ISSN: 2666-7649
DATA SCIENCE AND MANAGEMENT
KeAi
Available online at www.sciencedirect.com
ScienceDirect

Using machine learning to identify primary features in choosing electric vehicles based on income levels

Mingjun Ma, Eugene Pinsky

PII: S2666-7649(23)00044-9

DOI: https://doi.org/10.1016/j.dsm.2023.10.001

Reference: DSM 73

To appear in: *Data Science and Management*

Received Date: 23 February 2023

Revised Date: 1 October 2023

Accepted Date: 5 October 2023

Using Machine Learning to Identify Primary Features in Choosing Electric Vehicles Based on Income Levels

Mingjun Ma, Eugene Pinsky, PhD.

Department of Computer Science
Boston University Metropolitan College
United States

Author Information:
First author: Mingjun Ma Organization: Boston University Metropolitan College
Address: 1010 Commonwealth Ave, Boston, MA, 02215, USA
Email: mbm1027@bu.edu

Second author: Eugene Pinsky, PhD.
Organization: Boston University Metropolitan College
Office Address: 1010 Commonwealth Ave, Room 327, Boston, MA, 02215, USA
email: epinsky@bu.edu

Declarations: **Ethical Approval:** Not Applicable. No human and/ or

animal studies were performed in connection with this work.

Declaration of Interests: None

Competing Interests: The authors declare no conflict of interest. The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare relevant to this article's con- tent. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Funding: No funds, grants, or other support was received.

Availability of Data and Materials: No new data were created during

this study. The dataset used for analysis is the household data of electric vehicle ownership from the State of California. It is a publicly available dataset. Its url is:

https://datadryad.org/stash/dataset/doi:10.25338/B8P313

Data Availability Statement:
The data that support the findings of this study are openly available in ”Dryad” at

https://doi.org/10.25338/B8P313

Biographical Note:

Mingjun Ma:

Mingjun Ma is a master student in Boston University. He received the Bachelor's Degree in Chemical Engineering from Pennsylvania State University in May 2021, and now he has graduated with a master's degree in Applied Data Analytics in Boston University.

Eugene Pinsky:

Dr. Eugene Pinsky first joined the faculty of Boston University in 1986, after earning his doctorate at Columbia University. He was an assistant professor of computer science in the College of Arts & Sciences until 1993, when he left BU and gained extensive industry experience designing com- putational methods to analyze and monitor market risk at multiple trading and investment firms, including Bright Trading, F-Squared Investments, and Harvard Management Company. He has also applied his expertise in data mining, predictive analytics, and machine learning to video advertising at Tremor Video. Dr. Pinsky's areas of expertise include pattern recognition, clustering, regression, prediction, factor models, support vector machines, and other machine learning algorithms and methods for data analysis; multi-dimensional statistical data analysis of large time-series data; data mining and predictive analytics to uncover patterns, correlations, and trends; algo- rithmic trading, pricing models, financial modeling, risk, and portfolio anal- ysis; Python, R, C/C++, VBA, Weka, MATLAB, and MySQL; Big Data technologies and visualization (Hadoop/Hive, AWS, Tableau); design and implementation of software tools for quantitative analysis; and professional curriculum and course development.

Using Machine Learning to Identify Primary Features in Choosing Electric Vehicles Based on Income Levels

Abstract

An electric vehicle is becoming one of the popular choices when choosing a vehicle. People are generally impressed with electric vehicles' zero-emission and smooth drives, while unstable battery duration keeps people away. This study tries to identify the primary factors that affect the likelihood of owning an electric vehicle based on different income levels. We divide the dataset into three subgroups by household income - \$50,000 to \$150,000 or low-medium income level, \$150,000 to \$250,000 or medium-high income level, and \$250,000 or above - high-income level. We considered several machine learning classifiers, and Naive Bayes gave us a relatively higher accuracy than other algorithms in terms of overall accuracy and F_1 scores. Based on the probability analysis, we found that for each of these groups, one-way commuting distance is the most important for all three income levels.

Keywords: Unbabalced data; electric vehicle; machine learning; sampling with replacement; supervised learning; Naive Bayes

1. Introduction

An electric vehicle is one of the popular choices for vehicle owners. Zero emission and a smooth driving experience are the potential reasons for it. According to Majumder (Majumder, 2021), the mechanical efficiency of an electric vehicle is between 60% and 70%, while the efficiency of a vehicle with an internal combustion engine is 18% to 22% (Majumder, 2021). On the other side, short Environmental Protection Agency (EPA) range and charging issues keep people from electric vehicles - the average range of gasoline-powered vehicles is 300 miles (Bonges and Lusk, 2016). In comparison, the average range of an electric vehicle is 110 miles (Bonges and Lusk, 2016). The battery of an electric vehicle is not capable of supporting long-distance driving without charging stations. However, electric vehicles are popular in some regions regardless of traveling range.

A number of other authors (Sovacool et al., 2019; Carley et al., 2013; Axsen et al., 2016) suggested that high annual-income households show more interest and are more likely to have electric vehicles at home than low-income households. The price of an electric vehicle is generally more expensive than other types of vehicles. Furthermore, those earning more than 90,000 euros per year are more concerned about fuel efficiency, technical reliability, and safety, while households with less than 10,000 euros are concerned about financial savings (Sovacool et al., 2019). Those who earn 10,000 euros per year and own electric vehicles are also concerned about the public charging availability and charging speed (Sovacool et al., 2019). Therefore, the difference in income level enables people to have different thoughts about choosing an electric vehicle, which leads to two different decisions.

Our study aims to use classification machine learning models to predict the current ownership of electric vehicles based on different income levels in the United States. In the modeling process, we decided to apply different classification models, which include SVM, logistic regression, decision tree, and Naive Bayesian, to classify the ownership and identify the primary factor of different income levels in the United States. This research aims to find the primary factors that affect the likelihood of selecting an electric vehicle over other types of vehicles based on income level.

For our study, we took the data collected by researchers at the University of California, Davis, to study the ownership of electric vehicles (BEV) or fuel-celled vehicles (FCV) from the drivers of the State of California (Hardman, 2019). The dataset has 27,022 rows and 25 columns, 24 features plus Response ID. The dataset is recorded anonymously (Hardman, 2019). The top 5 rows of the original dataset are shown in Table 1. In the dataset, there are here are 921 rows for FCV owners and 13,123 rows for BEV owners. The remaining 12,978 rows correspond to participants without a response (Hardman, 2019). Such participants are excluded from our analysis. Furthermore, we reduced the number of features from 24 to 7 features. Our preliminary investigation using Logistic Regression has indicated that a number of features are not very significant; we chose the seven most significant features, and these are summarized in Table 2.

We are interested in analyzing this reduced dataset depending on the income levels. Fig. 1 shows the overall distribution of household income. Based on this empirical distribution, we split respondents into 3 income groups: $50,000 to $150,000 (Low-Medium, black left), $150,000 to $250,000 (Medium-High, grey), and $250,000 or above (High, black right).

Table 1. Top Five Columns

submitdate...	...	Gender...	# vehicles...	Annual VMT[1] Estimate	FCV[2], BEV[3]...
2017/06/02 11:30:57	...	0.0	2	14622.000000	0.0
2017/06/02 11:15:39	...	0.0	3	9197.142857	0.0
2017/06/02 10:51:59	...	1.0	4	15360.000000	0.0
NaN	...	NaN	1	NaN	NaN
2017/06/02 19:35:59	...	0.0	3	5082.352941	0.0

[1] VMT–means Vehicle Miles Travelled (Alvarez, 2023).
[2] FCV–means Fuel Cell Vehicle.
[3] BEV–means Battery Electric Vehicle.

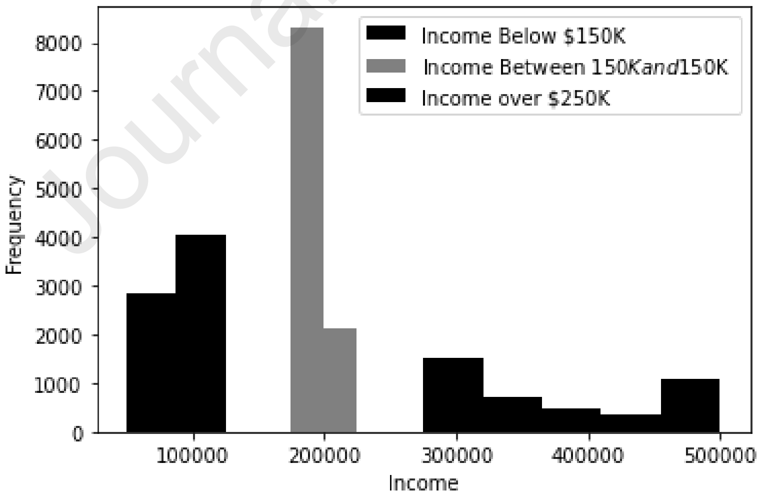


Fig. 1. Histogram of Income Level

2. Material And Methods

2.1. Data

After cleaning the null values, there are 93% of BEV owners with 7% from FCV owners from the original dataset, which is extremely unbalanced. Fuel Cell Vehicle (FCV) owners are classified as class 0, while the owners of Electric Vehicle (BEV) are classified as 1. We sample the 921 records of class 0 with replacements so that the data number of rows of class 0 is the same as the number of rows of class 1 so there will be enough data for us to train and test.

2.2. Assumptions and Data Selection

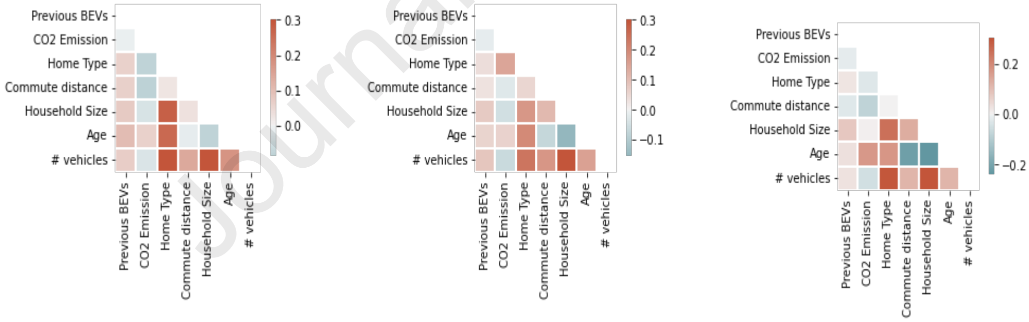
Our study aims to use simple and explainable machine learning tools to learn the primary factors of selecting a BEV from different income levels; we do not use complex methods such as Neural Networks in our study (Aggarwal, 2018). After sampling class 0 mentioned in Section 2.1, there are approximately 20,000 rows in the dataset. Therefore, the kNN algorithm is not a good candidate for analysis (Cover and Hart, 1967). In the modeling process, the following assumptions are made before the modeling process. First, the classification as an FCV vehicle (class 0) is assumed to be non-BEV, and the situation of Previous BEVs only affects personal decisions. Second, the selected variables for modeling apply to all places. The original data is collected from the State of California; the statistics differ from other places. In addition, there is no extreme one-commuting distance. The maximum value is above 2,000 miles, while the median of the one-way commute distance is 19 miles. No features are normally distributed. Therefore, for numeric variables, we use a standard practice in statistics and define outliers as values greater than $Q_3 + 1.5 \cdot (Q_3 - Q_1)$ or smaller than $Q_1 - 1.5 \cdot (Q_3 - Q_1)$ (Johnson and Kotz, 1970). Therefore, we remove such outliers from the dataset.

Next, we assume that the selected attributes are independent. This is supported by the low level of correlations as will be shown in Fig. 2a to Fig. 2c. We separate the dataset based on income levels. The household income is not presented for the prediction process. Therefore, based on these assumptions, we have selected seven attributes $f_1, f_2, \dots, f_7$ as the independent variables. The statistical data for seven independent variables is summarized in Table 2. The label "FCV, BEV Dummy" is the output. All seven input features are numeric. When suing some of the machine learning methods such as SVM, we need to scale the data.

Table 2. Statistical Data For Seven Input Features

Feature	Feature Name	Mean μ	Standard deviation σ
f_1	Previous BEVs	0.121	0.326
f_2	CO_2 Emission	1.540	1.588
f_3	Home Type	0.718	0.450
f_4	Commute Distance	14.827	10.045
f_5	Household size	2.918	1.241
f_6	Age	49.970	13.247
f_7	# vehicles	2.463	0.920

Let us now examine the correlations between features based on income levels. The correlation matrices are shown in Fig. 2. We see that for all three income levels from Fig. 2a to Fig. 2c, the absolute values for all correlations are lower than 0.5 and are typically around 0.1. Therefore, the features selected are independent or weekly correlated to each other. This weak correlation indicates that the dataset will be a good candidate for analysis by the Naive Bayes algorithm (Bishop, 2016; Hastle, 2018; Kroese et al., 2019).



(a) Low-Medium Income Level

(b) Medium-High Income Level

(c) High Income Level

Fig. 2. Correlation Matrix for Three Income Levels

3. Results

We applied several machine learning classifiers to predict the ownership class (Hastle, 2018). We tried two ways of classification, cross-validation by 10 splits and train-test split by 50% and 50% of the dataset. Our results are summarized in Table 3.

Table 3. Classifier Performance Accuracy

	Low		Medium		High	
Algorithm	Train	Cross-Val.	Train	Cross-Val.	Train	Cross-Val.
Naive Bayes	83%	86%	82%	85%	86%	88%
SVM	58%	58%	55%	55%	59%	60%
Logistic	56%	57%	55%	55%	59%	60%
Decision Tree	51%	97%	50%	96%	50%	97%
Random Forest	50%	78%	50%	72%	50%	85%

Table 4. Classifier Confusion Matrix And F_1 Scores

			Low		
Algorithm	TP	FP	FN	TN	F_1 Score
Naive Bayes	1,053	83	510	1,756	78%
SVM	1,139	1,003	424	836	61%
Logistic	758	699	805	1,140	50%
Decision Tree	1,385	28	178	1,811	93%
Random Forest	1,207	443	356	1,396	75%
			Medium		
Naive Bayes	3,087	404	1,177	3,918	80%
SVM	2,188	1,785	2,076	2,537	53%
Logistic	1,966	1,587	2,298	2,735	50%
Decision Tree	3,739	66	525	4,256	93%
Random Forest	2,612	735	1,652	3,587	69%
			High		
Naive Bayes	852	27	269	1,021	85%
SVM	672	442	449	606	60%
Logistic	690	457	431	591	61%
Decision Tree	1,034	13	87	1,035	95%
Random Forest	958	160	163	888	86%

We applied Decision Tree, Random Forest, Logistic Regression, SVM, and Naive Bayes (Bishop, 2016; Hastle, 2018; Kroese et al., 2019). Decision Tree builds a tree-structured algorithm to make predictions. Each internal node in the tree represents an input feature, and each branch formed by internal nodes represents a prediction outcome. Each leaf node in a decision tree represents the predicted outcome from each branch. The features that form a decision tree are selected by the lowest impurity level in the training process so that the data with the same class can be gathered - which is measured by entropy. Random Forest is a bagging algorithm that splits data into multiple independent batches and builds separate decision trees. The most-voted tree is selected for the prediction process. Logistic Regression predicts the probability of an outcome based on the input features, and the equation is expressed as

$$P(\text{class}) = \frac{1}{1 + \exp(-(b_0 + b_1 x_1 + ... + b_n x_n))} \tag{1}$$

where b_0 represents the intercept, x_n represents n^{th} input feature, and b_n represents logistic coefficient. SVM finds an optimal hyperplane to separate data points from different classes. Naive Bayes Algorithm applies Bayes' theorem and conditional probabilities to predict the probability of an event under all features, and each feature is assumed to be independent of other features, and the equation is expressed as (Bishop, 2016; Hastle, 2018; Kroese et al., 2019):

$$P(\text{class} \mid x_1, ..., x_n) = \frac{P(x_1, ..., x_n \mid \text{class}) \cdot P(\text{class})}{P(x_1, ..., x_n)} \tag{2}$$

Our research aims to find algorithms in an easy-computing and explainable way, so we don't consider complex algorithms, such as Neural Networks. Those algorithms predict with high accuracies, but the confounding issue is that those algorithms are hard to explain. We are looking for algorithms such as logistic regression. In our experiment, the Naive Bayes algorithm performs high accuracy, and "LIME" helps us find each class's probability under a particular variable (Ribeiro, 2016a,b).

Table 3 shows the overall performance of predicting current ownership of electric vehicles based on different machine learning models, which are evaluated based on overall accuracy, and the confusion matrix and F_1 scores are indicated in Table 4. According to Table 4, True Positive (TP) refers to the rows whose predicted and actual values are equal to 1. True Negative (TN) refers to the rows whose predicted and actual values equal 0. False Positive (FP) refers to predicted values of 1 with actual values of 0. False Negative (FN) refers to predicted values of 0 with actual values of 1. F_1 scores are calculated as:

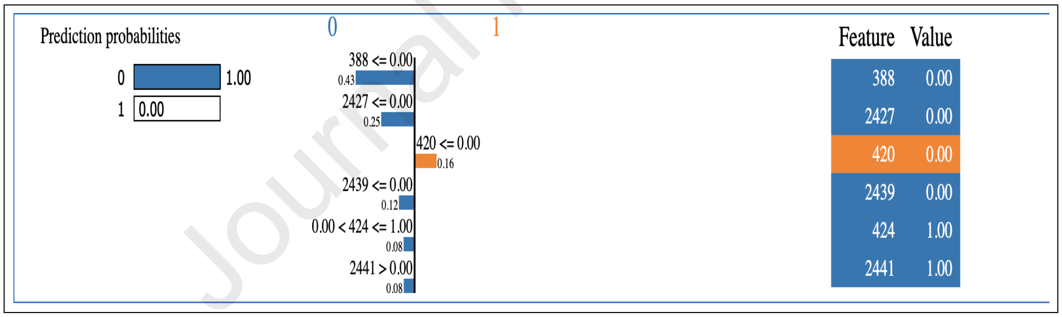
$$F_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

The seven features indicated in Section 2 are independent variables, and the feature "FCV, BEV Dummies" is the dependent variable. The overall result Table 3 shows that the Naive Bayes algorithm is able to correctly predict the ownership of the electric vehicle with accuracies all greater than 80% for all income levels by train - test split and cross-validation, so the Naive Bayes performs better across three different income levels. According to the correlation matrices in Fig. 2a, to 2c, features in each subgroup are under low correlations, which is the assumption of Naive Bayes (Huber, 2009; Upton and Cook, 1996; Hastle, 2018). This explains why the Naive Bayes algorithm performs best for all three groups. Regarding all algorithms and income levels, we observed that the cross-validation selection method predicts 0% to 5% higher accuracy than the train-test split training, which is indicated in Table 3.

The seven features mentioned in Section 2.2 are numeric variables, but none of the features are normally distributed. Therefore, we used the Multinomial Naive Bayes algorithm instead of the Gaussian Naive Bayes algorithm. For independent variables, every value from seven features in the dataset is treated as a tokenizer in text mining. The existing value from a single row is encoded to 1 if the value exists, 0 if not existing. We applied Local Interpretable Model-Agnostic Explanations, or "LIME" package (Ribeiro, 2016b), to compute the probability of being classified as 0 or 1 while the feature equals a specific value. According to Dr. Marco Tulio Ribeiro (Ribeiro, 2016a)- the developer of LIME, LIME explains the black-box algorithms such as Neural Networks in an explainable way and how the model behaved in every single case (Ribeiro, 2016a). Such analysis of our case is discussed in Figures 2, 3, and 4. We assume that the probability refers to the probability of class 0 or 1 when a feature is less than equal to the value. We have studied several algorithm-explaining tools, including ANCHOR, LIME, and SHAP (Science, 2019). All three tools are able to provide good analysis. However, LIME is the best tool to explain Naive Bayes because LIME is a well-developed tool with plenty of models than the others (Science, 2019).

Table 5. Index Number and Corresponding Feature Values

Low		Medium		High	
Value Index	Value	Value Index	Value	Value Index	Value
388	$f_2 = 2.65$	497	$f_2 = 2.66$	1,938	$f_4 = 36.89$
2,427	$f_5 = 5$	489	$f_2 = 2.58$	1,066	$f_4 = 11.23$
2,439	$f_6 = 75$	2,192	$f_4 = 17.7$	721	$f_4 = 5.92$
424	$f_3 = 0$	3,466	$f_4 = 36.07$	200	$f_2 = 0.93$
2,441	$f_7 = 1$	3,782	$f_5 = 8.0$	135	$f_2 = 0.02$
420	$f_2 = 2.97$	3,775	$f_5 = 1.0$	399	$f_2 = 2.99$
2,222	$f_4 = 32.65$	481	$f_2 = 2.5$	2,066	$f_6 = 35.0$
557	$f_4 = 2.0$	3,289	$f_4 = 32.65$	NaN	NaN
NaN	NaN	3,788	$f_6 = 35.0$	NaN	NaN



Feature	Value
388	0.00
2427	0.00
420	0.00
2439	0.00
424	1.00
2441	1.00

Fig. 3. LIME Analysis For Low - Medium Income Level (Class 0)

We randomly pick up 10 rows that are correctly predicted for each income level- 5 rows of data from class 1, and 5 rows of data from class 0. For each income level, We also picked one row of data on each predicted classification in which the probability of the predicted class is the highest among others (over 90%) in Fig. 3 to Fig. 8 that shows the probability of predicting as class 0 or 1 and the probability of being classified as 0 or 1 while the feature equals a specific value. The numbers labeled in Fig. 3 to Fig. 8 are the indices of every single value in the seven selected features as input, and the result from each income level has a different index. The index and the corresponding

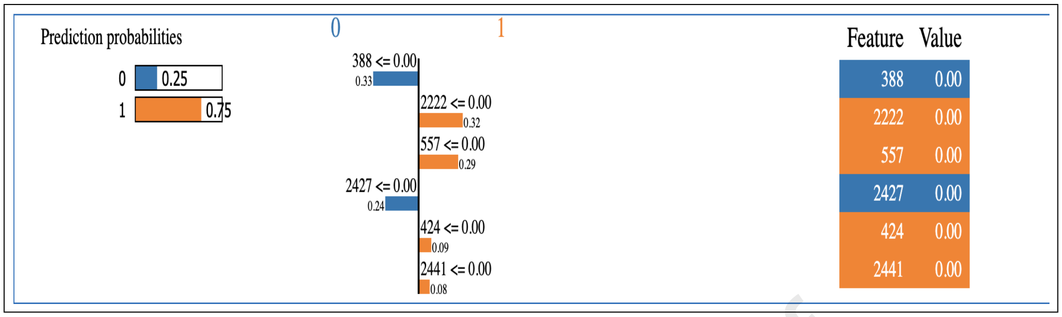


Fig. 4. LIME Analysis For Low - Medium Level (Class 1)

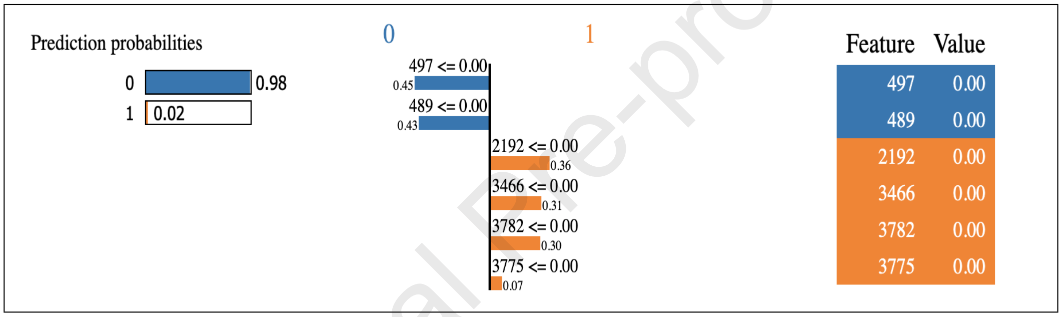


Fig. 5. LIME Analysis For Medium - High Income Level (Class 0)

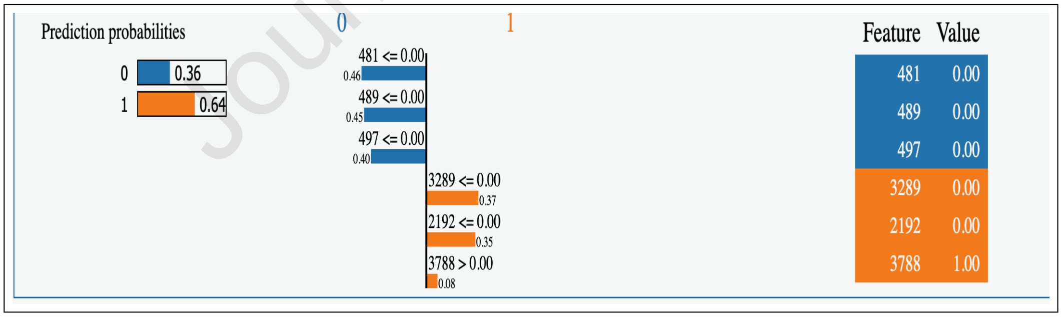


Fig. 6. LIME Analysis For Medium - High Income Level (Class 1)

values for each income level are recorded in Table 5. We find out that "One-way commute distance" is the most important factor for all income levels (for low-medium and high-income levels).

For drivers under low - medium income level (Household income level between $5,0000 and $150,000), 70% of people whose one-way commuting

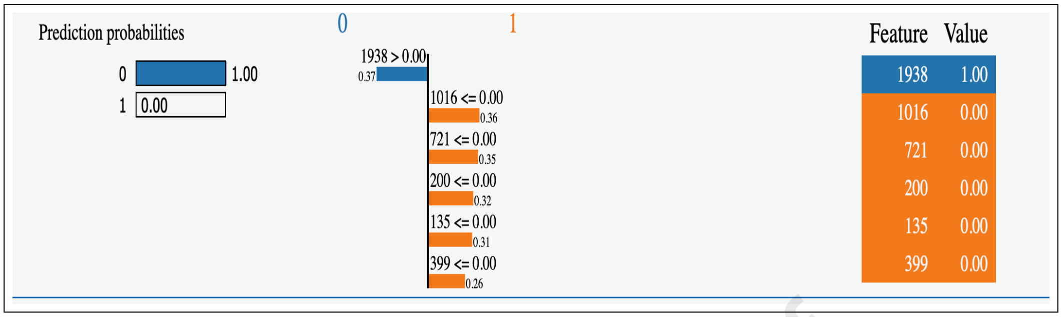


Fig. 7. LIME Analysis For High-Income Level (Class 0)

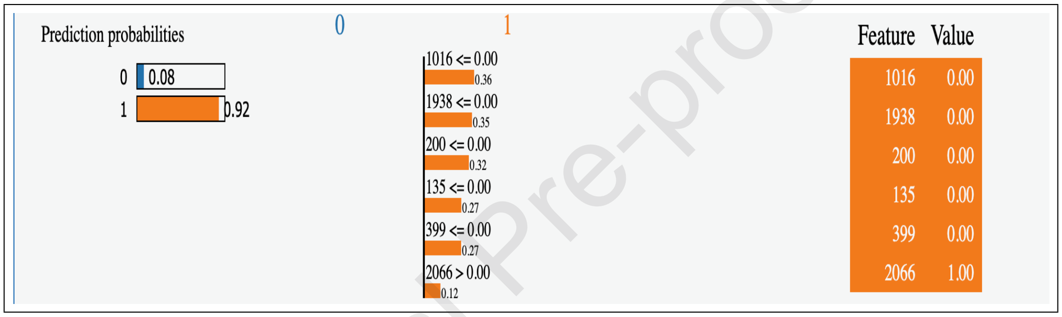


Fig. 8. LIME Analysis For High-Income Level (Class 1)

distance is less than or equal to 10 miles own Fuel Cell Vehicles (FCV), and the median and mean commuting distances for low-medium income levels are 12.97 miles and 13.85 miles. This indicates that people under the low-medium income level prefer to own a non-electric vehicle at short commuting distances, and people who need long commuting distances prefer electric vehicles over non-electric vehicles. This also applies to high-income-level car owners (Household income level greater than $250,000) whose average commuting distances are 16.14. 65% of drivers whose one-way commuting distances are less than 12 miles drive non-electric vehicles. The longest one-way commute distance for all three income levels is 44.71 miles; extreme one-way commuting is excluded from the classification process. For low - medium, and high-income level car owners, an electric vehicle is preferred when the commuting distance is above the average one-way commuting distance from the corresponding income level. For car owners under the medium-high income level (Household income level between $150,000 and $250,000), 63% of drivers will prefer non-electric vehicles if the one-way commute distance is less than 32 miles.

Age is another important factor for medium - high-income and high-income car owners. Drivers over the age of 35 for both income levels are more likely to choose the Fuel Cell Vehicle (FCV).

4. Related Literature and Discussion

Sovacool et. al. (Sovacool et al., 2019) studied the ownership of electric vehicle ownership and conducted a number of extensive surveys. However, no prediction model associated with the research was presented. According to their study, high-income households believe that electric vehicles bring a comfortable and smooth driving experience. Low-income electric vehicle owners would choose a vehicle with a lower price and a lower range. In addition, they surveyed rural and suburban drivers at random regardless of income level. It was surprising to see that there is higher ownership of electric vehicles in rural areas than in urban or suburban areas, even if the commuting distance for rural areas is longer than in urban or suburban areas (Sovacool et al., 2019). Drivers from both areas believe the driving experience is important, but the charging accessibility in suburban areas is lower than in urban areas. Rural areas have fewer driving or charging restrictions (Sovacool et al., 2019). For low-medium income and high-income levels, our result indicates that an electric vehicle is preferred when the traveling distance is higher than the average commuting distance of the corresponding income level. In the modeling process, we assume that the maximum one-way commute distance is less than 45 miles - no extreme commuting distance. Under this situation, the study confirms that electric vehicles bring a more comfortable driving experience than non-electric vehicles (Sovacool et al., 2019). However, the data we used for the experiment has no data for charging accessibility. In the future, we will extend our research on charging accessibility.

For all income levels, Sovacoola, Kesterb, Noel, and Rubens (Sovacool et al., 2018) have found that the ownership of people under higher ages - such as Government officials and retirees - is under increasing trend for all income levels, even if there is little experience with the electric vehicle around them; they believed that elder people had short commuting distances and higher budgets. In addition, for large households, large households had more experience in electric vehicles than smaller households and were not willing to own an electric vehicle (Sovacool et al., 2018).

Our prediction model is able to predict the ownership of electric vehicles with high accuracy, but our study has limitations. The current data that records the ownership of electric vehicles is from the State of California, which does not represent everywhere. For example, the household income in the State of California does the minimum income level of $50,000, which does not represent the real case. Next, the State of California is regarded as a warm region (n.d., n.d.) - which can be one of the reasons for the high ownership of electric vehicles. In addition, our dataset does not include the region's local policies and infrastructure, which is one of the primary concerns (Mpoi et al., 2023) for electric - vehicle owners. The recent study from Mpoi et al. (Mpoi et al., 2023) studied the primary factors of owning an electric vehicle in Greece under similar weather conditions with California (n.d., n.d.) but low ownership of electric vehicles (Mpoi et al., 2023), primarily because of overall costs, and deficient charging station (Mpoi et al., 2023; Giansoldati et al., 2020), and the average income of Greece is lower than other European nations, such as Germany and UK (Mpoi et al., 2023). The study aims to find primary factors that affect the altitude of electric vehicles after the government's encouragement policies - such as tax exemptions and financial incentives (Mpoi et al., 2023). The statistic values showed that purchase cost is the primary concern, followed by the driving range, maintenance cost, and charging time, and the installation of charging stations every 10 to 15 kilometers is the first incentive that is able to increase the electric vehicle's ownership (Mpoi et al., 2023). Under similar weather conditions (n.d., n.d.), commuting distance is not the only factor that affects personal decisions. The accessibility charging facility and the car expense can be another two factors that affect the decisions (Mpoi et al., 2023).

The electric vehicle's battery cannot supply power during cold weather. Zhou et. al. (Zhou et al., 2017) studied an electric vehicle's battery performance with the temperature change. The study has found out that the battery capacity is 90% at 0°C; the capacity of an electric vehicle decreases exponentially when the external temperature drops below 0°C (Zhou et al., 2017), while the battery performs in 100% between 20 °C and 45 °C (Zhou et al., 2017). In our dataset, there is no data on the location of the household, and this is our limitation in considering the weather condition that affects the decision of electric vehicle ownership.

5. Conclusion

This work focused on identifying the primary features of choosing an electric vehicle by income level. We emphasize explainable machine learning methods, so we do not use complex methods such as Neural Networks in our study (Aggarwal, 2018). We considered a number of such methods and found that the Naive Bayesian gives the highest accuracy for each income level. Our results suggest that the one-way commuting distance is the most important factor for low-medium, medium-high, and high-income levels. Our study has limitations on the local infrastructure, culture, and weather conditions. In the future, we hope to extend our work to more extensive datasets on electric vehicles.

Declarations:

Acknowledgment: We would like to thank the editor in - chief and the anonymous reviewers for valuable feedback.

Ethical Approval: Not Applicable. No human and/ or animal studies were performed in connection with this work.

Competing Interests: The authors declare no conflict of interest. The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare relevant to this article's content. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Funding: No funds, grants, or other support was received.

Availability of Data and Materials: No new data were created during this study. The dataset used for analysis is the household data of electric vehicle ownership from the State of California. It is a publicly available dataset. Its url is:

`https://datadryad.org/stash/dataset/doi:10.25338/B8P313`

Data Availability Statement: The data that support the findings of this study are openly available in "Dryad" at

`https://doi.org/10.25338/B8P313`

.

Aggarwal, C.C., 2018. Neural networks and deep learning. Springer.

Alvarez, L., 2023. What is vehicle miles traveled (vmt)? Https://urbanlogiq.com/what-is-vehicle-miles-traveled-vmt/.

Axsen, J., Goldberg, S., Bailey, J., 2016. How might potential future plug-in electric vehicle buyers differ from current "pioneer" owners? Transp. Res. D: Transp. Environ. 47, 357–370. doi:10.1016/j.trd.2016.05.015.

Bishop, C., 2016. Pattern Recognition and Machine Learning. Springer.

Bonges, H., Lusk, A., 2016. Addressing electric vehicle (ev) sales and range anxiety through parking layout, policy and regulation. Transp. Res. A: Policy Pract.. 83, 63–73. doi:10.1016/j.tra.2015.09.011.

Carley, S., Krause, R., Lane, B., et al., 2013. Intent to purchase a plug-in electric vehicle: A survey of early impressions in large us cites. Transp. Res. D: Transp. Environ. 18, 39–45. doi:10.1016/j.trd.2012.09.007.

Cover, T., Hart, 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27. doi:10.1109/TIT.1967.1053964.

Giansoldati, M., Monte, A., Scorrano, M., 2020. Barriers to the adoption of electric cars: Evidence from an italian survey. Energy Policy 146. doi:https://doi.org/10.1016/j.enpol.2020.111812.

Hardman, S., 2019. Sociodemographic data for battery electric vehicle owning households in California (From NCST Project "Understanding the Early Adopters of Fuel Cell Vehicles"). Dataset ICPSR05404-v1. Dryad. California. Https://doi.org/10.25338/B8P313.

Hastle, T., 2018. Elements of Statistical Learning. Pearson.

Huber, P.J., 2009. Robust Statistics. J. Wiley, New York.

Johnson, N., Kotz, S., 1970. Distributions in Statistics. J. Wiley, New York.

Kroese, D., Botev, Z., Taimre, T., et al., 2019. Data Science and Machine Learning. Chapman and Hall CRC Publishing.

Majumder, P., 2021. Data analysis and price prediction of electric vehicles. Https://www.analyticsvidhya.com/blog/2021/09/data-analysis-and-price-prediction-of-electric-vehicles/.

Mpoi, G., Milioti, C., Mitropoulos, L., 2023. Factors and incentives that affect electric vehicle adoption in greece. Int. J. Transp. Sci. Technol. doi:https://doi.org/10.1016/j.ijtst.2023.01.002.

n.d., n.d. What place in europe has the same climate as southern california? Https://www.touristmaker.com/blog/what-place-in-europe-has-the-same-climate-as-southern-california/.

Ribeiro, M., 2016a. Lime - local interpretable model-agnostic explanations. Https://homes.cs.washington.edu/ marcotcr/blog/lime/.

Ribeiro, M., 2016b. Local interpretable machine learning (lime). Https://lime-ml.readthedocs.io/en/latest/.

Science, O.O.D., 2019. Cracking the box: Interpreting black box machine learning models. Https://odsc.medium.com/cracking-the-box-interpreting-black-box-machine-learning-models-bc4bdb2b1ed2.

Sovacool, B.K., Kester, J., Noel, L., et al., 2018. The demographics of decarbonizing transport: The influence of gender, education, occupation, age, and household size on electric mobility preferences in the nordic region. Glob. Environ. Change 52, 86–100. URL: https://www.sciencedirect.com/science/article/pii/S095937801830030X, doi:https://doi.org/10.1016/j.gloenvcha.2018.06.008.

Sovacool, B.K., Kester, J., Noel, L., et al., 2019. Income, political affiliation, urbanism and geography in stated preferences for electric vehicles (evs) and vehicle-to-grid (v2g) technologies in northern europe. J. Transp. Geogr. 78, 214–229. URL: https://www.sciencedirect.com/science/article/pii/S0966692318302928, doi:https://doi.org/10.1016/j.jtrangeo.2019.06.006.

Upton, G., Cook, I., 1996. Understanding Statistics. Oxford University Press, Oxford.

Zhou, C., Guo, Y., Huang, W., et al., 2017. Research on heat dissipation of electric vehicle based on safety architecture optimization. Journal of Physics: Conf. 916. doi:10.1088/1742-6596/916/1/012036.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: