Classification of potential electric vehicle purchasers: A machine learning approach

Technological Forecasting & Social Change 168 (2021) 120759

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Classification of potential electric vehicle purchasers: A machine learning approach

Javier Bas[a,*], Cinzia Cirillo[b], Elisabetta Cherchi[c]

[a] Department of Economics, Universidad de Alcalá, 0.7 Facultad de Ciencias Económicas, Empresariales y Turismo, Alcalá de Henares, Madrid 28802 , Spain
[b] Department of Civil and Environmental Engineering, University of Maryland, 3250 Kim Bldg., College Park, MD 20742, United States
[c] School of Engineering, Newcastle University, Cassie Bldg., Newcastle upon Tyne, United Kingdom

ARTICLE INFO

Keywords:
Electric vehicle
Adoption
Machine learning
Prediction
Cluster analysis
Imputation

ABSTRACT

Among the many approaches towards fuel economy, the adoption of electric vehicles (EV) may have the greatest impact. However, existing studies on EV adoption predict very different market evolutions, which causes a lack of solid ground for strategic decision making. New methodological tools, based on Artificial Intelligence, might offer a different perspective. This paper proposes supervised Machine Learning (ML) techniques to identify key elements in EV adoption, comparing different ML methods for the classification of potential EV purchasers. Namely, Support Vector Machines, Artificial Neural Networks, Deep Neural Networks, Gradient Boosting Models, Distributed Random Forests, and Extremely Randomized Forests are modeled utilizing data gathered on users' inclinations towards EV. Although a Support Vector Machine with polynomial kernel slightly outperforms the other algorithms, all of them exhibit comparable predictability, implying robust findings. Further analysis provides evidence that having only partial information (e.g. only socioeconomic variables) has a significant negative impact on model performance, and that the synergy across several types of variables leads to higher accuracy. Finally, the examination of misclassified observations reveals two well-differentiated groups, unveiling the importance that the profiling of potential purchaser may have for marketing campaigns as well as for public agencies that seek to promote EV adoption.

1. Introduction

In the last few decades, the vehicle-miles traveled, as well as passenger-miles traveled, have increased in the United States (Bureau of Transportation Statistics, 2020). Such a rise leads to traffic congestion and, consequently, greater fuel consumption and pollution, in a country that is already the first oil consumer in the world (bp, 2020). Among the many approaches towards fuel economy, the adoption of alternative fuel vehicles, especially electric vehicles (EV), may have the greatest impact. The number of studies that have explored EV adoption is large, either taking the agent's perspective, or trying to predict penetration through more macroeconomic approaches. Although these studies point often to the same direction, they offer very different EV market evolution in terms of time and magnitude.

In this context, it might be worth exploring and testing new methodological perspectives. Machine Learning (ML) techniques are currently applied to an enormous variety of topics such as fraud detection (Bolton and Hand, 2002), robotics (Stone and Veloso, 2000), spam filtering (Guzella and Caminhas, 2009), translation services (Sagiroglu et al., 2007), preventive health care (Deo Rahul, 2015), computer vision (Oliver et al., 2000), as well as transportation, the field for which a literature review is developed in the next section. This has been possible thanks to the exponential growth of information brought about by electronic devices; an amount that will continue to expand due to the Internet of Things (Docherty et al., 2018). In the case of transportation, the smart use of the data generated by on-road vehicles presents an extraordinary opportunity to improve transportation systems. However, this task overcomes the capabilities of traditional data analysis and clearly points to ML as a solution. Congestion reduction, safety improvement, environmental impact mitigation, and energy consumption optimization are examples of the most common lines of research in which ML techniques have been applied.

However, there are other less explored fields of application, such as the classification of potential consumers into adopters/non-adopters. This is a topic that presents interesting challenges. Adoption is demand-driven, and demand roots into purchasers' behavior, beliefs and

* Corresponding author.
E-mail addresses: javier.bas@uah.es (J. Bas), ccirillo@umd.edu (C. Cirillo), elisabetta.cherchi@ncl.ac.uk (E. Cherchi).

https://doi.org/10.1016/j.techfore.2021.120759
Received 30 June 2020; Received in revised form 16 February 2021; Accepted 15 March 2021
Available online 15 April 2021

attitudes; elements that are intrinsically difficult to define and gather. Even if reliable information on these aspects is available, it is usually not in large quantities and, even less frequently, in conjunction with other variables of interest such as vehicle ownership, sociodemographic, vehicle attributes, or social characteristics. This is the context in which our work aims to shed light. The contribution of this paper is to use information on all these elements, collected through a survey specifically designed for this purpose, to compare the throughput of supervised ML algorithms when applied to the classification of individuals into EV adopters. Namely, we apply *Support Vector Machines* (SVM), *Artificial Neural Networks* (ANN), *Deep Neural Networks* (DNN), *Gradient Boosting Models* (XGBoost), *Distributed Random Forest* (DRF), and *Extremely Randomized Forest* (XRF). This exercise is relevant for several reasons. First, a correct identification of the key variables of potential purchasers profiling not only allows to improve predictions, but also to determine the drivers of the EV adoption process. This work helps in doing so by revealing the role played by aspects commonly left aside (social and attitudinal) in both parametric and non-parametric studies, and by showing that it is their synergy of information of different nature (about the individual and the vehicle itself) that produces a better classification. Secondly, exploring the techniques that work best in a case of this nature can pave the way for stakeholders interested in staying one step ahead of the complex decision process that leads to the adoption of an EV.

These are contributions from which industry and public agencies can benefit alike. Moreover, to the best of our knowledge, the particularities of this study make it novel. As we will develop in the following sections, we make use of heterogeneous microdata that combine the so-called *Revealed* and *Stated* preferences, collected via survey specifically designed to gather individuals' willingness to purchase an EV. We feed seven ML algorithms of varying complexity with this comprehensive dataset to predict the adoption of the EV, also conducting a study of what may occur to those individuals who are not correctly classified by the best of these techniques.

The rest of the paper is organized as follows. After a literature review on ML applications, Section 3 briefly presents the supervised ML techniques applied in this study. Section 4 introduces the data and the methodology followed, while Section 5 exhibits the results. Finally, Section 6 summarizes the main conclusions.

2. Literature review

Alternative fuel vehicles have been the subject of several ML applications, especially in topics such as battery estimation, energy consumption, or range estimation. ANN (Zahid et al., 2018) and SVM (Sheng and Xiao, 2015) have been used to estimate the state of health or the state of charge of batteries; as well as other less known approaches such as *fuzzy c-means clustering with backpropagation* (Hu et al., 2016). More recently, Fukushima et al. (2018) proposed the use of energy consumption predictive models to forecast the energy consumption of new EV in the absence of training data. To estimate vehicle's range, Yavasoglu et al. (2019) utilized an ANN with one hidden layer of 60 neurons in conjunction with a *Decision Tree* (DT) to estimate the road type when it is not known. Stop delivery times prediction (Hughes et al., 2019), traffic flow estimation (Liu et al., 2019), driving behavior recognition (Yi et al., 2019), or parking occupancy prediction (Yang et al., 2019) are other specific transportation issues to which ML techniques have been applied. However, examples on the adoption of ML techniques in transportation research using stated preference (SP) data are not abundant. Lee et al. (2019) applied Gradient Boosting Machines method to understand the user preference related to autonomous vehicle. They included attitudinal variables, such as pro-AV sentiments, the environmental concern, the interest in AV technology, and attitudes towards public transit in the study and evaluated their relative importance to AV preferences. Hernandez et al. (2016) applied a decision tree framework to obtain explainable results about the impact of transportation user perception and attitudes on their preferences. Zhao et al. (2020) used SP survey data to compare the results of ML models with those obtained from a logit one.

On the other hand, there exist several works that have carried out comparisons across different algorithms. Jahangiri and Rakha (2015) used data from cellphones' accelerometers and gyroscopes to predict transportation mode, comparing the prediction accuracy of SVM, DT methods and *k-nearest neighbors* (KNN). Results showed that RF and SVM had the best performance, although they have difficulties in differentiating between car mode and bus mode. Huang et al. (2011) discriminated driving conditions using speed and acceleration data, comparing the predicition throughputs of SVM, ANN, linear and quadratic classifiers, and *K-means clustering*. A similar work is that of Wang et al. (2018), who applied similar tecniques to driving style classification. One especially comprehensive work is that of Sun et al. (2019) who compared the results of *Multinomial Logistic Regression* (MLR), *Classification and Regression Trees* (*CART),* and *Gradient Boosting Decision Trees (GBDT)* for the prediction of electrical vehicle range. Results showed that GBDT could optimize predictions and reduce errors better than the other two techniques. Another comprehensive comparative study is the one carried out by Goebel and Plötz (2019) who estimated the utility factor (i.e. ratio of miles travelled with electric energy over the total number of miles travelled) for hybrid vehicles. Four different approaches were compared: *Regression Tree* (RT), RF, SVM and ANN, concluding that SVM and ANN gave the best estimation accuracy. More in line with the spirit of the present work are the studies of de Zarazua de Rubens (2019) and Jia (2019). The first uses *K-means clustering* to create six consumer segments around EV adoption. The second compares five machine learning techniques in the context of alternative fuel vehicles. Lee et al. (2014) also presents an interesting exercise that combines a Bass model with ML algorithms to explain the diffusion process of pre-launched products.

Finally, there exist two general reviews of classification techniques. Kotsiantis (2007) defined a score on relevant aspects for several methods. RF excels at speed of classification, handling all kind of attributes (discrete/continuous) and explanation ability, although accuracy is not one of its strengths. On the contrary, SVM are very accurate and fast, with high tolerance to irrelevant attributes, although its results are difficult to explain and its speed of learning increases significantly as the number of attributes grows. Finally, the performance of the ANN seems to be somewhere in between, with a dangerous tendency to overfitting. More recently, Singh et al. (2016) carried out a similar exercise in terms of pros and cons, which results coincide with those of Kotsiantis.

Although a comparison of methods has already been carried out in publications of other fields, this has not been the case in the field of transportation, specially using SP data. The aforementioned works of Lee et al. (2019) and Hernandez et al. (2016) do make use of SP, but do not have a comparative aim. The studies of Zhao et al. (2020) and Jia (2019) are similar in nature to our work; however the former is centered on mode choice and not on the adoption of a new technology, while the latter presents notable differences with ours. Specifically, Jia (2019) does not focus on EV, and it only considers the newest vehicle in the household. Moreover, it does not take into account the social component involved in the adoption of a new technology, the observations with missing information are removed from the dataset, and the algorithms applied are not among the most advanced. Our work, on the contrary, is specifically designed to a) gather individuals' willingness to purchase an EV, b) perform an advanced process for imputing unknown information, c) include social and attitudinal elements involved in the decision-making process, and d) compare several state-of-the-art algorithms used to predict adoption. Therefore, we consider that this work contributes significantly to the literature by proposing classification techniques for the adoption of new technology vehicles using unique data and the most recent ML methods.

3. Supervised machine learning techniques for classification

In this study we train ML models that can accurately classify whether a person is a potential buyer of an EV based on a variety of factors such as socioeconomic characteristics, information about social relationships, car ownership, trip information, and attitudes towards technology and the environment (see Section 4 below). This process is considered supervised learning, where a ML algorithm processes the data on a training subset to generate label predictions that are validated in a different testing subset. Both training and validation subsets are drawn from the same original data, thus we apply the *K-Fold Cross Validation* method to each ML model to avoid any possible unintentional selection bias when splitting the data. In general, these procedures are highly computationally intensive, especially as the number of data points and dimensionality grows. This, together with the elimination of irrelevant variables, makes the practice of carrying out a feature selection process common, which we conduct as described in Section 4.3.

On the other hand, a particular challenge when facing a ML project is the enormous diversity of algorithms that can be applied to the same problem. Although there may be some guidelines on which one should be applied to each case, the truth is that different approaches may lead to significant deviations of the level of performance. Additionally, depending on the complexity of the case at hand, relatively simple methods may perform better than more advanced ones. Therefore, for this work we decided to compare the performance of three families of techniques that comprises algorithms of different complexity: Support Vector Machines, tree-based methods, and neural networks. Concretely, we estimate Support Vector Machines (SVM; Schölkopf and Smola, 2018) with both radial and polynomial kernel for the first family; Extreme Gradient Boosting Machine (XGBM; Chen and Guestrin, 2016), Distributed Random Forests (DRF; Breiman, 2001), and Extremely Randomized Forests (XRT; Geurts et al., 2006), for the second; and Artificial Neural Networks (ANN; Haykin, 1994) and Deep Neural Networks (DNN; Liu et al., 2017), for the third family of models. Since describing these methods is not the ultimate goal of this article and it could eclipse its true objective, we refer the reader to the references indicated for a deeper understanding of them.

4. Data collection and methodology

The data used in this work was specifically collected to study the inclination of individuals towards the EV and the role played by their social structure in the choice of this type of vehicle (Bas et al., 2020). They were gathered in two phases via online surveys in the United States. A first pilot was developed to explore questionnaire consistency as well as to check if the information obtained from it obeyed to the object of the study. After minor improvements to the questions and its structure, a second version of the survey was released, which lasted about two months between December 2019 and February 2020. The main goal was to examine the adoption of the EV, thus, we considered that the target population should be holders of driver's license since they would have driving experience and be familiar with vehicle related aspects such as refueling and its costs. Concretely, the criterion for participation in the survey was to be over the age of 18, hold a driver's license, and reside in the State of Maryland, U.S. Additionally, whether or not people participated in the pilot was not a criterion for being excluded of the final survey, which was taken by 380 users (6 were removed due to inconsistencies in their responses). Each of them faced 6 or 9 different choice tasks (see Section 4.1 below), yielding a total of 3174 pseudo-observations. Completion time was between 10 and 20 min, depending on the number of vehicles owned and the number of choice tasks. The questionnaire was designed and operated with the survey platform Qualtrics (Qualtrics Research Core), and consisted of 5 sections:

- *Social Network*: The interviewee was asked to enter the number of members of different groups (close relatives, relatives, friends and acquaintances), as well as the number of individuals composing several subgroups from which it is possible to derive trust or affinity. Namely:

  How many of them would you leave a spare key to your house to?
  How many of them would you discuss important personal matters with?
  How many of them do you share hobbies with?
  How many of them have EV experience?
  How many of them would you talk to about EV technology?
  How many of them do you think that five years from now you will still have relationship with?

  The social component that this section elicits is relevant when it comes to adopting a new technology because people around us, such as family members, friends, colleagues, or even people that we do not know, influence our behavior and decisions, directly or indirectly (Cherchi, 2017). We all have some tendency to either yield to group pressures or to agree to the majority, which can happen because of the desire of being accepted, or because of the desire to do the right thing (Crutchfield, 1955). Either way, individuals tend to turn to members of their own group in order to gather information, which may involve a change in attitudes, beliefs or behavior. Screenshots of the questions on the social network of the respondent can be found in Figs. A1 and A2 in the Appendix. Finally, the name of a person of each group was also required in this question, for purposes related to the stated choice experiment described in Section 4.1.
- *Vehicle ownership*: The second section aimed to identify the vehicles owned in the household, and if the next purchase would be an additional one or on the contrary would replace one of them.
- *Stated Choice Experiment (SCE):* The third section consisted of a SCE pivoted around some of the values collected previously. The choice tasks included vehicle attributes as well as variables that allow to identify the effect of the feedback provided by members of the social network. The next subsection provides detail on the SCE.
- *Trip information:* The fourth block of questions collected information about the trips made by the respondent, in order to know about the possible use of the EV. It also included three questions to identify the patterns of use of carsharing and rideshare apps.
- *Attitudinal factors:* The last section was dedicated to gathering information about the attitudes of the user towards the environment, technology, and EV. The questions consisted of statements (3 about the environment; 4 about technology, in general; 5 about EVs, specifically) on which the interviewee had to show agreement on a Likert scale ranging from *Strongly disagree* to Strongly *agree*. We paid special attention in formulating the statements to make them sufficiently generic so that anyone, experienced in EVs or not, could answer them, while allowing us to disentangle their position with respect to these environmental and technological factors. This section also contains the socioeconomic questions. A screenshot of the question on attitudes, as well as a summary of the socioeconomic variables can be found in Table A1 the Appendix.

Two aspects related to this methodology should be highlighted. On the one hand, the approach presented can be considered common among the abundant studies that apply SCE, in terms of design and organization. Vehicle ownership and trip information are usual pieces of information on which to build parametric models for mode or route choice. So are social and attitudinal elements, although less frequently, and normally not in conjunction with those just mentioned. However, as discussed in Section 2, all these aspects may also contribute to the adoption of the EV. For example, ownership of an electric vehicle may indicate an inclination to purchase another one (maybe a generation upgrade). On the contrary, long commuting trips may be a clear disincentive to adopting this technology due to range issues. It is obvious as well that a greater concern for the environment and an interest in new

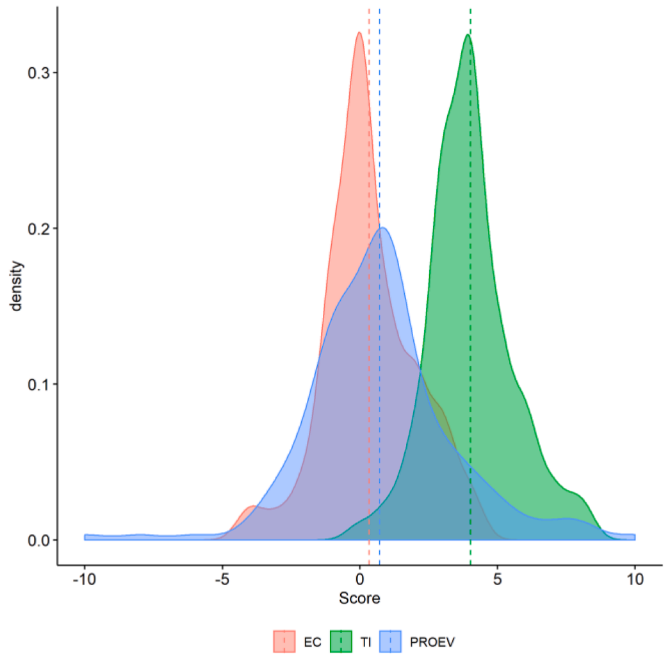


Fig. 1. Distribution of attitudinal scores.

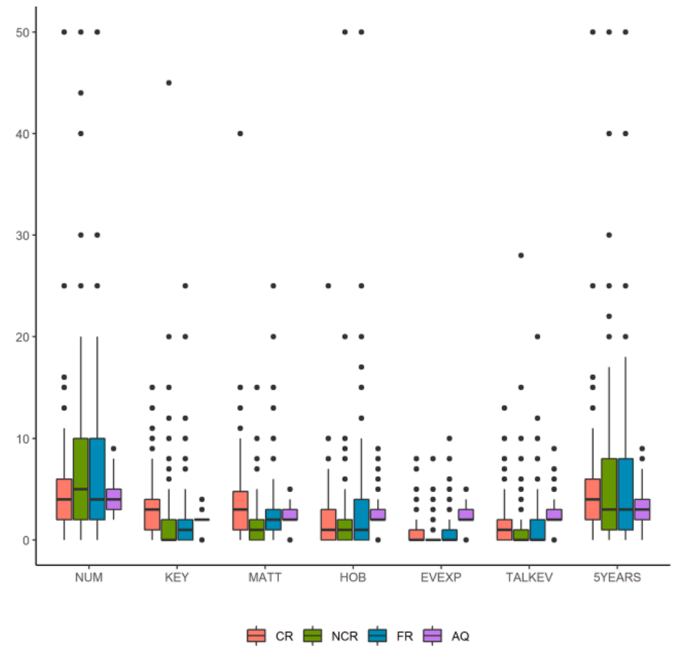


Fig. 2. Distribution of the number of people in each social subgroup.

technologies and, in particular, in the EV technology, are favorable elements for an individual to become an adopter. Therefore, we consider that each of these elements can play a key role in the adoption of the EV, and that this is the first work, to the best of our knowledge, that integrates them in the estimation of ML algorithms, as discussed in Section 2 above.

4.1. Stated choice experiment

The experiment designed for this study consisted of choices between electric and gasoline vehicle, including the option of choosing none of them. The last two were grouped together in one only class in order to reflect EV adopting/non-adopting behavior. The experimental design included 5 attributes (price, propulsion cost, range, charging/refueling time, and tax deduction amount), as well as a variable controlling for the effect of social influence (number of EV sold last month). The choice of the attributes was grounded on the comprehensive literature on the topic and previous experience on other surveys. Their levels were based on vehicles of reference, although adapted to cover all the choice spectrum and to avoid the dominance of a feature that could lead the user to always choose one of the alternatives. We followed an efficient design with Bayesian priors, uniformly distributed, with preliminary values obtained from (Jensen et al., 2016) and Cherchi (2017). We defined 24 choice situations, divided into 4 blocks, which allowed attribute level balance, ensuring estimation on the whole range of levels. This design was optimized for 3 categories of vehicles (Small, Mid-size, and Large) with specific values for each of them. Before the first section, the respondents were asked what the size of a new vehicle would be if they were to buy one, and then redirected to the survey branch corresponding to that case. This way they faced scenarios closer to their purchasing stated preferences, which contributed to more realistic choices.

An unique feature of this survey was that, once the six scenarios had been evaluated, the respondent was presented to three more scenarios randomly chosen from the ones that he had already seen and evaluated. However, a new piece of information was provided along with the level of attributes; a sentence that expressed positive or negative feedback on EVs, attributed to a person belonging the respondent's Social Network. It is important to note that the respondents were not notified about the fact that the scenarios with feedback were actually repeated. The feedback was given in the following form:

"Bruce thinks that having to change your activities because of driving an EV is annoying."

An analysis of the choices made by the individuals in this sample reveals that 19.19% of the responses given in the choice tasks correspond to *Adoption,* while 80.81% to *Non-adoption.* More interestingly, among the first group, 38.57% of the individuals received positive feedback; meaning that the effect of the information received from someone of one's social network is limited, a result in accordance to Bas et al. (2020)

On the other hand, since attitudes play a fundamental role in this study, we have carried out a more detailed study of the attitudinal profile of these individuals. For this purpose, we assigned a value from −2 (*Strongly disagree*) to 2 (*Strongly agree*) to each of the responses to the statements on the attitudinal question. We then added them up for each category (*Environmental Concern, Technology Inclined*, and *Pro-EV)* in order to compute a representative score. Their distributions are plotted in Fig. 1, where the dashed lines indicate the average value.

The *Technology Inclined* attitude scores the highest on average, above 4 (out of a maximum of 10 and a minimum of −10), meaning that respondents might be early adopters or, at least, that they have interest in new technologies. *Pro-EV* and *Environmental Concern* have lower average scores (considering that their scales range from −6 to 6 and −8 to 8, respectively). In addition, the three distributions are reasonably symmetric, with the *Pro-EV* one being flatter and having long tails, representing more dispersion of the sample in this matter. Therefore, it is possible to conclude that these are individuals inclined to technology, with no special interest in the environment, and equally in favor and against EVs.

As for the composition of the individuals' social network, the other novel element of this work, the average number of close relatives (CR) declared was 2, while the average number of non-close relatives (NCR), friends (FR), and acquaintances (AQ), was 3, 12, and 4, respectively. The reduced number of acquaintances reported is surprising, yet a consistent fact among the pilots and the final survey. Regarding the nature of these groups, respondents were also asked to reveal the number of individuals composing various subgroups from which it was possible to derive trust or affinity (see Fig. A2 in the appendix). Their composition can be seen in Fig. 2. For the *key* and the *matters discussion* questions, the average is

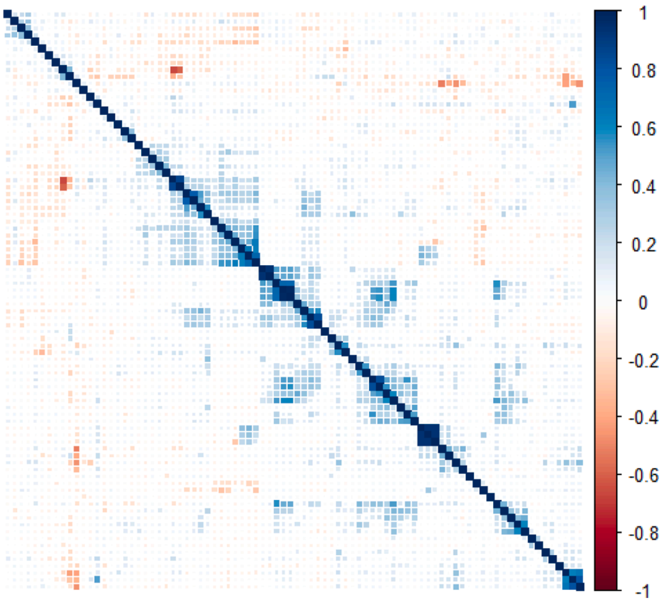


Fig. 3. Map of features correlations.

higher for CR than for the other groups; naturally, one would leave a spear key or talk about important issues to members of their family but not much to other people. On the other hand, the average number of persons with whom respondents can talk about EVs, or who actually have experience in driving EVs, is very reduced, as expected. Values are high for the *5 years* question, evidencing certain optimism of individuals regarding the future of their social relations. Also, although not shown in this figure, it is worth to mention that FR is the group with which the individuals in this sample have the most frequent contact (*Every day)*, followed by CR (*Once a week)*, NCR (*1–3 times a month)* and, lastly, AQ (*1–3 times a month*, too*)*.

4.2. Imputation of NA values

The social question included in the first section of this survey is a distinguishing feature of this data collection. It will help to identify whether the social network structure of the individuals is significant in adopting EV. However, the design of this inquiry involved a particularly inconvenient casuistry, i.e., the interviewee may not know the number of individuals in a group. That is, one may genuinely not know how many acquaintances she has or how many friends she can talk to about EV technology, for instance. Therefore, it was necessary to offer an *I don't know* option, which meant a missing value when selected. Since the ML techniques to be applied cannot handle missing values, it was necessary to impute them. For this task, we relied on the Multiple Imputation by Chained Equations (MICE) method, which, roughly, regresses a variable with missing values on other selected features, and replace *NAs* by simulated draws from its predictive distribution (for more information on MICE, see Buuren and Groothuis-Oudshoorn (2011)). In our case, we first imputed the number of members of the main social groups, where missing, using all the other information in the data. Then, we imputed in a second round the social subgroups using the same information plus the recently imputed one.

4.3. Feature selection

Working with a large set of predictors may actually be a drawback in the analysis as they are more likely to be correlated as their number grows. Fig. 3 shows a visualization of the correlation matrix among the features of our dataset. The intersection of each row and columns is colored according to the value of the correlation coefficient between these variables, following the legend coding. Some 'correlation clusters' can be identified, but they mostly respond to the social network variables, which columns are located next to each other for each subgroup. For instance, all the columns storing the information regarding the *Friends* group, are placed together and, obviously, the number of friends one shares hobbies with, which talks about personal matters, etc., is correlated to the total number of friends that one has.

Although non-parmetric classifiers, like the ones we use in this study, are not very sensitive to correlation, the inclusion of unnecessary variables may lead to the curse of dimensionality. However, non-parametric classifiers, like the ones we use in this study, are not very sensitive to correlation. However, the inclusion of unnecessary variables will lead to the *curse of dimensionality*; the larger the feature space, the sparser the data becomes. In other words, the amount of observations for each combination of feature values becomes insufficient for reliable estimations. On the other hand, a large number of features also increases the complexity of the models, which become prone to overfitting; they will fit the training data so well that they will not be able to correctly predict the classes of new observations. Fortunately, these issues may be overcome through dimension reduction techniques that reduce the number of variables, yet preserving, to a reasonable extent, the information that they keep. The approach followed in this study is the application of a preliminary Random Forest in order to identify the importance of each variable in the data set. Then, the top variables in terms of importance will be used in the models. We chose this technique over others, such as

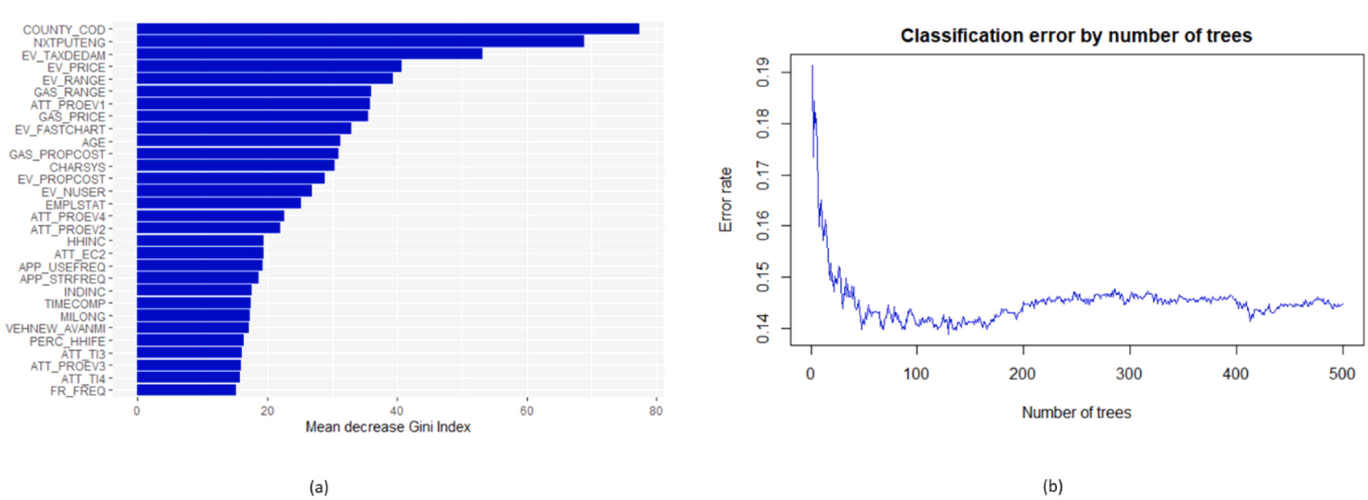


Fig. 4. Variable importance (a) and Classification error by number of trees (b).

Table 1
Model performance.

	SVM Radial		SVM Polynomial		ANN		XGBoost		DRF		XRT		DNN	
	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA
A	280	63	302	80	292	94	486	436	466	456	382	540	406	516
NA	95	515	73	498	83	484	37	1438	54	1421	26	1449	43	1432
Accuracy	0.8342		0.8395		0.8143		0.8027		0.7872		0.7639		0.7668	
Sensitivity	0.7467		0.8053		0.7787		0.9749		0.9634		0.9824		0.9708	
Specificity	0.891		0.8616		0.8374		0.5271		0.5054		0.4143		0.4403	

the widely used Principal Component Analysis, since it keeps variables in its original form, instead of building new constructs that are difficult to interpret. Fig. 4a shows the 30 most important variables (over a total of 84) when choosing the type of vehicle after running a Random Forest composed by 500 trees on the original dataset, number of trees for which the error rate stabilizes (Fig. 4b).

The analysis reveals that the most important features are: the county in which the user resides and the engine of the next purchase (*electric, gasoline, hybrid* or *other*). For the former, some of the counties in the State of Maryland are among the richest in the U.S. Thus, this variable may actually be reflecting a geographical high-income distribution. They are followed by; the amount of the income tax deduction associated to the EV purchase; its price and range; the range of the gasoline vehicle; and *ATT_PROEV1*, which reflects the respondents' level of agreement to the sentence *Electric vehicles should play an important role in our mobility systems*. The rest of the top 10 inputs are the price of the gasoline vehicle, the time of fast charging of the electric one, and the age of the respondent. *EV_NUSER*, which measures the effect of social conformity, is also ranked high, even above household income. Also, some of the variables that provide information on the individuals' attitudes towards environment (*ATT_EC2*), EV (*ATT_PROEV, ATT_PROEV2)*, and technological progress (*ATT_TI3, ATT_TI4)*. It is encouraging to confirm that the number of members of some social groups is also important (*FR_FREQ)*. On the opposite side, not shown in the figure, are; other sociodemographic variables such as gender or marital status; the size of the next vehicle to be purchased; who will drive it or for what purpose; and the structure of the outermost social group (Acquaintances).

5. Results

5.1. Models' performance

The results obtained by applying the ML classification techniques introduced in Section 3 are reported in Table 1. It shows the confusion matrices, as well as the averaged accuracy over all *k-folds*, the *Sensitivity* and the *Specificity*. These last two statistics provide the proportion of true positives (an adopter classified as such) and true negatives (a non-adopter classified as such) correctly identified.

The accuracy of all methods is similar, although that of the SVMs, Neural Network and XGBoost is slightly better. It is worth noting that the most complex methods (XRT and DNN) are those that performs the worst, which is natural considering the nature of these algorithms. Deep learning architectures incorporate several layers that learn by computing non-linear input-output mappings. This makes the algorithms capable of learning from high-level abstractions, which is more proper of audio, video, speech or images than of a case like ours, not particularly complex in mathematical terms. In any case, the accuracy is not lower than 0.766 and therefore we can safely affirm that more than 76% of the choices made by individuals were predicted correctly, no matter the technique used. In this regard, the confusion matrices at the top of the table present the actual choices (row) and the predictions (column). The values in the diagonals correspond to correct predictions, which are homogeneous among the first three methods but not among the other four. This disparity is evident in the very high Sensitivities that these methods offer, which contrast with the low Specificities. In other words, their predictive power is mainly based on correctly identifying the adopters, significantly misidentifying the potential non-adopters.

Therefore, attending to the statistics described, we can conclude that the methods exhibit comparable predictability –specially SVMs, ANN

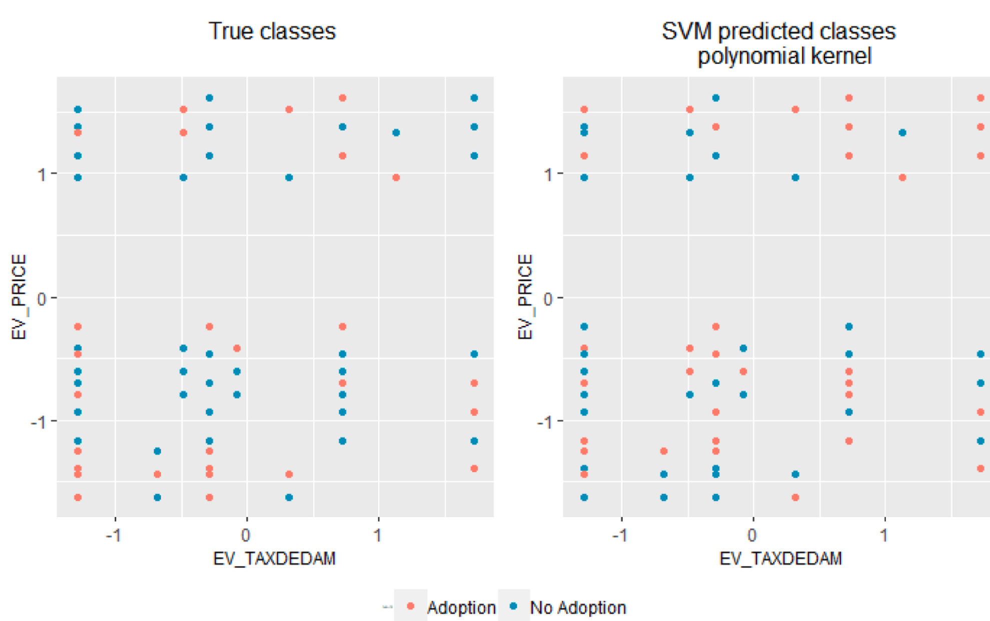


Fig. 5. Actual and predicted classes by the SVM with polynomial kernel.

Table 2
Classification of top variables.

Socioeconomic	Attitudinal and Social	Attributes
County Zip code	Pro-EV 1	Tax deduction when buying EV
Engine of next purchase	Pro-EV 2	EV price
Age	Pro-EV 4	EV range
Charging system at home	Environmental Concern 2	EV fast charging time
Employment status	Number of EV users	EV propulsion cost
Household Income		Gas vehicle price
Frequency of use of ridesharing		Gas vehicle range
Frequency of use of ridesharing with strangers		
Individual income		
Time dedicated to compulsory activities		

and XGBoost, implying robust and reproducible results no matter which of these popular ML techniques is used. Nevertheless, since SVM with polynomial kernel seems to have slightly higher capabilities in predicting the adoption of EV when compared to the other algorithms, Fig. 5 evidences, for illustrative purposes, the similarity of the pattern of the actual classes and the classes predicted by this method, plotted by two of the most relevant variables found in the preliminary analysis.

Now, the top variables depicted in Fig. 4 are of different nature, and they can be grouped into three well-differentiated areas: socioeconomic, attitudinal and social, and vehicle-related attributes. It is worth remembering that the attitudinal variables correspond to several indicators unveiling the inclination of individuals towards the environment, technology, and EV. Table 2 shows this classification.

Considering this very distinct groups, an interesting question is which of them represents the bulk of the predictive power. To answer this, we estimate again the best model found (Support Vector Machine with polynomial kernel), but separately for each group of variables. Results are shown in Table 3.

As expected, the accuracy with respect to the general model decreases in all cases. However, in Sub-models 2 and 3 the fall is dramatic; 17 and 24 percentage points are lost, respectively. Moreover, the decrease in the Sensitivity in sub-models 2 and 3 is especially notorious; it goes from 0.7467 in the general model to just 0.408 and 0.392. That is, if only attitudinal or only attributes-related variables are used, most of the *adopters* will be misclassified as *non-adopters*. In any case, Sub-model 3 is not statistically significantly different from assigning randomly the classes, as the *p-value* above 0.05 evidences.

To connect these results with the actual adoption of the EV, we can take a closer look at the regions with the highest diffusion of this technology. In doing so, we can identify a correspondence between the drivers of change concurring in them and the variables described above. In Norway, the European country with the highest proportion of EVs, these vehicles are exempt from registration fees as well as from a 25% value-added tax. Similar conditions exist in Denmark and Sweden, and it is no coincidence that in these countries environmental awareness is high, as is the average income compared to other parts of Europe. This is evidenced by the work of Haustein et al. (2021). It shows the relevance of certain vehicle attributes, income, and userś attitudes towards the environment and EVs when it comes to adopt this technology in Denmark and Sweden. This findings are likewise supported by Glerum et al. (2013), Jensen et al. (2014), Lee et al. (2019), and others. If we turn our attention to the American market, the States of California and Oregon show the largest adoption rate in the U.S (Lutsey et al., 2015). The actions adopted in these regions at the state and city levels are consistent with those just mentioned. This is, purchasing subsidies, tax benefits, environmental oriented policies, and policies aimed at the diffusion of the EV.

Therefore, the common denominator in regions that present high EV adoption seems to be composed of all the elements that gravitate around price (subsidies, deductions and exemptions), the characteristics of the vehicle itself (specially range), and a set of attitudes towards the environment and technology. These aspects coincide with the most important variables found in our dataset, which leads us to believe that our findings are in line with the observed reality.

5.2. Misclassified observations

The best model (SVM with polynomial kernel) does not correctly classify about 16% of the observations. An interesting question is whether these individuals share characteristics that make the algorithm fail when classifying them. In order to reveal these traits, we first carried out a cluster analysis of the misclassified observations to identify, if they existed, groups of individuals. Then, we performed an exploratory data analysis on all the variables incorporated to the model estimation.

Cluster analysis is a term that covers several procedures for finding subgroups of observations that are similar to each other in a data set. These subgroups may exist or not, therefore, the first step is to assess if the data is clusterable. In order to do so, the Hopkins' statistic (Lawson and Jurs, 2002) is calculated. It measures the probability that a given set of data is generated by a uniform distribution. In other words, it tests the randomness of the information. Specifically, if the observations are uniformly distributed the statistic would be 0.5. However, if clusters are present, the value is higher. A result above 0.75 indicates a clustering tendency at the 90% confidence level. In the case of our misclassified observations, the Hopkins' statistic is 0.799, therefore, this group of individuals is clusterable. Visual assessment is also possible relying in the algorithm of (Bezdek and Hathaway, 2002), which computes the dissimilarities between the observations of the data set and displays them in an image. Fig. 6 illustrates this visualization for our case. White or red points represent low dissimilarity between two observations. Therefore, the whiter or redder the image, the more clusterable the data set is. Attending to both Hopkins' statistic and the visual assessment, we can conclude that our misclassified individuals are subject to clustering.

The second step is to find out how many clusters the data should be divided into, since this is not known in advance. One approach to identify the groups is Hierarchical clustering, which provides a tree-based representation of the observations called *dendrogram* (for a complete description of this algorithm, we refer the reader to (James et al., 2013). Observations that merge at the bottom are very similar, while observations that fuse close to the top are different. The number of branches in which the dendrogram splits at the top of the tree indicate the optimal number of clusters the data may be split into. The

Table 3
SVM with polynomial kernel performance by group of top variables.

	Sub-model 1 (Socioeconomic)		Sub-model 2 (Attitudinal and Social)		Sub-model 3 (Attributes)	
	Adopting	**No Adopting**	**Adopting**	**No Adopting**	**Adopting**	**No Adopting**
Adopting	266	89	153	92	147	145
No adopting	109	489	222	486	228	433
Accuracy	0.79922		0.6705		0.6086	
Sensitivity	0.7093		0.4080		0.3920	
Specificity	0.8460		0.8408		0.7491	
p-value	0.00		0.00		0.4613	

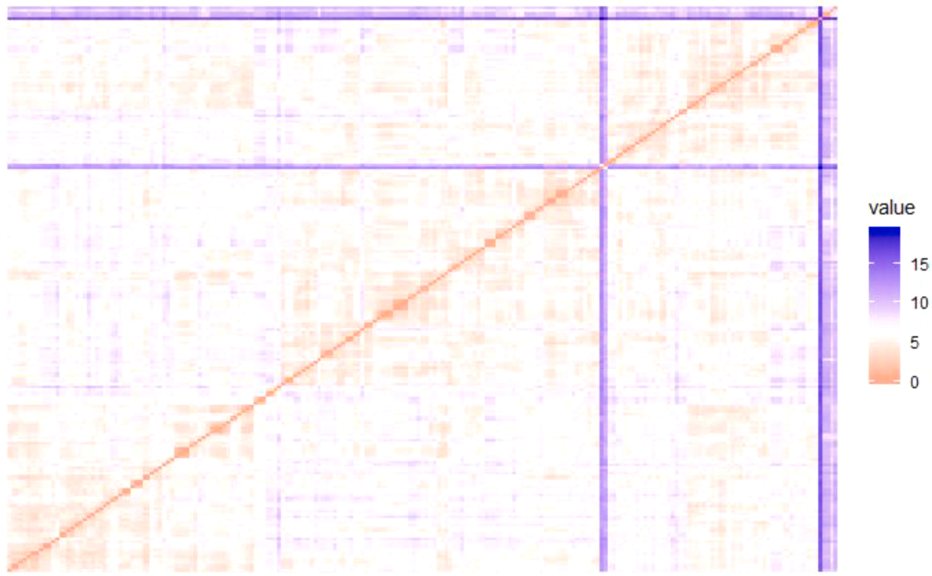


Fig. 6. Clustering tendency of the misclassified observations.

dendrogram in Fig. 7 shows how the misclassified observations are grouped into two clusters.

After identifying that the data is clusterable into two subgroups, the classification is performed. To do so, we opted for the K-means algorithm (MacQueen, 1967), which partitions the data set into *K* distinct, non-overlapping clusters seeking the smallest *within-cluster variation.*

It is possible to visualize the partitioning results for the chosen number of clusters (two, based on the preliminary analysis) drawing a scatter plot of data points colored by cluster. Since the data set contains more than two variables, a Principal Component Analysis has been performed to reduce the dimensionality (for a comprehensive description of this method see (Jolliffe, 1986). The clusters for the two main Princial Components are shown in Figure 8.

Now, in order to find the characteristics common to the members of each cluster, and the differences in-between clusters, an exploratory data analysis has been carried out. This required the examination of the main statistics of each variable as well as of their distribution. The results are summarized in Table 4

A high share of the individuals belonging to Cluster 1 are retired and live in the Prince George's county in Maryland (a low-income one, in comparison to the other counties), while those belonging to Cluster 2, are younger and live, predominantly in the Montgomery county (a high-income one). Cluster 2 seems to present an interesting infrequent use of ridesharing apps as well, either riding alone or with strangers. However, the most enlightening characteristic of the first group of potential adopters may be the fact that they show little environmental concern although they scored high in the pro-EV attitude evaluation. Moreover, the opposite occurs in Cluster 2, where individuals scored high in environmental concern, but low in Pro-EV attitude. This is to some extent contradictory since EVs contribute positively to reduce climate change, so a person that is environmental concerned usually has also a pro-EV attitude and frequently chooses EV. We think that this

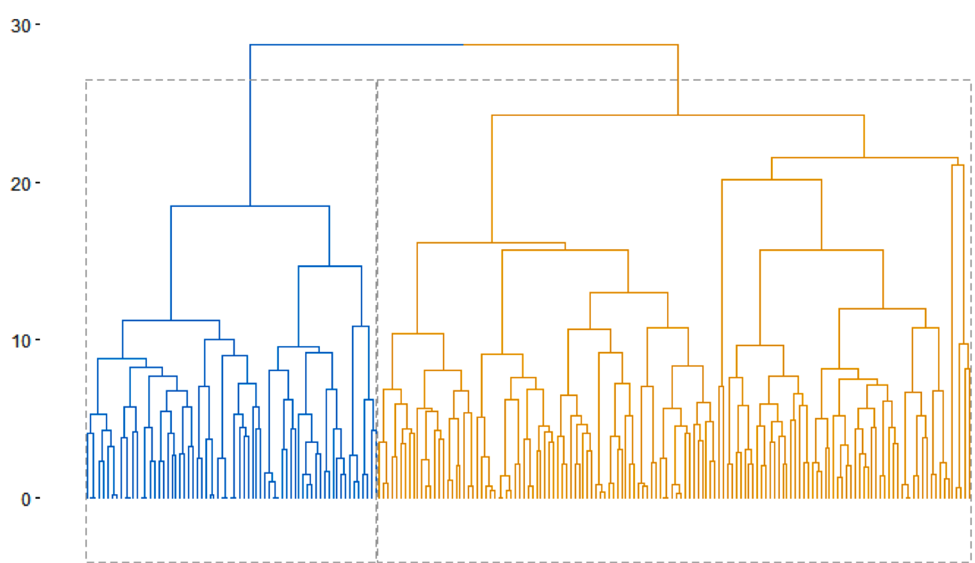


Fig. 7. Hierarchical clustering of the misclassified observations.

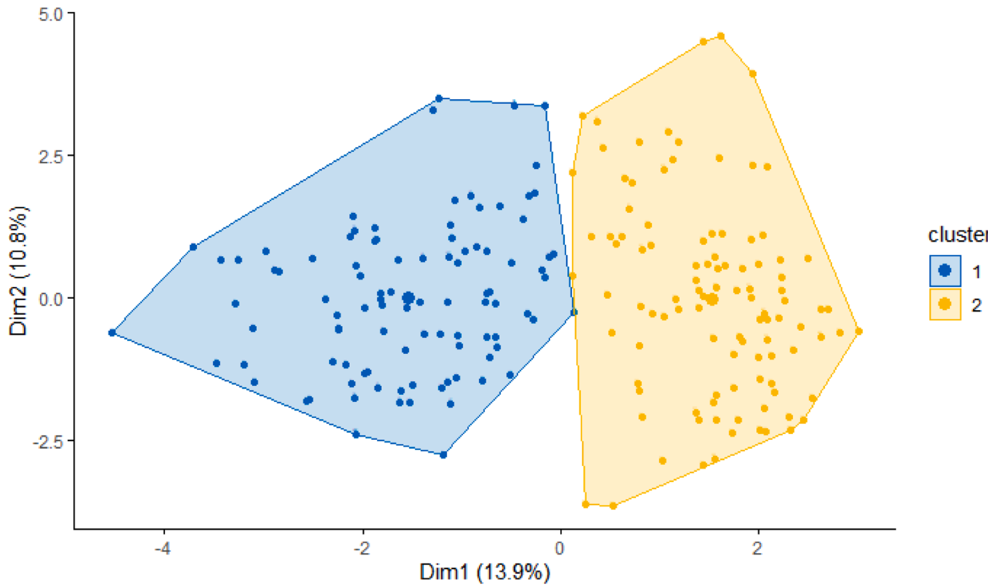


Fig. 8. Clustering results, two main Principal Components.

Table 4
Clusters characteristics.

Cluster 1	Cluster 2
High number of retired individuals	More presence of youngsters
Predominance of Prince George's county	Predominance of Montgomery county
Little concerned about the environment	Concerned about the environment
Pro-EV attitude	No Pro-EV attitude
	Infrequent use of ridesharing apps

contradiction is precisely what may be behind the misclassification; the algorithm might find trouble in labeling a person that expresses cares about the environment but does not have an inclination towards EVs, and vice versa.

The particular implications of misclassification depend on the case at hand. False positives and false negatives have obviously more impact in health-related experiments than they do in transportation or business matters. But in the specific approach that we are exploring, categorize a potential purchaser as an adopter when she is not, or vice versa, can severely affect expensive marketing campaigns that make use of socio-economic information as profiling factors. On the other hand, the case of unbalanced data (as is the case here since the percentage of adopters differ greatly from that of non-adopters) commonly yields skewed class distribution, i.e. prediction of the majority class. In this regard, classifying a non-adopter as an adopter is not desired, but less critical from a marketing perspective than classifying an adopter as a non-adopter since this may mean ignoring a consumer who is actually a potential purchaser. Therefore, algorithms based on user profiling that are capable of correctly allocate consumers in their respective clusters will be able to make better predictions and, therefore, provide important competitive advantages to those who develop and implement them. Of course, in the same way, and in another order of things, a correct identification and classification can impact policies that seek to stimulate the adoption of the electric vehicle.

6. Conclusions

In this paper we study two important dimensions of the EV adoption problem. First, we explore the most influencing factors in the adoption of the EV; second, we carry out a predictive analysis based on different machine learning methods. In addition, we analyze the structure of the observations that the best algorithm fails to classify, looking for the common characteristics of those individuals. This work is based on data collected through a stated choice survey specifically designed to gage the inclination of individuals towards EV. It pays special attention to the role played in their choices by the structure of their social network, as well as their attitudes towards the environment, technology, and EVs. Support Vector Machines (radial and polynomial kernel), Tree-based algorithms (Gradient Boosting Models, Distributed Random Forest, and Extremely Randomized Forest), and Neural Networks (one-layer Neural Network and Deep Neural Network) were estimated to classify individuals into adopters, and their respective throughputs were highlighted.

With respect to the first objective, the ML based analysis shows that when classifying individuals based on their propensity to adopt an EV, the most important factors are: the county in which the respondents live, the type of engine (electric or not) of the next vehicle to be acquired, vehicle characteristics, and both *PROEV* and *Technology Inclined* attitudes. Since there are no special differences among the counties of the State of Maryland in terms of power grid or charging infrastructure, we believe that this variable actually hides an income effect. Some of these counties are among the richest in the U.S. and they are evidence of a clear geographical income distribution. Among the vehicle characteristics, the most relevant seems to be the income tax deduction that the U.S. government provides when buying an EV. Considering that the fourth most important variable is the vehicle price, we can conclude that all the elements that gravitate around price are fundamental in the individuals' inclination to adopt this technology. However, other vehicle attributes are also crucial, such as the range and the time of fast charging, as well as the existence of charging infrastructure in the household. This is all valuable information for the automotive and power industry since these are precisely some of the barriers highlighted by users and researchers for a wide-scale implementation of the EV technology (Berkeley et al., 2017; Tran et al., 2012; Bonges et al., 2016; Dimitropoulos et al., 2013). The presence of attitudinal factors suggests that, beyond economic incentives, how users feel or behave towards certain concepts (environment, technology, EVs) is also of relevance. This is also an interesting finding for the public administration since, although these aspects are obviously inherent to each person, fostering social awareness about them would also boost the EV market. We consider these results to be in line with certain characteristic aspects of regions where there has been a more marked evolution of EV adoption. Northern European countries, as well as the states of California and Oregon, in the U.S., seem to have in common a society of high average income, committed to the environment, and with a marked interest in technology and EVs themselves (Glerum et al., 2013; Haustein et al., 2021; Jensen et al., 2014; Lutsey et al., 2015).

Concerning the second objective, the accuracy of all methods is similar, although that of the SVMs, Neural Network and XGBoost is slightly better. The most complex methods (XRT and DNN) are those that performs the worst, a result that we consider logical since these methods are more appropriate for problems of a higher mathematical complexity. In any case, the SVM with polynomial kernel yields an accuracy of 83.45%. Regarding the predictions that are correct (adopters and non-adopters classified as such), these are homogeneous among the three top methods (SVMs and ANN) but not among the other four, which present very high Sensitivities in contrast to low Specificities. In other words, their predictive power is mainly based on correctly identifying adopters, significantly misidentifying the potential non-adopters. This is of special importance since it is a main finding and unfortunately enough, it is expressed the other way around. Therefore, attending to the statistics described, we can conclude that the methods exhibit comparable predictability –specially SVMs, ANN and XGBoost, implying robust and reproducible results no matter which of these popular ML techniques is used.

Finally, we tried to identify characteristics common to the misclassified individuals. To do so, we carried out a cluster analysis followed by an exploratory data analysis. The results show that the observations incorrectly predicted belong to two well-differentiated groups. The first is characterized by retired persons that live in a low-income county and that do not care much about the environment but have a pro-EV attitude. The second cluster, in contrast, is composed by young potential customers that live in a high-income county, and that care about the environment although do not show special interest for the EV. This apparent contradiction might be the reason why the algorithm fails in classifying them. Misclassification may in fact affect all stakeholders involved in the EV adoption process. Categorizing a potential purchaser as an adopter when she is not, or vice versa, can severely impact expensive marketing campaigns that make use of socioeconomic information as profiling factors. Our study suggests that algorithms based on a variety of user and vehicle aspects will be more capable to correctly allocate consumers in their respective clusters and, therefore, make better predictions that will provide important competitive advantages to those who develop and implement them. In the same vein, a successful identification will improve the efficiency of any policy that seeks to stimulate the adoption of the EV. Focusing on the aspects that are most important for consumers, such as subsidies, operating cost exemptions, or the improvement of recharging infrastructures (Lutsey et al., 2015), will undoubtedly increase their chances of success.

This work is not without limitation. The results obtained are relative to the SP database available for the study and cannot be generalized. The quickly developing literature on AI methods offers vast opportunities to test different algorithms, classify customers, and predicts their choices. We hope to contribute to the combination of classical econometric analysis and ML techniques, which would help build comprehensive analysis tools for policy evaluations, and to support important decisions about investments in new and emerging technologies.

CRediT authorship contribution statement

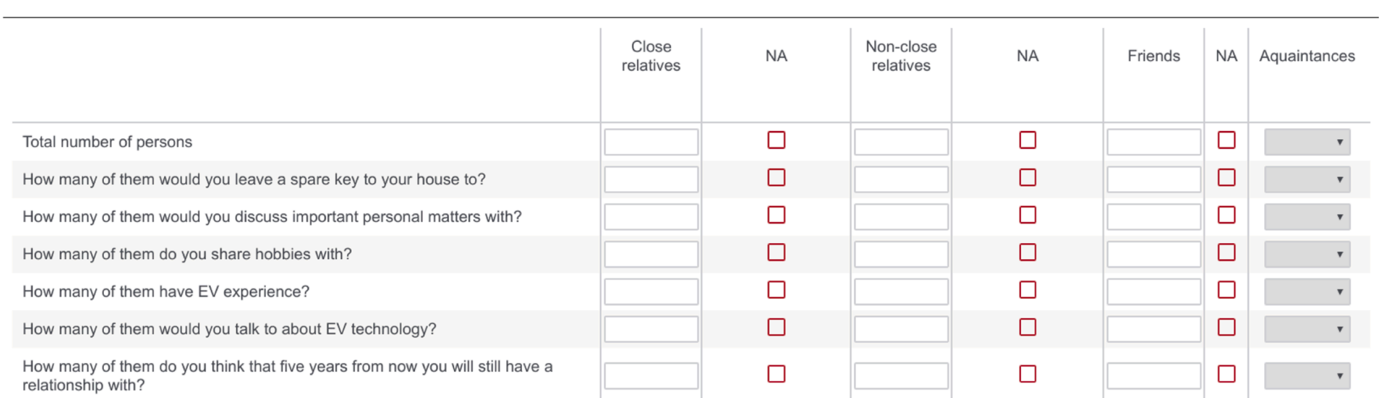
Javier Bas: Conceptualization, Data curtion, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Cinzia Cirillo:** Funding acquisition, Supervision, Writing – review & editing. **Elisabetta Cherchi:** Data curtion, Supervision.

Declaration of Competing Interest

None.

Appendix

Figs. A1 and A2, Table A1.



	Close relatives	NA	Non-close relatives	NA	Friends	NA	Aquaintances
Total number of persons							
How many of them would you leave a spare key to your house to?							
How many of them would you discuss important personal matters with?							
How many of them do you share hobbies with?							
How many of them have EV experience?							
How many of them would you talk to about EV technology?							
How many of them do you think that five years from now you will still have a relationship with?							

Fig. A1. Questionnaire question to collect information on the structure of the individuals' social network.

Please select a level of agreement to the following statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I do what I can to contribute to reduce global climate changes, even if it costs more and takes time (EC)	O	O	O	O	O
The authorities should not introduce legislation that forces citizens and companies to protect the environment (EC)	O	O	O	O	O
Electric vehicles should play an important role in our mobility systems (EC)	O	O	O	O	O
It is not important for me to follow technological development (TI)	O	O	O	O	O
I often purchase new technology products, even though they are expensive (TI)	O	O	O	O	O
I am optimistic about the future of shared mobility (such as carshare and rideshare) (TI)	O	O	O	O	O
New technologies create more problems than they solve (TI)	O	O	O	O	O
If I use an electric vehicle instead of a conventional vehicle, I would have to cancel some activities (ProEV)	O	O	O	O	O
Electric vehicles are more reliable than conventional vehicles (ProEV)	O	O	O	O	O
I am concerned that EVs are not powerful enough to make a safe takeover (ProEV)	O	O	O	O	O
When forced to change daily activity arrangement, I don't feel anxious. (ProEV)	O	O	O	O	O

Fig. A2. Questionnaire question to collect information on the attitudes of individuals.

Table A1
Summary of socioeconomic variables.

Age	
Min	18
Max	86
Ave	45
Female	60.88%
Married	48.63%
Employment status	
Government full time	7.25%
Government part time	0.84%
Private full time	35.91%
Private part time	7.16%
Self-employed	6.59%
Retired	17.24%
Student	5.84%
Unemployed	10.36%
Other	8.81%
Education degree	
Less than high school	1.97%
High school	14.32%
Graduate or professional degree	19.60%
Bachelor's degree	31.38%
Some college	32.70%
Individual gross income	
Min	$0
Max	$400,000
Ave	$53,724
Household gross income	
Min	$0
Max	$645,975
Ave	$86,187
% Income living expenses*	
Min	1
Max	99
Ave	60.71%

* Income share spent in Housing, Healthcare, Insurance, Food and Education.

References

Bas, J, Cirillo, C, Cherchi, E., 2020. A Stated Choice Experiment for consodering Social Conformity in the adoption of Electric Vehicles. ISCTSC 2020 - The 12th International Conference on Transport Survey Methods. In preparation.

Berkeley, N., Bailey, D., Jones, A., Jarvis, D., 2017. Assessing the transition towards Battery Electric Vehicles: A Multi-Level Perspective on drivers of, and barriers to, take up. Transportation Research part A: policy and practice 106, 320–332.

Bezdek, JC, Hathaway, RJ, 2002. VAT: a tool for visual assessment of (cluster) tendency. Proceedings of the 2002 International Joint Conference on Neural Networks (IJCC '02) 3, 2225–2230.

Bolton, R.J., Hand, D.J., 2002. Statistical fraud detection: a review. Stat. Sci. 17 (3), 235–249. JSTOR. Bp Statistical Review of World Energy 2020. (2020). 68.

Bonges, Lusk, A.C., 2016. Addressing electric vehicle (EV) sales and range anxiety through parking layout, policy and regulation. Transportation Research Part A: Policy and Practice 83, 63–73.
bp, 2020. Statistical Review of World Energy https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.
Buuren, S., Groothuis-Oudshoorn, K., 2011. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45 (3) https://doi.org/10.18637/jss.v045.i03.
Chen T., & Guestrin C., (2016). XGBoost: A Scalable Tree Boosting System. KDD'16 : Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and data mining, 785–794 . https://doi.org/10.1145/2939672.2939785.
Cherchi, E., 2017. A stated choice experiment to measure the effect of informational and normative conformity in the preference for electric vehicles. Transp. Res. A Policy Pract. 100, 88–104. https://doi.org/10.1016/j.tra.2017.04.009.
Crutchfield, R.S., 1955. Conformity and character. American Psychologist 10 (5), 191–198 doi.org/10.1037/h0040237.
Deo Rahul, C, 2015. Machine learning in medicine. Circulation 132 (20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
Dimitropoulos, A., Rietveld, P., Van Ommeren, J.N., 2013. Consumer valuation of changes in driving range: A meta-analysis. Transportation Research Part A: Policy and Practice 55, 27–45.
Docherty, I., Marsden, G., Anable, J., 2018. The governance of smart mobility. Transp. Res. A Policy Pract. 115, 114–125. https://doi.org/10.1016/j.tra.2017.09.012.
Fukushima, A., Yano, T., Imahara, S., Aisu, H., Shimokawa, Y., Shibata, Y., 2018. Prediction of energy consumption for new electric vehicle models by machine learning. IET Intell. Transp. Syst. 12 (9), 1174–1180. https://doi.org/10.1049/iet-its.2018.5169.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42. https://doi.org/10.1007/s10994-006-6226-1.
Glerum, A., Stankovikj, L., Thémans, M., Bierlaire, M., 2013. Forecasting the demand for electric vehicles: accounting for attitudes and perceptions. Transp. Sci. 48 (4), 483–499. https://doi.org/10.1287/trsc.2013.0487.
Goebel, D., Plötz, P., 2019. Machine learning estimates of plug-in hybrid electric vehicle utility factors. Transp. Res. D Transp. Environ. 72 (April), 36–46. https://doi.org/10.1016/j.trd.2019.04.008.
Guzella, T.S., Caminhas, W.M., 2009. A review of machine learning approaches to Spam filtering. Expert Syst. Appl. 36 (7), 10206–10222. https://doi.org/10.1016/j.eswa.2009.02.037.
Haustein, S., Jensen, A.F., Cherchi, E., 2021. Battery electric vehicle adoption in Denmark and Sweden: recent changes, related factors and policy implications. Energy Policy 149, 112096. https://doi.org/10.1016/j.enpol.2020.112096.
Haykin, S., 1994. Neural Networks: A Comprehensive Foundation, 1st ed. Prentice Hall PTR.
Hernandez, S., Monzon, A., de Oña, R., 2016. Urban transport interchanges: a methodology for evaluating perceived quality. Transp. Res. A Policy Pract. 84, 31–43. https://doi.org/10.1016/j.tra.2015.08.008.
Hu, X., Li, S.E., Yang, Y., 2016. Advanced machine learning approach for lithium-ion battery state estimation in electric vehicles. IEEE Trans. Transp. Electrif. 2 (2), 140–149. https://doi.org/10.1109/TTE.2015.2512237.
Huang, X., Tan, Y., He, X., 2011. An intelligent multifeature statistical approach for the discrimination of driving conditions of a hybrid electric vehicle. IEEE Trans. Intell. Transp. Syst. 12 (2), 453–465. https://doi.org/10.1109/TITS.2010.2093129.
Hughes, S., Moreno, S., Yushimito, W.F., Huerta-Cánepa, G., 2019. Evaluation of machine learning methodologies to predict stop delivery times from GPS data. Transp. Res. C Emerging Technol. 109, 289–304. https://doi.org/10.1016/j.trc.2019.10.018.
Jahangiri, A., Rakha, H.A., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans. Intell. Transp. Syst. 16 (5), 2406–2417. https://doi.org/10.1109/TITS.2015.2405759.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Introduction. In: James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), *An Introduction to Statistical Learning: With Applications in R*. Springer, pp. 1–14 https://doi.org/10.1007/978-1-4614-7138-7_1.
Jensen, A.F., Cherchi, E., de Dios Ortúzar, J., 2014. A long panel survey to elicit variation in preferences and attitudes in the choice of electric vehicles. Transp. Amst. 41 (5), 973–993. https://doi.org/10.1007/s11116-014-9517-6.
Jensen, A.F., Cherchi, E., Mabit, S.L., Ortúzar, J.de D., 2016. Predicting the Potential Market for Electric Vehicles. Transportation Science 51 (2), 427–440 https://doi.org/10.1287/trsc.2015.0659.
Jia, J., 2019. Analysis of alternative fuel vehicle (AFV) adoption utilizing different machine learning methods: a case study of 2017 NHTS. IEEE Access 7, 112726–112735. https://doi.org/10.1109/ACCESS.2019.2934780.
Jolliffe, I.T., 1986. Principal components in regression analysis. In: *Principal component analysis*. Springer, New York, NY, pp. 129–155.
Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. Informatica 31, 249–268.
Lee, D., Mulrow, J., Haboucha, C.J., Derrible, S., Shiftan, Y., 2019. Attitudes on autonomous vehicle adoption using interpretable gradient boosting machine. Transp. Res. Rec. 2673 (11), 865–878. https://doi.org/10.1177/0361198119857953.
Lawson, R.G., Jurs, P.C., 2002, May 1. *New index for clustering tendency and its application to chemical problems* (world) [Research-article]. American Chemical Society https://doi.org/10.1021/ci00065a010.
Lee, H., Kim, S.G., Park, H., Kang, P., 2014. Pre-launch new product demand forecasting using the Bass model: a statistical and machine learning-based approach. Technol. Forecast. Soc. Change 86, 49–64. https://doi.org/10.1016/j.techfore.2013.08.020.
Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26. https://doi.org/10.1016/j.neucom.2016.12.038.
Liu, Z., Liu, Y., Meng, Q., Cheng, Q., 2019. A tailored machine learning approach for urban transport network flow estimation. Transp. Res. C Emerging Technol. 108, 130–150. https://doi.org/10.1016/j.trc.2019.09.006.
Bureau of Transportation Statistics, 2020. https://www.bts.gov/topics/national-transportation-statistics. (Accessed February 2020).
Lutsey, N., Searle, S., Chambliss, S., Bandivadekar, A., 2015. Assessment of leading electric vehicle promotion activities in United States cities. White paper of the Intrenational Council of Clean. Transportation (ICTT)., 1–60.
MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14,, 281–297.
Oliver, N.M., Rosario, B., Pentland, A.P., 2000. A Bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), 831–843. https://doi.org/10.1109/34.868684.
Sagiroglu, S., Yavanoglu, U., Guven, E.N., 2007. Web based machine learning for language identification and translation. In: Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA), pp. 280–285. https://doi.org/10.1109/ICMLA.2007.27.
Schölkopf, B., Smola, A.J., 2018. Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press. https://doi.org/10.7551/mitpress/4175.001.0001.
Sheng, H., Xiao, J., 2015. Electric vehicle state of charge estimation: nonlinear correlation and fuzzy support vector machine. J. Power Sources 281, 131–137. https://doi.org/10.1016/j.jpowsour.2015.01.145.
Singh, A., Thakur, N., Sharma, A., 2016. A review of supervised machine learning algorithms. In: Proceedings of the 10th INDIACom; 3rd International Conference on Computing for Sustainable Global Development 2016, pp. 1310–1315.
Stone, P., Veloso, M., 2000. Multiagent systems: a survey from a machine learning perspective. Auton. Rob. 8 (3), 345–383. https://doi.org/10.1023/A:1008942012299.
Sun, S., Zhang, J., Bi, J., Wang, Y., Moghaddam, M.H.Y, 2019. A machine learning method for predicting driving range of battery electric vehicles. J. Adv. Transp. https://doi.org/10.1155/2019/4109148.
Tran, M., Banister, D., Bishop, J.D.K., McCulloch, M.D., 2012. Realizing the electric-vehicle revolution. Nature Climate Change 2 (5), 328–333 https://doi.org/10.1038/nclimate1429.
Wang, Z., Wang, Y., Zeng, R., Srinivasan, R.S., Ahrentzen, S., 2018. Random forest based hourly building energy prediction. Energy Build. 171, 11–25. https://doi.org/10.1016/j.enbuild.2018.04.008.
Yang, S., Ma, W., Pi, X., Qian, S., 2019. A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources. Transp. Res. C Emerging Technol. 107, 248–265. https://doi.org/10.1016/j.trc.2019.08.010.
Yavasoglu, H.A., Tetik, Y.E., 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, D., Su, J., Liu, C., Quddus, M., Chen, W.-.H., 2019. A machine learning based personalized system for driving state recognition. Transp. Res. C Emerging Technol. 105, 241–261. https://doi.org/10.1016/j.trc.2019.05.042.
Zahid, T., Xu, K., Li, W., Li, C., Li, H., 2018. State of charge estimation for electric vehicle power battery using advanced machine learning algorithm under diversified drive cycles. Energy 162, 871–882. https://doi.org/10.1016/j.energy.2018.08.071.
Zarazua de Rubens, G., 2019. Who will buy electric vehicles after early adopters? Using machine learning to identify the electric vehicle mainstream market. Energy 172, 243–254. https://doi.org/10.1016/j.energy.2019.01.114.
Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. Soc. 20, 22–35. https://doi.org/10.1016/j.tbs.2020.02.003.

Javier Bas Vicente: Javier Bas is a Visiting Professor at Universidad de Alcalá de Henares, where he teaches Statistics and Econometrics. His background in Economics and Econometrics allowed him to carry out studies in welfare measures associated to changes in toll policy. He has also conducted research on the prediction of the diffusion of the electric vehicle considering Social Conformity. He currently works in Machine Learning methods associated to the diffusion of new technologies, and to the forecast of financial assets. He has participated and continue participating in several transportation projects in the State of Maryland (U.S). Prior his academic career, he worked in the consultancy industry for several years, developing international projects in Economics and Business. His specialty areas are in transport modeling, machine learning, advanced econometrics, and survey methods.

Cinzia Cirillo: Cinzia Cirillo is a Full Professor at the University of Maryland, Department of Civil and Environmental Engineering. Her specialty areas are in transport modeling techniques and survey instruments, application of advanced statistical and econometrics methods, and on the analysis of their results to predict consumer demand and behavior for various transportation options. She is conducting projects in the US, in Europe and in the Middle East, she has also taught advanced classes in China and Taiwan. Her current research is about dynamic discrete choice models, discrete-continuous models, and their application to the market penetration of new technology vehicles. Recently, one of her PhD students won the Eric Pas prize with a thesis on social interactions in activity and travel behavior models.

Elisabetta Cherchi: Elisabetta Cherchi is Professor of Transport at the School of Engineering, Newcastle University (UK) and Adjunct Professor, School of Economics & Management, Beijing Jiaotong University (China). She currently Chairs the International Association of Travel Behaviour Research (IATBR), is Co-Editor in Chief of Transportation Research Part A: Policy and Practice, and past Associate Editor of Transportation.