IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 28, 2020, accepted August 4, 2020, date of publication August 7, 2020, date of current version August 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014851

Predicting the Regional Adoption of Electric Vehicle (EV) With Comprehensive Models

JIANMIN JIA[1], BAIYING SHI[1], FA CHE[2], AND HUI ZHANG[1]
[1]School of Transportation Engineering, Shandong Jianzhu University, Jinan 250101, China
[2]Zibo Transportation Service Center, Zibo 255000, China

Corresponding author: Fa Che (zbsglgljyanghuke@zb.shandong.cn)

This work was supported in part by the National Natural Science Funding, China, under Grant 41901396, and in part by the Doctoral Funding Program from the Shandong Jianzhu University under Grant X18052Z.

ABSTRACT Adoption of electric vehicles (EVs) has been regarded as one of the most important strategies to address the issues of energy dependence and greenhouse effect. Empirical reviews demonstrate that wide acceptance of EV is still difficult to achieve. This research proposes to investigate the factors that might trigger the wide usage of EVs to support the energy policy. The real-world owners of EV were extracted from the 2017 National Household Travel Survey (NHTS), which provides large-scale individual characteristics. NHTS dataset was processed to establish the comprehensive estimation model for EV adoption with considering vehicle, personal and household factors. Besides the commonly social-economic factors, the gasoline price and car sharing program were found to be significant for EV adoption. Additionally, since the EV owners are only 1.29% of all vehicle owners, this article introduced the imbalanced dataset technique, which was seldom considered in existing researches. Subsequently, several machine learning methods were utilized to build the prediction model, and the model performance analysis indicates the Decision Tree (DT) model outperforms other models. A regional EV penetration map was also generated for the U.S. to validate the proposed approach. Implications for further research, transport policy and EV market are discussed.

INDEX TERMS EV adoption, socio-economic factors, 2017 NHTS, imbalanced dataset, comprehensive models.

I. INTRODUCTION

Transportation has been considered as one of the major sources for greenhouse effect, since it generates over a quarter of the greenhouse gas [1]. Consequently, electric vehicle (EV), consuming clean energy, are generally believed to promote the sustainable transportation system and becoming increasing popular. However, the usage of EV is still low. In 2018, there are over 17 million automobiles sold in U.S., while the EV sales, including the battery electric vehicle (BEV) and plug-in hybrid electric vehicle (PHEV), only reached 361,307 units, occupying 2.09% of the auto sales market [2], [3]. Even in China, EV sales, over 1.2 million units, only accounts for 5.15% of the auto sales in 2018 [4]. Tremendous effort is still needed to promote the EV adoption around the world.

It is necessary to obtain the accurate estimation of EV usage to conduct the regional EV planning on sales market, charging infrastructure, etc. [5]. Multiple sources are utilized to infer the EV usage in recent researches. Some of them has conducted the analysis on charging infrastructure with considering the assumed traffic flow [6], [7], electric taxi or bus fleet travel [8], [9], which can't address the private EV usage. On the other hand, some other literatures have investigated the variables related to vehicle usage [10], [11], which attracts the vehicle manufacturers and governor' attention. However, Bjeerkan *et al.* [12] and Han *et al.* [13] argued that existing researches were mainly based on the stated preference (SP) survey with a few respondents, which can't illustrate the real EV market penetration.

Therefore, the 2017 National Household Travel Survey (NHTS), including over 200,000 real-world respondents, was utilized to explore the influencing factors for regional EV adoption. To build the prediction model, different machine learning methods are employed and compared. The implications for transport policy and EV market are also discussed. The rest of the paper is organized as follows. Section 2 provides the literature review. Subsequently, data sources and

The associate editor coordinating the review of this manuscript and approving it for publication was Chintan Amrit.

variables are addressed in section 3, followed by the method section. Section 5 and 6 provide the results and conclusion, respectively.

II. LITERATURE REVIEW

An extensive of literature has discussed the adoption behavior of innovations, which may be affected by the technology attributes, adopter's characteristics and social-economic environment. For EV adoption, the bahavioral response to purchase and use is commonly explored [14], [15]. The prediction models, involving economic factors, environmental factors, demographic factors, etc., is widely utilized to investigate the EV adoption behavior [16]–[18].

As introduced by the Ajzen [19], the theory of planned behavior (TPB) model focuses on the intended behaviour and the model is commonly utilized to interpret and explain the EV adoption bahavior in a few literatures. Among them, Moons and De Pelsmacker [20] explored the adoption of EVs through collecting consumer's attitude on EV price and performance. Egbue and Long [21] found that the battery capacity is still the most important factor when compared to the environmental factors for the consumer. Additionally, personal attribute, such as experience and knowledge, is also important in the TPB model. For instance, the consumer was found to be aware of environment issue with the increasing education level. Ziegler [22] concluded that the respondents are more willing to use the sustainable vehicle it they are concerned of the environment issues, based on the SP survey. Similarly, Daziano and Bulduc [23] investigated consumers' attitude towards the vehicle price and its environment performance. The consumer even wanted to spend more for the EV if they have environment concerns. Moreover, to explore the EV adoption, Wang *et al.* [24] also established an extended TPB model in terms of individual attitudes and sustainable factors.

On the other hand, Everett [25] proposed the diffusion of innovations (DOI) theory to explore the technology diffusion. Especially, the diffusion model is also introduced for vehicle adoption, which involves two categories [26]. The first category employed the traditional diffusion model. In terms of alternative fuel vehicle sales in German, The Bass diffusion model was developed and utilized by Massiani and Gohs [27]. The results indicated that the innovation coefficient was highly affected by the market scale. Through the SP survey, Cordill [28] investigated respondents's attitude on EV adoption. The EV price, gasoline price and gasoline consuming were found to be the three most important factors. In the research from Jensen *et al.* [29], the diffusion model was built in terms of the lag time for market share. The second category introduced the integration with agent-based and discrete choice model. For instance, the discrete choice model was combined with the diffusion model in Boston area [30]. The highest level of EV market share in 2030 was estimated to 22%. Additionally, the agent-based model was also used by McCoy and Lyons [31] to predict the EV diffusion. The agents were generated in terms of the socio-economic characteristics and environment attitude from the detailed survey microdata. Similarly, to investigate the regional EV market penetration, another agent-based model for was proposed by Noori and Tatari [32] with considering the government impacts. The results indicated that the government support plays an important role in promoting the EV usage.

In addition, some scholars also attempted to find the factors that have impacts on individual's adoption behavior. Li *et al.* [10] reviewed the related researches and summarized three groups of potential factors. Firstly, situational factors reflect the vehicle performance, like the vehicle price, battery range and vehicle emission [10], [33]. Besides the driving range and cost of EV, the environmental performance was also found to attract consumers [34]. Secondly, demographic factors describe the personal characteristics. The young male consumers were found to be more willing to use the EVs [35], [36]. Psychological factors are believed to affect the consumer's attitude directly. For instance, living experience of consumer are usually affecting their adoption decision other than the vehicle price [37].

However, aforementioned literatures are mainly based on the small-scaled SP surveys, in which the respondent is assumed to be EV user in terms of the response. It is hard to validate the purchasing behaviour even if the respondent intends to use the EV. Therefore, researchers started to conduct the analysis with the real-world EV users. For instance, Sang and Bekhat [38] conducted the regression analysis on EVs usage in terms of the real-world EV drivers. The results provide recommendations for the government policy and EV market penetration. Moreover, Javid and Nejat [39] developed a logistic regression model with 2012 California Household Travel Survey (CHTS). Prediction results were validated with the real EV penetration data.

Thus, the comprehensive model in terms of the real-world EV usage data to investigate the influencing factors for EV adoption are still meaningful. This article investigated the 2017 NHTS, involving over 200,000 respondents. The imbalanced dataset issue, rarely considered by existing studies, was also addressed in the study. Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) were utilized to build the comprehensive prediction model. The details of the methodology and assumptions are described in the following section.

III. DATA SOURCES AND VARIABLES

A. 2017 NHTS

Ranging from April 2016 to April 2017, the national travel survey, 2017 NHTS, includes four datasets to explain the demographics, household information, travel information and vehicle characteristic. In the 2017 NHTS dataset, over 250 thousand vehicles with various type are included. Especially, due to the collection on different day, the trip-dataset was removed from the dataset. The other three datasets were combined to conduct the analysis for adoption behaviour of EV or conventional vehicle (CV). CV mainly refers to the

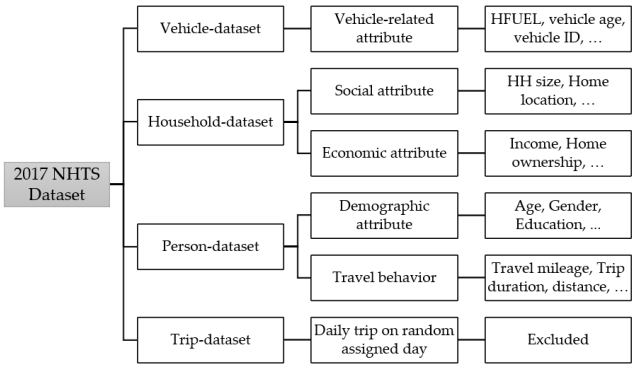


FIGURE 1. 2017 NHTS framework.

gasoline vehicle and diesel vehicle. Fig. 1 shows the detailed structure of the NHTS, while the trip-dataset is excluded from the analysis. Subsequently, the potential factors affecting EV adoption are explored in terms of the other three datasets.

B. VEHICLE-RELATED VARIABLES

As mentioned above, the 2017 NHTS was used to explore the EV adoption behavior. The respondents in the survey are assumed to give the accurate information. Additionally, the new vehicle users, purchasing the vehicle recent years, in the U.S. were selected to investigate the latest trend for EV adoption. The vehicle category was obtained from the variable ''HFUEL'', which presents the vehicle energy description in this article. In order to distinguish the vehicle category, EV and conventional vehicle (CV) are defined. The EV includes the BEV and PHEV that use the battery, while the CV is defined as the vehicles consume gasoline and diesel.

Nevertheless, one respondent may have more than one vehicle, which can cause mistake the classification analysis for the influencing factors. To address this issue, it is assumed that the respondent is the owner for the latest vehicle, while the information for other vehicles are discarded. Moreover, the vehicle age is also restricted within three years to illustrate the vehicle adoption trend and potential vehicle market. Additionally, the respondent would be removed from the dataset, if the essential variables are missing, such as the basic individual characteristics, vehicle category and family characteristics. Ultimately, through the data cleansing process, a total of 31,322 respondents were kept for the following analysis. This study didn't adjust the sample, since the sample bias correction has already been involved in the NHTS.

C. HOUSEHOLD-RELATED VARIABLES

The household-related variables can be categorized into two groups: economic and social variable. In several empirical researches, the ''Household Income'' variable was found to affect the EV sales [40], [41], while it was believed to be ineffective in the research from Sierzchula *et al.* [42]. In this study, the annual household income is defined as the categorical variable, which consist of five groups: 1 = 25,000$ or less, 2 = 25,001$ to 50,000$, 3 = 50,001$ to 75,000$, 4 = 75,001$ to 100,000$, 5 = 100,001$ or more. The ''Homeown'' variable, revealing the economic status of one household, is the other economic variable involved in the analysis.

For social variables, the ''Household size'' defines the number of family members in the household. As this variable may be correlated to the vehicle size decision and seat usage, it is believed to have impacts on vehicle category choice. Additionally, the variable ''Young child'' can also be related to the vehicle choice, since the young child under 4 years requires the baby chair. The variable ''Household vehicle'', defining the total vehicles count for all the family members, was believed to have impacts on the vehicle adoption [43]. Moreover, ''Urban rural'' is another household-related variable, which describes the impact from the adjacent environment and transportation infrastructure nearby. The variable ''Population density'' was selected to address the issue whether the vehicle adoption behaviour is affected by the population around the family. Eight categories of the population density are described by categorical variables: 1 = 1∼100, 2 = 101∼500, 3 = 501∼1,000, 4 = 1,001∼2,000, 5 = 2,001∼4,000, 6 = 4,001∼10,000, 7 = 10,001∼25,000, 8 = more than 25,000.

In contribution to this research, several questions correlated to respondent's attitude are also included. The attitudes are believed to affect the choice of travel pattern. One of them is the ''Price'' variable, defining whether the gasoline price has impacts on the EV adoption or not. It is a categorical variable containing five categories: 1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree. The survey also collects respondent's attitude on travel expense based on the variable ''Place''. Similarly, the same category definition is provided to variable ''Price''.

D. PERSON-RELATED VARIABLES

The demographic characteristics and travel information for each household member are defined with the person-related variables. Among them, gender and age are the widely used variables to present the individual characteristic. Thus, statistical test on gender and age difference was conducted in this article. ''Education'', defining the individual characteristic, is believed to have impacts on EV adoption [42]. Considering the various level of education, it is defined as the categorical variable with five groups containing primary school, high school, college, bachelor and graduate level. Moreover, this article also attempts to investigate whether the ''RACE'' variable have impacts on the daily travel patterns. Similarly, Race is defined as the categorical variables. In order to demonstrate the working or employing status for each person, the ''Multi-job'' and ''Occupation'' are selected. ''Multi-job'' variable is related to the number of jobs for each respondent and ''Occupation'' variable explains the job characteristic with 5 groups: 1 = service, 2 = government, 3 = factory or farming, 4 = professional, 5 = not employed.

Subsequently, this reseach also explore the variables affecting travel pattern. One of the variables is ''Car sharing'',

TABLE 1. Statistics of the variables.

Variable	Mean	St. Dev.	Min	Max
EV	0.012	0.199	0	1
Household income	3.824	1.304	1	5
Home own	0.851	0.356	0	1
Household size	2.447	1.173	1	12
Young child	0.110	0.391	0	4
Household vehicle	2.426	1.185	1	12
Urban rural	0.761	0.426	0	1
Population density	3.843	1.301	1	8
Price	2.984	1.283	1	5
Place	3.003	1.052	1	5
Age	53.372	15.81	18	92
Gender	0.536	0.498	0	1
Education	3.639	1.069	1	5
Race	1.392	1.196	1	7
Multi-job	0.055	0.227	0	1
Occupation	3.835	1.325	1	5
Car sharing	0.005	0.074	0	1
Time to work	14.772	25.276	0	600
Year mile	10955.48	11951.90	0	200,000
Sample number	31,322			

which explains the frequency for the respondent to attend the car sharing program. It is found that the car sharing programme may affect the people's travel pattern and vehicle choice [39]. Another variable is the ''Time to work'', which describes the average daily travel time for commute trip. The similar variable ''Trip distance'' is not selected, since it is hard to predict the accurate gasoline usage [44]. ''Year mile'' is the other variable related to the travel pattern. It describes the annual driving mileage of the respondent and demonstrates the vehicle performance on battery capacity and gasoline usage.

After the data cleansing process for the NHTS dataset, 31,322 samples are kept. Table 1 presents the variables summary in this article. According to the 2017 NHTS, only 1.29% of the vehicles are EV. On average, there are 2.4 vehicles and 2.5 family members for a household. The annual average family income is over 70,000$. The mean number of young children is 0.11 for each household. Interestingly, the average value for attitude variable Price and Place are both close to the third category, which illustrates the balanced attitude between the respondents. Additionally, there is no significant gender difference for the number of female and male respondents in the sample. Moreover, the medium and old person are more likely to adopt a vehicle, as 53 is the mean age of the vehicle owner. For the travel pattern analysis, the respondents that have ever used the car sharing program only occupies 0.5%, which is a low percentage. On the other hand, it takes about 14.7 minutes to the work place, and the annual driving mileage is over 10,000 miles.

IV. METHODOLOGY

It is generally a classification problem to distinguish the vehicle type for the adoption. Aforementioned variables, involving discrete categorical variables and continuous numerical variables, were explored by the machine learning approaches. Additionally, the imbalanced dataset problem and corresponding adjustment are also discussed.

A. LOGISTIC REGRESSION (LR) MODEL

First defined in 1960s, logistic regression model (LR) is widely used to deal with the discrete choice problem [45]. LR can deal with the classification problem through incorporating multiple independent variables. Therefore, it was employed to explore the EV adoption behavior. In this study, there are two vehicle categories, namely EV and CV, which are presented by Y. The independent variables are defined by X, which was expressed as

$$ln\left(\frac{p(Y_i = m|X)}{p(Y_i = 1|X)}\right) = \beta_m + \sum_{j=1}^{n} \beta_{mj} X_{mj} = Z_{mi} \quad (1)$$

Thus, the equations utilized to generate the probability of the vehicle usage can be denoted by

$$p(Y_i = 1|X) = \frac{1}{1 + \sum_{h=2}^{M} exp(Z_{hi})} \quad (2)$$

$$p\,(Y_i = m\,|\,X) = \frac{exp(Z_{mi})}{1 + \sum_{h=2}^{M} exp\,(Z_{hi})} \quad m = 2, \cdots, M \quad (3)$$

where, M is the vehicle type number. $X = X_1, X_2, \cdots, X_n$ represents the influencing factors, while n is the number of factors. β_0 denotes interception condition, whicle β is the coefficients.

B. NAÏVE BAYES (NB) CLASSIFIER

Naïve Bayes (NB) classifier is also commonly utilized for the classification problems [46]. Through computing the prior probability of the category, the NB classifier is capable to infer the most likely class. In this study, the prior probability for category Y_i can be expressed by

$$p\,(Y_i\,|\,X) = \frac{p\,(X\,|\,Y_i)\,p\,(Y_i)}{p\,(X)} > p((Y_k\,|\,X)\,,\ \forall 1 \le i \ne k \le m \quad (4)$$

Furthermore, the NB classifier could be simplified through maximizing the $p\,(X\,|\,Y_i)\,p\,(Y_i)$. Therefore, the calculation of prior probability can be converted to

$$Y_{NBC} = \underset{Y_i \in Y}{argmax}\, p(Y_i) \prod_{j=1}^{n} p(X_j|Y_i) \quad (5)$$

where, Y_i defines the dependent variables. $X = X_1, X_2, \cdots, X_n$ defines the independent variables. $p(Y_i)$ defines the prior probability of the vehicle class Y_i.

C. SUPPORT VECTOR MACHINES (SVM) MODEL

To distinguish different classes, Vapnik proposed the SVM to search the optimal hyper-plane [47]. Let $X=(X_1, X_2, \cdots, X_n)$ be the independent variables, while the vector $Y = (Y_1, Y_2)$ presents the vehicle type. Thus, the classification function can be denoted by

$$f_x = sign[\sum_{i=1}^{n} \alpha_i Y_j * k\,(X, X_i) + c] \quad (6)$$

where, c is the offset from the origin of the hyper-plane. n presents the independent variables number. α_i defines the

TABLE 2. Description of statistical indexes.

Metric	Method	Explanation
True Positive (TP)	Obtained directly	Number of EV samples classified with true value
True Negative (TN)	Obtained directly	Number of CV samples classified with False value
False Positive (FP)	Obtained directly	Number of EV samples classified with false value
False Negative (FN)	Obtained directly	Number of CV samples classified with True value
True Positive Rate (TPR)	$TPR = TP/(TP + FN)$	The proportion of EV adoption that are classified correctly as "EV".
True Negative Rate (TNR)	$TNR = TN/(TN + FP)$	The proportion of CV adoption that are classified correctly as "CV".
Accuracy (ACC)	$ACC = \frac{TP + TN}{TP + FP + TN + FN}$	The proportion of EV and CV adoption that are classified correctly, which indicates the predictive capability.
Receiver Operating Characteristic (ROC)	AUC value	AUC defines the score of model performance in terms of the area under the ROC curve

positive constant. $k(X, X_i)$ is the kernel function. For EV and CV classification, the Equation 6 can be solved in terms of

$$Y_j\left[\omega^T\varphi(X_i)+c\right] \Longleftrightarrow \begin{matrix}\omega^T\varphi(X_i)+c \geq 1, if\, Y_j = +1(AFV)\\ \omega^T\varphi(X_i)+c \leq 1, if\, Y_j = -1(CV)\end{matrix} \tag{7}$$

where $\varphi(X_i)$ is a nonlinear function to divide the space. ω presents the weight.

D. DECISION TREE (DT) MODEL

As the non-parametric supervised approach, Decision Tree (DT) is usually utilized to solve the prediction and classification problems [48]. In this article, the DT model is utilized for EV classification. Establish and prune are the two steps modelling of DT. It is built to produce a largest-sized tree and conduct self-prunes after sensing the ideal pruning threshold. Sequentially, the classification for each sub-node is based on the gain-ratio. It can be computed with following equations [49].

$$GainRatio(X, T) = \frac{Gain(X, T)}{SplitInfo(X, T)} \tag{8}$$

$$Gain(X, T) = Entrophy(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} Entrophy \tag{9}$$

$$SplitInfo(X, T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} log_2 \frac{|T_i|}{|T|} \tag{10}$$

where, T is the training dataset, while $T_i(i = 1, 2, \cdots, n)$ is the subset. X presents the influencing factor.

E. RANDOM FOREST (RF) MODEL

The RF model consists of a bunch of decision trees [50]. In this article, the classification tree is established in terms of the EV adoption samples. Through identifying the prediction variables, each node within the tree is built. Subsequently, the optimal split is determined through maximizing the gain-ratio mentioned above. Additionally, to calculate the factor impurity belonging to each category, the Gini-Index is used to select the factor. It can be computed by following equation.

$$GiniIndex = \sum\sum_{j \neq i} \left(\frac{f(Y_i, T)}{|T|}\right)\left(\frac{f(T_j, T)}{|T|}\right) \tag{11}$$

where, T presents the training dataset. $\frac{f(Y_i,T)}{|T|}$ presents the probability belonging to category Y_i.

TABLE 3. VIF analysis.

Variable	VIF	1/VIF
Household income	1.7646	0.5666
Home own	1.4503	0.6895
Household size	2.5583	0.3909
Young child	1.6560	0.6039
Household vehicle	1.7603	0.5681
Urban rural	2.8048	0.3565
Population density	3.2112	0.3114
Price	2.4172	0.4137
Place	2.2513	0.4442
Age	2.4904	0.4015
Gender	1.1082	0.9023
Education	1.4539	0.6878
Race	1.3002	0.7691
Multi-job	1.0661	0.9380
Occupation	1.4690	0.6807
Car sharing	1.0113	0.9888
Time to work	1.2879	0.7764
Year mile	1.2173	0.8215

F. DATASET ADJUSTMENT AND MODEL PERFORMANCE EVALUATION

According to the statistics summary of NHTS, the EV only occupies 1.29% of the surveyed vehicle, which is an extremely imbalanced dataset. The imbalanced distribution issue that might lead to the biased classification have been rarely addressed in the literature related to the vehicle adoption. Zheng *et al.* [51] reviewed the techniques to deal with the imbalanced dataset and found that oversampling approach and undersampling approach are commonly used to generate the adjusted dataset.

Generally, the goal of oversampling approach is to gain the samples belonging to the minority category and keep the majority category samples. The gained samples are generated through duplicating the original samples. Nevertheless, the classification model will be overfitted with the constructed samples. On the contrary, reducing the sample number is the basic rule for undersampling approach. Similarly, the discarded samples are selected randomly. The disadvantage of undersampling approach is the sample waste of the original dataset.

Some other improved sampling approaches are also developed based on the original oversampling and undersampling algorithm. For instance, Chawla *et al.* [52] defined Synthetic Minority Over-sampling Technique (SMOTE) as a heuristic sampling approach, which is also utilized in vehicle usage

TABLE 4. Variables coefficients.

Variable	B	Std. Error	p	Exp(B)
Household income	0.269	0.032	0.000	1.309
Home own	0.382	0.169	0.024	1.465
Household size	0.125	0.043	0.004	1.133
Young child	0.181	0.105	0.085	1.199
Household vehicle	0.268	0.058	0.000	1.307
Urban rural	0.266	0.215	0.215	1.305
Population density	0.240	0.039	0.000	1.271
Price	0.146	0.051	0.004	1.157
Place	0.136	0.058	0.020	1.146
Age	0.001	0.004	0.879	1.001
Gender	-0.512	0.104	0.000	0.599
Education	0.524	0.061	0.000	1.689
Race	0.077	0.037	0.039	1.080
Multi-job	0.380	0.186	0.041	1.463
Occupation	0.045	0.047	0.344	1.046
Car sharing	0.340	0.086	0.045	0.000
Time to work	0.002	0.002	0.161	1.002
Year mile	0.000	0.000	0.011	1.000
Constant	-10.023	0.413	0.000	0.000

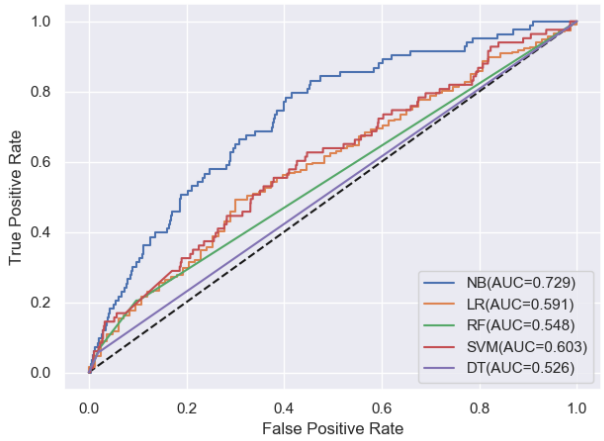


FIGURE 2. AUC value for the prediction models.

researches [53]. The advantage of SMOTE is to generate the new sample other than duplicate the existing sample, which is believed to well fit the prediction model. The generated samples can be denoted by

$$S_{new} = S_i + \omega(S' - S_i) \tag{12}$$

where, S_i presents the samples belonging to minority category (EV). S' is the selected sample close to S_i. ω defines the weight.

Moreover, the performance of prediction models should be evaluated and compared for both training and testing datasets. Pham *et al.* [46] listed many qualitative and quantitative methods used to validate the performance of classification models. In this article, the prediction model is built for EV and CV adoption among the respondents. Adoption of EV is defined as the true value with 1, while adoption of CV is defined as the false value with 0. The detailed description of statistical index is presented in Table 2.

TABLE 5. Model performance with imbalanced testing subset.

Parameter	NB	LR	RF	SVM	DT
TP	0	0	0	0	2
FP	3	0	36	0	330
FN	248	248	248	248	226
TN	6014	6017	5981	6017	5687
TPR	0.0000	0.0000	0.0000	0.0000	0.0088
TNR	0.9995	1.0000	0.9940	1.0000	0.9452
ACC	0.9599	0.9604	0.9547	0.9604	0.9110

V. RESULTS AND DISCUSSION

A. EXPLORING THE VARIABLES FOR PREDICTION MODELS

According to the section 3, 18 variables were selected as the potential variables to establish the prediction model for EV adoption. Nevertheless, each variable contributes differently in the model, which requires a statistical test to check the variable significance and discard the insignificant ones.

Variance Inflation Factor (VIF) is a commonly used measurement for the multicollinearity test. It is easy to calculate the VIF and the high value means the high potential of collinearity between variables. Javid and Nejat [39] explained the procedure to compute the multiple correlation coefficients for variables and VIFs can be expressed as

$$VIF_j = \frac{1}{1 - R_j^2} \tag{13}$$

where, R_j^2 presents the multicollinearity coefficients.

Generally, R_j ranges from 0 to 1, while 0 means there exists no multicollinearity issue for variable x_j. Similarly, the value of VIF_j changes with the R_j. It indicates potential multicollinearity problem if the value of VIF is higher than 10. As presented in Table 3, all the VIFs are lower than 10, which means no multicollinearity among them.

On the other hand, the significance test of the potential influencing factors in the prediction model was

TABLE 6. Model performance using balanced testing subset.

Parameter	Training subset					Testing subset				
	NB	LR	RF	SVM	DT	NB	LR	RF	SVM	DT
TP	4	5	1458	23	1763	0	0	151	2	188
FP	48	46	59	0	0	6	0	17	0	67
FN	2017	2021	562	1997	257	260	260	110	258	72
TN	46012	46008	46000	46059	46059	5999	6005	5988	6005	5937
TPR	0.0019	0.0000	0.7217	0.0114	0.8727	0.0015	0.0000	0.5788	0.0092	0.7218
TNR	0.9990	1.0000	0.9987	1.0000	1.0000	0.9991	1.0000	0.9972	1.0000	0.9888
ACC	0.9571	0.9580	0.9871	0.9585	0.9947	0.9576	0.9585	0.9799	0.9588	0.9777

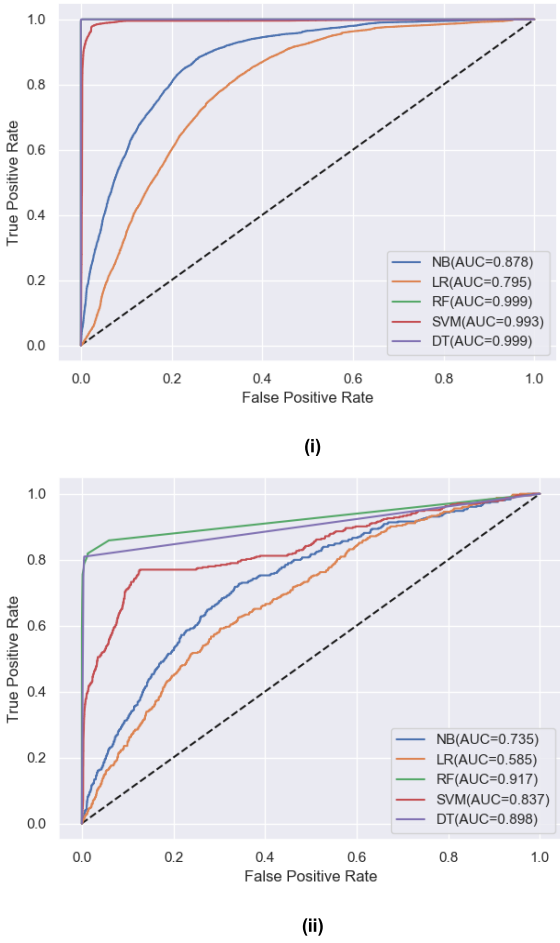


FIGURE 3. AUC value of prediction model using adjusted (i)training dataset, (ii)testing dataset.

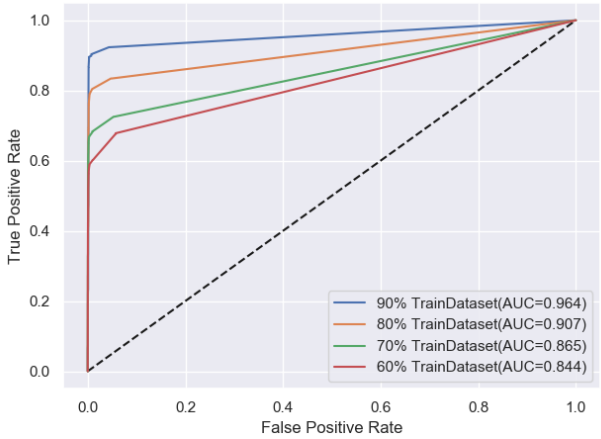


FIGURE 4. AUC value for different training dataset.

also conducted. As a widely employed approach for factor analysis, the backward elimination (BE) approach was used to select the variables in this article. There are two steps for BE. Firstly, the contribution of each factor is calculated. And secondly, the insignificant factors that contribute least to build the model will be removed. It is a repeated process until all the factors are within the criterion. In this article, 0.05 is set as the entry and removal criteria for the p value of variable. In terms of logistic regression analysis, the BE approach was used to check the combination of variables. The variables coefficients are presented in Table 4. In the table, B, Std. Error, p and exp(B) represents the coefficients, stand error, p value and exponential value of coefficients, respectively.

Moreover, the table suggests that the variables, such as young child, urban rural, age, occupation and time to work, should be removed from the prediction model, as their p values are higher than 0.05. Especially, besides the commonly used social-economic variables, the Car sharing and Price variable were found to be significant in the prediction model, which provide the evidence for the future policymaking Subsequently, 13 variables are selected for the analysis in the following section.

B. PREDICTING BASED ON THE ORIGINAL IMBALANCED DATASET

Original dataset described in data source section was separated to testing dataset and training dataset. The samples proportion between them is 20% and 80%. As described above, the prediction model for EV adoption is built with the training dataset, while the testing dataset is for validation.

The partition of original dataset is a completely random process and the distribution between EV and CV for the subsets is consistent with the original dataset. Thus, there are 321 owned EVs among 25,057 samples within the training subset. Similarly, there are 83 owned EVs among 6,265 samples within the testing subset. In order to construct the prediction model, Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Tree and Random Forest approaches were utilized in terms of the training subset. The proposed models were validated with the testing dataset. Various statistical indexes were used to measure and compare the model performances, which are presented in Figure 2 and Table 5.

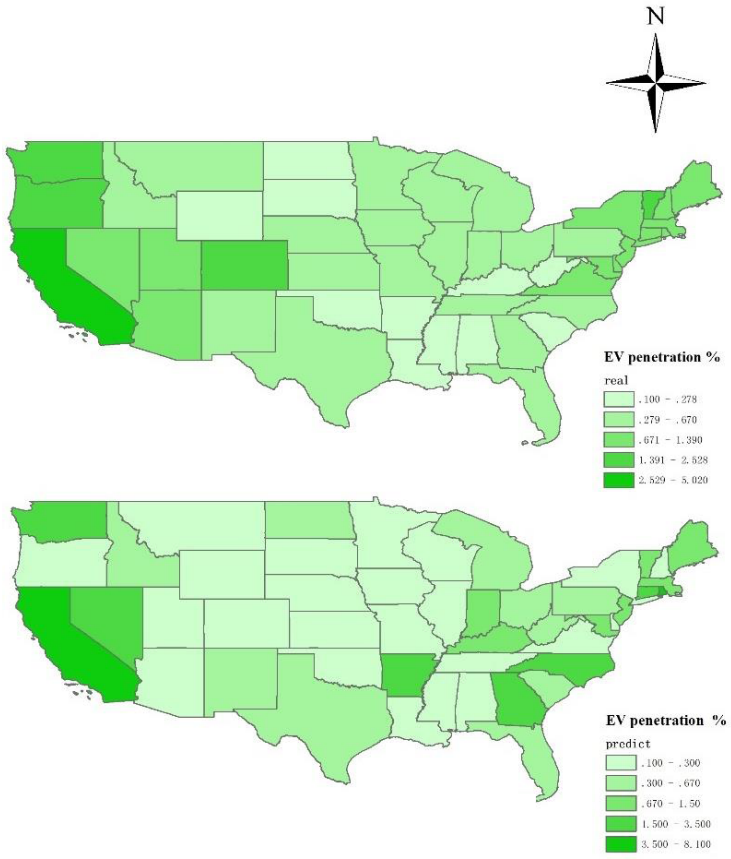


FIGURE 5. The EV penetration map derived from the real-world and prediction model.

The results suggest that five models all have a well accuracy (ACC) score higher than 0.9. However, these models are not applicable as the TP value, that is the true prediction of EV adoption, is really low. Additionally, the AUC value in Figure 2 also prove the unacceptable performance of prediction models, since the AUC value ranges from 0.526 to 0.729. It is believed that the imbalanced distribution of the original dataset leads to the lack of data to build the prediction model.

C. PREDICTING BASED ON THE ADJUSTED DATASET

As described in methodology section, the training subset was increased to 37,208 samples and 12,472 samples were defined to adopt the EV in terms of the SMOTE approach. The proportion between EV user and CV user is 34% and 66%, which indicates a relatively balancing distribution of the training subset.

Same to the imbalanced dataset, five prediction models for EV adoption were built in terms of the adjusted training subset. Similarly, the statistical indexes utilized to measure and compare the model performances are presented in Figure 3 and Table 6. It can be found that the ACC score for the proposed models are still higher than 0.9 except for the NB and LR models with the adjusted training dataset. Besides, the RF, SVM and DT models have well prediction with the TP and applicable TPR. In the model testing, SVM, DT and RF model have the high ACC score, nevertheless DT model have the best TP estimation. Moreover, Figure 3 indicates that the DT and RF model performs well with both training dataset and testing dataset. However, DT model provides the higher TPR, which indicates the better prediction capability for EV adoption.

Subsequently, a sensitivity analysis was also conducted for the proposed DT model. A combination of various proportion, 60%, 70%, 80% and 90%, of the original dataset was used to generate the balanced training subset to explore the impacts of sample number. Same to the model performance analysis, AUC value of the prediction model was computed and compared, which is shown in Figure 4. The results indicate that the prediction model could have better performance if more samples are provided. When 90% of the original dataset was selected to establish the prediction model, the AUC is the highest, that is 0.94.

D. VISUALIZATION OF PREDICTED EV PENETRATION

In order to illustrate the prediction results of the proposed DT model, the ArcGIS was utilized to generate the visual EV penetration map. Especially, the regional EV penetration analysis can be conducted if the regional demographics, social-economical and vehicle related information are provided. As described above, there are 31,322 samples in the original dataset, while 404 of them have EVs. Through the proposed DT prediction model, there are 396

estimated samples owning EVs, which demonstrates a well estimation.

To the best of the author's knowledge, the 2017 EV sales data [54], providing the real-word EV penetration, was compared with the estimation based on the 2017 NHTS. Figure 5 presents the national EV penetration model derived from the real world and prediction model for EV adoption. It indicates the well performance of the proposed prediction model due to the similar distribution between the EV penetration maps. Moreover, it can be found that the California state is more willing to adopt the EV in terms of the EV penetration level.

VI. CONCLUSION

In order to reduce the pollutants emission in the transportation sector, EV has been regarded as an ideal solution. In this article, the large-scale 2017 NHTS was utilized to explore the potential factors that are deemed to be associated with EV adoption and the proposed approach is believed to support the government and EV manufacturer to promote the sustainable transportation. Firstly, the real-world EV users other than the intended users were extracted from the NHTS. To determine the predicting variables, the BE regression analysis was conducted to measure the contribution of each variables and discard the insignificant variables. Therefore, only 13 variables were kept to establish the prediction model for EV adoption. Besides the factors investigated in previous researches, this article proposed to explore the attitude's impacts on the car sharing program and gasoline price, which was innovatively involved.

In addition, 31,322 samples were extracted from the 2017 NHTS dataset, while only 1.29% of the samples own an EV. It is an extremely imbalanced distribution dataset that may lead to the biased classification for EV adoption, which was proved by the results section. Thus, this research proposed to adjust the original dataset with SMOTE approach. Subsequently, Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Tree and Random Forest approaches were utilized to build the prediction models based on the training subset. Model performance analysis indicated that the DT model is the best prediction model with both high True Positive Rate (TPR) and AUC value. Additionally, the proposed model was utilized to output a national EV penetration map, which illustrates the trend of the regional EV usage.

From a policy perspective, this article signifies the social-economical and normative influencing factors on EV users. The proposed approach is meaningful for governors and manufactures to understand regional EV adoption and make policies to promote the EV usage. Besides the commonly considered variables, Car sharing and Price were found to be significant for EV usage. This knowledge is important when developing polices for decreasing greenhouse gas emissions and promoting the sustainable transportation. For instance, policy makers could encourage the car sharing programme to enhance individual's environment concern. The fuel costs advantage for EV technology can also be emphasized in the policy to make vehicle consumers to adopt the EV.

Due to the lack of data, it still takes a huge effort to explore the impacts from government regional policy and personal psychographics on EV adoption in the future researches. For instance, the incentives and tax policy between different states can be measured and compared in terms of the prediction model for EV adoption. Furthermore, the factors correlated to the PHEV and BEV adoption may not be the same. More analysis should be conducted to investigate the two vehicle categories separately. The recent developed machine learning approach, such as XGBOOST [55] and LightGBM [56], can also be considered in the future work.

REFERENCES

[1] U. S. Environmental Protection Agency (EPA). *Fast Facts on Transportation Greenhouse Gas Emissions*. Accessed: Apr. 29, 2020. [Online]. Available: https://www.epa.gov/greenvehicles/fast-facts-transpor tation-greenhouse-gas-emissions

[2] U. S. Department of Energy. *U.S. HEV Sales by Model (1999-2017)*. Accessed: Apr. 29, 2020. [Online]. Available: https://afdc.energy.gov /data/

[3] GreenTechMedia. *US Electric Vehicle Sales Increased by 81% in 2018*. Accessed: Apr. 29, 2020. [Online]. Available: https://www. Greentechmedia.com/articles/read/us-electric-vehicle-sales-increase-by-81-in-2018#gs.afzlad

[4] Detroit Free Press. *China Threatens to Sideline U.S. Automakers in Booming EV Tech*. Accessed: Apr. 29, 2020. [Online]. Available: https://www.freep.com/ story/money/cars/mark-phelan/2019/03/27/china-electric-vehicles-production/3217195002/

[5] J. Jia, C. Liu, and T. Wan, "Planning of the charging station for electric vehicles utilizing cellular signaling data," *Sustainability*, vol. 11, no. 3, p. 643, Jan. 2019.

[6] S.-Y. Ge, L. Feng, H. Liu, and L. Wang, "The planning of electric vehicle charging stations in the urban area," in *Proc. 2nd Int. Conf. Electron. Mech. Eng. Inf. Technol.*, 2012, pp. 1598–1604.

[7] C. Wu, C. Li, and L. Du, "A method for electric vehicle charging infrastructure planning," *Automat. Electr. Power Syst.*, vol. 34, no. 24, pp. 36–39. 2010.

[8] N. Shahraki, H. Cai, M. Turkay, and M. Xu, "Optimal locations of electric public charging stations using real world vehicle travel patterns," *Transp. Res. D, Transp. Environ.*, vol. 41, pp. 165–176, Dec. 2015.

[9] W. Tu, Q. Li, Z. Fang, S. Shaw, B. Zhou, and X. Chang, "Optimizing the locations of electric taxi charging stations: A spatial–temporal demand coverage approach," *Transp. Res. C, Emerg. Technol.*, vol. 65, pp. 172–189, Apr. 2016.

[10] W. Li, R. Long, H. Chen, and J. Geng, "A review of factors influencing consumer intentions to adopt battery electric vehicles," *Renew. Sustain. Energy Rev.*, vol. 78, pp. 318–328, Oct. 2017.

[11] G. Cecere, N. Corrocher, and M. Guerzoni, "Price or performance? A probabilistic choice analysis of the intention to buy electric vehicles in European countries," *Energy Policy*, vol. 118, pp. 19–32, Jul. 2018.

[12] K. Y. Bjerkan, T. E. Nørbech, and M. E. Nordtømme, "Incentives for promoting battery electric vehicle (BEV) adoption in norway," *Transp. Res. D, Transp. Environ.*, vol. 43, pp. 169–180, Mar. 2016.

[13] L. Han, S. Wang, D. Zhao, and J. Li, "The intention to adopt electric vehicles: Driven by functional and non-functional values," *Transp. Res. A, Policy Pract.*, vol. 103, pp. 185–197, Sep. 2017.

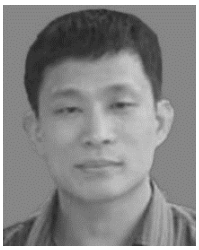
[14] B. M. Al-Alawi and T. H. Bradley, "Review of hybrid, plug-in hybrid, and electric vehicle market modeling studies," *Renew. Sustain. Energy Rev.*, vol. 21, pp. 190–203, May 2013.

[15] Z. Rezvani, J. Jansson, and J. Bodin, "Advances in consumer electric vehicle adoption research: A review and research agenda," *Transp. Res. D, Transp. Environ.*, vol. 34, pp. 122–136, Jan. 2015.

[16] M. Coffman, P. Bernstein, and S. Wee, "Electric vehicles revisited: A review of factors that affect adoption," *Transp. Rev.*, vol. 37, no. 1, pp. 79–93, Jan. 2017.
[17] G. H. Broadbent, D. Drozdzewski, and G. Metternicht, "Electric vehicle adoption: An analysis of best practice and pitfalls for policy making from experiences of Europe and the US," *Geography Compass*, vol. 12, no. 2, 2018, Art. no. e12358.
[18] B. Yildiz, "Assessment of policy alternatives for mitigation of barriers to EV adoption," Ph.D. dissertation, Dept. Eng. Technol. Manage., Portland State Univ., Portland, OR, USA, 2018.
[19] I. Ajzen, "The theory of planned behavior. Organizational Behav," *Human Decis. Processes*, vol. 50, pp. 179–211, Jan. 1991.
[20] I. Moons and P. De Pelsmacker, "Emotions as determinants of electric car usage intention," *J. Marketing Manage.*, vol. 28, nos. 3–4, pp. 195–237, Mar. 2012.
[21] O. Egbue and S. Long, "Barriers to widespread adoption of electric vehicles: An analysis of consumer attitudes and perceptions," *Energy Policy*, vol. 48, pp. 717–729, Sep. 2012.
[22] A. Ziegler, "Individual characteristics and stated preferences for alternative energy sources and propulsion technologies in vehicles: A discrete choice analysis for germany," *Transp. Res. A, Policy Pract.*, vol. 46, no. 8, pp. 1372–1385, Oct. 2012.
[23] R. A. Daziano and D. Bolduc, "Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model," *Transportmetrica A: Transp. Sci.*, vol. 9, no. 1, pp. 74–106, Jan. 2013.
[24] S. Wang, J. Fan, D. Zhao, S. Yang, and Y. Fu, "Predicting consumers' intention to adopt hybrid electric vehicles: Using an extended version of the theory of planned behavior model," *Transportation*, vol. 43, no. 1, pp. 123–143, Jan. 2016.
[25] M. R. Everett, *Diffusion of Innovations*. New York, NY, USA: Springer, Dec. 1995.
[26] M. Lavasani, X. Jin, and Y. Du, "Market penetration model for autonomous vehicles on the basis of earlier technology adoption experience," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2597, no. 1, pp. 67–74, Jan. 2016.
[27] J. Massiani and A. Gohs, "The choice of bass model coefficients to forecast diffusion for innovative products: An empirical investigation for new automotive technologies," *Res. Transp. Econ.*, vol. 50, pp. 17–28, Aug. 2015.
[28] A. Cordill, "Development of a diffusion model to study the greater PEV market," Ph.D. dissertation, Dept. Civil Eng., Univ. Akron, Akron, OH, USA, 2012.
[29] A. F. Jensen, E. Cherchi, S. L. Mabit, and J. D. D. Ortúzar, "Predicting the potential market for electric vehicles," *Transp. Sci.*, vol. 51, no. 2, pp. 427–440, May 2017.
[30] M. Brown, "Catching the PHEVer: Simulating electric vehicle diffusion with an agent-based mixed logit model of vehicle choice," *J. Artif. Societies Social Simul.*, vol. 16, no. 2, p. 5, 2013.
[31] D. McCoy and S. Lyons, "Consumer preferences and the influence of networks in electric vehicle diffusion: An agent-based microsimulation in ireland," *Energy Res. Social Sci.*, vol. 3, pp. 89–101, Sep. 2014.
[32] M. Noori and O. Tatari, "Development of an agent-based model for regional market penetration projections of electric vehicles in the united states," *Energy*, vol. 96, pp. 215–230, Feb. 2016.
[33] J. Dumortier, S. Siddiki, and S. Carley, "Effects of providing total cost of ownership information on consumers' intent to purchase a hybrid or plug-in electric vehicle," *Transp. Res. A, Policy Pract.*, vol. 72, pp. 71–86, Feb. 2015.
[34] E. H. Noppers, K. Keizer, J. W. Bolderdijk, and L. Steg, "The adoption of sustainable innovations: Driven by symbolic and environmental motives," *Global Environ. Change*, vol. 25, pp. 52–62, Mar. 2014.
[35] N. Prakash, R. Kapoor, and A. Kapoor, "Gender Preferences for alternative energy transport with focus on electric vehicle," *J. Social Sci.*, vol. 10, no. 3, pp. 22–114, 2014.
[36] S. Carley, R. M. Krause, B. W. Lane, and J. D. Graham, "Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cites," *Transp. Res. D, Transp. Environ.*, vol. 18, pp. 39–45, Jan. 2013.
[37] I. Lai, Y. Liu, and X. Sun, "Factors influencing the behavioural intention towards full electric vehicles: An empirical study in Macau," *Sustainability*, vol. 7, no. 9, pp. 12564–12585, 2015.
[38] Y.-N. Sang and H. A. Bekhet, "Modelling electric vehicle usage intentions: An empirical study in Malaysia," *J. Cleaner Prod.*, vol. 92, pp. 75–83, Apr. 2015.
[39] R. J. Javid and A. Nejat, "A comprehensive model of regional electric vehicle adoption and penetration," *Transp. Policy*, vol. 54, pp. 30–42, Feb. 2017.
[40] K. S. Gallagher and E. Muehlegger, "Giving green to get green? Incentives and consumer adoption of hybrid vehicle technology," *J. Environ. Econ. Manage.*, vol. 61, no. 1, pp. 1–15, Jan. 2011.
[41] G. Tal and M. A. Nicholas, "Studying the PEV market in California: Comparing the PEV, PHEV and hybrid markets," in *Proc. World Electr. Vehicle Symp. Exhib. (EVS)*, Nov. 2013, pp. 1–10.
[42] W. Sierzchula, S. Bakker, K. Maat, and B. van Wee, "The influence of financial incentives and other socio-economic factors on electric vehicle adoption," *Energy Policy*, vol. 68, pp. 183–194, May 2014.
[43] S. Musti and K. M. Kockelman, "Evolution of the household vehicle fleet: Anticipating fleet composition, PHEV adoption and GHG emissions in Austin, Texas," *Transp. Res. A, Policy Pract.*, vol. 45, no. 8, pp. 707–720, Oct. 2011.
[44] M. A. Tamor, C. Gearhart, and C. Soto, "A statistical approach to estimating acceptance of electric vehicles and electrification of personal transportation," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 125–134, Jan. 2013.
[45] K. E. Train, *Discrete Choice Methods With Simulation*, vol. 30. Cambridge, U.K.: Cambridge Univ. Press, Jun. 2009, pp. 688–692.
[46] B. T. Pham, B. Pradhan, D. Tien Bui, I. Prakash, and M. B. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environ. Model. Softw.*, vol. 84, pp. 240–250, Oct. 2016.
[47] V. Vapnik, "The nature of statistical learning theory," in *Springer Science & Business Media*, vol. 29. Springer, Jun. 2013, Art. no. 409409.
[48] H. A. Nefeslioglu, E. Sezer, C. Gokceoglu, A. S. Bozkir, and T. Y. Duman, "Assessment of landslide susceptibility by decision trees in the metropolitan area of istanbul, turkey," *Math. Problems Eng.*, vol. 2010, pp. 1–15, 2010.
[49] D. Tien Bui, B. Pradhan, O. Lofman, and I. Revhaug, "Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and Naïve bayes models," *Math. Problems Eng.*, vol. 2012, pp. 1–26, 2012.
[50] J. Dou, A. P. Yunus, D. Tien Bui, A. Merghadi, M. Sahana, Z. Zhu, C.-W. Chen, K. Khosravi, Y. Yang, and B. T. Pham, "Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the izu-oshima volcanic island, japan," *Sci. Total Environ.*, vol. 662, pp. 332–346, Apr. 2019.
[51] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Comput. Inform.*, vol. 34, no. 5, pp. 1017–1037, 2016.
[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[53] J. Jia, "Analysis of alternative fuel vehicle (AFV) adoption utilizing different machine learning methods: A case study of 2017 NHTS," *IEEE Access*, vol. 7, pp. 112726–112735, 2019.
[54] EVAdoption. *EV Market Share by State*. Accessed: Apr. 29, 2020. [Online]. Available: https://evado ption.com/ev-market-share/ev-market-share-state/
[55] T. Q. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
[56] G. L. Ke, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 1–9.

JIANMIN JIA was born in Heze, Shandong, China, in 1988. He received the B.S. degree in information and computing science and the M.S. degree in transportation engineering from Shandong University, in 2010 and 2013, respectively, and the Ph.D. degree in transportation engineering from Florida International University, Miami, FL, USA, in 2017.

From 2013 to 2017, he was a Research Assistant with the Lehman Center for Transportation Research (LCTR), Florida International University. Since 2018, he has been a Lecturer with Shandong Jianzhu University, China. He is the author of more than ten articles. His research interests include traffic simulation and modeling, big data analysis, and travel demand forecasting.



BAIYING SHI received the Ph.D. degree in transportation engineering from Tongji University. He is currently an Associate Professor with the School of Transportation Engineering, Shandong Jianzhu University. His research interests include traffic management, traffic operation, and so on.

FA CHE is currently pursuing the Ph.D. degree in engineering. He is working with the Zibo Transportation Service Center. He has studied urban road operation analysis and road safety operation technology research for a long time.

HUI ZHANG received the B.S. degree in applied mathematics from Qingdao Agricultural University, in 2009, and the M.S. degree in system science and the Ph.D. degree in transportation planning and management from Beijing Jiaotong University, in 2012 and 2016, respectively.

Since 2016, he has been a Lecturer with the School of Transportation Engineering, Shandong Jianzhu University. His research interests include transportation engineering and transportation planning.