Unveiling the Heterogeneity of Vehicle Purchasing Choices among Car-owning Households: A Comprehensive Analysis Using Machine Learning and Logit Models

Lingyun Zhong[a], Taewhan Ko[a], Meiting Tu[b,*], Dominique Gruyer[c], Tongtong Shi[b]

[a] *Department of Civil and Environmental Engineering, Univerisity of Michigan, 2350 Hayward St, Ann Arbor, 48109, Michigan, United States*
[b] *College of Transportation Engineering, Tongji University, The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 4800 Cao'an road, 201804 Shanghai, PR China*
[c] *IFSTTAR, University Gustave Eiffel, 77420 Champs sur Marne, France*

Abstract

Many nations have set goals to increase electric vehicle (EV) sales and even surpass that of internal combustion engine vehicles (ICEV). When promoting EVs in the market, vehicle purchase behavior analysis is highly important, which requires careful analysis of consumer heterogeneity. In this study, we use the Panel Study of Income Dynamics dataset to study new car purchasing behaviors of car-owning households. First, We use a hybrid sampling method combined with Kmeans-undersampling and SMOTE to alleviate class imbalance. Then, we use a multinomial logit model to gain a general idea of the population's tendencies. Finally, we used LightGBM and Tree Explainer to add a more detailed behavioral analysis. The results show that households with higher income, older vehicles, married couples, younger members, higher transportation expenditures, and EV loyalty are more inclined to buy EVs. In general, this study provides a new perspective on examining the heterogeneity of vehicle purchase decisions by car-owning households. Using the Logit model and SHAP, the interaction effect of variables across different demographics is explored, providing more detailed insights into consumer behaviors to help improve the penetration rate of EVs.

Keywords: Vehicles purchase, Consumer heterogeneity, Data Resampling, Logit Model, TreeExplainer

---

*Corresponding author

1. Introduction

In recent years, electric vehicles (EVs) have gained great interest due to their significant impact on improving environmental sustainability. Many nations across the world have implemented comprehensive policies to incentivize the widespread adoption of EVs (Qadir et al. (2024)). However, despite these policies, the current EV market penetration remains below expected levels (Jia (2019)). This shortage can be attributed, in part, to the innate vehicle purchase behavior of consumers.

Within the United States, over 90 percent of households already own one or more vehicles (Molloy et al. (0)). According to the research conducted by Smart and Klein (2015), for most families, being carless is only a temporary situation, only 5% do not own a car across all seven waves of data examined in their analysis. As such, to grow the EV penetration rate, analyzing the vehicle purchase behavior of car-owning households is important. However, many existing studies work on a national macro scale when studying car purchases (Yang et al. (2017), Le Vine and Polak (2019)). Although many researchers have recently begun to realize that vehicle purchase decision is a complex behavior on the household level, they still fail to focus on the car-owning households (Blumenberg et al. (2020)). In addition, this type of studies often focus on one specific time point with cross-sectional data and does not consider changes in car purchase decisions from the time dimension (Le Vine et al. (2018), de Jong and Kitamura (2009), Klein and Smart (2019), Oakil et al. (2016)). When modeling and analyzing traffic and consumer behavior logit models are often used (Anderson et al. (1988), Wiginton (1980)). Despite their relatively simple form, logit models have gained popularity among researchers due to their excellent interpretability. Yet, they have weak fitting capabilities; in contrast, machine learning models have strong fitting capabilities, but very weak in interpretability.

In this study, we use the Panel Study of Income Dynamics (PSID) data spanning up to 10 years (2011-2021) (Pan (2021)) to fill the above research gaps. Our study aims to construct a comprehensive user profile for the vehicle market by exploring the heterogeneous factors that influence people's future vehicle purchase decisions, which can help refine future strategies to promote EV adoption.

The main contributions of this work are as follows:

- We utilize the PSID dataset spanning from 2011 to 2021, to analyze the novel topic of heterogeneity in new car purchasing decisions among car-owning households.
- We propose the Hybrid data resampling method which combines Kmeans-undersampling with SMOTE-oversampling to address the serious data imbalance in our original dataset
- The interpretable machine learning tool Tree Explainer is used to conduct a more comprehensive SHAP Value-based micro-behavior analysis of the respondents. In contrast, the multinomial logit(MNL) model performs general disaggregate behavioral explanations to verify basic user trends.

The remainder of this paper is organized as follows. We first review the related literature in Section 2. We then provide descriptions for the data and variables of our model in Section 3. Next, we describe the logit models and TreeExplainer in detail in Section 4. We follow this with the model results in Section 5. Finally, we discuss our conclusions and limitations in Section 6 and 7.

2. Related Work

2.1. Vehicle Purchase Behavior Analysis

The vehicle purchase decision is an extremely complex decision-making process due to the many influencing factors. According to past studies, it can be affected by vehicle attributes and demographic attributes (Shende (2014)). The overall objective here was to examine the vehicle purchase behavior of a population within a specific time cross-section and analyze the significant factors influencing decision-making behavior. For demographic attributes, total household income is one of the most important attributes for vehicle purchase (Dargay (2001)). In general, **households with higher incomes** are more willing to replace their vehicles. Also, many studies (Sharma (2015), Vrkljan and Anaby (2011)) have found that people of different age groups and genders show great differences in vehicle purchase behavior. At the same time, Bhardwaj and Bishnoi (2023) found that consumers' education level and employee attributes have a significant impact on car purchasing behavior, **employees** and **people with a high level of education** are more willing to buy cars. Examining from the family level, Monga et al. (2012), Peters et al. (2015) found that family demands play an important role in car purchase decisions, such as children in households, and marital status, which indicate that larger families, families with multiple children, and married families are more likely to purchase cars. As for vehicle attributes, the fuel type of the vehicle (Sanitthangkul et al. (2012)) affects the consumers' attitudes toward vehicle selection and final decision. At the same time, Hensher (2013) has identified the effect of mileage on consumer decisions, experiments show that **longer vehicle mileage** will make drivers more inclined to change vehicles. Absent the direct data on mileage, vehicle age can be used as a surrogate for it.

Although vehicle purchase decision-making has always been an area of focus for researchers, there is no perfect comprehensive dataset. Part of this is due to privacy. Most respondents are unwilling to disclose too much sensitive data at the household level (Muti and Yıldız (2023)). On the other hand, the sample size and consistency also cause concerns. The primary obstacle lies in the need to observe a single family over an extended period of time, resulting in a lack of data sources. For instance, in 2014, Zhang et al. (2014) was able to conduct a detailed life history survey in Japan, but only on 1,000 households. The findings revealed that changes in family employment and education significantly impact the family's vehicle purchase decisions. However, experiments only on such a small dataset make the final results less convincing. Therefore, large-scale datasets sampled on the national level can be a good solution for vehicle purchase, such as the Panel Study of Income Dynamics (PSID) (Li (2024)), the China Household Finance Survey (CHFS) (Li (2023)), etc.

2.2. Data Augment Methods

Imbalanced data distribution is quite common in traffic scenarios. For instance, the most severe traffic accidents often represent only a small fraction of the overall accident data (Parsa et al. (2019)). Additionally, data imbalance frequently occurs in behavioral analysis fields, such as among different travel behaviors (Chen and Cheng (2023)). Such data imbalance causes the model to favor the characteristics of majority class samples while ignoring minority samples during classification, which is detrimental to behavior analysis. To address this issue, the basic idea of the imbalanced data processing method is to change the sample distribution of the original data set, reducing or eliminating the imbalance.

Undersampling reduces the imbalance of the data set by deleting old majority-class samples from the original dataset. Random undersampling (Mishra (2017)) is a representative type of undersampling algorithm. It primarily achieves data balance by randomly selecting and deleting samples from the majority class. Pozo et al. (2021) combined random undersampling and decision tree model to identify the service level of parking areas in Spain. However, randomly deleting samples can change the distribution of the original data, leading to poor model performance. Therefore, Lin et al. (2017) consider combining random undersampling with clustering algorithms such as k-means to form several majority class sample clusters. By sampling majorities within each cluster, the representativeness of the retained data points is improved without changing the distribution of the original data. Based on this idea, Zheng et al. (2021) proposed a method for selecting representative samples, which effectively improved the accuracy of accident data classification.

Oversampling reduces the imbalance of the data set by adding new minority-class samples. The SMOTE algorithm proposed by Chawla et al. (2002) is the most representative oversampling method, which generates new minority class samples between minority class samples through linear interpolation. As a representative algorithm of oversampling, researchers have proposed a large number of variants based on SMOTE, such as Borderline-SMOTE (Han et al. (2005)), Kmeans- SMOTE (Xu et al. (2021)), etc. Wei and Pan (2021) used SMOTE to oversample data points of EV purchase intention to improve the performance of LightGBM model. Similarly, Jia (2019) used SMOTE to oversample the data of alternative fuel vehicle users based on the 2017 National Household Travel Survey (NHTS) data and effectively improved the prediction accuracy of the Random Forest model.

Regarding the issue of imbalanced data for multiple behaviors, few studies mention sampling methods. Based on the literature, combining two kinds of basic sampling methods could be a feasible approach.

2.3. Behavioral Modeling Methods

The key to analyzing consumer behavior is choosing an appropriate model to capture the heterogeneity of consumers. The most classic and widely used one is undoubtedly the logit model (Anderson et al. (1988), Wiginton (1980)), which owes its popularity to its simple mathematical form and strong interpretability. The logit model is based on the stochastic utility maximization theory (Anas (1983)). The coefficients of the fitted model can be well explained as changes in odd ratios. The logit model was employed in the analysis of vehicle purchase behavior as early as 1998. McCarthy and Tay (1998) utilized a nested logit model to characterize consumers' propensities towards purchasing energy-saving cars. Many subsequent researchers continued along similar lines. For instance, Ling et al. (2021) delved into the influence of vehicle fuel attributes on vehicle purchase decisions using survey data collected in Beijing. Cirillo et al. (2017) employed nine years of survey data to observe the dynamics of vehicle and fuel prices, investigating how their changes and demographic attributes impact consumer behavior.

However, most of the above studies rely on analyzing the coefficients of the logit model. These analyses tend to focus on the group characteristics of the entire sample, without exploring the impact of the individual sample and individual characteristics on decision-making. Meanwhile, we often see large feature spaces with non-linear features in the field of traffic surveys (Ding et al. (2021)), which the logit model has trouble dealing with. This makes researchers eager to introduce a more efficient data-driven machine learning model for behavior analysis. For example, Bas et al. (2021) utilized a large number of machine learning models such as support vector machines, random forests, gradient boosting trees, and deep neural networks. However, the conclusive analysis consistently lacks a detailed exploration of the user behavior of specific groups. This final analysis is constrained by the inherent challenge of interpretability in machine learning. The improvement in classification accuracy brought by machine learning is obvious, but its shortcomings cannot be ignored. This loss of interpretability makes its analysis of consumer behavior inferior to the logit model, which is unacceptable to transportation scientists.

In the pursuit of enhancing the interpretability of machine learning, a growing number of researchers are actively exploring interpretability tools to unravel the intricacies of complex models, aiming for more exhaustive and reliable results. For instance, the SHAP (SHapley Additive exPlanations) tool (Lundberg and Lee (2017)) emerged as a robust solution. Rooted in game theory's Shapley values, SHAP excels in providing interpretability not only for the overarching global model insights but also for the localized interpretations specific to individual samples. This dual interpretive capacity holds considerable promise, especially when applied to the nuanced analysis of consumer behavior.

To that end, the TreeExplainer has been seeing use to study consumer behavior on a relatively micro level. Ahmed and Roorda (2022) employed both random forest and the logit model to analyze commercial vehicle purchase decisions on an enterprise level. Their findings demonstrated the superior predictive accuracy of the random forest models compared to a basic logit model. However, in terms of interpretability, the analysis lacked feature importance rankings and did not leverage the full potential of the SHAP value, which can provide detailed local explanations about feature interactions. Currently, a comprehensive study that serves as a noteworthy reference is research conducted by Jin et al. (2022) on vehicle disposal behavior within households. This research is based on the public PSID dataset (Pan (2021)), effectively integrating vehicle attributes and demographic data in the cross-sectional dimension, along with changes in household characteristics over time. This study is expected to provide valuable insights and inspiration for our research endeavors.

3. Data

3.1. Data Source

To investigate the new vehicle purchase decisions of car-owning families, we require a dataset that satisfies two different dimensions. First, in the cross-sectional dimension, we need information on vehicle attributes and demographic details. Secondly, in the longitudinal dimension, we need data for individual households over time, specifically capturing changes in household attributes to explore the driving forces of vehicle purchase decisions. Therefore, we have chosen the Panel Study of Income Dynamics (PSID) data (Pan (2021)) as our primary data source. The PSID is a household survey directed by faculty from the University of Michigan. The survey began in 1968 and has public data available for at least every 2 years from the start to 2021. In recent years, the survey began adding questions about hybrid vehicle (EV) or EV ownership to the household vehicle questionnaire. More specifically, these questions were introduced from 2011. Hence, we take the survey data, which is available every 2 years, for 2011, 2013, 2015, 2017, 2019, and 2021. Throughout those years, the survey received 8203, 8355, 8378, 8793, 8783, and 8029 responses, respectively. Of those responses, 6403, 6401, 6433, 6869, 6944, and 5889 responses were from households that already owned a vehicle. While the survey responses do differentiate HV and EV, due to the extremely low number of both, we consolidate all variables related to HV and EV into one and represent them as EV variables. Table 1 depicts the vehicle purchase decisions over the years.

Table 1: Vehicle Purchase Decisions per Year

Year	ICEV Purchase	EV Purchase	No Purchase	Total
2011	693	15	5695	6403
2013	691	37	5673	6401
2015	833	41	5559	6433
2017	866	38	5965	6869
2019	769	48	6127	6944
2021	707	71	5111	6889

Even after combining the EV and HV numbers together, the numbers for their purchase are quite low. Therefore, in order to achieve statistical significance, we aggregated the data over the years into one dataset.

3.2. Description of Explanatory Variables

From the PSID dataset, we select 8 variables. The variable descriptions are detailed below. Table 2 shows the full list of variables along with their mean values for each decision.

- **Number of Children**: The number of minors in the household.

- **Age**: Age of the respondent taken as the reference age of the household.
- **Completed Education (year)**: Education level in years (0-17) of the respondent taken as the reference education level of the household.
- **Marital Status**: The marital status of the household. To simplify the data, we categorize divorced and widowed as unmarried.
- **Annual Income($) (log)**: Total annual income of the household. We assume that excessively high income does not significantly increase the household's purchase behavior. Therefore, we take the log values.
- **EV Ownership**: Whether the household have previously purchased electric vehicles.
- **Number of Owned Vehicles**: Total number of vehicles owned by the household prior to the survey year.
- **Expenditure Ratio (%)**: The proportion of travel costs incurred by non-private car travel to the annual income of the household.

Table 2: Descriptive Statistics of Variables

Variable	Population Mean	ICEV Purchase	EV Purchase	No Purchase
Number of Children	0.79	1.16	1.24	0.74
Age	46.60	43.14	42.56	47.09
Completed Education (year)	13.73	13.45	13.57	13.77
Marital Status	0.53	0.67	0.68	0.50
Annual Income ($)(log)	10.95	11.21	11.33	10.91
Last Vehicle Age (year)	7.61	8.61	8.71	7.47
EV Ownership	0.06	0.03	0.32	0.06
Number of Owned Vehicles	1.79	1.64	1.58	1.81
Expenditure Ratio (%)	21.94	24.95	23.95	21.52

All variables that refer to a specific individual (age, education years, marital status) refers to the reference person (i.e., respondent). All other variables refer to the family unit. The annual income variable was taken as the natural log value of the original to better deal with outliers. Marital status and EV ownership are binary variables that indicate whether they are married and whether they previously owned an EV. Last vehicle age refers to the age of the latest vehicle that the family purchased. Both EV ownership and number of owned vehicles do not include the possible vehicle purchase in the survey year. The expenditure ratio refers to the percentage of annual transportation expenditure over the total annual expenditure. Figure 1 illustrates the distribution of the non-binary variables.

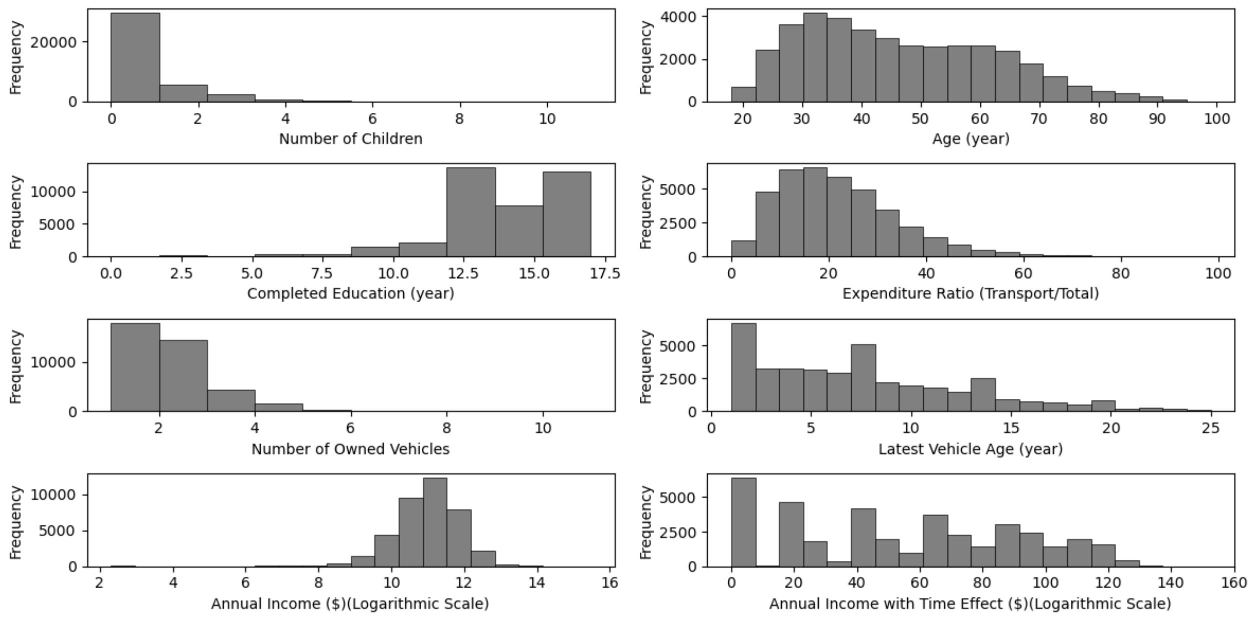


Figure 1: Variable Distributions

Furthermore, when we aggregate the data from different years, we add in a time-effect for the annual income variable to capture the potential effect of time. Based on the concepts provided by Liu Liu et al. (2023), we formulate the new variable.

$$a_n = Annual\ Income_n * I_{time} \tag{1}$$

Where a_n is the new variable for household n and the time effect I_{time} is 0, 2, 4, 6, 8, and 10 for 2011, 2013, 2015, 2017, 2019, and 2021, respectively.

3.3. Hybrid Data Resampling

As shown above, there is a great imbalance in the number of data points for each category. More specifically, most samples belong to the No Purchase decision, while very few belong to the EV Purchase section. Such an imbalance hampers our ability to properly run models on this data. To deal with this, we undersample the No Purchase group and oversample the EV Purchase group with SMOTE (Chawla et al. (2002)). In more detail, we use the KMeans-undersampling method (Kumar et al. (2014)). This approach forms multiple clusters of No Vehicle Purchase behaviors and then performs random sampling within each cluster. This ensures that the majority class, after sampling, retains most of its important information. Our proposed hybrid data resampling methods are shown in Figure 2

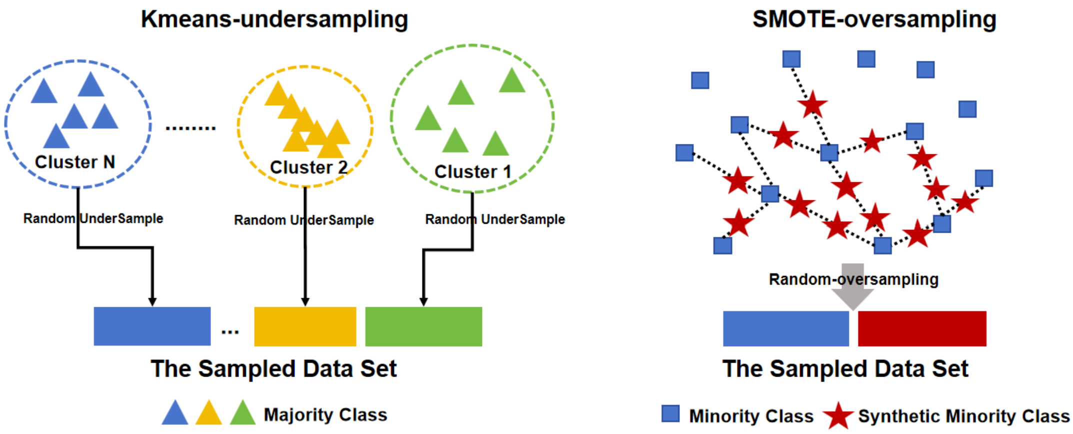


Figure 2: Hybrid Data Resampling Methods (Kmeans-undersampling and SMOTE)

First, we use K-means clustering to divide the No Purchase group into 6 clusters. Then, data points are evenly sampled from each cluster until the total number of data points matches that of ICEV purchase. Afterwards, the SMOTE method is used to create synthetic samples of the EV purchase group to match the ICEV Purchase numbers. In short, we end up with an even number of samples for all three decision groups. Figure 3 depicts a summary of how the data resampling is done.

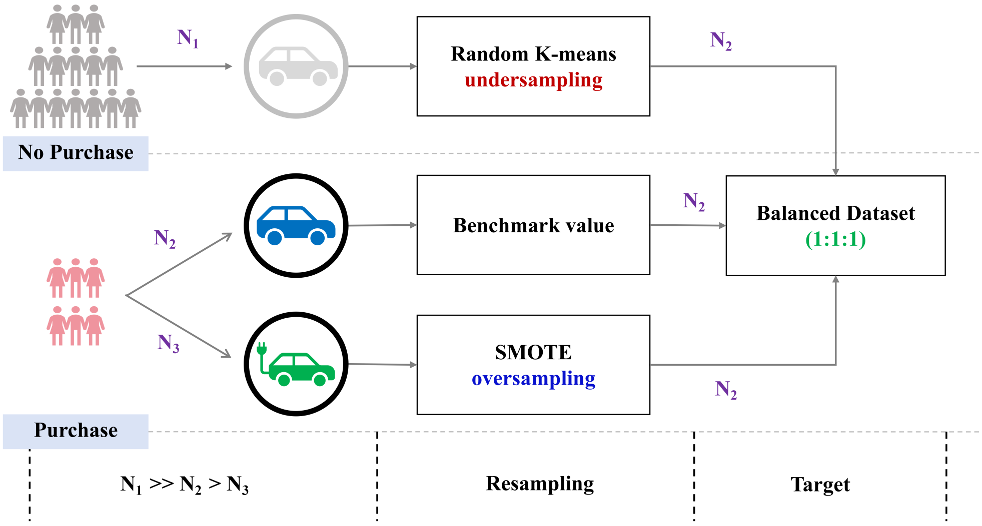


Figure 3: Data Resampling Process

4. Methodology

In this section, we present our model framework, which consists of three components: Data processing; Behavior model construction; and Behavior heterogeneity analysis.

- To process the data sample, we selected socio-demographic data from 2011 to 2021 with the corresponding vehicle attribute information from the PSID dataset. At the same time, we use a hybrid data resampling method combining Kmeans-undersampling and SMOTE to alleviate the class imbalance in the dataset.

- To build a consumer behavior analysis model, we use the multinomial Logit model and LightGBM to fit the sample data separately.

- To analyze the behavior heterogeneity, we analyze the macro trends and heterogeneity of the data set based on the logit model coefficients and then evaluate each data point and feature in the sample based on the Tree Explainer and SHAP value indicators to provide a more detailed consumer behavior portrait.

Figure 4 illustrates the entire process. We explain the three components in detail as follows.

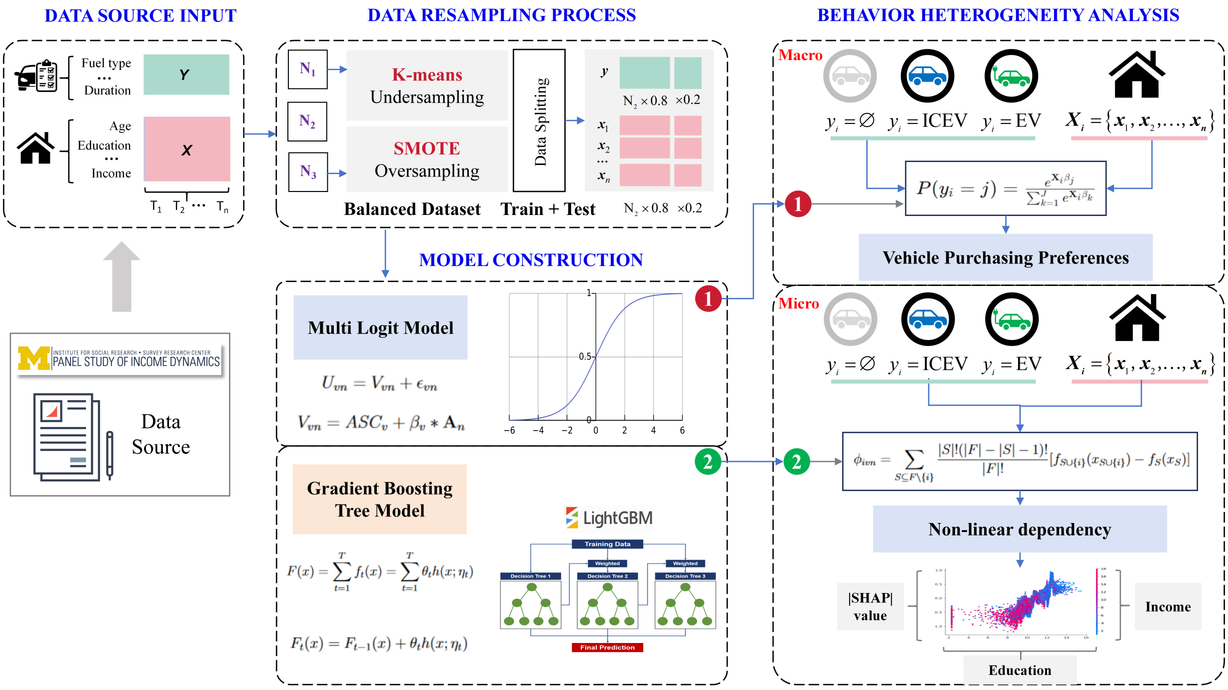


Figure 4: Model Framework

4.1. Multinomial Logit Model

We construct a multinomial logit (MNL) model of the vehicle purchase decision based on Ahmed and Roorda (2022). The MNL model is based on the utility maximization approach where the utility function U for the vehicle purchase choice v of household n is defined as follows:

$$U_{vn} = V_{vn} + \epsilon_{vn} \tag{2}$$

Where V_{vn} is the systematic component and ϵ_{vn} is the unobserved component of utility.

We further formulate V_{vn} as follows.

$$V_{vn} = ASC_v + \beta_v * \mathbf{A}_n \tag{3}$$

Where ASC_v and β_v are the alternative specific constant and the set of coefficients for each variable for choice v, and $\mathbf{A}_n$ is the set of variable values for household n.

The unobserved component is assumed to be extreme value (Type I) distributed independently and identically across alternatives v, household n. Then, the probability of decision v being chosen by household n becomes:

$$P_{vn} = \frac{e^{V_{vn}}}{\sum_{k \in K} e^{V_{vn}}} \tag{4}$$

Where K is the set of vehicle purchase choices (No purchase = 0, ICEV purchase = 1, EV purchase = 2).

The log-likelihood function becomes:

$$log(L(\beta)) = \sum_{n \in N} \sum_{k \in K} y_{vn} log(P_{vn}) \tag{5}$$

where $y_{vn} = 1$, if household n makes vehicle purchasing decision v and zero otherwise.

4.2. LightGBM Model

LightGBM model was introduced by Ke et al. (2017). It is a type of gradient boost decision tree model (GBDT) (Friedman (2001)), making it an ensemble algorithm. Ensemble-based algorithms create several classifiers (mostly decision trees) and combine the outputs to reduce error. As for Boosting, there is a correlation between the various base classifiers. During training, each base classifier gives a higher weight to the samples that were misclassified by the previous base classifier. The final result is obtained based on the weighting of the results of each of the classifiers.

We directly adopt the GBDT model from Peng et al. (2023). The objective of GBDT is to minimize the loss function by setting an approximation function as a linear combination of addictive decision trees. The loss function is adopted as Log Loss : $L(y, F(x)) = -\sum_{i=1}^{N} y_i \cdot \log(\frac{e^{F(x)}}{\sum_{l=1}^{N} e^{F_i(x)}})$ in this study. And $F(x)$ will be given by the following:

$$F(x) = \sum_{t=1}^{T} f_t(x) = \sum_{t=1}^{T} \theta_t h(x; \eta_t) \tag{6}$$

where T is the number of trees, η_t is the set of parameters for the t-th tree $h(x;\eta_t)$; θ_t is the weight of $h(x;\eta_t)$ and can be estimated by minimizing the loss function. The training framework will follow several steps:

Step 1: Initialize the model with a constant value:

$$F_0(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i;\theta) \tag{7}$$

where N is the number of instances.

Step 2: Compute so-called pseudo-residuals, which is calculated for each data sample i in iteration round t:

$$r_{t,i} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{t-1}(x)} \tag{8}$$

Step 3: (x_i, r_t, i),$i = 1, 2, ..., N$ is used to fit the t th $(t = 1, 2, ..., T)$ decision tree $h(x;\eta_t)$ and get the terminal region $R_{t,j}$,$(j = 1, 2, ..., J_t)$, where J_t is the size of the tree. Then, compute the multiplier θ_t by solving the following one-dimensional optimization problem:

$$\theta_t = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i; F_{t-1}(x) + \theta h(x;\eta_t)) \tag{9}$$

Step 4: Update the model:

$$F_t(x) = F_{t-1}(x) + \theta_t h(x;\eta_t) \tag{10}$$

LightGBM is one of the most efficient methods in ensemble-based algorithms, with higher prediction accuracy, faster training speed, and more efficient processing of massive data. Therefore, this study chose to participate it in behavioral analysis.

4.3. SHAP (SHapley Additive exPlanations)

We supplement our analysis with an interpretable machine learning method. Traditional behavior analysis most often directly assesses the influence of the variables on the final choice probability. However, we would like to know the local explanations of how each variable contributes to each specific choices. As such, we adopt the TreeExplainer to interpret our LightGBM results.

The TreeExplainer was introduced in 2020 (Lundberg et al. (2020)) and has since greatly helped in the field of behavioral analysis. Diverging from the global interpretability feature importance ranking in traditional tree models, the TreeExplainer introduces a novel interpretability tool based on SHAP values, offering both global and local explanations. This innovative approach allows for a more comprehensive interpretability of the model. The SHAP value is calculated as follows. It represents the sequential impact on the model's output of observing each input feature averaged over all possible subset variable orderings (Jin et al. (2022)):

$$\phi_{ivn} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{11}$$

Where ϕ_{ivn} is the SHAP value of i-th variable of outcome decision v for household n; F represents the set of all features; $S \subseteq F$ represents a subset of features; $i \in F$ represents a single features; f_S and $f_{S \cup \{i\}}$ represent models trained on feature set S and $S \cup \{i\}$, respectively; x_S and $x_{S \cup \{i\}}$ represent the values of the features in the set S and $S \cup \{i\}$, respectively.

5. Result

5.1. Multinomial Logit Model

Table 3 represents the results from our MNL model. We take the No Purchase option as our reference. In our experiment, we will use Statsmodel in Python (Seabold and Perktold (2010)) as a tool to fit the logit model.

Table 3: Multinomial Logit Model (Reference: No vehicle purchase)

	ICEV Purchase					
Variable	Coefficient	Std Error	z	P> $\|z\|$	[0.025	0.975]
Constants	-11.2662	0.320	-35.261	0.000	-11.892	-10.640
Age	-0.0129	0.001	-10.117	0.000	-0.015	-0.010
Number of Children	0.1622	0.014	11.711	0.000	0.135	0.189
Annual Income	0.8387	0.028	29.431	0.000	0.783	0.895
Completed Education	-0.0655	0.007	-9.030	0.000	-0.080	-0.051
Marital Status	0.8619	0.041	20.786	0.000	0.781	0.943
Expenditure Ratio	0.0419	0.001	29.639	0.000	0.039	0.045
Last Vehicle Age	0.0996	0.004	27.722	0.000	0.093	0.107
EV Ownership	-0.5316	0.090	-5.876	0.000	-0.709	-0.354
Number of Owned Vehicles	-0.6182	0.025	-25.148	0.000	-0.666	-0.570
Annual Income * Time Effect	0.0016	0.000	3.646	0.000	0.001	0.003
	EV Purchase					
Variable	Coefficient	Std Error	z	P> $\|z\|$	[0.025	0.975]
Constants	-16.9801	1.205	-14.096	0.000	-19.341	-14.619
Age	-0.0152	0.005	-2.916	0.004	-0.025	-0.005
Number of Children	0.2149	0.051	4.211	0.000	0.115	0.315
Annual Income	1.0325	0.107	9.616	0.000	0.822	1.243
Completed Education	-0.0708	0.028	-2.565	0.010	-0.125	-0.017
Marital Status	0.8803	0.161	5.465	0.000	0.565	1.196
Expenditure Ratio	0.0446	0.005	8.201	0.000	0.034	0.055
Last Vehicle Age	0.1300	0.013	9.666	0.000	0.104	0.156
EV Ownership	2.1109	0.146	14.459	0.000	1.825	2.397
Number of Owned Vehicles	-0.8245	0.102	-8.098	0.000	-1.024	-0.625
Annual Income * Time Effect	0.0074	0.002	4.065	0.000	0.004	0.011

Compared to the reference category (i.e., No Vehicle Purchase), we find that certain **family demographic** attributes make families more inclined to purchase a vehicle, which includes being young households, having many children, being high-income, having low to medium education levels, being married, and having high travel expenses. Regarding **vehicle attributes,** households that use old cars or have no or few cars are also more likely to buy a car.

Compared with two different purchasing behaviors, we find that households with many children, high incomes, old cars, and few or no cars are more likely to buy EVs rather than ICEVs. Additionally, households who are loyal users of electric cars are more likely to choose electric vehicles when replacing their current ones, which, to some extent, illustrates the path dependence of consumers when buying cars; loyal users of ICEVs and EVs do not easily change their consumption preferences. Furthermore, the willingness of high-income individuals to purchase electric cars has increased in recent years according to the variable **Annual Income*Time Effect**.

5.2. Performance of Behavioral Models with Hybrid Data Resampling

we compare the Logistic Regression, Decision Tree, Naive Bayes, and LightGBM models in terms of Accuracy, Precision, Recall, and F1-score. The results indicate that the hybrid data resampling method effectively improves the classification performance in the presence of class imbalance. Additionally, the LightGBM model demonstrates strong classification performance according to the following Figure 4.

Table 4: Performance Measures Comparison between Different Models

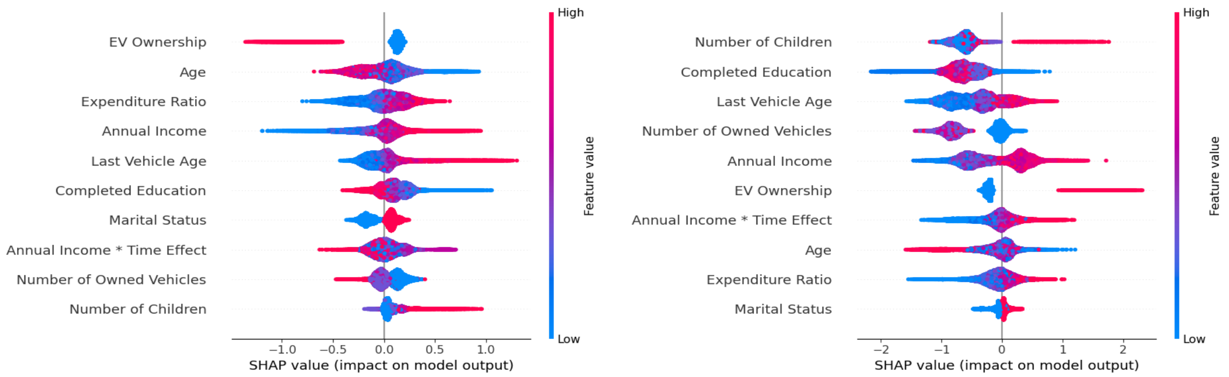
Models	**Accuracy**	**Precision**	**Recall**	**F1**
Logistic Regression (with data resampling)	52.7%	53.5%	52.6%	52.8%
Decision Tree (with data resampling)	70.8%	70.7%	70.5%	70.7%
Naïve Bayes (with data resampling)	53.1%	57.9%	53.1%	53.0%
LightGBM (no data resampling)	88.9%	53.1%	39.5%	41.5%
LightGBM (with data resampling)	80.2%	80.5%	78.9%	80.2%

5.3. Tree Explainer

While the MNL model does provide us with some valuable insights as above, it is limited in giving us a more detailed segmentation of the consumer groups. In other words, its coefficients only explain the direct impact of the variables on general consumer behavior.

Therefore, we have introduced the LightGBM model along with the Tree Explainer (Lundberg et al. (2020)), a machine learning interpretation tool to address these limitations. This tool allows us to conduct quantitative analysis on the impact of each variable by incorporating the SHAP values. Its advantage over the MNL model lies in its ability to provide an analytical function for each data point in the dataset, representing each consumer. With this, we can look at the consumer behavior on a more micro level.

In Figure 5, we show the SHAP summary plots for 2 kinds of vehicle purchase behaviors. The plots show the global impact of each independent variable on household behavior. Most of the results are highly consistent with the MNL model. The importance of variables in the summary plots is sorted vertically from large to small. We can find that, among which high-income families have a stronger tendency to buy cars. And we can see that whether or not a household has purchased an EV in the past (**EV Ownership**) has a very strong impact on the subsequent decisions. This again reveals that there is evident path dependency in choosing to buy an EV.



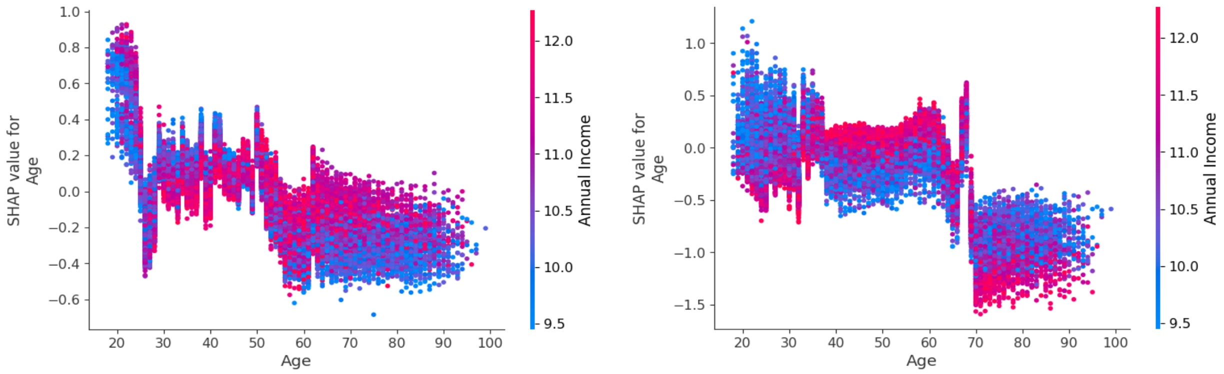
(a) ICEV purchase

(b) EV purchase

Figure 5: SHAP Value Importance

The Tree Explainer additionally provides an evaluation tool for the interaction SHAP value between two variables. Leveraging this feature, we can get further insights into consumer behavior.

Figure 6 illustrates the interaction SHAP values between age and income. Recall in the MNL model, higher age meant less likely to purchase a vehicle. In the SHAP model, we can find that consumers in different age groups show completely different distributions. The crowd can be roughly divided into groups of age 0-25, 25-45, 45-65, and above 65. The younger crowd is more aggressive and has a stronger willingness to purchase a new vehicle; the crowd over 65 years old is more conservative and has a weaker willingness to buy a new vehicle. When it comes to the high-income crowd, the 25-65 crowd is more inclined to get an EV, while the older group prefers ICEVs.

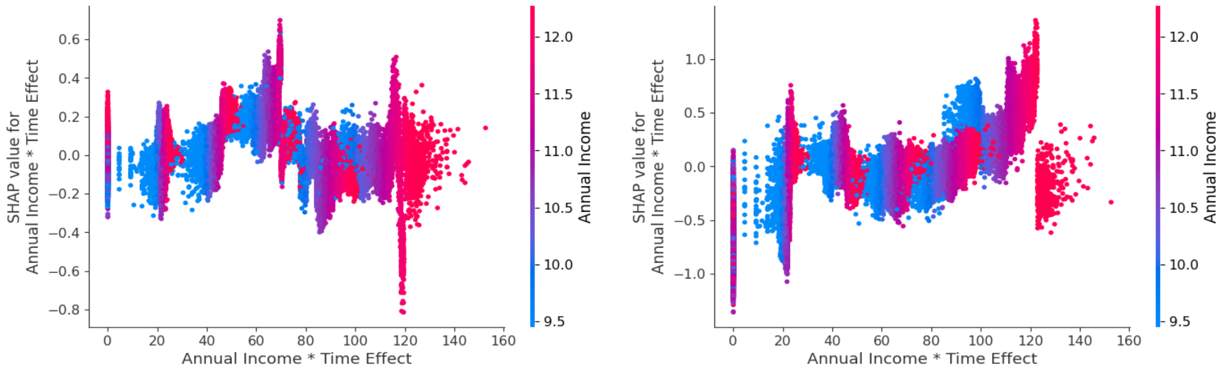


(a) Interaction SHAP Value (ICEV)

(b) Interaction SHAP Value (EV)

Figure 6: Interaction SHAP Value (Age v.s. Family Income)

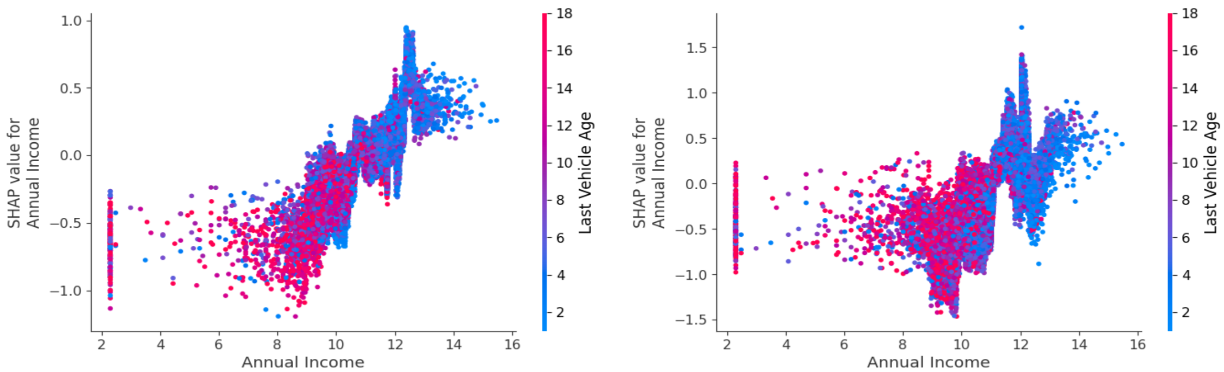
Figure 7 explores the time effect for the household income. As mentioned before, the new variable **Annual Income * Time Effect** captures the effect of time over income. We can see that as time goes by, the tendency to buy an EV gradually increases, especially in recent years (i.e. 2021). While ICEV purchase tendency grew from 2011 to 2015, the SHAP value for ICEV purchase returned to 0 as EVs became more prominent, indicating a decline in tendency. Also, high-income households are significantly more likely to purchase EVs over ICEVs.



(a) Interaction SHAP Value (ICEV) (b) Interaction SHAP Value (EV)

Figure 7: Interaction SHAP Value (Time Effect of Family Income)

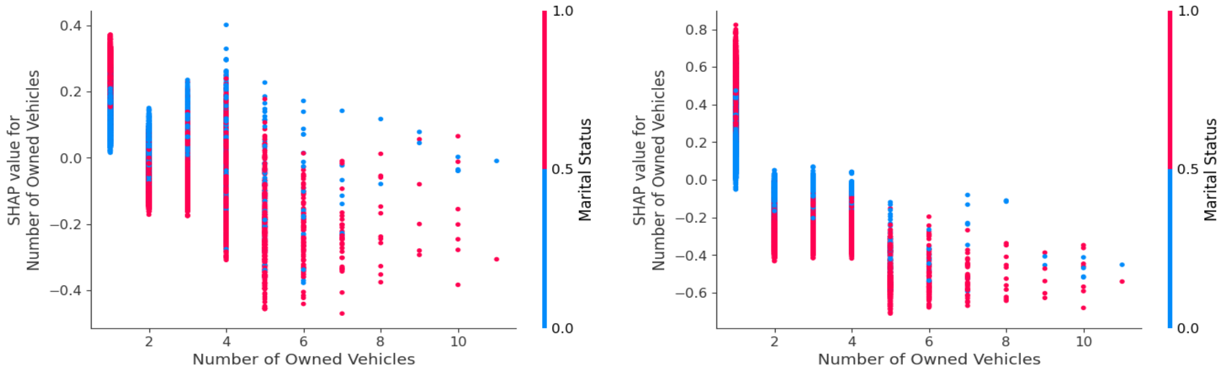
Figure 8 shows the interaction between income and vehicle age. The trend is generally in line with our intuition. The family is more likely to buy a new vehicle as the total household income increases. At the same time, households can be divided into above and below the average income. The above-average group is more likely to buy a vehicle if their latest vehicle is old. On the other hand, the below-average group is actually less likely to buy a vehicle the older their previous vehicle is. We attribute this phenomenon to affordability, the lower income households are stuck with their old cars. At the same time, when it comes to EV purchases, even many of the lower income households were willing to get a new one over their old car.



(a) Interaction SHAP Value (ICEV) (b) Interaction SHAP Value (EV)

Figure 8: Interaction SHAP Value (Annual Income v.s. Vehicle Age)

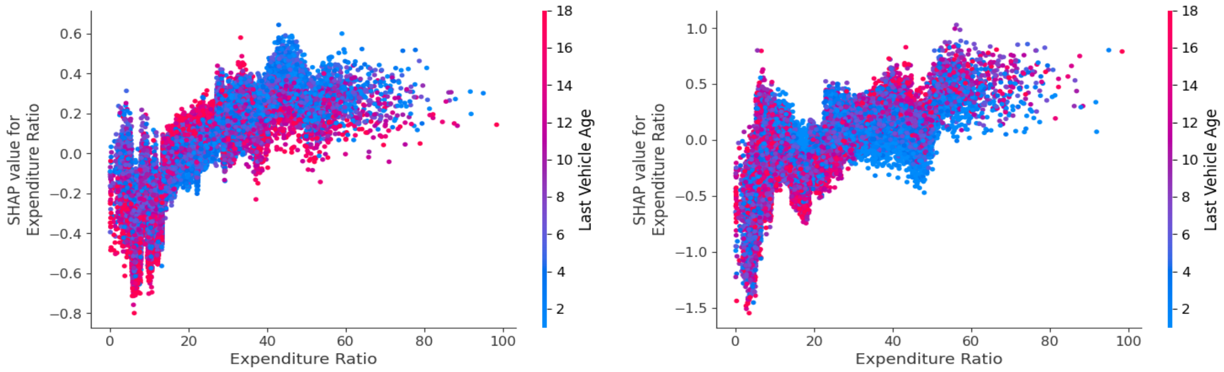
We then look at the relationship between the number of owned vehicles and the marital status with Figure 9. It is worth noting that the number of owned vehicles does not count the new vehicle if a new purchase is made, even if there were two purchases in the same year. if the household only had a single vehicle, married couples are more likely to get another car than unmarried individuals. Interestingly, when it comes to getting three or more cars, unmarried households show a higher likelihood than married households. We conclude that this is because 2 cars are usually a necessity for a married couple but anything more is most likely a luxury for an enthusiast with no family obligations. Also, those third or later cars are usually an ICEV.



(a) Interaction SHAP Value (ICEV)   (b) Interaction SHAP Value (EV)

Figure 9: Interaction SHAP Value (Number of Vehicle v.s. Marital Status)

When looking at the vehicle age and expenditure ratio (**Expenditure Ratio**) with Figure 10, we see a diminishing marginal utility. While the value of the SHAP value increases with the increase of the expenditure ratio, the gradient drops rapidly after about 25% mark. At the same time, according to the color distribution of the variable **Last Vehicle Age**, households with older vehicles are more sensitive to the expenditure ratio. The slope of their SHAP value is greater than that of households with newer cars.



(a) Interaction SHAP Value (ICEV) (b) Interaction SHAP Value (EV)

Figure 10: Interaction SHAP Value (Expenditure Ratio v.s. Model Age)

6. Conclusion

This study employs a combination of a traditional MNL model and an interpretable machine learning tool TreeExplainer to provide an explanatory analysis of vehicle purchase behavior within households. The PSID national dataset is taken to provide household information over a decade. From the PSID data, we utilized K-means-undersampling on the No Purchase group and SMOTE on the EV Purchase group to make data balance. The MNL model gave us a general idea of the population's tendencies and proved the path dependency of the consumers. It also helped us select variables that can reliably explain the outcome of the decision. The TreeExplainer allowed for a much more detailed analysis of consumer behavior based on different groups. In particular, we were able to see that people in different age groups or different income groups show wildly different patterns based on other independent variables.

Most people in general still preferred to get an ICEV, while households with higher income or more children leaned towards EV. When grouped by age, younger people were more likely to buy an EV when they have high income, whereas older generations are more likely to get an EV if they have a low income. When divided based on income, the higher income group is more willing to get a new car, especially if their old one is old. Most of the lower income group stayed with their old car, except the ones who bought an EV, possibly due to better deals with EV subsidies. Also, most households tend to get 2 vehicles and stop. This behavior is a lot more evident with married couples, as most vehicle purchase after the second one were from unmarried individuals.

Based on the given results, we provide pathways to bolster the EV market penetration. Seeing that even lower-income households are willing to get a new EV, we can see that subsidies do have a positive effect. As such, future subsidies can specifically target the younger age group to increase EV sales among the younger low-income demographic. Also, it would be important to incentivize married couples to get an EV as their second vehicle, as that is the biggest portion of the market.

In conclusion, this study analyzes the novel topic of heterogeneity in new vehicle purchase behaviors among car-owning households. The combination of Kmeans-undersampling and SMOTE algorithms helps us overcome the issues of severe imbalance in our data. The behavior of consumers is thoroughly analyzed both globally and locally via the MNL model and the TreeExplainer SHAP value results. The final result will help us examine vehicle purchasing behavior from a more comprehensive perspective and provide strong support for the future promotion of EVs on the market.

7. Limitation

One limitation of our work is that we do not follow individual households during the survey periods. While the time-effect variable is introduced to capture the potential effect of time, the surveys are aggregated into one large dataset. Future work could include conducting panel regressions to extensively explore how individual household's behavior changes over time.

Moreover, while we validated the reliability of the responses as much as we could during the data-cleaning step, unreliability still exists. The PSID questionnaire is a very long survey that not all respondents will answer truthfully and thoroughly. As a result, we often found inconsistencies between different respondents or even within a single response, resulting in a limited selection of variables, which depends on the future gradual improvement of the PSID data set.

8. ACKNOWLEDGMENTS

This study was supported by Nature Science Foundation of China [Grant number: 52302441].

9. AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: study conception and design: Lingyun Zhong, Taewhan Ko, Meiting Tu; data collection: Lingyun Zhong, Taewhan Ko; analysis and interpretation of results: Lingyun Zhong, Taewhan Ko, Meiting Tu, Dominique Gruyer; draft manuscript preparation: Lingyun Zhong, Taewhan Ko, Meiting Tu, Tongtong Shi. All authors reviewed the results and approved the final version of the manuscript.

References

S. A. Qadir, F. Ahmad, A. M. A. Al-Wahedi, A. Iqbal, A. Ali, Navigating the complex realities of electric vehicle adoption: A comprehensive study of government strategies, policies, and incentives, Energy Strategy Reviews 53 (2024) 101379.

J. Jia, Analysis of alternative fuel vehicle (afv) adoption utilizing different machine learning methods: a case study of 2017 nhts, IEEE Access 7 (2019) 112726–112735.

Q. Molloy, N. Garrick, C. Atkinson-Palombo, A new approach to understanding the impact of automobile ownership on transportation equity, Transportation Research Record 0 (0) 03611981231174444. URL: https://doi.org/10.1177/03611981231174444. doi:10.1177/03611981231174444. arXiv:https://doi.org/10.1177/03611981231174444.

M. J. Smart, N. J. Klein, A longitudinal analysis of cars, transit, and employment outcomes (2015).

Z. Yang, P. Jia, W. Liu, H. Yin, Car ownership and urban development in chinese cities: A panel data analysis, Journal of Transport Geography 58 (2017) 127–134.

S. Le Vine, J. Polak, The impact of free-floating carsharing on car ownership: Early-stage findings from london, Transport Policy 75 (2019) 119–127.

E. Blumenberg, A. Brown, A. Schouten, Car-deficit households: determinants and implications for household travel in the us, Transportation 47 (2020) 1103–1125.

S. Le Vine, C. Wu, J. Polak, A nationwide study of factors associated with household car ownership in china, IATSS research 42 (2018) 128–137.

G. C. de Jong, R. Kitamura, A review of household dynamic vehicle ownership models: holdings models versus transactions models, Transportation 36 (2009) 733–743.

N. J. Klein, M. J. Smart, Life events, poverty, and car ownership in the united states, Journal of Transport and Land Use 12 (2019) 395–418.

A. T. M. Oakil, D. Manting, H. Nijland, Dynamics in car ownership: the role of entry into parenthood, European Journal of Transport and Infrastructure Research 16 (2016).

S. P. Anderson, A. De Palma, J.-F. Thisse, A representative consumer theory of the logit model, International Economic Review (1988) 461–466.

J. C. Wiginton, A note on the comparison of logit and discriminant models of consumer credit behavior, Journal of Financial and Quantitative Analysis 15 (1980) 757–770.

Panel study of income dynamics(2021), Public Use Data Produced and Distributed by the Survey Research Center, 2021. URL: https://psidonline.isr.umich.edu/guide/default.aspx, accessed April 2021.

V. Shende, Analysis of research in consumer behavior of automobile passenger car customer, International Journal of Scientific and Research Publications 4 (2014) 1–8.

J. M. Dargay, The effect of income on car ownership: evidence of asymmetry, Transportation Research Part A: Policy and Practice 35 (2001) 807–821.

H. Sharma, A study of demographic characteristics influencing consumer behaviour regarding premium car brands, Pacific Business Review Internationa 8 (2015) 17–30.

B. H. Vrkljan, D. Anaby, What vehicle features are considered important when buying an automobile? an examination of driver preferences by age and gender, Journal of safety research 42 (2011) 61–65.

R. Bhardwaj, V. K. Bishnoi, Steering preferences: Investigating demographic influences on car buying external cues, Tuijin Jishu/Journal of Propulsion Technology 44 (2023) 1390–1402.

N. Monga, B. Chaudhary, S. Tripathi, Car market and buying behavior: A study of consumer perception, International Journal of Research in Management, Economics and Commerce 2 (2012) 44–63.

A. Peters, P. de Haan, R. W. Scholz, Understanding car-buying behavior: Psychological determinants of energy efficiency and practical implications, International Journal of Sustainable Transportation 9 (2015) 59–72.

J. Sanitthangkul, A. Ratsamewongjan, W. Charoenwongmitr, J. Wongkantarakorn, Factors affecting consumer attitude toward the use of eco-car vehicles, Procedia-Social and Behavioral Sciences 40 (2012) 461–466.

D. A. Hensher, Dimensions of automobile demand: a longitudinal study of household automobile ownership and use, Elsevier, 2013.

S. Muti, K. Yıldız, Using linear regression for used car price prediction, International Journal of Computational and Experimental Science and Engineering 9 (2023) 11–16.

J. Zhang, B. Yu, M. Chikaraishi, Interdependences between household residential and car ownership behavior: a life history analysis, Journal of Transport Geography 34 (2014) 165–174.

S. Li, Vehicle ownership over the life course among older americans: a longitudinal analysis, Transportation 51 (2024) 247–270.

Y. Li, Vehicle ownership, sustainable mobility and well-being in rural china, Environment, Development and Sustainability (2023) 1–24.

A. B. Parsa, H. Taghipour, S. Derrible, A. K. Mohammadian, Real-time accident detection: Coping with imbalanced data, Accident Analysis & Prevention 129 (2019) 202–210.

H. Chen, Y. Cheng, Travel mode choice prediction using imbalanced machine learning, IEEE Transactions on Intelligent Transportation Systems 24 (2023) 3795–3808.

S. Mishra, Handling imbalanced data: Smote vs. random undersampling, Int. Res. J. Eng. Technol 4 (2017) 317–320.

R. F. Pozo, A. B. R. González, M. R. Wilby, J. J. V. Díaz, M. V. Matesanz, Prediction of on-street parking level of service based on random undersampling decision trees, IEEE Transactions on Intelligent Transportation Systems 23 (2021) 8327–8336.

W.-C. Lin, C.-F. Tsai, Y.-H. Hu, J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, Information Sciences 409 (2017) 17–26.

M. Zheng, T. Li, X. Zheng, Q. Yu, C. Chen, D. Zhou, C. Lv, W. Yang, Uffdfr: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification, Information Sciences 576 (2021) 658–680.

N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (2002) 321–357.

H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International conference on intelligent computing, Springer, 2005, pp. 878–887.

Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, X. Han, A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data, Information Sciences 572 (2021) 574–589.

Q. Wei, Y. Pan, Research on the purchase intention for electric vehicles via machine learning methods, in: 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), IEEE, 2021, pp. 249–253.

A. Anas, Discrete choice theory, information theory and the multinomial logit and gravity models, Transportation Research Part B: Methodological 17 (1983) 13–23.

P. S. McCarthy, R. S. Tay, New vehicle consumption and fuel efficiency: a nested logit approach, Transportation Research Part E: Logistics and Transportation Review 34 (1998) 39–51.

Z. Ling, C. R. Cherry, Y. Wen, Determining the factors that influence electric vehicle adoption: A stated preference survey study in beijing, china, Sustainability 13 (2021) 11719.

C. Cirillo, Y. Liu, M. Maness, A time-dependent stated preference approach to measuring vehicle type preferences and market elasticity of conventional and green vehicles, Transportation Research Part A: Policy and Practice 100 (2017) 294–310.

C. Ding, X. Cao, B. Yu, Y. Ju, Non-linear associations between zonal built environment attributes and transit commuting mode choice accounting for spatial heterogeneity, Transportation Research Part A: Policy and Practice 148 (2021) 22–35.

J. Bas, C. Cirillo, E. Cherchi, Classification of potential electric vehicle purchasers: A machine learning approach, Technological Forecasting and Social Change 168 (2021) 120759. URL: https://www.sciencedirect.com/science/article/pii/S0040162521001918. doi:https://doi.org/10.1016/j.techfore.2021.120759.

S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

U. Ahmed, M. J. Roorda, Modeling freight vehicle type choice using machine learning and discrete choice methods, Transportation Research Record 2676 (2022) 541–552.

L. Jin, A. Lazar, C. Brown, B. Sun, V. Garikapati, S. Ravulaparthy, Q. Chen, A. Sim, K. Wu, T. Ho, et al., What makes you hold on to that old car? joint insights from machine learning and multinomial logit on vehicle-level transaction decisions, Frontiers in Future Transportation 3 (2022) 894654.

J. Liu, C. Wu, S. Le Vine, S. Jian, Panel data analysis of chinese households' car ownership and expenditure patterns, Transportation Research Part D: Transport and Environment 123 (2023) 103915.

N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

N. S. Kumar, K. N. Rao, A. Govardhan, K. S. Reddy, A. M. Mahmood, Undersampled k-means approach for handling imbalanced distributed data, Progress in Artificial Intelligence 3 (2014) 29–38.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).

J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

B. Peng, Y. Zhang, C. Li, T. Wang, S. Yuan, Nonlinear, threshold and synergistic effects of first/last-mile facilities on metro ridership, Transportation Research Part D: Transport and Environment 121 (2023) 103856.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature machine intelligence 2 (2020) 56–67.

S. Seabold, J. Perktold, Statsmodels: econometric and statistical modeling with python., SciPy 7 (2010).