety happened only when SOC falling below certain level, namely the comfortable range threshold, from fully charge during driving (Franke et al., 2012, Guo et al., 2018). Inspired by this concept, Xu et al. proposed the most related study by assuming that as the decreasing of SOC from fully charge, range anxiety of an EV driver would remain at 0. Then when SOC dropped below the comfortable range threshold, range anxiety would convexly increase with the increase of SOC following a polynomial function from 0 to a maximum value (Xu et al., 2020). But this explicit model formation between anxiety and SOC was just for simplicity and still needed further justification. Hence, how to understand range anxiety via BEV SOC, and does it impact charging decision are yet to be answered, preferably with real-world BEV usage data.

To address these research gaps from existing studies, in the next sections, with the initial SOC value before charging as the target variable, we focus on modeling the EV drivers' charging behavior decisions, including the timing and context in which charging events occur, with the goal of improving the understanding of range anxiety and its associated contributing factors.

Learning model

Prior to the discussion of technical details, an assumption is made that one EV has only a single driver, so that the charging behavior observed from the field is consistently attributed to one electric vehicle driver. Based on such assumption, we then move on to defining the features.

Feature definition

Eighteen features are defined and extracted from the multi-source data, which cover information on the following six categories: 1) driver profile, 2) vehicle profile, 3) station profile, 4) traffic profile, and 5) temporal context, as well as 6) spatial context information of charging events. Based on our domain knowledge and literature review, these features are believed to play a role in the driving range, and are thus defined in the modeling process. One thing to note here is that some variables are numerical data, with continuous or discrete values, while some other variables are categorical data with a limited number of values. Table 1 below lists definitions of the categories and some of their features.

Driver profile

The first set of features that we extract is the drivers' profile, including 1) daily charging frequency DCF_j, 2) existence of near-home charging stations $NHCS_j$, 3) distance between charging station and driver's home DCH_{ij}, and 4) if the charging event is a near-home charging $NHCG_i$. In the above notations, i is the index of charging events and j is the index for the driver.

Daily charging frequency DCF_j: When predicting the SOC level for charging events, we suspect a driver's profile, identified from historical data, may play an important role. Here, a driver's risk-taking attitude is measured by daily charging frequency. In this way, other things being equal, drivers who charge their EVs less often, in a given time period, are said to be more risk-taking and aggressive. Thus, the hypothesis becomes that such a personal behavior preference may impact their SOC level before charging.

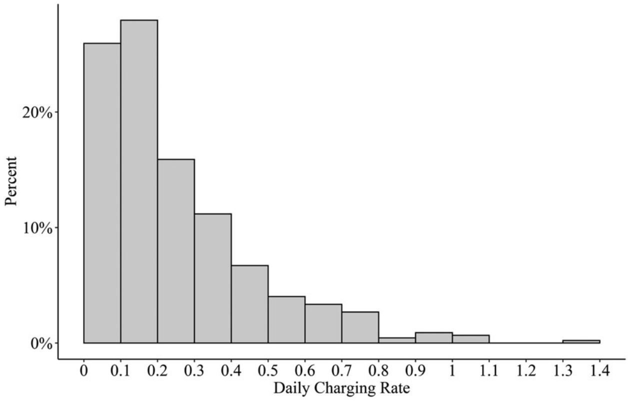

Figure 1. Distribution of driver's daily charging frequency.

Daily charging frequency DCF_j is calculated via Eq. (1) below, where $count(CE_{i,j})$ denotes the total number of charging events from driver j, while $Min(date(CE_{i,j}))$ and $Max(date(CE_{i,j})$ represent the first and last dates that the driver was observed to use the system, and thus $Max(date(CE_{i,j}) - Min(date(CE_{i,j}))$ becomes the number of days this driver utilizes the charging systems.

$$DCF_j \; = \frac{count(CE_{i,j})}{Max(date(CE_{i,j}) - Min(date(CE_{i,j}))} \qquad \text{Eq. (1)}$$

Following this definition, our preliminary analysis discovered that driver characteristics vary significantly. In Figure 1 below, X-axis is the daily charging frequency and Y-axis is the percentage of drivers in each category. An exponential decrease pattern can be observed, with about 50% of drivers charging their EVs 0-0.2 times per day (i.e., at most, once every 5 days), while fewer than 5% percent of EV users charge more than once a day. By observing the data, it is obvious that a clear variance exists in the daily charging rates of EV drivers, which would justify usage of the DCF_j feature in the learning model.

Existence of a near-home charging station near driver j $NHCS_j$ is defined as a binary variable to indicate if there is at least one charging station located in the same zip code area of the driver's home address. It is expected that EV users, with public charging facilities near home, may be more likely to charge at that particular facility instead of at another location since they are familiar with the surrounding environment and it is more convenient. $NHCS_j$ can be calculated as Eq. (2) below.

$$NHCS_j \; = \begin{cases} 1 \;\; if \;\; ZPDR_j \in \;\; ZPCSS \\ \quad 0 \;\; Otherwise \end{cases} \qquad \text{Eq. (2)}$$

In which $ZPDR_j$ is the zip code of vehicle j's registered address, while $ZPCSS$ represents the set of zip codes of all charging stations.

The distance between charging event i and the home location of driver j DCH_{ij}. It is suspected that, when EV drivers drive further away from home, the range anxiety may become more intense, as the battery level becomes lower and the distance to home is longer. The Haversine formula (shown below) is used to calculate the great circle distance between two points on the Earth. The latitude and longitude of the charging event i (lat_1, lon_1) are directly available in the dataset, while those of the driver j's home (lat_2, lon_2) is approximated with the centroid of the zip code. R is the radius of the Earth.

$$\begin{gathered} d_{lon} = lon_2 - lon_1 \\ d_{lat} = \;\; lat_2 - lat_1 \\ a = (\sin(d_{lat}/2))^2 + \cos(lat_1)* \; \cos(lat_2)*(\sin(d_{lon}/2))^2 \\ c = \;\; 2* \; atan2(sqrt(a), \; sqrt(1-a)) \\ DCH_{ij} = R*c \end{gathered} \qquad \text{Eq. (3)}$$

Neighborhood charging $NHCG_i$ is defined as a binary variable to indicate if charging event i happens in the same zip code area of the driver's home address. We suspect that the initial SOC for the charging events could be significantly different with different $NHCG_i$ values.

$$NHCG_i \;\; = \begin{cases} 1 \;\; if \;\; ZPCD_i = ZPDR_{ij} \\ \quad 0 \;\; Otherwise \end{cases} \qquad \text{Eq. (4)}$$

In which $ZPCD_i$ is the zip code of the station where charging event i happens, and $ZPDR_{ij}$ is the zip code of vehicle j's registered address.

Vehicle profile

Vehicle battery capacity C_j is measured in units of kWh. The hypothesis is that, with a larger-capacity battery, the EV driving range is longer and the driver's range anxiety will be reduced. C_j is not explicitly available in the dataset, so Eq. (5) -Eq. (6) is developed to infer its value.

$$\begin{gathered} MAXE_j = Max(E_i) \\ \text{s.t.} \\ U_j = \{i\}, where \;\; u_i = j \\ count(U_j) \geq THRH_b \end{gathered} \qquad \text{Eq. (5)}$$

Where $MAXE_j$ represents the maximum amount of energy that vehicle j can be charged, and is a lower bound of battery size C_j. E_i is the energy that is charged into the vehicle in event i. U_j is a set of charging event ID, and the constraint $u_i = j$ allows us to extract all records for vehicle j. To reasonably approximate battery size, $THRH_b$ is defined as a threshold to determine if a vehicle has sufficient observations in the system. In other words, if a vehicle has been charged for a significant number of times, i.e. $count(U_j) \geq THRH_b$, the lower bound $MAXE_j$ is considered to be approaching the vehicle's battery size C_j. $THRH_b$ takes the value of 10 in this manuscript.

$$\begin{gathered} C_j = CCAR_Z \\ \text{s.t.} \\ Z = \arg\min_z \left(CCAR_z - MAXE_j\right) \\ and\ CCAR_z > MAXE_j \end{gathered} \qquad \text{Eq. (6)}$$

$MAXE_j$ is then compared with the battery size of popular vehicle models $CCAR_z$ in the market, such as Nissan Leaf, BMW i3, Chevrolet Bolt, Hyundai Kona, and Ford Focus Electric, to find the vehicle model with the closest battery size. In Eq. (6), we find the vehicle model Z that has a larger, but closest battery size when compared with the estimated $MAXE_j$, and then use its battery size $CCAR_Z$ to approximate that of the vehicle j.

Station profile

The density of charging stations in each zip code area DN_k is considered as an indicator of the charging infrastructure supply, and is calculated by Eq. (7) below.

$$DN_k = CSNUM_k / ARSZ_k \qquad \text{Eq. (7)}$$

In which $CSNUM_k$ is the total number of charging stations in the zip code where station k is located, and $ARSZ_k$ is the area size of the zip code with a unit of square miles.

Traffic profile

Annual average daily traffic $AADT_k$, is defined as the average traffic volume on the nearby road where station k is located. In Eq. (8), $AADTSUM_k$ is the summation of daily traffic volume on every roadway segment in the zip code area where station k is located, and $ARSZ_k$ is the zip code area in square miles.

$$AADT_k = AADTSUM_k / ARSZ_k \qquad \text{Eq. (8)}$$

It should be noted that $AADT_k$ includes, not only the vehicles that originate from, or drive into a TAZ, but also the bypass traffic that may not stop at the TAZ at all. As such, the usage of $AADT_k$ may lead to an over-estimation of the charging demand, especially for the areas that have freeways or highways, that mainly serve bypass traffic.

Daily traffic production TP_k is defined as the daily traffic volume that originates from a particular TAZ, to overcome the above-mentioned issue with $AADT_k$. With such properties, TP_k focuses on the meaningful traffic flow that is more relevant with the studied TAZ.

$$TP_k = TPSUM_k / ARSZ_k \qquad \text{Eq. (9)}$$

In Eq. (9), $TPSUM_k$ is the summation of traffic production that originates from the TAZ.

Temporal context information

A total of six features represent the temporal context information of the charging event i. They are all defined as binary variables that take values of either 1 or 0. *Work-day feature* WD_i checks if the charging event i happens on a typical workday.

$$WD_i = \begin{cases} 1\ if\ \ event\ \ i\ \ happens\ \ on\ \ workday \\ \quad 0\ Otherwise \end{cases} \qquad \text{Eq. (10)}$$

Time of day feature TD_i checks if the charging event i happens in the night time.

$$TD_i = \begin{cases} 1\ \ ifevent\ \ i\ \ time \in [10pm, 7am] \\ \quad 0\ \ Otherwise \end{cases} \qquad \text{Eq. (11)}$$

Season features, $SN1_i$, $SN2_i$, $SN3_i$, $SN4_i$, represent spring, summer, fall, and winter, respectively, with $SN1_i + SN2_i + SN3_i + SN4_i = 1\ \forall i$.

Spatial context information

Land use features $LDU1_i$, $LDU2_i$, $LDU3_i$, $LDU4_i$, represent residential, industrial, business, and airport land-use types, respectively. The hypothesis is that different land-use types indicate different trip purposes and, subsequently, impact EV driver's charging behaviors. Since charging stations are usually built in a parking lot/space that is affiliated with some point of interests (such as a shopping mall), to simplify the data analysis process, the land use feature is determined by the nearest point of interest.

SOC values and imputation

In the dataset, there is an attribute column explicitly named "SOC", which allows us to directly pull the data. A low SOC indicates that the drivers do not charge their vehicles until the remaining battery percentages become low, while a high SOC indicates that the drivers choose to charge their vehicles, even though the battery levels are still high. Data records with missing SOC values are not uncommon due to a variety of reason, and depend on the plug type of charging stations that are used. Whether end state of charge is known or not depends on the plug type of charging stations (which is recorded by the charging devices). In the dataset, plug types of "J1772" and "NEMA 5-20 R" indicate that end SOC is not available, while plug types of "Combo" indicate that the

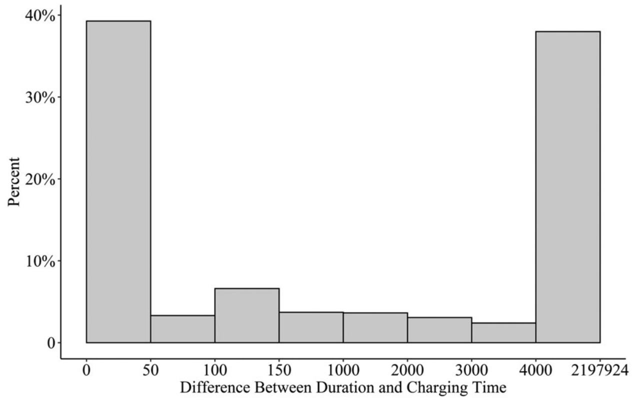

Figure 2. Difference between TC_i and D_i.

end SOC is available. To deal with this issue, a SOC imputation method is developed to infer the value of initial SOC $ISOC_i$.

Scenario 1: end SOC is available

End SOC represents the battery level after charging. When available, the initial SOC value can be directly inferred by backtracking how much energy is charged into the EV during charging event i.

$$ISOC_i = ESOC_i - EC_i/C_i \qquad \text{Eq. (12)}$$

In which $ISOC_i$ is the value of the initial SOC of charging event i, $ESOC_i$ is the value of the end SOC. EC_i is the amount of energy charged into the vehicle, and C_j is the battery capacity of the vehicle.

Scenario 2: end SOC not available

In a case when the end SOC is also missing, Eq. (12) is not applicable. For the data records in this scenario, we focus on the data records of vehicles that are fully charged, which effectively equals scenario 1 with $ESOC_i$=100%.

To identify which vehicles are fully charged, we first look at the difference between the time spent on charging for event TC_i and the duration of charging event D_i.

$$FULC_i = \begin{cases} 1 \;\; if \;\; D_i \geq \;\; TC_i + THRH_a \\ \quad 0 \;\; Otherwise \end{cases} \qquad \text{Eq. (13)}$$

Eq. (13) is designed to detect a scenario where a vehicle is fully charged, and is automatically disconnected by the system, but remains parked at the station for a significant length of time after the charging event. $FULC_i$ is a binary variable indicating if an EV is fully charged in charging event i. D_i is the charging duration, and TC_i is the time spent on actual charging. $THRH_a$ is a threshold.

The time difference between TC_i and D_i of all charging records in the dataset is visualized in Figure 2 below, with X-axis being the time difference in seconds and Y-axis being the percentage of data in each category. Two clear peaks are observed. The peak on the left shows that about 40% EVs leave the charging station within 50 seconds after the charging event is terminated, while the peak on the right shows that another 40% EVs remain parked at the charging stations for over an hour. As such, 150 seconds (i.e., 2.5 minutes) are used as the threshold $THRH_a$ to flag the vehicles that are fully charged.

For these charging records, Eq. (12) becomes Eq. (14)

$$ISOC_i = 1 - EC_i/C_i \qquad \text{Eq. (14)}$$

Ensemble learning model

In this manuscript, following the approach of the boosting-based ensemble method, especially the Extreme Gradient Boosting (Chen & Guestrin, 2016), an efficient tree boosting-based regression model is developed to predict initial SOC levels based on the above-mentioned 18 features. In essence, the model builds on the basis of gradient boosting by adding regularization to combat over fitting along with multiple other additions. At the beginning, a simple regression tree model was trained to capture the nonlinear interactions between selected features and the initial SOC value. Then the developed ensemble method sequentially fits new trees to new residuals of the prediction of previous tree and, then, minimizes the loss when adding the latest prediction. Such strategy would create an ensemble model from numerous weak regression trees, and thus these weak learners would be converted to a strong predictor. In the training, unlike other ensemble models, the objective function that includes a training loss component and a regulation component would be minimized simultaneously, so that the model is designed to achieve good training results while controlling the complexity of the model to avoid overfitting.

Assume that our training dataset is $I = \{(x_1, y_1), (x_2, y_2), \ldots (x_i, y_i) \ldots (x_n, y_n)\}$, in which x_i is an array with 18 elements representing the 18 features, defined in Table 1 and calculated via Eq. (1)~Eq. (11). y_i is the initial SOC value from the dataset, which is either directly retrieved from the system or inferred by Eq. (12)~Eq. (14). The representation can be expressed by Eq. (15) and Eq. (16).

$$x_i = [x_{i1}, x_{i2}, \ldots x_{i18}] = [DCF_j,\; NHCS_j,\; DCH_{ij}, NHCG_i, C_j, DN_k, AADT_k, TP_k, WD_i, TD_i,\; SN1 \sim 4_i,\; LDU1 \sim 4_i] \qquad \text{Eq. (15)}$$

$$y_i = ISOC_i \qquad \text{Eq. (16)}$$

The objective function is shown in Eq. (17) to minimize the summation of a training loss component l and a regularization component Ω. Optimizing training loss component l that encourages predictive models into fitting well in the training data, can at least get you close to training data that is hopefully close to the underlying distribution. On the other hand, optimizing regularization component Ω encourages simpler models which tend to have smaller variances in future predictions, making predictions stable. As such, the model is designed to achieve good training results while controlling the complexity of the model to avoid overfitting and to improve computational efficiency.

$$Min\ \ L(\Theta) = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad \text{Eq. (17)}$$

In which $\widehat{y}_i$ is the predicted initial SOC of sample i, i.e., $\widehat{ISOC}_i$, y_i is its observed value $ISOC_i$, and n is the sample size. $l(\widehat{y}_i, y_i)$ denotes the training loss, which takes the square loss form and is calculated via Eq. (18).

$$l(\widehat{y}_i, y_i) = (y_i - \widehat{y}_i)^2 \qquad \text{Eq. (18)}$$

In Eq. (17), f_k is the k^{th} base regression tree out of the total K trees, with $\Omega(f_k)$ representing the complexity of the k th tree and taking the following form as Eq. (19).

$$\Omega(f_k) = \gamma T + \frac{1}{2}\ \lambda \|\omega\|^2 \qquad \text{Eq. (19)}$$

In Eq. (19), T is the number of leaf nodes of a base tree. γ is a penalty parameter to control the complexity of the tree structure. ω represents the weight of leaf nodes, while λ is a parameter to control the regularization degree of f_k.

In Eq. (17), Θ is the set of all regression trees f_k that have been built, as shown in Eq. (20), in which K is the total number of trees.

$$\Theta = \{f_1, f_2, \ldots f_K\} \qquad \text{Eq. (20)}$$

In the end, the system outputs predicted initial SOC value of event i $\widehat{y}_i$ via Eq. (21) below.

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad \text{Eq. (21)}$$

As opposed to a bagging-based approach, such as the random forest model, the boosting-based ensemble method generates base learners sequentially that are dependent on each other. Thus, if we use $\widehat{y}_i^{(t-1)}$ to denote the predicted value at $t-1$ iteration, when running the model for another iteration, the new predicted value $\widehat{y}_i^{(t)}$ becomes:

$$\widehat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \sum_{k=1}^{t-1} f_k(x_i) + f_t(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i)$$

Eq. (22)

Eq. (17), when combined with Eq. (18) and Eq. (22), can be converted into Eq. (23).

$$\begin{aligned} Min\ \ L(\Theta)^{(t)} &= \sum_{i=1}^{n} l(y_i, \widehat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \\ &= \sum_{i=1}^{t} l(y_i, \widehat{y}_i^{(t)} + f_t(x_i)) + \sum_{k=1}^{t} \Omega(f_k) \\ &= \sum_{i=1}^{t} ((\widehat{y}_i^{(t-1)} + f_t(x_i) - y_i)^2) + \sum_{k=1}^{t-1} \Omega(f_k) \\ &\quad + \Omega(f_t) \\ &= \sum_{i=1}^{t} (f_t(x_i)^2 + 2(\widehat{y}_i^{(t-1)} - y_i) * f_t(x_i)) \\ &\quad + \Omega(f_t) + \sum_{i}^{t} {y_i}^2 + \sum_{k=1}^{t-1} \Omega(f_k) \end{aligned}$$

Eq. (23)

Note at iteration t, the third and fourth terms in Eq. (23), $\sum_i^t {y_i}^2 + \sum_{k=1}^{t-1} \Omega(f_k)$, are constants and, thus, can be skipped. Next, we regroup the objective by the leaf:

$$\begin{aligned} Min\ \ L(\Theta)^{(t)} &= \sum_{i=1}^{t} (f_t(x_i)^2 + 2(\widehat{y}_i^{(t-1)} - y_i) * f_t(x_i)) + \Omega(f_t) \\ &= \sum_{i=1}^{t} \left({\omega_j^{(t)}}^2 + 2\left(\widehat{y}_i^{(t-1)} - y_i\right) * \omega_j^{(t)} \right) + \gamma T + \frac{1}{2} \lambda {\omega_j^{(t)}}^2 \\ &= \sum_{i=1}^{t} \left(\left(1 + \frac{\lambda}{2}\right) * {\omega_j^{(t)}}^2 + 2\left(\widehat{y}_i^{(t-1)} - y_i\right) * \omega_j^{(t)} \right) + \gamma T \end{aligned}$$

Eq. (24)

So the objective function becomes a quadratic function of leaf nodes $\omega_j^{(t)}$. For $L(\Theta)^{(t)}$ to take the minimum value, let's make the first derivative $\frac{\partial L(\Theta)^{(t)}}{\partial \omega_j^{(t)}} = 0$

$$\frac{\partial L(\Theta)^{(t)}}{\partial \omega_j^{(t)}} = \sum_{i=1}^{t} (2 + \lambda) * \omega_j^{(t)} + 2\left(\widehat{y}_i^{(t-1)} - y_i\right) \ = 0$$

Eq. (25)

So that we can solve

$$\omega_j^{(t)*} = \frac{\sum_{i=1}^{t} 2\left(\widehat{y}_i^{(t-1)} - y_i\right)}{\sum_{i=1}^{t} (2 + \lambda)} \qquad \text{Eq. (26)}$$

And the optimal solution becomes

$$L(\Theta)^{(t)*} = \frac{\sum_{i=1}^{t} 4*\left(\widehat{y}_i^{(t-1)} - y_i\right)^2}{\sum_{i=1}^{t}(2+\lambda)} + \gamma T \qquad \text{Eq. (27)}$$

Data description

In this section, both charging even log data and travel demand model data are introduced. We argue that such a large-scale real-world transaction dataset and the travel demand data, when combined, would provide rich information on, not only the charging activities that have occurred, but also the context of these events. The entire dataset is divided into two subsets, including 90% data for training purposes and the remaining 10% data for testing purposes.

Charging event log data

The data were collected from 468 public charging stations in Kansas City, Missouri (KCMO), between January 2014 and December 2019. The dataset includes a total of 208,187 charging records from 3,096 users. Therein, 16,286 charging records had explicit initial SOC value before charging. Most of the stations are concentrated in the downtown area of KCMO. The spatial distribution of charging stations is shown in Figure 3 below, in which (a) shows an overview and (b) zooms in to the downtown area. Figure 4 is a heatmap to show the spatial distribution of charging events.

Table 2 shows sample data from the dataset, in which only the most critical and relevant information is displayed.

Figure 5 below presents the distribution of drivers by the number of charging stations used. The number of charging stations visited by most of the drivers was below 10, which showed that each EV driver tended to have their own regular charging location preference. However, only about five percent of drivers charged their EVs at one single station.

Figure 6 below presents some statistical characteristics of the charging events. (a) shows the number of charging events by month, and (b) shows the pattern of charging activities in a week. The observation is that summer months, and weekdays, have more charging events than winter months, or weekends, which are consistent with the general activity pattern of travel behaviors.

In terms of the charging start time, (c) shows the there are three periods of time (around 8 am, between noon and 2 pm, from 5 pm to 8 pm) with higher number of charging records than the others. These patterns are explained by the fact that morning is when commuters leave their homes, and plug in to charge their vehicles once they arrive at their destinations. Over lunch time, they may also charge their vehicles in front of a restaurant and, during the evening hours, the vehicles may be charged again after they leave work and charge their vehicles at their destinations, which may or not be their homes. The end time of the charging records, however, shows a very different pattern. As seen in (d), there are also three periods of time (4-6pm, 11am-1pm, around 10 pm) with higher number of records, but at different times of a day. These patterns demonstrate that, while the plug-in events happen most during the morning peak-hours, some EVs do not leave the charging station after they are fully charged.

Travel demand model data

The travel demand model data was provided by Mid-America Regional Council (MARC), the Metropolitan Planning Organization (MPO) for the bi-state Kansas City region. The data provided includes a daily Origin-Destination demand matrix and a road network of Kansas City. The Origin-Destination (OD) demand matrix is a matrix in which each cell represents the number of trips from origin traffic analysis zones (TAZ) (row) to the destination TAZ (column). The entire Kansas City area is divided into 2,510 TAZ, which are shown in Figure 7 below. Each polygon represents a TAZ. Therein, the gray part denotes the region of KCMO where charging stations were located.

Consequently, there are a total of 2,510 * 2,510 cells in the OD demand table given by MARC, and each cell denotes the number of daily trips from one TAZ to another in Kansas City. Then, by summarizing the cells by row, we can obtain the trip production of each TAZ, while the trip attraction of each TAZ can be calculated by summarizing by column. A portion of OD matrix is shown in Table 3 below.

Next, with this OD matrix and the location of each TAZ, the travel demand is assigned onto the road network of Kansas City. Fortunately, this work has been completed and calibrated by MARC, so that the retrieved travel demand model matches with real-world traffic conditions. In the given road network given, there are a total of 10, 533 nodes and 24, 601 links. On each link, the traffic volume is thus derived. Figure 8 below shows the visualization of traffic

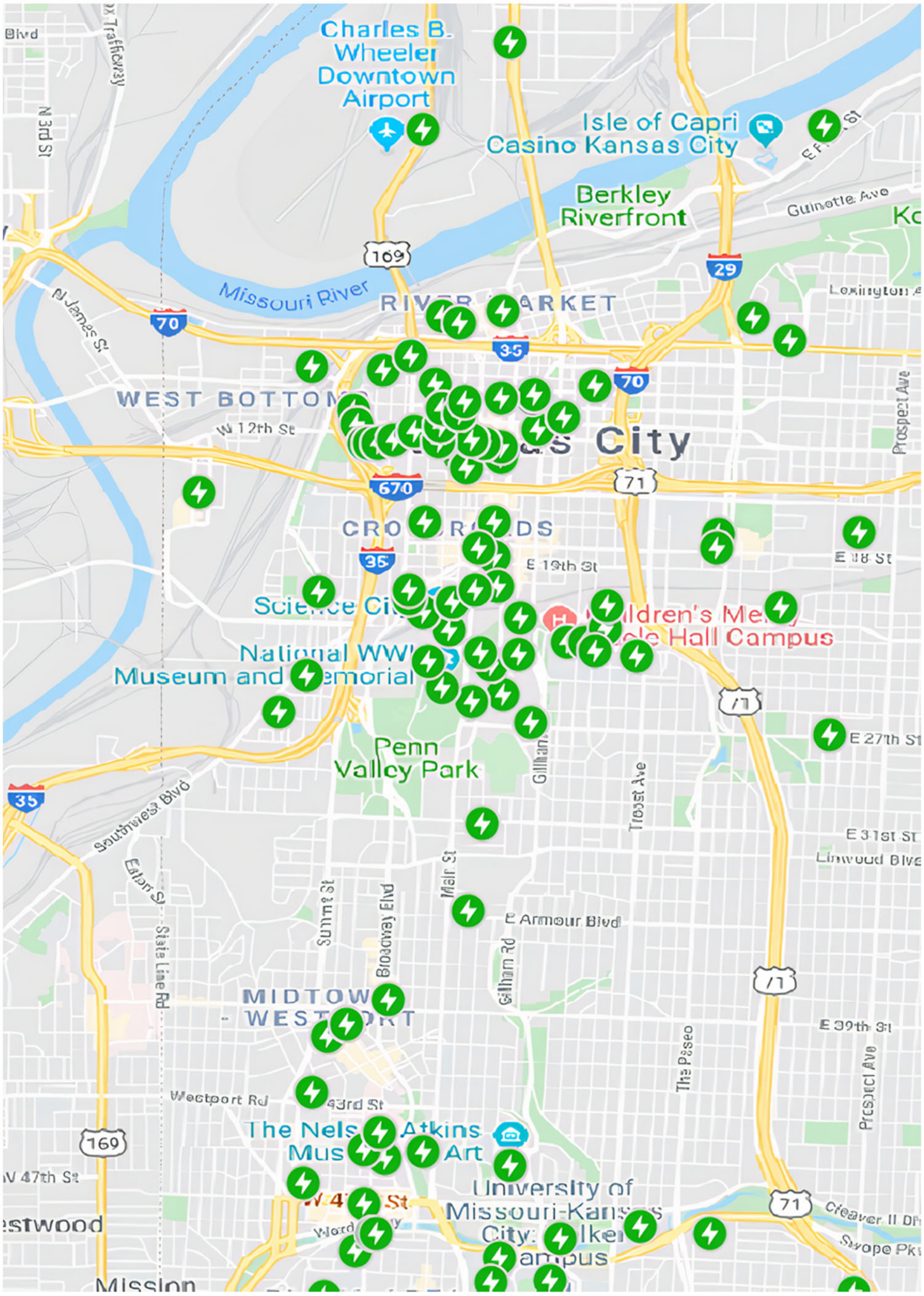

Figure 3. (a) overview; (b) downtown.

volume on each link, in which the dark colors represent heavy traffic, while lighter colors indicate lighter traffic conditions.

Modeling results

Prediction accuracy

Table 4 presents the sample sizes and modeling performance for the three datasets: training, validation, and testing. Three performance measures were used, including R square, root mean squared residuals (RMSE), and mean absolute error (MAE). It was found that R Square values ranged between 0.5391 ~ 0.5988, RMSE was between 0.1362 and 0.1438, and MAE was between 0.1037 and 0.1114. When comparing testing and training datasets, the result suggested the performance was similar in all three measures and, in particular, the RMSE and MAE were almost identical. This finding was consistent with the design principal of the proposed model, that by minimizing an objective function which included a training loss component and a regulation

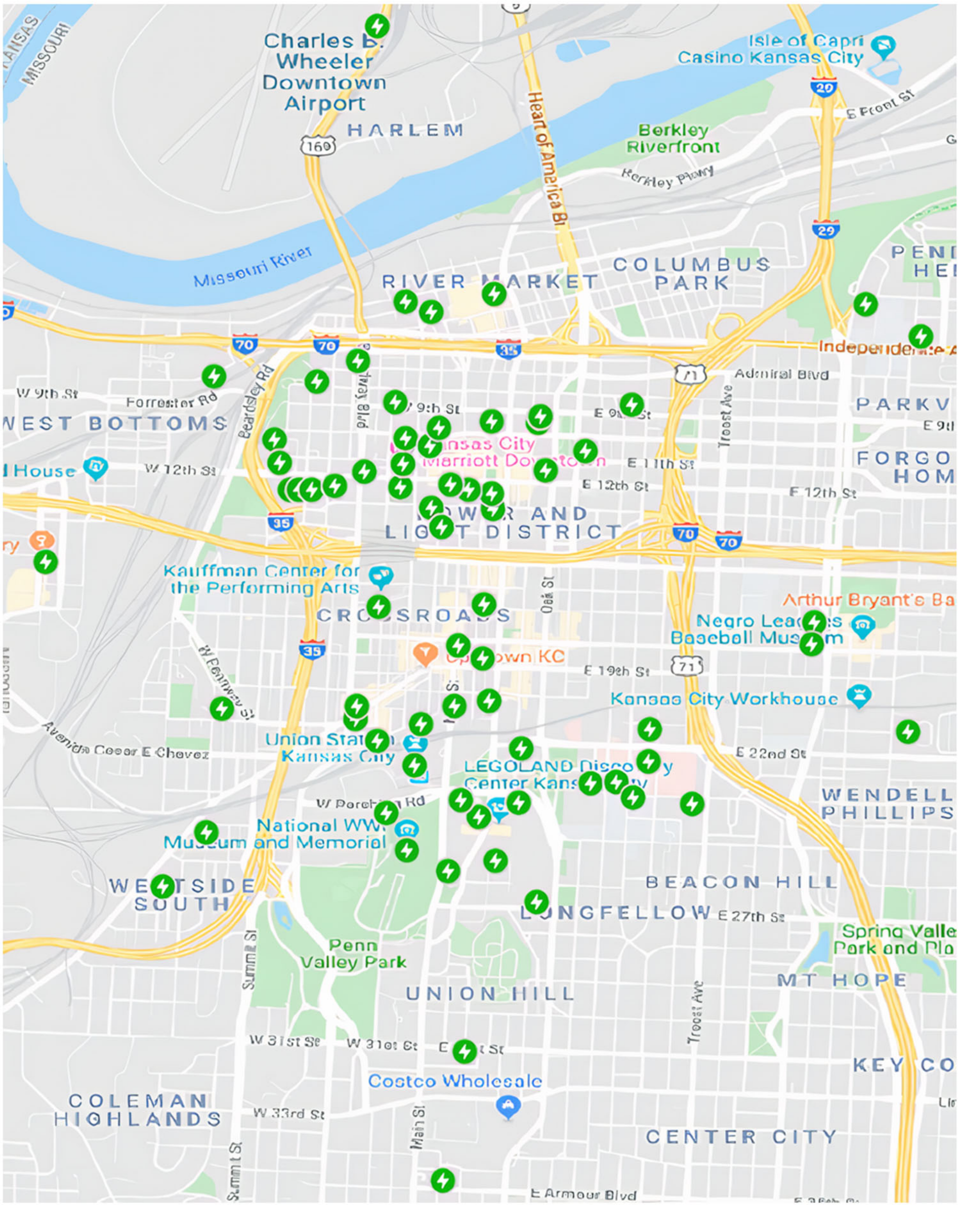

Figure 3. Continued.

component, the model was designed to achieve good training results, while controlling the complexity of the model to avoid overfitting. Thus, the performance on the testing dataset was also satisfactory.

Next, the distribution of modeling accuracy was examined to gain a further understanding of the model performance. From Figure 9, it was clear that the residuals of testing samples were concentrated in the middle and dropped sharply on each side. In other words, the most frequently observed data bins were those with low residuals (i.e., with best prediction accuracy) and, as the residual increased, the frequency dropped sharply. The plot was nearly evenly distributed on the negative and positive sides.

The model's performance, with regard to different SOC values, was also investigated. Figure 10a suggested that the prediction error was the highest at 0.23, when the initial SOC was the lowest, and then dropped rapidly to below 0.10, and remained stable with a higher SOC value. This result was initially surprising, as it suggested the model produced a biased prediction under certain circumstances. To investigate the underlying reasons, data size in different data bins were examined in Figure 10b. The plot showed the data size [0.0,0.1] for the bin was extremely low, indicating that most EV drivers (if not all) charged their vehicles before the battery ran extremely low, which was consistent with our general

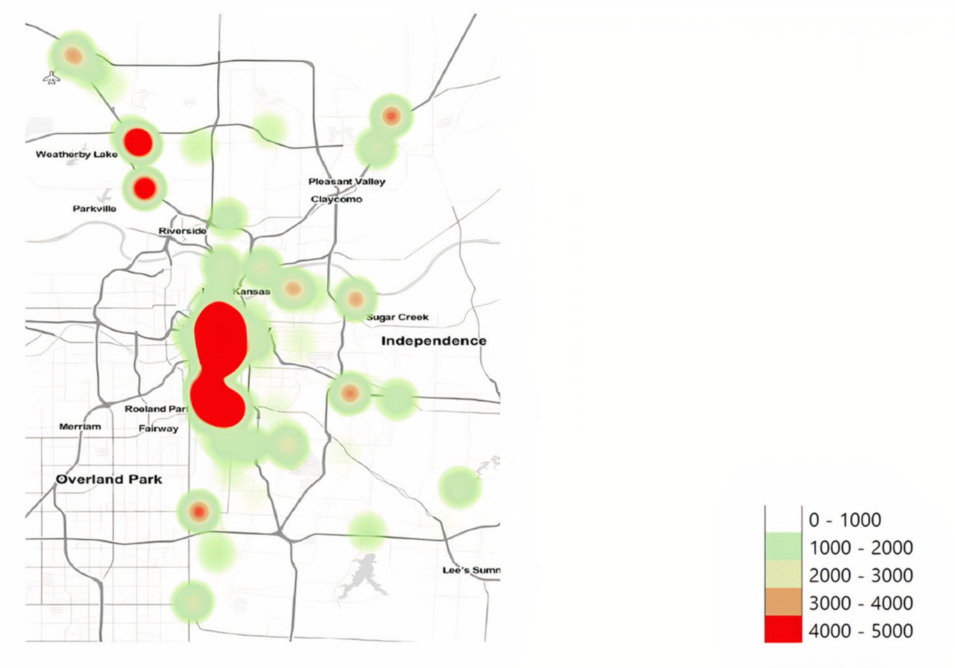

Figure 4. Distribution of charging events in KCMO.

Table 2. Sample charging event log data.

Station Name	Start Date	End Date	Duration (hh:mm:ss)	Charging Time (hh:mm:ss)	Energy (kWh)	Latitude	Longitude	User ID
KCPL / @JE DUNN PG125C	9/18/2018 14:27	9/18/2018 15:50	1:23:15	1:23:02	5.934	39.1011	94.5763	756927
KCPL / @WOLF PG −129 A	9/18/2018 6:57	9/18/2018 15:39	8:42:04	2:51:18	6.766	39.0999	94.5791	1533491
KCPL / @LOOSE PRK-121A	9/18/2018 15:09	9/18/2018 15:10	0:00:30	0:00:00	0	39.0346	94.5947	598567

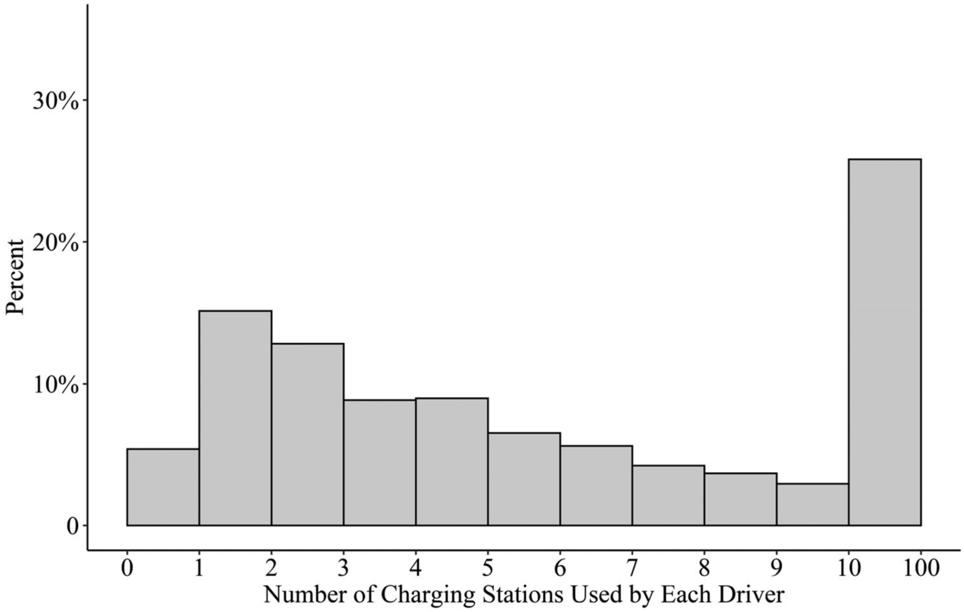

Figure 5. Distribution of drivers by number of charging stations used.

understanding. As such, the low data size led to relatively poor performance. This observation was also consistent with the lower data size for SOC intervals [0.1-0.2] and [0.9-1.0], where the data sizes were also low and prediction accuracies were not as good as others.

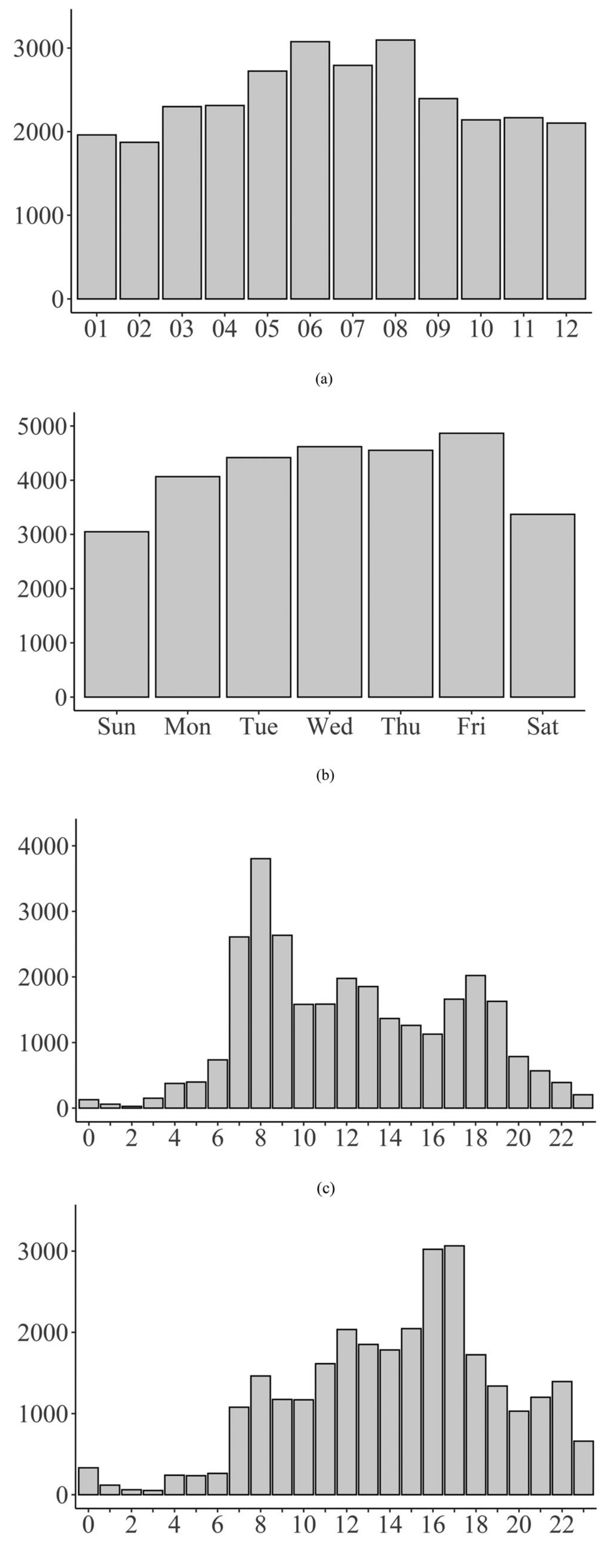

Figure 6. (a) Number of charging records by month; (b) Charging records day of the week distribution; (c) Charging records start time distribution; (d) Charging records end time distribution.

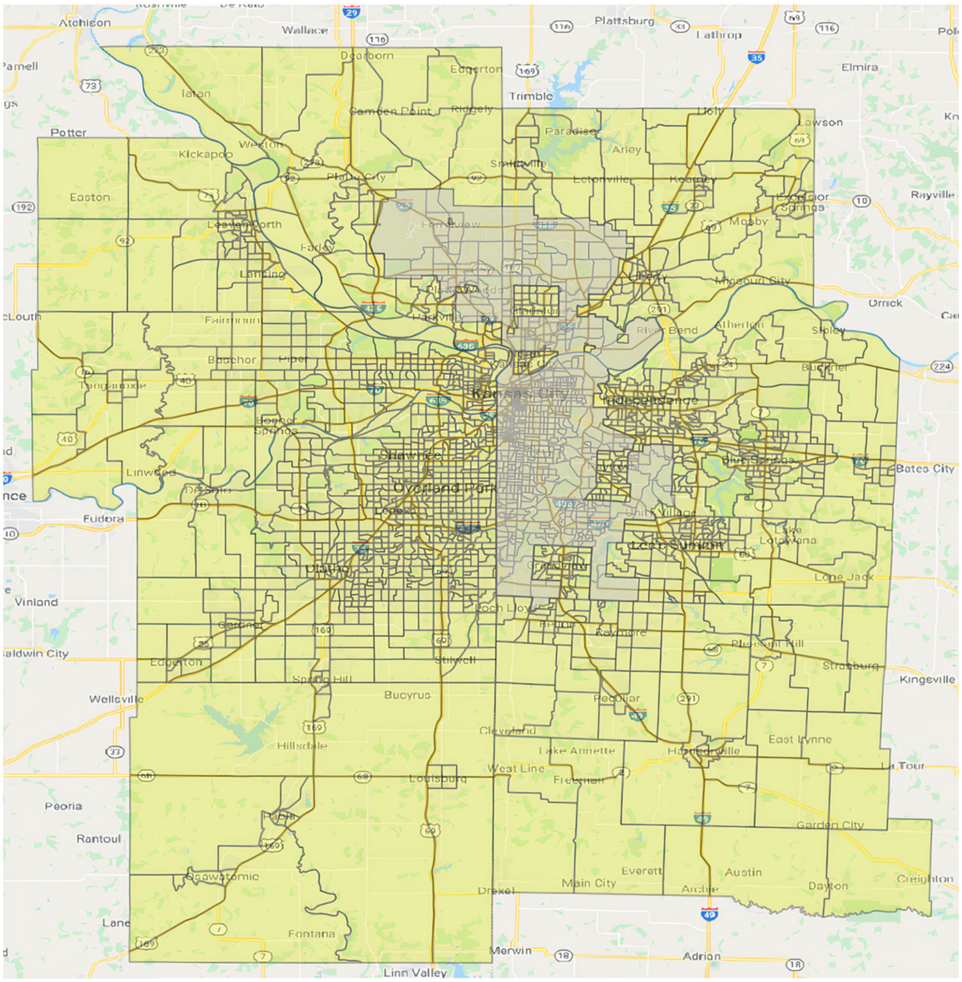

Figure 7. 2510 Traffic analysis zones in Kansas City.

Table 3. OD demand table sample.

TAZ	101	102	103	…	9525	Production
101	0	0	12.52	…	0.98	12901.64
102	0	0	0	…	0.24	3126.32
103	12.52	0	0	…	0.14	3322.37
…	…	…	…	…	…	…
9525	0.98	0.24	0.14	…	0	4171.30
Attraction	12901.64	3126.32	3322.37	…	4171.30	

Feature importance ranking and interpretation

To understand the explicit impact of the 18 defined features for SOC prediction, the importance of each feature was measured by changes in the residual sums of squares (RSS) as they were added. For easy comparison, the obtained values were scaled to have a maximum value of 100. The top ten predictors, with the highest degree of importance, that included six numerical and four binary ones, are presented in Table 5.

Table 5 suggests that, for the proposed ensemble learning model, seven features apparently had a more significant impact than the others. Their relative importance values were all over 25 and corresponded to 93.64% of the total contribution. Two driver-profile features: the driver's daily charging frequency (DCF_j) and the distance between the charging station and the driver's home (DCH_{ij}), were suggested as being the most important. The vehicle profile feature, battery capacity of vehicle (C_j), was shown to be next. Traffic conditions and station profiles were also important but to a lesser degree. This ranking suggested that the driver's charging decisions were mostly affected by endogenous features (i.e., those from driver profile and vehicle profile), although context information, such as traffic or charging station availability, also played a role.

Based on Table 5, some suggestions on charging infrastructure locations can be made to reduce the driving range anxiety of EV users. First, daily traffic prediction $\mathrm{TP_k}$ was found to be more influential than average annual daily traffic $\mathrm{AADT_k}$. In other words, the charging behavior decisions were influenced more by the trips that either originate from, or drive to at a nearby location, and were less related with the bypass traffic that may not stop at the TAZ at all. As such,

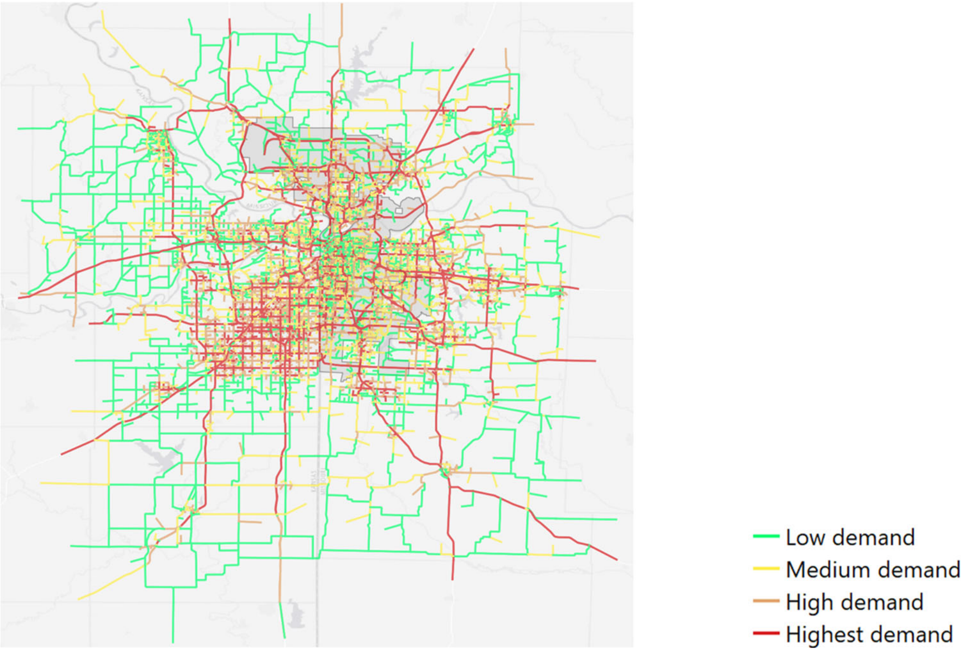

Figure 8. Traffic distribution on 24, 601 links of Kansas City.

Table 4. Modeling accuracy analysis.

Measures	Training	Validation	Testing
Sample Size	24066	6017	3343
R Square	0.5988	0.5372	0.5391
RMSE	0.1362	0.1460	0.1438
MAE	0.1037	0.1114	0.1098

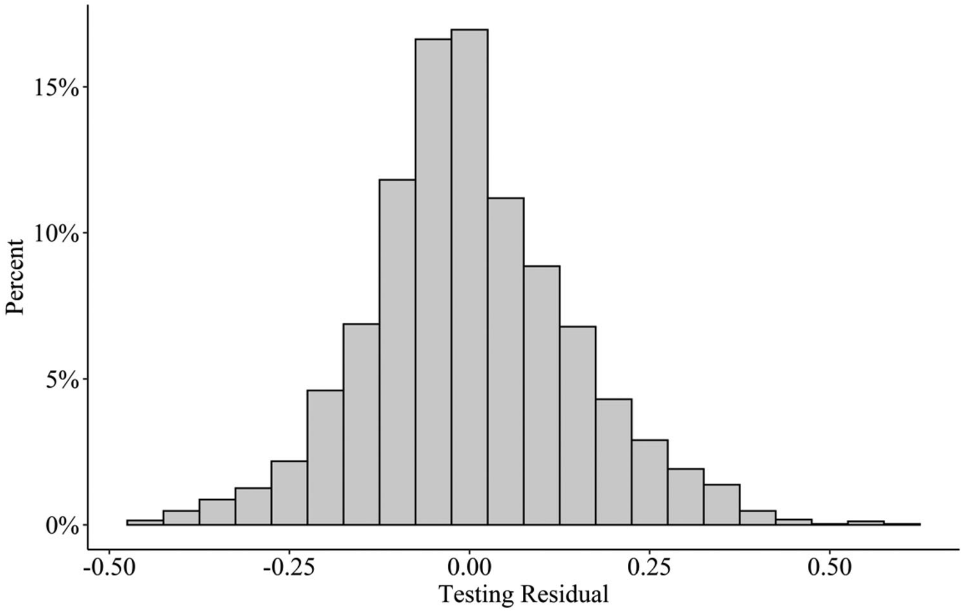

Figure 9. Distribution of prediction residuals.

when building new charging infrastructures, priorities should be given to the areas that generates or attracts higher traffic volume, rather than those with overall traffic volume which may include large amount of bypass traffic. Next, as DCH_{ij} was suggested to be the second most important feature, it was suggested that priorities should be given to the areas that are relatively further away from residential areas when

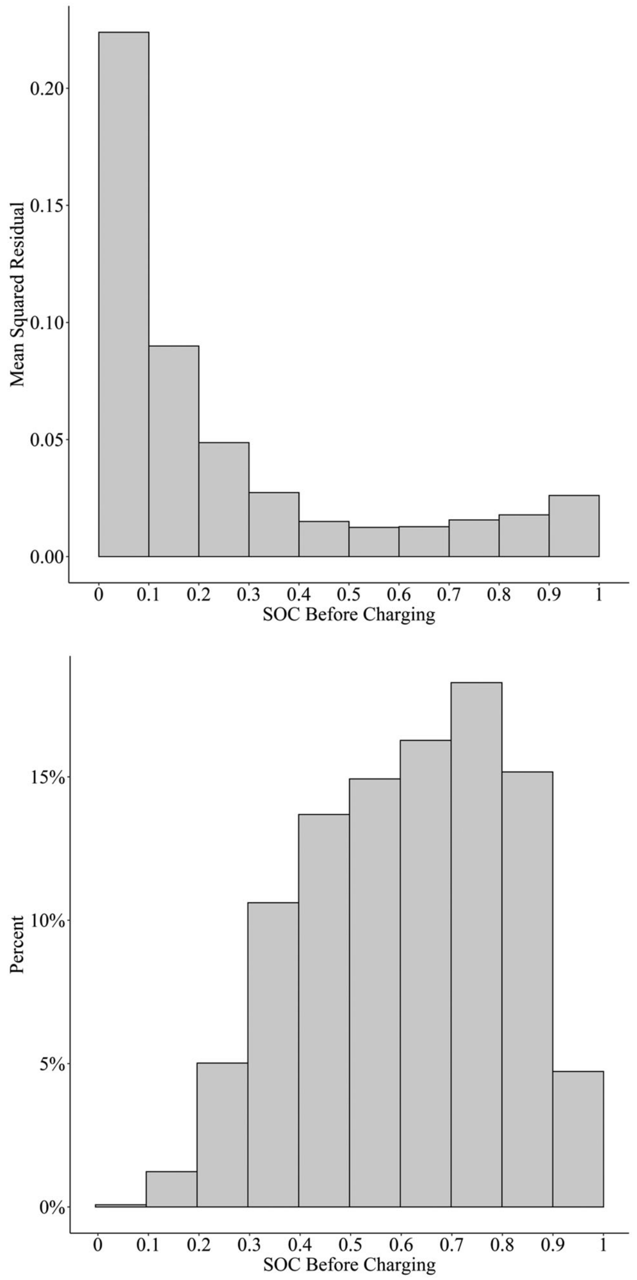

Figure 10. (a) Mean squared residual of testing samples; (b) SOC level of training samples.

building new charging infrastructures. Last but not the least, considering the positive relationship between driver's daily charging frequency and battery capacity of vehicle to initial SOC value, areas with more EV users who were considered conservative should be equipped with more charging facilities.

Comparison with benchmark models

In this section we compared the performance of the developed model with two popular prediction models. The first was a simple parametric non-ensemble model, multiple linear regression (MLR) model, and the other was a bagging-based ensemble learning model, random forest model (RFM). The comparisons were made from two aspects: computational efficiency and modeling accuracy.

Table 5. Importance of top ten features of the ensemble learning model.

Features	Importance
Driver's daily charging frequency (DCF_j)	100.00
Distance between charging location and home address (DCH_{ij})	94.40
Battery capacity of vehicle (C_j)	85.65
Daily traffic production (TP_k)	57.25
Average annual daily traffic ($AADT_k$)	56.11
Density of charging stations (DN_k)	33.72
Charging at night (TD_i)	25.74
Existence of charging station near driver's home ($NHCS_j$)	6.82
If the charging event happened near driver's home ($NHCG_i$)	3.58
Charging happened in business area ($LDU3_i$)	3.52

Computational efficiency

The numerical analysis was performed with R Statistical Software on a desktop computer with Intel (R) Core (TM) i9-9920X CPU @ 3.50 GHz. The computational efficiency of each model can be measured by the total setting up time, i.e., the sum of parameter tuning time and model training time with optimal parameter set. For the former, all the three models were tuned to identify their best possible performances to support fair comparison. For each model, we used 90% data as the training dataset and 10% data as the testing dataset. The non-ensemble MLR model was tuned by best subset selection to determine the optimal number of variables. In this case, when 15 features were selected, the MLR model generated the best possible prediction accuracy. With regards to the bagging-based ensemble RFM model, two essential parameters were tuned by five-fold cross validation with RMSE as a performance measure, including *mtry*, which was the value of the predictors randomly selected at each node, and *ntree*, which was the number of the bootstrap samples for each regression tree. It turned out that $mtry = 4$ and $ntree = 3100$ would allow the RFM to generate the best possible prediction accuracy. When it comes to the proposed boosting-based ensemble model, with five-fold cross validation, a total of four parameters tested, including 1) maximum depth–the maximum height of each tree in the learning algorithm, 2) minimum loss reduction for splitting, or gamma–the minimum loss reduction required to make a further partition on a leaf node of the tree; 3) learning rate–sometimes called shrinkage rate that represents how fast the algorithm moves in

Table 6. Computational efficiency of three models.

Model	Parameters	Range to test	Optimal value	Tuning time	Training time	Total setting up time
MLR	Number of variables	Min = 1, max = 17, step = 1	15	0.17 sec	0.10 sec	0.27 sec
RFM	mtry	Min = 1, max = 17, step = 1	4	26.7 min	9.21 min	614.04 min
	ntree	Min = 100, max = 5000, step = 100	3100	578.13 min		
Proposed	Max_depth	Min = 2, max = 10, step = 2	8	2.59 min	0.37 min	26.33 min
	gamma	Min = 0, max = 0.04, step = 0.01	0.01	2.57 min		
	eta	Min = 0.005, max = 0.025, step = 0.005	0.005	20.80 min		
	nrounds	Min = 100, max = 5000, step = 100	4300			

one step; and 4) number of trees–nrounds, a parameter to specify the number of trees to build in the learning process. Therein, the optimal learning rate and number of trees were jointly searched. After parameter tuning, each model can be trained to get its best performance. Table 6 also gives the setting up time of the three models.

Table 6 suggests that the MLR model ran the fastest in terms of both tuning and training, due to its simplicity in modeling. When comparing the two ensemble learning models, the proposed model was able to finish parameter tuning in less than 30 minutes, whereas the RFM took over 600 minutes. Moreover, the proposed model only needed less than one minute for training, yet the RFM costed over nine minutes. In total, the proposed model was demonstrated to run 23 times faster than the RFM, highlighting its high efficiency in the setting up process. Such difference in computational speed between the two ensemble learning models reveals their different way to train each weak learner. The bagging-based RFM model trained each regression tree independently and finally took the average of all the predictions from different trees. The boosting-based proposed model, on the other hand, sequentially fitted new sub-learner based on the residuals of the prediction of previous regression tree and, then, minimized the loss when adding the latest prediction. The fitting of consecutive trees could always reduce training loss utilizing the descending direction obtained from the prior tree. Thus, the proposed model converged faster than the RFM model. Also, since the proposed boosting-based model ensured that the accuracy could always be improved from the prior tree, its learning curve was smoother than RFM as Figure 11. In Figure 11a, as the increase of base trees of RFM, the training loss, i.e., RMSE, dropped overall, yet sometimes increased slightly when some poor trained trees were aggregated. However, in Figure 11b, the training loss of the boosting-based model always kept reducing without fluctuations until reaching convergence, when 4300 base learners were generated. Therefore, compared with bagging-based RFM, the proposed boosting-based model provided a faster and more stable training process toward convergence.

Modeling accuracy

After all three models were tuned with the optimal parameters, we firstly compared their prediction accuracy on the testing dataset. Similar to Section "Prediction accuracy", again we focused on the R square, RMSE and MAE values, and the results were shown in Table 7.

Table 7 suggested that the proposed model outperformed the other two models in all three aspects. Its R square value 0.5391 was the highest among the three, and was much higher than MLR model at 0.2623, marking an over 27% improvement. Its RMSE value was the lowest at 0.1438 which was much better than MLR model at 0.1450 and slightly better than RFM model at 0.1450. Its MAE value was again the lowest at 0.1098 which was much better than MLR model at 0.1506, and was slightly better than RFM model at 0.1107. These results consistently suggested that the simple parametric MLR, although simple and easy to implement, generated the worst performance, while the two non-parametric ensemble learning models were proven to be advantageous in prediction, and the proposed model slightly outperformed RFM model.

Conclusions

This manuscript develops a gradient boosting-based ensemble learning approach to model electric vehicle driver's range anxiety, and to understand at what battery percentages do EV drivers charge their vehicles, and what are the associated contributing factors. The real-world charging event log data from public charging stations in Kansas City, Missouri, and the macroscopic travel demand model maintained by the MPO are used. A total of 18 features are extracted from the multi-source data, and among them, seven features apparently had a more significant impact than the others. The model suggested that the driver's charging decisions were mostly affected by endogenous

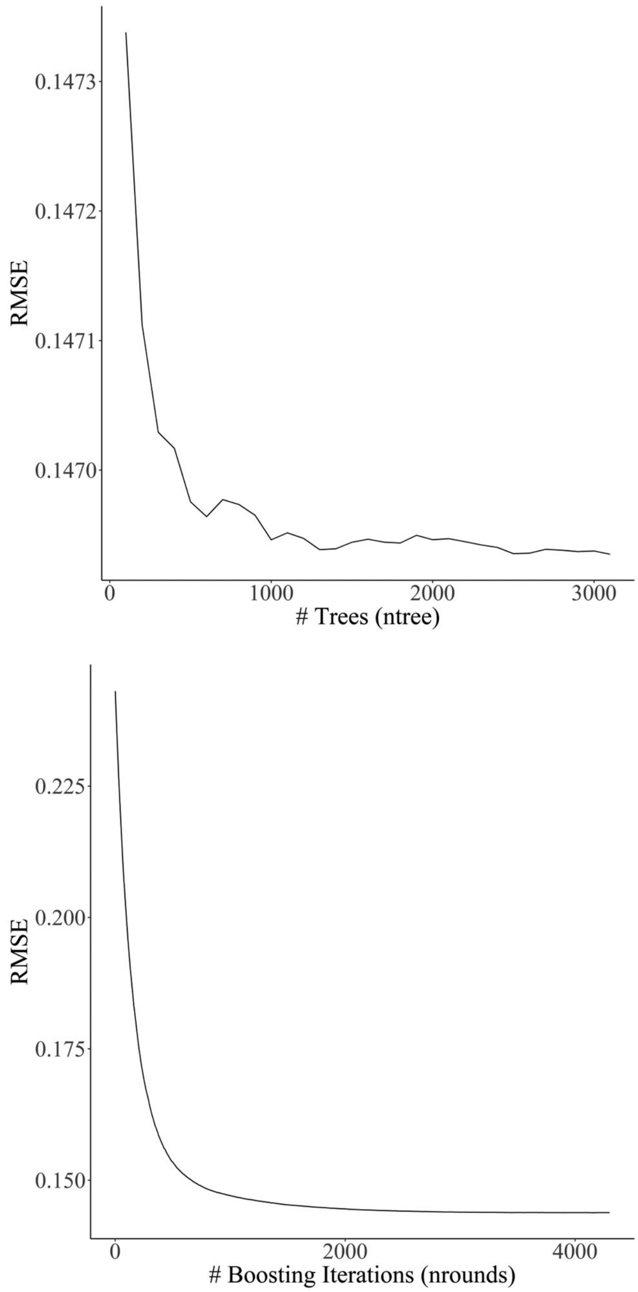

Figure 11. (a) Bagging-based RFM; (b) The proposed boosting-based model.

features, although context information, such as traffic or charging station availability, also played a role. The results showed that the proposed model could give a satisfactory result with a R Square value of 0.54 and root mean square error of 0.14, both better than multiple linear regression model and random forest model. The model also runs efficiently when compared with RFM, resulting in a 96% reduction in training time.

Table 7. Comparison among three models.

	R Square	RMSE	MAE
MLR	0.2623	0.1820	0.1506
RFM	0.5319	0.1450	0.1107
Proposed	0.5391	0.1438	0.1098

In addition to the offering of a technical approach to understand EV driver's range anxiety, the modeling results also lead to some suggestions on charging infrastructure locations to reduce the driving range anxiety of EV users. Three types of areas are suggested as the priorities, when it comes to the new charging infrastructure development. The first is the areas that are high in daily traffic prediction, i.e. the areas that generate or attract higher traffic volume, and they are suggested to be more important than those with overall traffic volume which may include large amount of bypass traffic. The second is the areas that are relatively further away from residential areas, as it is suggested to be the second most important feature, and finally, consistent with our general understand, the areas with more EV users who are considered conservative should be equipped with more charging facilities.

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This material is based upon work supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy's (EERE) Vehicle Technologies Office under the Award Number DE-EE008474. The authors are also thankful for support from the Metropolitan Energy

Center (MEC), the City of Kansas City Missouri (KCMO), Lilypad, Mid-America Regional Council (MARC), and Evergy (formerly Kansas City Power and Light Company (KCP&L)).

References

Ahmadi, P. (2019). Environmental impacts and behavioral drivers of deep decarbonization for transportation through electric vehicles. *Journal of Cleaner Production*, *225*, 1209–1219. https://doi.org/10.1016/j.jclepro.2019.03.334

Amini, M. H., Kargarian, A., & Karabasoglu, O. (2016). ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation. *Electric Power Systems Research*, *140*, 378–390. https://doi.org/10.1016/j.epsr.2016.06.003

Bonsu, N. O. (2020). Towards a circular and low-carbon economy: Insights from the transitioning to electric vehicles and net zero economy. *Journal of Cleaner Production*, *256*, 120659. https://doi.org/10.1016/j.jclepro.2020.120659

Chen, T. D., Kockelman, K. M., & Khan, M. (2013). Locating electric vehicle charging stations: Parking-based assignment method for Seattle, Washington. *Transportation Research Record: Journal of the Transportation Research Board*, *2385*(1), 28–36. https://doi.org/10.3141/2385-04

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). San Francisco, CA. Association for Computing Machinery.

Chung, Y.-W., Khaki, B., Li, T., Chu, C., & Gadh, R. (2019). Ensemble machine learning-based algorithm for electric vehicle user behavior prediction. *Applied Energy*, *254*, 113732. https://doi.org/10.1016/j.apenergy.2019.113732

Franke, T., Neumann, I., Bühler, F., Cocron, P., & Krems, J. F. (2012). Experiencing range in an electric vehicle: Understanding psychological barriers. *Applied Psychology*, *61*(3), 368–391. https://doi.org/10.1111/j.1464-0597.2011.00474.x

Guo, F., Yang, J., & Lu, J. (2018). The battery charging station location problem: Impact of users' range anxiety and distance convenience. *Transportation Research Part E: Logistics and Transportation Review*, *114*, 1–18. 18. https://doi.org/10.1016/j.tre.2018.03.014

Hao, X., Wang, H., Lin, Z., & Ouyang, M. (2020). Seasonal effects on electric vehicle energy consumption and driving range: A case study on personal, taxi, and ridesharing vehicles. *Journal of Cleaner Production*, *249*, 119403. https://doi.org/10.1016/j.jclepro.2019.119403

Jabeen, F., Olaru, D., Smith, B., Braunl, T., & Speidel, S. (2013). Electric vehicle battery charging behaviour: Findings from a driver survey. In *Proceedings of the Australasian Transport Research Forum*.

Khwaja, A. S., Venkatesh, B., & Anpalagan, A. (2020). Short-term individual electric vehicle charging behavior prediction using long short-term memory networks. In *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*.

Kim, S., Yang, D., Rasouli, S., & Timmermans, H. (2017). Heterogeneous hazard model of PEV users charging intervals: Analysis of four year charging transactions data. *Transportation Research Part C: Emerging Technologies*, *82*, 248–260. https://doi.org/10.1016/j.trc.2017.06.022

Morrissey, P., Weldon, P., & O'Mahony, M. (2016). Future standard and fast charging infrastructure planning: An analysis of electric vehicle charging behaviour. *Energy Policy*, *89*, 257–270. https://doi.org/10.1016/j.enpol.2015.12.001

Neubauer, J., & Wood, E. (2014). The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility. *Journal of Power Sources*, *257*, 12–20. https://doi.org/10.1016/j.jpowsour.2014.01.075

Nilsson, M. (2011). *Electric vehicles: The phenomenon of range anxiety*. E. Consortium.

Oliver, E. (2010). Diversity in smartphone energy consumption. In *Proceedings of the 2010 ACM Workshop on Wireless of the Students, by the Students, for the Students*.

Shepero, M., & Munkhammar, J. (2018). Spatial Markov chain model for electric vehicle charging in cities using geographical information system (GIS) data. *Applied Energy*, *231*, 1089–1099. https://doi.org/10.1016/j.apenergy.2018.09.175

Sun, X.-H., Yamamoto, T., & Morikawa, T. (2015). Charge timing choice behavior of battery electric vehicle users. *Transportation Research Part D: Transport and Environment*, *37*, 97–107. https://doi.org/10.1016/j.trd.2015.04.007

Xu, M., Meng, Q., Liu, K., & Yamamoto, T. (2017). Joint charging mode and location choice model for battery electric vehicle users. *Transportation Research Part B: Methodological*, *103*, 68–86. https://doi.org/10.1016/j.trb.2017.03.004

Xu, M., Yang, H., & Wang, S. (2020). Mitigate the range anxiety: Siting battery charging stations for electric vehicle drivers. *Transportation Research Part C: Emerging Technologies*, *114*, 164–188. https://doi.org/10.1016/j.trc.2020.02.001

Yang, Y., Yao, E., Yang, Z., & Zhang, R. (2016). "Modeling the charging and route choice behavior of BEV drivers. *Transportation Research Part C: Emerging Technologies*, *65*, 190–204. https://doi.org/10.1016/j.trc.2015.09.008

Yavasoglu, H., Tetik, Y., & Gokce, K. (2019). Implementation of machine learning based real time range estimation method without destination knowledge for BEVs. *Energy*, *172*, 1179–1186. https://doi.org/10.1016/j.energy.2019.02.032

Yi, Z., Liu, X. C., Wei, R., Chen, X., & Dai, J. (2021). "Electric vehicle charging demand forecasting using deep learning model. *Journal of Intelligent Transportation Systems*, 1–14. https://doi.org/10.1080/15472450.2021.1966627