future transportation

MDPI

Article

Understanding Post-COVID-19 Household Vehicle Ownership Dynamics Through Explainable Machine Learning

Mahbub Hassan [1,*], Saikat Sarkar Shraban [2], Ferdoushi Ahmed [3], Mohammad Bin Amin [4,*] and Zoltán Nagy [5]

[1] Department of Civil Engineering, Faculty of Engineering, Chulalongkorn University, Pathumwan, Bangkok 10330, Thailand
[2] School of Applied Sciences and Technology, Shahjalal University of Science and Technology, Sylhet 3100, Bangladesh; 2019333535@student.sec.ac.bd
[3] Faculty of Economics, Prince of Songkla University (PSU), Hat Yai, Songkla 90110, Thailand; ferdoushi.a@psu.ac.th
[4] Doctoral School of Management and Business, Faculty of Economics and Business, University of Debrecen, Böszörményi Street 138, 4032 Debrecen, Hungary
[5] Doctoral School of Management and Business Administration, John von Neumann University, 6000 Kecskemét, Hungary; nagy.zoltan@nje.hu
* Correspondence: mahbub.hassan@ieee.org (M.H.); binamindu@gmail.com (M.B.A.)

Abstract

Understanding household vehicle ownership dynamics in the post-COVID-19 era is critical for designing equitable, resilient, and sustainable transportation policies. This study employs an interpretable machine learning framework to model household vehicle ownership using data from the 2022 National Household Travel Survey (NHTS)—the first nationally representative U.S. dataset collected after the onset of the pandemic. A binary classification task distinguishes between single- and multi-vehicle households, applying an ensemble of algorithms, including Random Forest, XGBoost, Support Vector Machines (SVM), and Naïve Bayes. The Random Forest model achieved the highest predictive accuracy (86.9%). To address the interpretability limitations of conventional machine learning approaches, SHapley Additive exPlanations (SHAP) were applied to extract global feature importance and directionality. Results indicate that the number of drivers, household income, and vehicle age are the most influential predictors of multi-vehicle ownership, while contextual factors such as housing tenure, urbanicity, and household lifecycle stage also exert substantial influence highlighting the spatial and demographic heterogeneity in ownership behavior. Policy implications include the design of equity-sensitive strategies such as targeted mobility subsidies, vehicle scrappage incentives, and rural transit innovations. By integrating explainable artificial intelligence into national-scale transportation modeling, this research bridges the gap between predictive accuracy and interpretability, contributing to adaptive mobility strategies aligned with the United Nations Sustainable Development Goals (SDGs), particularly SDG 11 (Sustainable Cities), SDG 10 (Reduced Inequalities), and SDG 13 (Climate Action).

Keywords: vehicle ownership; explainable machine learning; SHAP; post-COVID-19 mobility; sustainable development goals

Received: 21 August 2025
Revised: 18 September 2025
Accepted: 25 September 2025
Published: 2 October 2025

Citation: Hassan, M.; Shraban, S.S.; Ahmed, F.; Amin, M.B.; Nagy, Z. Understanding Post-COVID-19 Household Vehicle Ownership Dynamics Through Explainable Machine Learning. *Future Transp.* **2025**, *5*, 136. https://doi.org/10.3390/futuretransp5040136

1. Introduction

Household vehicle ownership remains a foundational element of contemporary transportation systems, serving as a critical determinant of individual mobility, accessibility, and modal choice. In automobile-oriented societies such as the United States, vehicle availability

not only influences travel behavior but also reflects deeper structural dynamics, including socioeconomic inequality, urban–rural spatial asymmetries, and land-use patterns [1]. From a public policy perspective, household vehicle ownership informs decisions related to infrastructure investment, climate mitigation, and equity-sensitive mobility planning, each of which aligns with the broader objectives of the United Nations Sustainable Development Goals (SDGs), particularly SDG 11 (Sustainable Cities and Communities), SDG 13 (Climate Action), and SDG 10 (Reduced Inequalities) [2–4].

Traditionally, the modeling of vehicle ownership has relied on econometric techniques such as discrete choice models, count regressions (e.g., Poisson, negative binomial), and linear regression frameworks [5,6]. These approaches, while offering interpretability and statistical rigor, are constrained by assumptions of linearity, homoscedasticity, and error independence. As a result, they often fall short in capturing the nonlinear, interactive, and heterogeneous behaviors that increasingly characterize household transportation decision-making, particularly in dynamic post-crisis contexts.

The COVID-19 pandemic has significantly reshaped the landscape of urban mobility, prompting a reevaluation of long-standing assumptions. With the widespread uptake of remote work, a surge in e-commerce, declining public transit ridership, and a reconfiguration of spatial routines, household travel behavior has undergone structural change [7]. While some of these impacts reflect temporary disruptions, such as short-term declines in trip frequency during lockdowns or initial risk aversion toward shared modes, others indicate longer-term structural shifts in travel behavior. Emerging evidence points to sustained increases in telecommuting, suburban residential relocation, and continued caution toward public transit use, all of which alter baseline household mobility needs. Distinguishing between transient shocks and enduring transformations is therefore essential, and this study positions household vehicle ownership as a structural indicator of post-pandemic mobility change. These shifts have affected not only how frequently people travel, but also how they perceive and manage transportation risk, how they allocate capital to vehicles, and how they interact with the built environment. Consequently, understanding household vehicle ownership in the post-COVID-19 era requires analytical tools capable of capturing these evolving patterns. The 2022 National Household Travel Survey (NHTS) provides a uniquely situated empirical foundation for this inquiry. As the first nationally representative travel dataset collected after the onset of the pandemic, it captures critical shifts in household structure, remote work adoption, travel frequency, income volatility, and access disparities [8]. Compared to earlier NHTS waves (2001, 2009, 2017) [9], the 2022 data reflect a new equilibrium of behavioral, economic, and spatial conditions, making it methodologically and policy-relevant for understanding current ownership trends [10–14].

In parallel, recent advances in machine learning (ML) have opened new frontiers in travel behavior modeling, enabling the analysis of complex, high-dimensional, and nonlinear data structures [15,16]. ML methods such as Random Forest, XGBoost, Support Vector Machines, and Naïve Bayes classifiers have demonstrated superior performance across a wide range of transport applications including demand forecasting, mode choice prediction, and traffic safety analysis [17,18]. However, despite their predictive power, these models often function as "black boxes", providing limited insight into the underlying drivers of predicted outcomes [19]. This opacity presents a major challenge for public-sector decision-making, where transparency, trust, and accountability are paramount, especially in domains with distributional implications, such as household vehicle ownership. To meet these interpretive demands, the field has increasingly turned to explainable artificial intelligence (XAI), a growing body of methods designed to render machine learning models more transparent and diagnostically informative [20,21].

Among the most influential tools within the XAI domain is SHAP (SHapley Additive exPlanations), which attributes importance to each feature by decomposing model predictions using cooperative game theory [22]. SHAP values allow researchers and practitioners to understand not just which variables influence predictions, but how their impact varies across the dataset, enabling both global and local interpretability [23]. Other techniques, such as LIME, offer instance-level approximations but may lack the consistency and theoretical robustness required for policy-grade insights [24]. In the context of household vehicle ownership where decisions intersect with income, geography, lifecycle stage, and access to alternatives, XAI methods provide a critical bridge between computational models and policy needs [25]. Importantly, these tools also align with SDG 16 (Peace, Justice and Strong Institutions) by enhancing transparency and institutional legitimacy in AI-driven governance. By revealing the drivers of model predictions in an interpretable manner, XAI fosters stakeholder trust, supports evidence-based decision-making, and helps address ethical concerns associated with algorithmic bias and exclusion.

Despite these methodological advances, the integration of XAI into vehicle ownership modeling remains limited, with most studies emphasizing prediction accuracy at the expense of interpretability. Particularly in the wake of COVID-19, where new structural dynamics govern household mobility decisions, a transparent and behaviorally informed modeling framework is urgently needed to inform sustainable and inclusive transport policy.

This study addresses this research gap by proposing a comprehensive, interpretable, and empirically grounded machine learning application to model household vehicle ownership using the 2022 NHTS. The overarching objective is to examine how socioeconomic, spatial, and lifecycle variables shape vehicle ownership behavior in the post-pandemic context, while employing SHAP to generate actionable and interpretable insights. Specifically, this study makes four key contributions:

- Data Innovation: Leverages the first post-COVID-19 wave of the National Household Travel Survey to assess contemporary vehicle ownership patterns.
- Modeling Comparison: Applies and evaluates multiple supervised machine learning algorithms (e.g., XGBoost, Naïve Bayes, Random Forest, SVM) to determine predictive performance and behavioral consistency.
- Interpretability Integration: Uses SHAP to deliver both global and disaggregated feature attributions, thereby enabling transparent, explainable, and policy-aligned diagnostics.
- Policy and Sustainability Relevance: Translates model findings into insights that support SDG-linked priorities, including equity (SDG 10), urban sustainability (SDG 11), climate mitigation (SDG 13), and algorithmic governance (SDG 16).

By aligning advanced computational techniques with interpretability frameworks and SDG-driven planning goals, this research contributes a novel, policy-relevant approach to post-pandemic mobility analytics. In doing so, it supports the evolution of transportation modeling from a purely predictive science to a transparently interpretable, socially responsive, and sustainability-oriented discipline.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on vehicle ownership prediction, with emphasis on both traditional statistical approaches and emerging machine learning techniques. Section 3 outlines the methodology, including data processing, model selection, and explainability protocols. Section 4 presents empirical results, including model performance metrics and SHAP-based interpretations. Section 5 discusses the implications for transportation planning, sustainability, and algorithmic governance. Finally, Section 6 concludes with key findings and outlines avenues for future research.

2. Literature Review

Household vehicle ownership has long been central to transportation research because it shapes urban form, infrastructure investment, environmental externalities, and equity outcomes [26]. Methods have evolved from aggregate regressions and discrete choice models to machine learning and explainable artificial intelligence. This shift reflects both the rising complexity of travel behavior and the growing need for predictive frameworks that are accurate, transparent, interpretable, and relevant for policy.

Early empirical work used aggregate regressions to estimate ownership trends. For example, Sillaparcharn [27] analyzed provincial data for Thailand, and Bhat and Eluru [28] developed a discrete continuous copula-based framework that jointly modeled ownership, type choice, and usage. These econometric approaches offered clear behavioral interpretations, but they relied on restrictive assumptions about linearity, independence, and parametric form.

Recent studies increasingly adopt machine learning to capture complex, nonlinear, and high-dimensional relationships without prespecified functional forms. Ha [29] and colleagues and Bas [30] and colleagues used ensemble methods such as Random Forest, Gradient Boosting, and Neural Networks to predict household ownership and electric vehicle adoption, often achieving higher predictive accuracy than traditional models, especially when behavioral and attitudinal variables were available. Zambang [31] and colleagues confirmed these gains in a comparative study of nine supervised classifiers.

At the same time, machine learning has been criticized for its limited interpretability. Global importance metrics such as the Gini index or permutation importance identify which variables matter but not how they shape predictions for specific households or subgroups [32]. This opacity constrains usefulness in policy settings that demand accountability and transparency. In addition, much of the literature concentrates on electric vehicle adoption or purchase decisions, often using regional or pre-pandemic samples, with less attention to equity, spatial heterogeneity, or mechanisms of behavior. The present study addresses these gaps by applying SHAP-based interpretability to a national-scale framework with the 2022 National Household Travel Survey, thereby combining predictive accuracy with transparent, policy-relevant insight on household ownership transitions in the post-pandemic context.

Explainable methods have begun to close the interpretation gap. SHAP provides a rigorous allocation of feature contributions grounded in cooperative game theory. For instance, Naseri [33] and colleagues combined SHAP with XGBoost to study income and environmental factors in adoption decisions, while Ma and Pinsky [34] used LIME with Naive Bayes to examine heterogeneity across income groups. These studies demonstrate the value of pairing predictive strength with diagnostic insight, yet most are limited to electric vehicle outcomes and rarely use national data at scale.

Deep learning has also been explored for ownership modeling and can yield strong accuracy, but interpretability remains a challenge for public sector decision-making [35]. Ali [36] and colleagues compared machine learning with Multinomial Logit and concluded that while machine learning often improves predictive metrics, discrete choice models retain advantages for interpretability in strategic planning.

Important gaps remain. First, few studies integrate SHAP-based interpretability within national models that capture household-level heterogeneity in general ownership rather than only adoption of a specific technology. Second, many machine learning studies rely on pre-pandemic data and therefore cannot reflect structural changes such as greater telecommuting, residential relocation to suburbs, and substitution away from public transit. Third, equity analysis is often limited, with little attention to variation across demographic groups, geographic contexts, or tenure status. Fourth, behavioral attributes such as work-

from-home frequency, trip purpose, and modal resilience are commonly treated as ancillary rather than central explanatory factors. Finally, only a small number of studies explicitly link ownership modeling to broader sustainability and planning frameworks such as the Sustainable Development Goals. For example, Xu [37] and colleagues used Bayesian-optimized Support Vector Machines for sustainable user targeting but offered limited transparency and limited guidance for practice.

Although much of the post-pandemic literature centers on higher-income countries, an emerging body of work from the Global South provides valuable comparative perspectives on modal shifts, equity, and access. Examples include studies of constraints faced by women in South Asia [38], and biosecurity-driven mode shifts among students in Iran [39]. Research from East Asia shows that epidemic prevention concerns have reshaped perceptions of transit safety and comfort [40]. Work on ride-sharing willingness after the pandemic further illustrates how safety, reliability, and social interaction shape the adoption of shared options, with heterogeneity across demographic and attitudinal groups [41]. These insights underscore the continued importance of equity and context when interpreting ownership decisions.

To respond to the gaps above, the present study makes four contributions.

- Methodological advancement: We apply and compare ensemble-based classifiers XGBoost and Random Forest with Support Vector Machine and Naive Bayes in a unified and interpretable framework that uses SHAP for both global and local attributions.
- Empirical relevance: We use the 2022 National Household Travel Survey, which provides a nationally representative post-pandemic sample and captures emerging mobility behavior under new structural constraints.
- Equity and spatial sensitivity: We examine heterogeneity in ownership decisions across income, lifecycle stage, household structure, and spatial typologies including urban and rural settings.
- Policy and sustainability alignment: We translate model findings into actionable insights aligned with SDG 11 on sustainable cities, SDG 13 on climate action, and SDG 10 on reduced inequality, thereby linking machine learning evidence to the needs of transportation governance in the post-COVID-19 environment.

By integrating robust predictive modeling with transparent interpretability, this study bridges the divide between behavioral realism and data-driven computation. It advances vehicle ownership modeling toward an equitable, sustainable, and policy-aligned paradigm that can inform the mobility needs of diverse households in a rapidly changing world.

To position this study relative to recent machine learning and explainability work in transport, Table 1 contrasts scope, data scale, outcome focus, interpretability strategy, and policy orientation across representative studies and the present work.

Table 1. Recent studies related to transport ML and explainability research.

Study	Context and Data	Outcome Focus	Modeling Approach	Explainability	What is Different Here
[27]	Thailand, provincial data	Aggregate ownership levels	Aggregate regression	Not applicable	Moves from aggregate inference to household-level prediction with national microdata
[28]	Mixed contexts, microdata	Joint ownership, type, usage	Discrete choice with copula	Model coefficients	Extends beyond parametric structure to flexible machine learning with post hoc interpretability
[29]	Regional or study-specific samples	Household ownership or EV intention	Ensemble machine learning	Limited global importance	Uses national post-pandemic data and local interpretability for household-level insight

Table 1. *Cont.*

Study	Context and Data	Outcome Focus	Modeling Approach	Explainability	What is Different Here
[30]	Regional or study-specific samples	EV adoption	Ensemble and Neural Networks	Often minimal	Adds transparent SHAP analysis and an equity-oriented reading of results
[31]	Comparative machine learning	Classifier comparison for ownership	Nine supervised classifiers	Generic importance	Provides a unified framework with SHAP across multiple models using national data
[33]	EV adoption, study-specific	Adoption drivers	XGBoost	SHAP summaries	Shifts from EV adoption to general ownership at national scale in a post-pandemic setting
[34]	Purchase behavior, study-specific	Consumer heterogeneity	Naive Bayes	LIME	Applies SHAP across tree- and margin-based models and links findings to policy and the SDGs
[35]	Study-specific	Ownership prediction	Deep Neural Networks	Typically, none	Prioritizes interpretability with SHAP over black box accuracy for policy relevance
[36]	Comparative machine learning and MNL	Ownership prediction	Machine learning and discrete choice	Model coefficients for discrete choice	Combines machine learning accuracy with SHAP to recover behavioral structure at scale
[37]	Sustainable user targeting	Segmentation	Bayesian-optimized SVM	Limited	Embeds explainability and equity framing with national post-pandemic data
This study	United States, national sample, post-pandemic	General household ownership, single versus multiple vehicles	XGBoost, Random Forest, SVM, Naive Bayes in a unified framework	SHAP summary and violin plots for global and local attribution	Focus on post-pandemic ownership rather than EV only, consistent SHAP across models, equity and spatial sensitivity, and policy links to feasibility and the SDGs

3. Methodology

3.1. Data Source and Sample Selection

This study employs data from the 2022 wave of the National Household Travel Survey (NHTS), conducted by the Federal Highway Administration, USA. The NHTS is the most comprehensive source of household-level travel behavior in the United States, capturing information on vehicle ownership, trip characteristics, socio-demographics, and changes in mobility patterns in response to exogenous events such as the COVID-19 pandemic [42].

The 2022 wave of the NHTS is particularly important for examining post-pandemic mobility changes, as it includes variables related to telecommuting, shifts in trip-making behavior, and changes in attitudes due to health and mobility concerns [14]. For this study, the analytical sample was restricted to households that reported owning at least one vehicle, excluding those with zero vehicles to focus specifically on the factors influencing the ownership of additional vehicles beyond the initial acquisition. The NHTS dataset is organized into four main tables, household, person, trip, and vehicle, each containing detailed records on various aspects of travel behavior. For this study, data from the household and vehicle tables were used. The household table includes demographic and household-level data, while the vehicle data provides information on vehicle ownership and characteristics.

To integrate the vehicle and household datasets, the data were merged using the common HOUSEID identifier through the inner join function in R's "dplyr" package [43]. This process created a unified dataset that combined vehicle-specific data with corresponding household attributes, resulting in a final dataset containing responses from 7308 households. Survey weights from the NHTS were used in the descriptive statistics to ensure that the sample was representative of key demographic characteristics. However, these weights were not applied in the machine learning models due to limitations in incorporating them

into ensemble-based predictive algorithms. Comparisons were made to verify that the unweighted sample still represented the population well in terms of demographic factors. The methodological framework for this study comprises four primary phases: (i) data collection and preparation, (ii) data preprocessing, (iii) model selection and training, and (iv) model interpretability. The complete workflow is outlined in Figure 1.

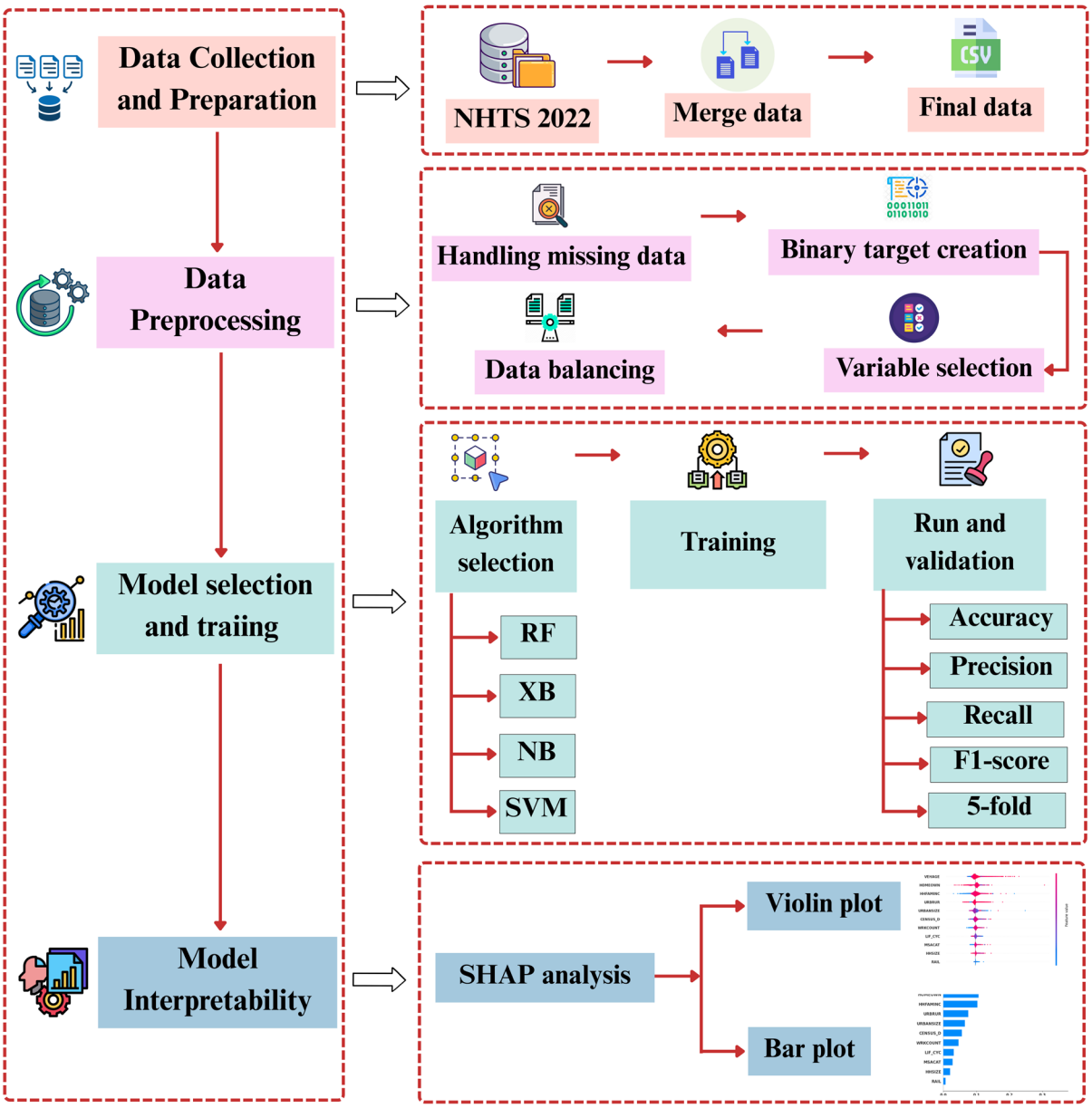


Figure 1. Overview of the methodological framework for vehicle ownership prediction using the NHTS 2022 dataset. The process comprises four primary phases: (i) data collection and preparation including data merging and formatting; (ii) data preprocessing involving missing data handling, binary target creation, variable selection, and class balancing; (iii) model selection and training utilizing four machine learning algorithms (RF, XB, NB, SVM) with 5-fold cross-validation and performance evaluation through accuracy, precision, recall, and F1 score; and (iv) model interpretability using SHAP analysis to visualize feature importance via violin and bar plots.

Spatial variation was captured through regional and urban–rural indicators provided in the NHTS. These variables allow the models to account for broad geographic differences, though they do not capture localized heterogeneity or spatial dependence effects. It should be noted that the 2022 NHTS does not include attitudinal or psychographic measures such

as environmental concern, perceived health risks, or lifestyle preferences. As such, the analysis is limited to socio-demographic, economic, and geographic predictors available in the dataset.

3.2. Data Cleaning and Preprocessing

The raw NHTS 2022 dataset required rigorous preprocessing to ensure internal validity and reproducibility. The cleaning procedure followed established transportation survey standards [44]. Variables with more than 20 percent missing responses, primarily attitudinal fields outside the scope of household vehicle ownership, were excluded. Household income, a variable known for high non-response, was replaced with the multiple imputed values provided within the official NHTS release, which apply hot-deck imputation techniques benchmarked to census records. For vehicle age, missingness was under two percent; these cases were imputed using the median vehicle age within the same census division and urban–rural stratum, ensuring contextual plausibility while minimizing distortion of distributional properties. Binary variables such as homeownership and urban–rural classification had less than two percent missing cases, which were addressed through listwise deletion to avoid introducing artificial classes. Continuous variables were examined for extreme values using both the interquartile range (IQR) criterion and visual inspection of histograms and boxplots [45]. Household incomes above USD 1 million and vehicle ages exceeding 50 years, each comprising fewer than 0.5 percent of cases, were winsorized at the 99th percentile. This approach mitigates leverage effects without discarding valid but extreme households. In contrast, households with seven or more members or with five or more licensed drivers were retained, as prior NHTS research indicates that such cases reflect multi-generational or extended family structures rather than measurement error.

All categorical predictors were systematically recoded for model readiness. Binary variables (e.g., homeownership, rail access, urban–rural classification) were dummy-coded as 0/1. Multinomial predictors such as census division, lifecycle stage, and metropolitan statistical area category were one-hot-encoded to prevent spurious ordinality. Rare categories representing fewer than one percent of households were collapsed into an "Other" category to preserve statistical stability. This approach maintains interpretability while avoiding model overfitting to sparse categories. Household-level and vehicle-level data were merged through the HOUSEID key using an inner join. Records with conflicting or logically inconsistent information (e.g., households reporting zero vehicles in the household file, but positive vehicle counts in the vehicle file) were excluded (<0.3 percent of cases). The final analytic dataset included 7308 households, representing the core vehicle-owning population. Descriptive distributions of household size, income, age composition, and regional location were compared between the cleaned analytic sample and weighted NHTS benchmarks. The cleaned sample aligned closely with population benchmarks, with deviations under three percentage points, confirming that the preprocessing did not introduce systematic bias.

Continuous predictors were standardized (z-score normalization to mean 0, standard deviation 1) to facilitate comparability across scales and improve model convergence, particularly for Support Vector Machines [46]. Categorical predictors were encoded prior to partitioning to avoid information leakage [47]. The dataset was randomly divided into 80 percent training and 20 percent testing subsets using stratified sampling to preserve the proportion of single- and multi-vehicle households. This structured preprocessing pipeline ensures that the analytic dataset is free from missing data bias, undue leverage from outliers, and spurious collinearity from categorical sparsity. By adhering to best practices in survey-based modeling, it establishes a robust foundation for subsequent machine learning estimation and interpretability analysis.

3.3. Dependent Variable Construction

The outcome of interest in this study is the household's vehicle ownership intensity, captured through a binary classification of whether a household owns exactly one vehicle or more than one vehicle. The dependent variable is defined as a binary indicator distinguishing single-vehicle households from multi-vehicle households. Zero-vehicle households were deliberately excluded from the analytic sample. This decision reflects both methodological and substantive considerations. Methodologically, zero-vehicle households constitute a structurally distinct group whose ownership outcomes are not on the same behavioral continuum as households that already own at least one vehicle. Their lack of vehicles is often a function of external structural factors such as high-density urban residence, strong public transit accessibility, parking constraints, or socioeconomic disadvantage rather than incremental household decision-making about acquiring additional vehicles. Substantively, our research objective was to isolate the determinants of transitioning from single-vehicle to multi-vehicle ownership, a decision context where factors such as household size, income, and residential environment operate differently than in zero-vehicle households. Including zero-vehicle households within the same model risks conflating fundamentally different mobility regimes and obscuring the behavioral mechanisms underlying multi-vehicle ownership. Thus, the analytic focus of this study is restricted to households already owning at least one vehicle. This approach diverges from the traditional binary modeling of vehicle access (zero vs. one or more vehicles) and instead focuses on the behavioral and structural factors that influence the acquisition of additional vehicles among already motorized households.

The rationale for this operationalization stems from emerging literature in post-pandemic travel behavior, which highlights significant heterogeneity among vehicle-owning households in terms of mobility needs, spatial accessibility, and modal resilience. For example, the COVID-19 pandemic prompted increased telecommuting, a decline in public transit usage, and growing reliance on private vehicles, particularly in multi-worker households [48,49]. These shifts likely impacted the marginal decision to acquire an additional vehicle rather than the baseline choice of becoming motorized.

The dependent variable, denoted as y, was derived from the HHVEHCNT field, representing the total number of household vehicles. Accordingly, the binary dependent variable y is defined as in Equation (1):

$$y = \begin{cases} 0, if\ HHVEHCNT\ =\ 1 \\ 1, if\ HHVEHCNT\ >\ 1 \end{cases} \quad (1)$$

This binary representation of vehicle ownership intensity enables the application of classification models to predict the probability that a household transitions from single- to multiple-vehicle ownership, conditional on a set of socio-demographic, behavioral, and spatial factors. Households with missing or non-binary vehicle count values were filtered out.

3.4. Explanatory Variables and Multicollinearity Assessment

The explanatory variables used in this study reflect a theoretically grounded and empirically supported set of factors associated with household vehicle ownership decisions. These include variables capturing household structure, income, housing tenure, lifecycle stage, urban form, and access to rail infrastructure. The selection of variables was informed by prior literature in travel behavior modeling and transportation economics, with special attention to pandemic-induced changes in mobility and location preferences [50–52]. Twelve independent variables were retained based on both theoretical justification and empirical support from prior studies.

To evaluate the presence of multicollinearity among the independent variables, we computed the Variance Inflation Factor (VIF) for each predictor. VIF quantifies the degree to which the variance of an estimated regression coefficient increases due to collinearity with other predictors [53]. While there is no universal threshold, values exceeding 5.0 (or, in some applications, 3.0) are typically considered indicative of problematic multicollinearity [54]. In this study, all variables exhibited VIF values well below these thresholds, suggesting a lack of collinearity that would adversely affect model estimation or interpretation. The VIF formula is provided in Equation (2):

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where R_i^2 represents the proportion of variance in the i-th independent variable explained by the other predictors in the model, while the tolerance is calculated as $1 - R_i^2$, indicating the variance not shared with other variables.

Table 2 presents the list of variables used in the model, along with their descriptions and associated VIF scores. Among the predictors, the highest VIF score was observed for the variable MSACAT (3.20), which remains well below the conservative threshold of 5.0. This variable was retained in the model due to its substantive role in capturing urban hierarchy and regional accessibility. Additionally, variables such as DRVRCNT, WRKCOUNT, and HHSIZE were carefully evaluated for potential multicollinearity given their conceptual proximity. However, each of these variables represents distinct behavioral dimensions, thereby justifying their inclusion in the final specification. The overall VIF analysis confirms that multicollinearity is not a substantial concern, thereby validating the use of all selected variables in subsequent model estimation.

Table 2. Explanatory variables and VIF scores.

Variables	Description	VIF Score
CENSUS D	Census division classification for home address	1.06
DRVRCNT	Number of drivers in the household	2.28
HHFAMINC	Household income (imputed)	1.48
HHSIZE	Total number of people in household	1.91
HOMEOWN	Whether home owned or rented	1.18
LIF_CYC	Lifecycle classification for the household	1.40
VEHAGE	Age of vehicle, based on model year	1.08
URBANSIZE	Urban area size where home address is located	1.64
URBRUR	Household in urban/rural area	1.97
MSACAT	MSA category for the HH home address	3.20
WRKCOUNT	Count of workers in household	1.99
RAIL	MSA heavy rail status for household	2.69

To enhance reproducibility and transparency, we provide an overview of the variable selection process. The candidate predictors were drawn from the NHTS 2022 household and vehicle files and screened in three stages. First, theoretical relevance was established based on prior transportation and travel behavior studies, with emphasis on variables repeatedly shown to affect vehicle ownership (e.g., household income, size, drivers, housing tenure, urban form). Second, empirical diagnostics were applied: variables with more than 20 percent missingness, limited variance, or inconsistent reporting were excluded (e.g., attitudinal items and rarely used categorical classes). Third, multicollinearity was assessed using Variance Inflation Factors (VIFs), and all retained predictors were found to be within acceptable thresholds. This structured process resulted in the inclusion of

twelve explanatory variables capturing household structure, socioeconomic status, and spatial context.

In summary, variables were retained when they satisfied both (i) theoretical justification in the literature and (ii) empirical suitability within the NHTS dataset. Variables excluded from the final models were primarily those with high missingness, weak behavioral grounding, or redundancy with included predictors.

3.5. Exploratory Correlation Analysis

To complement the multicollinearity diagnostics provided in Table 2 through Variance Inflation Factor (VIF) scores, an exploratory correlation analysis was conducted to assess bivariate relationships among the explanatory variables. The resulting Pearson correlation matrix is presented in Figure 2, where each cell depicts the correlation coefficient between a pair of variables. The strength and direction of association are color-coded, ranging from dark blue (strong positive correlation) to light shades near zero, and lighter hues for negative associations.

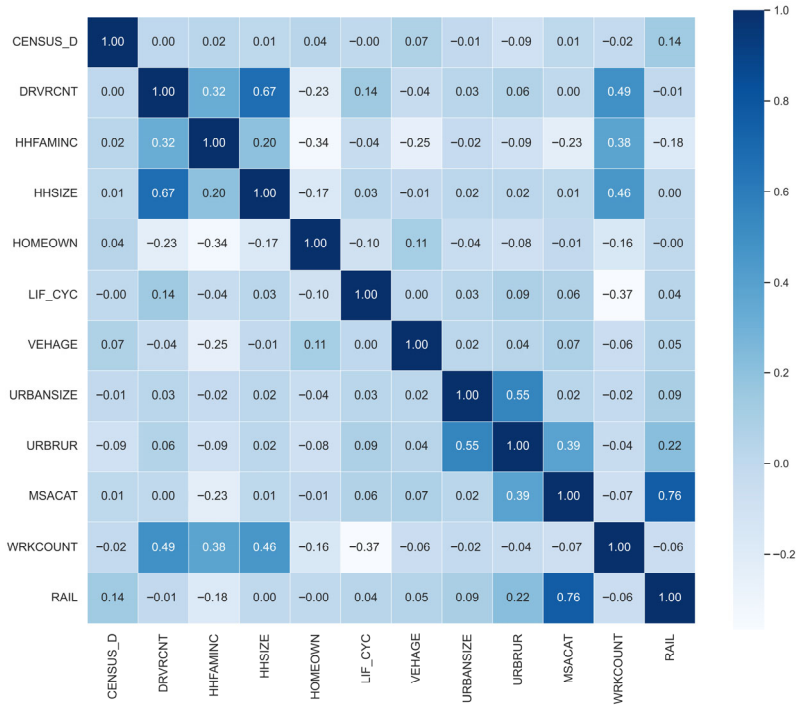


Figure 2. Pearson correlation matrix of the explanatory variables.

As shown in Figure 2, most variables exhibit weak to moderate pairwise correlation, with the majority of coefficients falling below $|0.50|$. The highest observed correlation is between DRVRCNT (number of drivers in the household) and HHSIZE (household size), with r = 0.67. This relationship is behaviorally plausible, as larger households tend to contain more eligible drivers. Similarly, a moderate correlation is observed between WRKCOUNT and HHSIZE (r = 0.46), indicating overlapping labor force dynamics within larger households.

A notable spatial pairing is the correlation between RAIL (rail access status) and MSACAT (metropolitan statistical area classification), where r = 0.76. Although this value approaches conventional thresholds for collinearity, the two variables capture distinct spatial dimensions, transportation infrastructure versus urban hierarchy, and are theoretically justified for inclusion. Their high correlation reflects structural alignment in metropolitan areas with heavy rail investment, rather than redundancy in measurement.

Importantly, no pairwise correlation exceeds 0.80, the heuristic threshold typically used to flag problematic collinearity in regression and machine learning applications. Additionally, spatial and urban form indicators (e.g., URBRUR, URBANSIZE, and CENSUS_D) exhibit low to moderate inter-correlations, affirming the empirical distinctiveness of geographic variables included in the model.

The combined evidence from the correlation matrix and VIF analysis confirms that multicollinearity is not a substantive concern in this study. All twelve explanatory variables are retained for subsequent machine learning modeling and SHAP-based interpretability, with sufficient statistical independence to ensure robust parameter learning and meaningful post hoc explanation.

3.6. Machine Learning Models and Training Procedures

The modeling process in this study followed a structured and replicable machine learning pipeline. The objective was to classify households as owning either a single vehicle or multiple vehicles based on a range of socio-demographic, spatial, and behavioral attributes. The workflow integrates feature selection, data preprocessing, model training, evaluation, and interpretability analysis.

To mitigate the class imbalance between households owning a single vehicle and those owning multiple vehicles, this study employed the Synthetic Minority Over-sampling Technique (SMOTE), a widely used resampling method introduced by Chawla et al. [55]. SMOTE synthesizes new minority class instances by interpolating between existing samples and their nearest neighbors in the feature space, thereby enhancing the representational density of the minority class without merely duplicating existing observations.

To preserve model integrity and prevent information leakage, SMOTE was applied exclusively to the training subset after the dataset was partitioned. Stratified random sampling was used to divide the data into 80% training and 20% testing subsets, ensuring proportional representation of the target classes in both sets [56]. The interpolation used in SMOTE is defined as Equation (3):

$$x_{new} = x_i + (x - x_i)\delta \tag{3}$$

where x_{new} represents the newly generated synthetic sample, x_i is the sample in the minority class, x is the random neighbor among the k-nearest neighbors, and δ is a random number ranging from 0 to 1.

This procedure yields a balanced training set, enabling the learning algorithms to more effectively capture decision boundaries associated with underrepresented class instances. Given the sensitivity of Support Vector Machines (SVM) to feature scaling, all predictors were standardized using z-score normalization via the StandardScaler transformation, resulting in zero-mean and unit-variance features.

To assess comparative predictive performance, the study implemented a diverse suite of machine learning classifiers:

- Random Forest;
- EXtreme Gradient Boosting (XGBoost);
- Support Vector Machine (SVM);
- Naïve Bayes.

Each model was trained and evaluated using consistent preprocessing pipelines to ensure fair comparisons across performance metrics and interpretability outcomes.

- Random Forest: The Random Forest classifier, introduced by Breiman, is an ensemble learning algorithm that combines multiple decision trees trained on bootstrapped subsamples of the data [57]. By aggregating the predictions of diverse trees, it reduces

variance and improves generalization compared to a single tree. In this study, Random Forest is applied to predict household vehicle ownership as a binary classification task.

The model was implemented using the Scikit-Learn library and configured with 100 trees, which provided a balance between computational efficiency and predictive performance. The Gini Index was used as the impurity measure for node splitting, and final predictions were obtained through majority voting across the ensemble [32].

The Gini Index is represented by Equation (4):

$$GiniIndex = \sum\sum_{j \neq i} \left(\frac{f(Y_i, T)}{|T|} \right) \left(\frac{f(T_j, T)}{|T|} \right) \quad (4)$$

where T represents the training dataset, and $f(Y_i,\ T)$ denotes the probability of belonging to a category Y_i.

The final prediction in the Random Forest model is determined by majority voting, as represented by Equation (5):

$$\hat{y} = majority\ vote(T_1(x), T_2(x), \ldots, T_n(x)) \quad (5)$$

where $\hat{y}$ is the predicted class label, $T_1(x), T_2(x), \ldots T_n(x)$ are the predictions of the base classifiers on input x, n is the number of base models, and the majority vote function selects the most frequent prediction among them.

- XGBoost: XGBoost was employed in this study because of its strong performance in capturing complex, nonlinear relationships and its proven predictive accuracy across a wide range of applications. As a gradient boosting method, XGBoost constructs decision trees sequentially, with each tree focusing on correcting the residual errors of the previous ones. This iterative process allows the model to effectively capture higher-order interactions and variable importance [58].

In this study, the XGBClassifier implementation in Python 3.13 was used. The model was tuned to optimize predictive performance for household vehicle ownership, with socio-demographic and behavioral variables serving as key predictors. XGBoost's combination of accuracy, efficiency, and interpretability of variable influence makes it especially well suited to the present research context.

Equations (6) and (7) represent the XGBoost model.

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \ldots \hat{y}_N \quad (6)$$

$$L(\theta) = \sum_{i=1}^{n} l(\hat{y}_l, y_j) + \sum_{m=1}^{M} \Omega(f_m) \quad (7)$$

where $\hat{y}$ is the predicted value, f_m is the output from the m-th, M is the total number of trees, $\hat{y}_l, y_j$ is the loss function, and $\Omega(f_m)$ is the regularization term that controls the complexity of each tree.

- Support Vector Machine: SVM was used in this study as a benchmark for binary classification of household vehicle ownership. SVM is effective for distinguishing between classes by constructing an optimal decision boundary, and with the use of kernel functions, it can capture nonlinear relationships between predictors and outcomes [59].

In this study, the radial basis function (RBF) kernel was applied to model complex, nonlinear interactions among socio-demographic features. Because SVM is sensitive to

feature scaling, the input data were standardized using Scikit-Learn's StandardScaler before training. The SVM classifier was implemented in Python with the LinearSVC module to ensure computational efficiency and robust performance.

The classification function is represented by Equation (8):

$$f(x) = sign\left[\sum_{i=1}^{n}\left(\alpha_i Y_j \times k(x, x_i)\right) + b\right] \tag{8}$$

where b is the offset from the origin of the hyperplane, n represents the number of independent variables, α_i defines the positive constant, and $k(x, x_i)$ is the kernel function that measures the similarity between x and x_i.

- Naïve Bayes: Naïve Bayes was included in this study as a baseline classifier due to its simplicity, efficiency, and suitability for binary prediction tasks. As a probabilistic model, it estimates the likelihood of class membership based on Bayes' theorem. Although its assumption of conditional independence among features is rarely met in practice, the algorithm often performs competitively, particularly on large datasets [60].

In this study, Naïve Bayes provided a useful point of comparison for more complex ensemble methods, allowing us to assess the trade-off between model simplicity and predictive accuracy in vehicle ownership classification.

The fundamental formula is shown in Equation (9):

$$P(y \mid x) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x)} \tag{9}$$

where $P(y \mid x)$ is the posterior probability of class y given the input features x, $P(y)$ is the prior probability of class y, $P(x_i \mid y)$ is the likelihood of feature x_i given class y, and $P(x)$ is the marginal probability of the feature vector x.

3.7. Model Evaluation and Performance Metrics

3.7.1. Confusion Matrix

The confusion matrix is a widely used tool for evaluating the performance of classification models, especially for binary classification tasks. It provides a breakdown of how well the model's predicted values match the actual class labels [61]. For binary classification tasks, such as predicting vehicle ownership, the confusion matrix helps categorize predictions into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This categorization enables the calculation of key performance metrics such as accuracy, precision, recall, and F1 score [62]. In the context of vehicle ownership, the confusion matrix for binary classification is structured as Equation (10):

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{10}$$

The terms used in the confusion matrix are defined as follows: TP refers to the number of correctly predicted instances in which households own more than one vehicle, while TN denotes the number of correctly predicted instances where households own exactly one vehicle. FP represents cases where the model incorrectly predicts that a household owns more than one vehicle when it actually owns only one. Conversely, FN represents instances in which the model incorrectly predicts that a household owns exactly one vehicle when it, in fact, owns more than one.

3.7.2. Performance Metrics

In evaluating the performance of machine learning models for binary classification tasks, several performance metrics are commonly used. These metrics help assess how well the model performs in predicting both the positive and negative classes. Accuracy, precision, recall, and F1 score are widely employed to evaluate model effectiveness, especially when evaluating the predictive performance in classification tasks [63].

Accuracy: Accuracy is the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances. It provides a general measure of the model's ability to make correct predictions. It is calculated as Equation (11):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

Precision: Precision measures the proportion of true positive predictions out of all predicted positive instances. It reflects the ability of the model to avoid false positives. It is expressed as Equation (12):

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

Recall: Recall calculates the proportion of actual positive instances that were correctly identified by the model. It shows how well the model can detect positive instances. It is shown as Equation (13):

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It balances both metrics and is especially useful in situations with class imbalance. It is expressed as Equation (14):

$$F1Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{14}$$

Together, these metrics provide a comprehensive framework for evaluating model performance, enabling comparative assessment across models with differing sensitivities to class imbalance and misclassification costs. Model performance was evaluated using accuracy, precision, recall, and F1 score, which together provide a balanced assessment of predictive capability under conditions of class imbalance. These measures directly capture both Type I and Type II classification errors and are widely used in transportation and machine learning research [64]. Although metrics such as AUC ROC or AUC PR provide additional perspectives, the inclusion of precision, recall, F1 score, and confusion matrices already offers a robust and interpretable evaluation of model performance in the context of this study.

3.8. Five-Fold Cross-Validation

To further validate generalizability, 5-fold cross-validation was performed for each model. In this approach, the dataset was randomly divided into five equal parts. For each iteration, the model was trained on four folds and validated on the remaining fold. This process was repeated five times so that every fold served as a test set once. The average performance across the five folds was then calculated to provide a more reliable estimate of the model's generalizability and reduce the risk of overfitting [65]. It is calculated as Equation (15):

$$CV_5 = \frac{1}{k} \sum_{i=1}^{k} M_i \tag{15}$$

where k is the number of folds, and M_i is the evaluation metric (accuracy, precision, etc.) obtained in the i-th fold.

3.9. Model Explainability and Interpretation

While machine learning models offer enhanced predictive power over traditional parametric methods, their complexity often renders them opaque to stakeholders seeking to understand the drivers behind specific outcomes. To address this issue, the present study adopts a suite of post hoc explainability techniques, with a particular focus on SHapley Additive exPlanations (SHAP), to interpret model predictions and enhance transparency.

SHAP values are grounded in cooperative game theory and provide a theoretically principled approach to attributing the contribution of each input feature to a specific prediction [66]. Formally, for a machine learning model f and a given instance x, the SHAP value $\varnothing_j$ for feature j is calculated as the average marginal contribution of feature j across all possible subsets of features, as expressed in Equation (16):

$$\varnothing_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (16)$$

where F denotes the full set of features, and S is a subset excluding j. The SHAP value $\varnothing_j$ thus represents the contribution of feature j to the difference between the prediction for x and the expected prediction across all instances.

In this study, SHAP was applied to the best-performing model, identified through cross-validation and confirmed via evaluation on the holdout test set. The focus was on generating both global and local interpretations of feature influence:

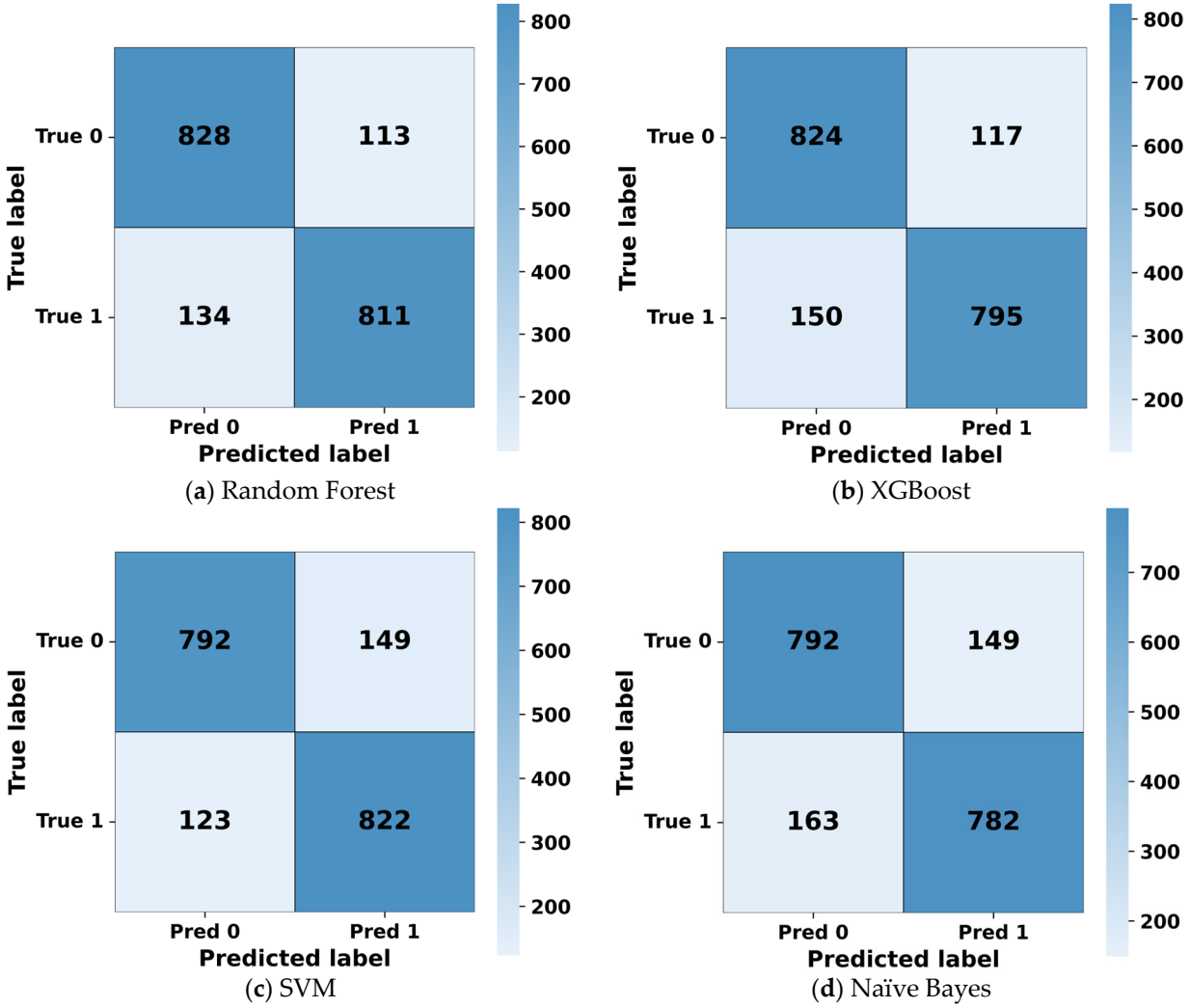
- Global interpretation was achieved by aggregating the mean absolute SHAP values across all observations in the dataset. This process yielded a ranked list of predictor variables, providing insights into the most influential socio-demographic and spatial factors contributing to the likelihood of owning multiple vehicles.
- To visualize these contributions, two complementary SHAP summary plots were employed:
    - A bar plot of mean absolute SHAP values, which illustrates global feature importance.
    - A violin (or beeswarm) plot, which simultaneously conveys feature importance, directionality, and distributional patterns across individual predictions.

To ensure the robustness of SHAP-based interpretability, we validated the outputs through three complementary strategies. First, cross-model consistency was assessed by comparing SHAP feature rankings across Random Forest, XGBoost, SVM, and Naïve Bayes; convergence on core predictors such as number of drivers, household income, and vehicle age confirmed stability across algorithmic families. Second, alignment with domain theory was examined by benchmarking SHAP-derived patterns against well-established behavioral findings in the vehicle ownership literature (e.g., higher driver counts increase vehicle ownership; rural households exhibit higher car dependence). This step ensured that SHAP outputs were not only statistically robust but also behaviorally plausible. Third, distributional checks were performed using violin and dependence plots to verify that the direction and magnitude of SHAP values were consistent across subgroups (e.g., high-income vs. low-income households, urban vs. rural households). These procedures demonstrate that the SHAP results provide reliable and interpretable insight into the determinants of vehicle ownership, rather than artifacts of a single algorithm or sample partition.

4. Results

4.1. Confusion Matrix-Based Performance Assessment

Figure 3 presents the confusion matrices for all four machine learning models: Random Forest, XGBoost, Support Vector Machine (SVM), and Naïve Bayes. These matrices summarize the number of correct and incorrect predictions made by the models in terms of actual and predicted classes.



(**a**) Random Forest (**b**) XGBoost

(**c**) SVM (**d**) Naïve Bayes

Figure 3. Confusion matrices of the four machine learning models: (**a**) Random Forest, (**b**) XGBoost, (**c**) SVM, and (**d**) Naïve Bayes. Each matrix summarizes the model's classification performance by showing the distribution of actual versus predicted class labels. Diagonal values indicate correct predictions, while off-diagonal values represent misclassification.

As shown in Figure 3, the Random Forest model achieved high classification performance, correctly identifying 828 households in class 0 and 811 in class 1. It had 113 false positives and 134 false negatives. The XGBoost model also performed well, with 824 true negatives and 795 true positives, but showed slightly more false negatives (150) compared to Random Forest. The SVM model predicted 822 instances of class 1 correctly (true positives) and had the lowest number of false negatives (123), although it misclassified 149 class 0 instances (false positives). The true negatives for SVM were 792, indicating that the model correctly identified 792 class 0 instances. Interestingly, Naïve Bayes showed an identical number of true negatives (792) and false positives (149) but had the highest number of false

negatives (163) and true positives (782), indicating its relative weakness in identifying class 1 instances.

These confusion matrices indicate that Random Forest and SVM (RBF) were more effective at distinguishing between the two classes, with fewer misclassifications overall. The differences in false positives and false negatives among the models provide insight into each model's strength in identifying specific ownership categories.

4.2. Model Performance Comparison

To evaluate the classification efficacy of the proposed machine learning framework in distinguishing single-vehicle- from multi-vehicle-owning households, four widely adopted classifiers were implemented: Random Forest, XGBoost, Support Vector Machine (SVM) with a radial basis function kernel, and Naïve Bayes. The models were trained on the preprocessed and balanced dataset and assessed on a holdout test set using four key performance metrics: accuracy, precision, recall, and F1 score. A comparative visualization of these metrics across models is provided in Figure 4.

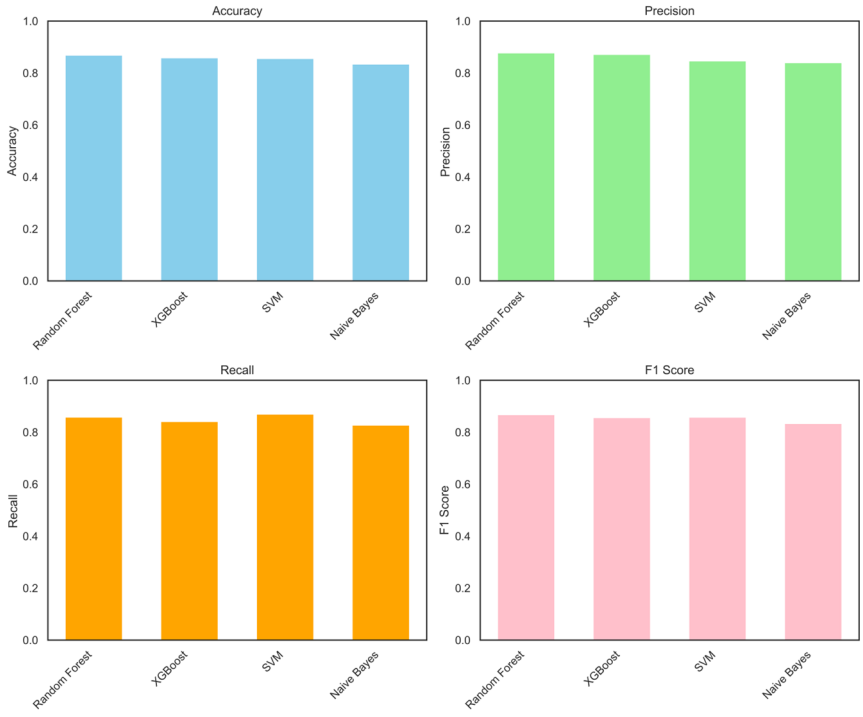


Figure 4. Comparative performance evaluation of four supervised machine learning models—Random Forest, XGBoost, Support Vector Machine (RBF kernel), and Naïve Bayes—across four classification metrics.

The Random Forest classifier emerges as the most balanced and robust performer, achieving the highest accuracy (0.8690), precision (0.8777), and F1 score (0.8678). These results underscore its capacity to generalize effectively across varying household typologies while maintaining high discriminative power. Its ensemble structure, rooted in bootstrapped aggregation and randomized feature selection, appears particularly well suited for modeling the nonlinear and heterogeneous patterns embedded in post-COVID-19 mobility behavior. The XGBoost model follows closely, with comparable performance across all four metrics: accuracy (0.8584), precision (0.8717), recall (0.8413), and F1 score (0.8562). Its iterative boosting mechanism enhances its ability to capture second-order interactions and localized data nuances, rendering it an effective alternative when fine-grained pattern recognition is prioritized.

Notably, the SVM classifier records the highest recall (0.8698) among all models, indicating a strong ability to correctly identify households with multiple vehicles. This performance is indicative of the SVM's strength in maximizing the decision margin within high-dimensional feature spaces. In policy-sensitive domains where false negatives carry higher costs such as under-identifying auto-dependent households, SVM may offer strategic value despite its lower precision (0.8465). By contrast, the Naïve Bayes classifier registers the lowest scores across all evaluation metrics, including accuracy (0.8346) and F1 score (0.8337). This performance reflects the algorithm's underlying assumption of conditional independence among features, which is unlikely to hold in the context of complex socio-spatial interactions. Nevertheless, its simplicity, interpretability, and computational efficiency make it a practical baseline and a candidate for hybrid ensemble integration in large-scale deployments.

In summary, the comparative evaluation confirms that ensemble tree-based methods, particularly Random Forest, provide the most reliable and interpretable balance between classification accuracy and generalizability in modeling post-pandemic vehicle ownership behavior. While SVM offers niche advantages in recall-sensitive applications, and Naïve Bayes delivers speed and transparency, ensemble models emerge as the most suitable for contexts requiring both predictive rigor and behavioral insight. These findings validate the suitability of interpretable ensemble learning for national-scale transport behavior modeling and substantiate its integration with post hoc explainability tools presented in subsequent sections.

4.3. Cross-Validation and Model Stability

To assess the generalizability and robustness of each model, a stratified 5-fold cross-validation procedure was conducted. Table 3 summarizes the average accuracy scores across folds for each model.

Table 3. Five-fold cross-validation mean accuracy.

Model	Mean CV Accuracy
Random Forest	0.8710
XGBoost	0.8626
SVM	0.8592
Naïve Bayes	0.8394

The Random Forest model showed the highest average accuracy during 5-fold cross-validation (0.8710), confirming its consistent performance across different data splits. XGBoost also performed well, with an average accuracy of 0.8626, making it a strong ensemble-based alternative. The SVM model followed closely, achieving an average accuracy of 0.8592. Its performance remained stable across folds, ranging from 0.8542 to 0.8659, which indicates good reliability.

The choice of SVM with a radial basis function kernel is particularly well-justified in the context of household vehicle ownership prediction. The nonlinear relationships inherent in the data, such as those between household income, urban classification, and the number of drivers, are effectively modeled by the RBF kernel, which maps input features into a higher-dimensional space. This capability enables the SVM classifier to identify intricate decision boundaries that may not be captured by linear or tree-based models alone. Naïve Bayes had the lowest average accuracy (0.8394) among the four models. However, it remains useful because it is fast, simple to implement, and provides probabilistic outputs.

The superior performance of the Random Forest classifier can be attributed to both the structure of the NHTS dataset and the algorithm's strengths. Household vehicle ownership

is influenced by nonlinear interactions among socio-demographic, spatial, and household composition variables (e.g., the combined effect of household size, number of drivers, and income). Random Forests excel at capturing such interactions because they aggregate predictions from multiple decorrelated decision trees, thereby accommodating heterogeneity without overfitting to individual subgroups. In addition, the dataset includes a mix of categorical and continuous predictors (e.g., census division, household income, vehicle age), and Random Forest is well suited to handle such mixed data types with minimal preprocessing. Compared to Support Vector Machines, which tend to over-rely on a few highly discriminative variables, and Naïve Bayes, which is constrained by independence assumptions, Random Forest strikes a balance between flexibility and generalization. XGBoost produced competitive results but was marginally less effective, likely due to its sensitivity to hyperparameter tuning and its iterative boosting structure, which can overweight rare or noisy patterns in survey-based data. Collectively, these characteristics make Random Forest particularly appropriate for modeling the complex, nonlinear, and high-dimensional decision processes underlying household vehicle ownership.

Overall, the results suggest that ensemble and kernel-based models are better suited to understanding the diverse and changing factors behind household vehicle ownership, especially in the post-pandemic period.

4.4. Model Interpretation Through SHAP Explainability

To better understand how different variables influence household vehicle ownership, SHAP (SHapley Additive exPlanations) analysis was applied to four machine learning models. The following summary bar plots show the average impact of each feature on the model's prediction for Random Forest, XGBoost, SVM with RBF kernel, and Naïve Bayes.

4.4.1. Feature Importance Analysis Across Models

Figure 5 presents the SHAP summary bar plots for each model, showing the average magnitude of each feature's influence on model predictions. In the Random Forest model, the number of drivers in a household (DRVRCNT) emerges as the most influential factor, far surpassing other variables. This dominance aligns with behavioral expectations, as a higher number of drivers typically correlates with a greater need for multiple vehicles. VEHAGE (age of household vehicles) is also highly ranked, suggesting that households with older vehicles may retain them longer or replace them gradually. Other top-ranking features include HHFAMINC (household income) and HOMEOWN (homeownership status), indicating that financial stability and tenure status support multi-vehicle ownership. Built environment indicators such as URBRUR (urban–rural classification), URBANSIZE (urban area size), and CENSUS_D (census division) appear lower in the ranking, reflecting a stronger emphasis on socio-demographic over spatial factors. Additional variables like MSACAT (metropolitan area category), HHSIZE (household size), LIF_CYC (lifecycle stage), WRKCOUNT (number of employed household members), and RAIL (availability of rail transit) contribute less prominently.

In XGBoost, DRVRCNT remains the top feature, but the importance is more evenly distributed across other variables. HHFAMINC, VEHAGE, and HOMEOWN retain high influence, highlighting the model's capacity to integrate economic and asset-based factors. Spatial attributes such as URBANSIZE, URBRUR, and CENSUS_D occupy mid-tier positions, suggesting that urban context contributes meaningfully, though not dominantly. The inclusion of LIF_CYC, WRKCOUNT, and MSACAT further emphasizes the model's ability to incorporate nuanced household dynamics. The SVM model demonstrates a sharply skewed importance profile, with DRVRCNT overwhelmingly dominant. All remaining features, including VEHAGE, HHFAMINC, and HHSIZE, display minimal contribution.

This pattern reflects the model's sensitivity to a few highly discriminative features, often at the expense of broader contextual interpretation. Spatial and socioeconomic variables contribute marginally, indicating a narrowed scope of explanatory power.

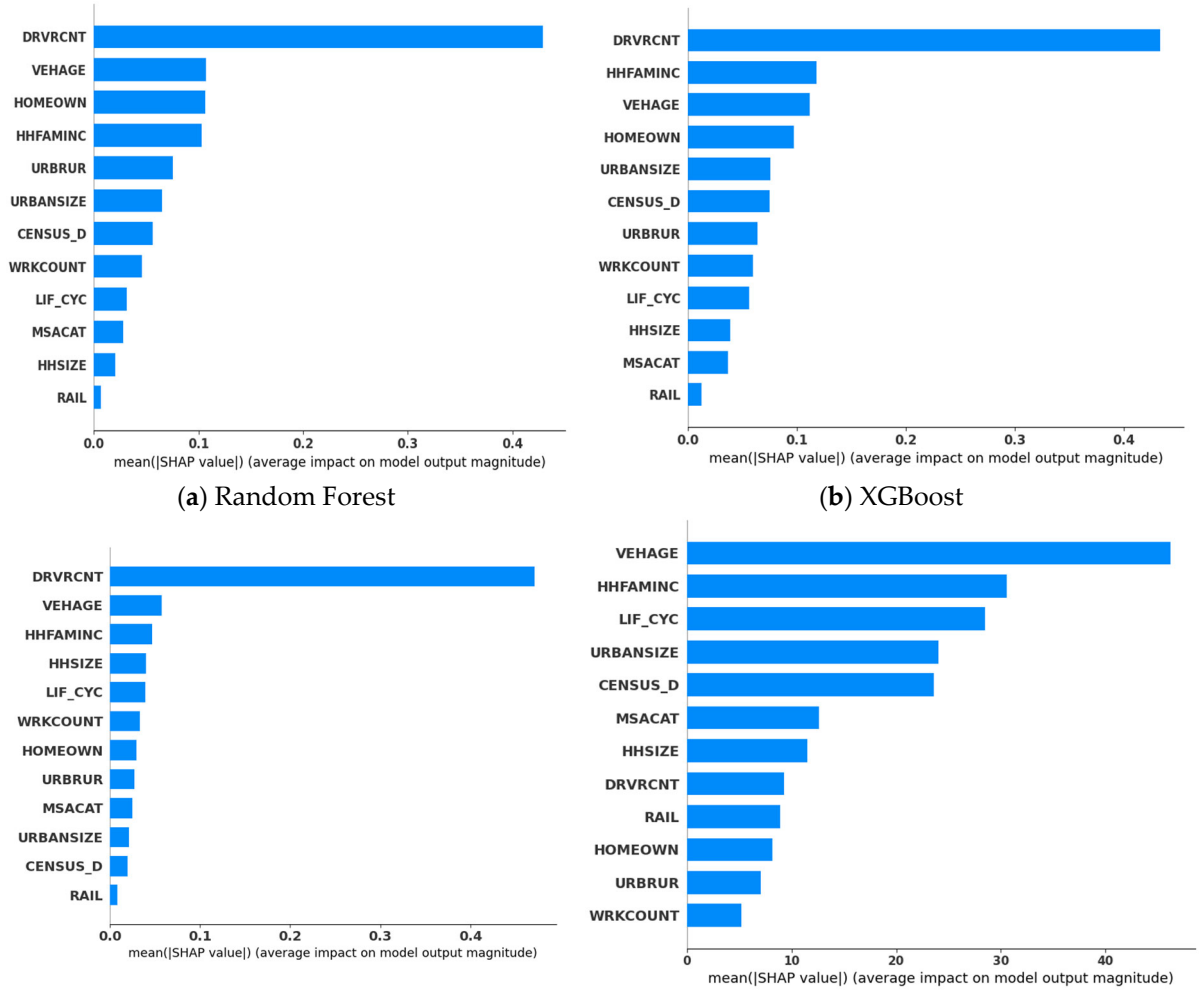


Figure 5. SHAP summary bar plots for the four models: (**a**) Random Forest feature importance plot, (**b**) XGBoost feature importance plot, (**c**) SVM feature importance plot, and (**d**) Naïve Bayes feature importance plot. Bars represent the mean absolute SHAP values, reflecting each feature's average contribution to model predictions. Higher values denote greater influence on distinguishing between single- and multi-vehicle households.

In contrast, the Naïve Bayes model presents a flatter importance curve. VEHAGE leads the ranking, followed by HHFAMINC and LIF_CYC, suggesting a preference for broad socio-demographic signals. DRVRCNT appears in a lower position, illustrating the model's limitations in capturing feature interactions. Spatial features such as URBANSIZE and URBRUR are present but exert limited influence. Overall, the Naïve Bayes model provides a baseline level of interpretability but lacks the flexibility to fully reflect the joint impact of multiple household and environmental factors.

In summary, Random Forest and XGBoost offer richer interpretability, recognizing both dominant and context-dependent features. In contrast, SVM and Naïve Bayes exhibit limited capacity to capture the full behavioral complexity of vehicle ownership, either by over-relying on a single predictor or failing to accommodate nonlinear interactions among features.

4.4.2. Comparative Model Explainability

The comparative SHAP analysis across Naive Bayes, Random Forest, SVM, and XGBoost models provides robust insight into the key drivers of multi-vehicle ownership and reveals how different algorithmic structures prioritize these drivers. Despite varying in model architecture and complexity, all four models converge on a small set of features as primary predictors, most notably DRVRCNT, HHFAMINC, VEHAGE, and HOMEOWN. Naive Bayes, while computationally efficient, displayed a flatter distribution of SHAP values and weaker representation of interaction effects. It highlighted variables such as VEHAGE and HHFAMINC, suggesting that it captures direct effects but fails to exploit joint or nonlinear relationships particularly among geographic and behavioral predictors like URBANSIZE and LIF_CYC.

In contrast, the Random Forest model demonstrated a more nuanced and balanced feature importance structure. While DRVRCNT was dominant, other variables such as VEHAGE, HOMEOWN, and URBRUR also contributed meaningfully. Random Forest's ability to model higher-order interactions and nonlinearities likely enabled it to integrate the influence of spatial and demographic context more effectively than Naive Bayes. The SVM model produced the most top-heavy SHAP profile. It relied overwhelmingly on DRVRCNT, with a sharp drop-off in the contributions of other features. This indicates that the kernelized decision boundary was formed primarily around one highly predictive feature, potentially limiting the interpretive value of the model. While this behavior explains the model's strong recall performance, it raises concerns about over-reliance on a single variable and underutilization of multi-dimensional contexts. Finally, the XGBoost model offered a hybrid profile. It preserved the dominance of DRVRCNT but allowed for more graduated influence from other predictors, such as HHFAMINC, VEHAGE, URBANSIZE, and CENSUS_D. This balance reflects the model's additive gradient boosting mechanism, which enables iterative refinement of predictor influence and typically yields superior generalization and interpretability.

Collectively, these findings demonstrate both convergence and divergence in model behavior. The convergence on core socio-demographic variables, especially DRVRCNT, HHFAMINC, and VEHAGE, underscores their structural importance in vehicle ownership decisions. The divergence in secondary features and the depth of their influence across models highlights the methodological trade-offs between model accuracy and interpretability. From a policy and behavioral modeling standpoint, ensemble methods such as XGBoost and Random Forest offer a favorable compromise. They achieve high predictive accuracy while preserving transparency via post hoc interpretability tools like SHAP. These properties make them especially useful in applications where both predictive performance and policy-relevant insights are necessary such as forecasting multi-vehicle ownership trends in a post-pandemic environment marked by volatility in travel behavior, telecommuting patterns, and household location choices.

4.5. SHAP Violin Plot Analysis: Influential Predictors of Vehicle Ownership Across Models

To uncover and critically examine the internal logic of machine learning predictions regarding multi-vehicle ownership, this section utilizes SHapley Additive exPlanations (SHAP) applied to four distinct classification models: Random Forest, XGBoost, Support Vector Machine (SVM), and Naïve Bayes. SHAP values offer a model-agnostic framework grounded in cooperative game theory, enabling both local explanations (instance-level) and global insights (feature-level aggregates). Figure 6 presents SHAP summary plots, violin plots that illustrate both the magnitude and direction of each feature's influence, as well as the distribution of feature values.

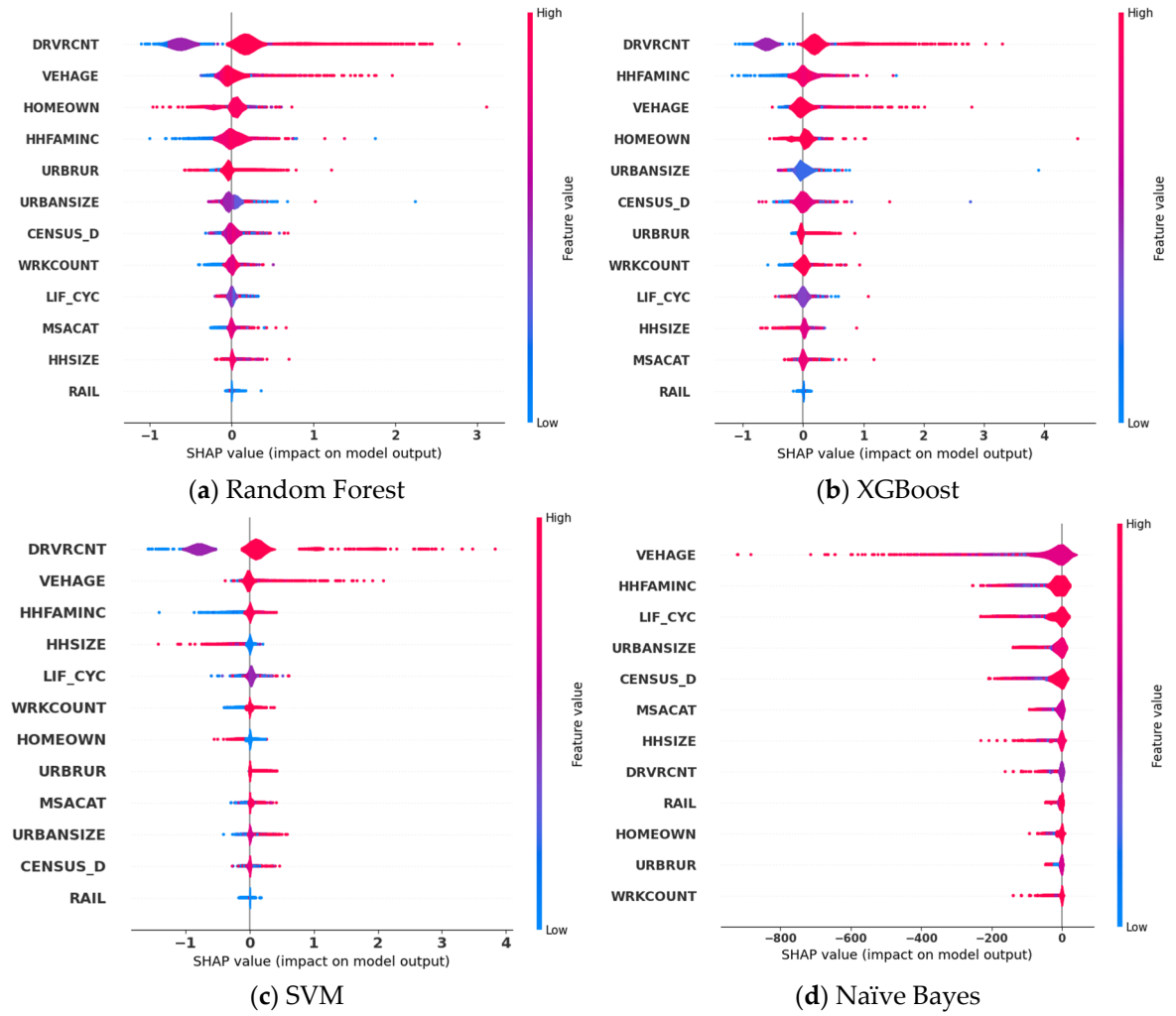


Figure 6. SHAP summary violin plots for the four models: (**a**) Random Forest, (**b**) XGBoost, (**c**) SVM (RBF), and (**d**) Naïve Bayes. These plots illustrate the distribution of SHAP values per feature, with color indicating the feature value (red for high, blue for low). The horizontal axis represents the impact on model output, showing how features influence the prediction of multi-vehicle ownership.

4.5.1. Random Forest: Interpretive Breadth and Behavioral Plausibility

In the Random Forest model, the number of licensed drivers (DRVRCNT) overwhelmingly emerges as the dominant predictor. Its strong positive SHAP values across a wide range of observations indicate a stable behavioral relationship: as intra-household mobility needs increase (proxied by DRVRCNT), the probability of owning multiple vehicles correspondingly rises. This is consistent with household utility maximization frameworks in travel behavior theory, where vehicle availability serves to reduce scheduling conflicts and increase accessibility.

The age of household vehicles (VEHAGE) exhibits a bimodal distribution in its SHAP profile, a particularly insightful result. Older vehicles tend to be positively associated with multi-vehicle ownership, potentially due to households retaining older cars for secondary or utilitarian purposes. In contrast, newer vehicles are generally associated with lower probabilities of owning additional vehicles, suggesting that new-car households may optimize usage or substitute with shared mobility. This bimodality captures longitudinal vehicle acquisition behavior, which static models often fail to represent.

HOMEOWN and HHFAMINC proxies for financial capital and residential stability show strong positive SHAP contributions, reinforcing empirical findings in transportation economics that link wealth accumulation and homeownership with private vehicle

accumulation. Importantly, spatial–contextual features such as URBRUR and URBANSIZE exhibit clear effects: residents of rural and low-density urban areas are more likely to own multiple vehicles, likely due to modal substitution effects in car-dependent geographies.

Variables like MSACAT, RAIL, WRKCOUNT, HHSIZE, and CENSUS_D offer second-order effects, but Random Forest's non-parametric nature allows these variables to interact in complex, nonlinear ways, capturing latent heterogeneity across geographic and socio-demographic strata.

4.5.2. XGBoost: Enhanced Sensitivity to Interaction and Nonlinear Effects

The XGBoost model, a gradient-boosted decision tree ensemble, confirms many of the insights from Random Forest but provides a finer delineation of interaction-sensitive effects. Once again, DRVRCNT is the top-ranked feature, but its influence appears more conditioned on interactions especially with HHFAMINC and URBANSIZE. This suggests that mobility need alone is not sufficient; rather, it interacts with economic resources and spatial structure to determine whether additional vehicles are acquired.

VEHAGE and HHFAMINC are highly dispersed in SHAP space, indicating strong but context-dependent effects. For example, older vehicles in high-income households may reflect vehicle retention, while in low-income households, they may reflect constrained substitution options. Similarly, high-income renters may behave differently from high-income homeowners in their vehicle ownership decisions effects that are captured in XGBoost's additive boosting framework.

XGBoost also uncovers nuanced influences of features like RAIL, MSACAT, and CENSUS_D, which are often considered weak predictors in traditional models but here emerge as moderators of more dominant variables. This reinforces the argument that ensemble models not only increase accuracy but also deepen interpretive resolution, especially in high-dimensional behavioral systems.

4.5.3. SVM: Focused Predictive Structure with Reduced Context Sensitivity

The SVM model, despite comparable predictive accuracy, demonstrates a narrow interpretive focus. The SHAP analysis reveals that most of the model's predictive variance is concentrated in just a few features, chiefly DRVRCNT, followed by VEHAGE and HHFAMINC. These features maintain a positive association with multi-vehicle ownership, but the model's insensitivity to broader spatial or household composition variables suggests a lack of contextual responsiveness.

This reflects the high-dimensional projection behavior of the RBF kernel, which excels at creating complex decision boundaries but often obscures feature interactions unless explicitly modeled. As a result, while SVM performs well in classification accuracy, it contributes relatively little to behavioral inference, limiting its utility in transportation planning contexts where policy relevance requires nuanced causal explanations.

4.5.4. Naïve Bayes: Simplicity at the Expense of Structural Realism

The Naïve Bayes classifier, based on the assumption of conditional independence among predictors, produces a notably centralized SHAP distribution. Only VEHAGE, HHFAMINC, and to a lesser extent LIF_CYC exhibit discernible influence. Most other variables including those known to be behaviorally significant (e.g., URBRUR, WRKCOUNT, MSACAT) cluster around zero, indicating that Naïve Bayes is unable to capture joint effects or multi-feature dependencies. While its low computational cost and probabilistic outputs make it useful for benchmarking or probabilistic modeling, its inability to accommodate complex interdependencies renders it less appropriate for nuanced inference in socio-spatial systems like post-COVID-19 vehicle ownership.

Table 4 provides a comparative overview of SHAP-based feature importance scores, illustrating model-specific sensitivity to both household-level and spatial variables.

Table 4. SHAP-based mean feature importance across models.

Feature	Random Forest	XGBoost	SVM (RBF)	Naïve Bayes
DRVRCNT (Drivers)	0.231	0.245	0.265	0.098
HHFAMINC (Income)	0.187	0.195	0.128	0.087
VEHAGE (Vehicle Age)	0.156	0.168	0.115	0.073
HOMEOWN (Homeownership)	0.141	0.134	0.062	0.065
URBRUR (Urban/Rural)	0.119	0.107	0.041	0.058
URBANSIZE	0.102	0.103	0.037	0.051
LIF_CYC (Lifecycle)	0.097	0.092	0.032	0.047
WRKCOUNT (Workers)	0.091	0.089	0.030	0.043
RAIL (Rail Access)	0.083	0.081	0.028	0.041
MSACAT (MSA Category)	0.078	0.075	0.026	0.039
HHSIZE (Household Size)	0.072	0.070	0.023	0.037
CENSUS_D (Division)	0.066	0.061	0.020	0.032

Overall, the SHAP analysis offers valuable insights into the roles that various socio-demographic and spatial factors play in shaping household vehicle ownership. Key features such as the number of drivers, vehicle age, household income, and homeownership consistently emerge as important influences. Built environment characteristics and household composition also contribute meaningfully, though to a lesser extent. This interpretive transparency enhances our understanding of how different variables shape mobility choices in the post-pandemic context. While SHAP interaction values can formally quantify pairwise feature interactions, our study emphasizes distributional and dependence-based SHAP visualizations, such as summary and violin plots. These outputs inherently reveal interaction and nonlinear effects, including threshold behaviors and context-dependent influences, in a transparent and policy-relevant manner.

4.5.5. Comparative Insights from SHAP Violin Plot Analysis

The SHAP summary violin plots across the four models provide valuable comparative insights into how each algorithm captures the distributional behavior of explanatory variables in shaping household vehicle ownership classification. While the feature importance bar plots identify average magnitude, the violin plots reveal how feature values are distributed across decision boundaries, thereby offering a richer, behaviorally grounded interpretation. Naive Bayes displays the most compressed and symmetric SHAP distributions, with limited dispersion across most features. Although it correctly identifies key predictors such as VEHAGE and HHFAMINC, the distributions are narrowly centered and fail to reflect nonlinearities or interaction effects. The inability to differentiate high and low value impacts for features like LIF_CYC and URBANSIZE reflects the model's restrictive assumption of feature independence.

By contrast, the Random Forest model exhibits greater SHAP variance and asymmetry, especially for DRVRCNT, VEHAGE, and HOMEOWN. It effectively distinguishes the directional influence of high-value observations, such as the increased likelihood of multi-vehicle ownership in high-income, suburban, or multi-driver households. The Random Forest violin plot reveals rich, multimodal distributions, indicating that the model captures conditional heterogeneity and non-monotonic effects, hallmarks of ensemble tree-based modeling.

The SVM model shows extreme SHAP skewness toward DRVRCNT, with other features contributing minimally. While high feature values for driver count sharply increase model confidence, other features exhibit tight distributions around zero. This top-heavy pattern highlights the SVM's margin-maximization bias, which tends to overfit to dominant

dimensions while ignoring potentially valuable, context-dependent variability, making it less suitable for nuanced policy interpretation.

Finally, the XGBoost model offers the most balanced violin distribution. It captures both the dominant influence of DRVRCNT and the distributed, second-order effects of variables like HHFAMINC, VEHAGE, URBRUR, and WRKCOUNT. The model's ability to register subtle shifts in SHAP values across a wide spectrum of features implies that XGBoost detects joint effects and thresholds (e.g., income saturation points or age-of-vehicle tipping points), which are critical for post-COVID-19 behavioral modeling. In synthesis, the violin plots affirm several key conclusions:

- Models differ not only in what features they prioritize, but in how they distribute predictive influence across populations and contexts.
- Random Forest and XGBoost offer the most interpretable and policy-relevant patterns, identifying both core and contextual determinants of vehicle ownership.
- Naive Bayes underestimates feature variance due to structural assumptions, while SVM over-concentrates on high-margin predictors.
- Only ensemble methods capture the combined influence of demographic, spatial, and behavioral variables, making them better suited for explanatory modeling in a post-pandemic transportation context.

These findings provide empirical support for the argument that explainability must go beyond feature ranking. Distributional interpretation, as afforded by SHAP violin plots, reveals not only which features matter but also how and under what conditions they exert influence. This level of insight is essential for informing transportation policy, infrastructure investment, and urban mobility planning in an increasingly behaviorally complex landscape.

Beyond these comparative strengths and limitations, the violin plots also revealed patterns that are not immediately intuitive yet hold important behavioral implications. Vehicle age (VEHAGE) displayed a dual influence: households retaining older cars were more likely to acquire additional vehicles as a precautionary strategy, while households with newer cars also exhibited a heightened probability of multi-vehicle ownership, reflecting aspirational upgrading. Similarly, while rural households predictably leaned toward multi-vehicle ownership, the plots also showed elevated probabilities among certain metropolitan households. This indicates that high-income or multi-worker families in transit-poor neighborhoods maintain strong incentives for additional vehicles despite living in dense urban contexts. These patterns underscore how SHAP-based distributional analysis uncovers hidden heterogeneity and threshold effects that refine our understanding of household mobility behavior.

5. Discussion

The findings of this study contribute both empirical clarity and theoretical advancement to the understanding of household vehicle ownership in the post-COVID-19 era. By combining multiple machine learning classifiers with SHAP-based interpretability, the analysis not only achieved high predictive performance but also offered insights into the behavioral mechanisms underlying vehicle acquisition. This dual emphasis on predictive accuracy and interpretive depth represents a methodological step forward in travel behavior research, particularly given the pandemic's disruption to established mobility patterns.

Across all four models, Random Forest, XGBoost, Support Vector Machine, and Naïve Bayes, a core set of features consistently emerged as the strongest predictors: number of licensed drivers, household income, and vehicle age. These findings reaffirm long-standing principles in transportation economics, where mobility demand is shaped by household resources, intra-household activity patterns, and capital replacement cycles. The

prominence of number of drivers highlights a fundamental behavioral truth: additional licensed drivers create structural pressure for additional vehicles. Similarly, household income provides the financial capacity to sustain multiple vehicles, while vehicle age reflects both capital replacement and adaptive strategies under uncertainty.

Beyond this convergence, the models diverged in their treatment of second-tier variables. Ensemble approaches, particularly Random Forest and XGBoost, identified residential stability, urban context, and lifecycle stage as significant predictors, while Support Vector Machine and Naïve Bayes attenuated these effects. This divergence is not purely technical; it reflects how different model architectures capture behavioral complexity. Ensemble methods, by accommodating interaction effects and collinearity, were better able to reflect the heterogeneity of household decision-making, while probabilistic and kernel-based models offered more constrained behavioral representations.

The SHAP analysis revealed that post-pandemic ownership decisions are not driven by economic rationality alone. Variables such as homeownership, urban–rural context, and household lifecycle gained predictive prominence, suggesting that risk aversion, residential relocation, telecommuting, and reduced trust in shared transport have reshaped household mobility strategies. The retention of older vehicles, indicated by the rising influence of vehicle age, further suggests adaptive responses to uncertainty, aligning with behavioral economics literature on bounded rationality under disruption.

These results carry direct implications for policy. Conventional instruments such as transit subsidies, vehicle taxation, or parking management often assume households respond primarily to marginal cost changes. The findings of this study suggest otherwise: vehicle acquisition is increasingly embedded in household structure, residential immobility, and risk perception. Effective interventions will therefore need to be geographically and demographically differentiated. Examples include bundling multimodal accessibility with telework incentives, designing family-oriented car-sharing programs, or providing targeted support for caregivers whose mobility needs amplify auto-dependence.

The diagnostic strength of SHAP lies in its ability to reveal not only which variables matter but also the conditional contexts under which transitions from single to multiple vehicles are most likely. This diagnostic value enhances both policy precision and public legitimacy. By moving from aggregate strategies toward household-specific interventions, planners can design more equitable and acceptable policies.

Methodologically, the study demonstrates the value of comparative explainability as a deliberate strategy. Performance metrics such as accuracy or F1 score, while important, are insufficient for policy-sensitive research. Structural transparency and behavioral fidelity must also be considered. The comparative analysis showed that ensemble models provided richer interpretive granularity, while probabilistic and kernel-based classifiers offered more limited behavioral expressiveness. This underscores that model selection is not only a statistical decision but also an epistemic one that shapes the narratives extractable from data.

The broader implications of these findings align closely with the United Nations Sustainable Development Goals [67]. The growth of multi-vehicle households in suburban and rural areas connects directly to SDG 11 on sustainable cities and communities, emphasizing the need for rural mobility innovations and context-sensitive transit strategies [68]. The influence of vehicle age highlights the climate relevance of SDG 13, calling for equity-sensitive scrappage and fleet renewal programs. Household income and homeownership point to challenges under SDG 10 on reducing inequalities, where affluent households consolidate resilience while vulnerable groups remain transit-dependent. The integration of explainable machine learning also supports SDG 16 on strong institutions by advancing transparency and accountability in policy modeling. Finally, lifecycle and caregiving variables under-

score the relevance of SDG 3 on health and well-being and SDG 5 on gender equality, where household mobility decisions intersect with family structure and care responsibilities. Taken together, the findings provide a diagnostic framework that translates predictive insights into policy instruments. Table 5 illustrates how the most influential predictors can inform targeted strategies, ranging from employer-supported mobility programs to rural microtransit pilots and progressive vehicle taxation. Each policy pathway carries potential benefits and constraints, underscoring the need for adaptive design, multilevel coordination, and mitigation strategies to address political and institutional barriers.

Table 5. Policy instruments informed by SHAP insights, associated trade-offs, and mitigation strategies.

Policy Instrument (Derived from Findings)	Expected Benefit	Likely Constraint/Trade Off	Potential Mitigation Strategy
Employer-supported mobility programs, car-sharing, and commuter benefits targeting multi-driver households (DRVRCNT)	Reduced incremental vehicle demand; support for multi-worker families	Requires coordination with employers; uneven adoption across firms	Integrate with existing commuter tax benefits; incentivize employer participation through subsidies or credits
Income-sensitive scrappage incentives and targeted EV subsidies (VEHAGE)	Fleet modernization; lower emissions; improved equity	Risk of subsidies disproportionately favoring affluent households; fiscal burden	Calibrate incentives by household income and tenure; prioritize low- and middle-income multi-vehicle households
Progressive vehicle taxation, congestion pricing, and road pricing (HOMEOWN, HHFAMINC)	Reduces over-motorization among affluent households; generates revenue for sustainable mobility	Politically contentious; regressive risk if poorly designed	Use revenues to subsidize transit and shared mobility in low-income neighborhoods; exemptions or rebates for vulnerable households
Rural MaaS, demand-responsive microtransit, and last-mile connectivity (URBRUR, URBANSIZE)	Improved access in car-dependent rural and suburban areas; reduced equity gap	High service delivery cost; institutional fragmentation across jurisdictions	Pilot programs; regional coordination; flexible on-demand service models
Lifecycle-sensitive transit passes, childcare travel support, and family-responsive mobility programs (LIF_CYC, HHSIZE, WRKCOUNT)	Enhances household equity; addresses caregiving and commuting needs	Administrative complexity; cross-agency coordination required	Integrate with employer benefits; align with social policy programs and family support initiatives

Overall, the results demonstrate that vehicle ownership in the post-COVID-19 era is stratified along socioeconomic and spatial lines, reinforcing the risk of a two-tier mobility system. Addressing these inequities requires embedding behavioral insights into equity-oriented policy design. By leveraging explainable machine learning, this study offers both methodological advancement and actionable guidance for achieving sustainable and inclusive mobility futures.

6. Conclusions

This study examined the determinants of household vehicle ownership in the post-COVID-19 context by analyzing data from the 2022 National Household Travel Survey and applying a suite of interpretable machine learning models. Framing vehicle ownership as a binary classification problem distinguishing single-vehicle from multi-vehicle households, the analysis uncovered both established and emerging behavioral dynamics shaping acquisition decisions during a period of systemic disruption.

The results confirmed the enduring influence of household income, number of licensed drivers, and vehicle age, consistent with classical theories of travel demand. At the same time, spatial characteristics, lifecycle stage, and housing tenure emerged as increasingly relevant factors, reflecting pandemic-related changes in work location, residential mobility, and perceived risks associated with shared transport modes.

Methodologically, the use of SHAP-based interpretability enhanced both predictive performance and explanatory transparency. Ensemble models such as Random Forest and XGBoost captured interaction effects and behavioral heterogeneity more effectively than kernel-based or probabilistic baselines. By demonstrating how feature importance varies across household types and regional contexts, the models bridged the gap between advanced computation and actionable policy insight.

Despite these contributions, several limitations should be acknowledged. First, the analysis excluded zero-vehicle households. While justified analytically, this decision limits generalizability and weakens equity assessment, since such households are disproportionately low-income, immigrant, or urban core residents. Future studies should adopt multinomial ownership frameworks or parallel models that include zero-vehicle states to more fully capture structural and behavioral differences. Second, reliance on cross-sectional data prevents causal inference and masks temporal dynamics. Although associations identified are behaviorally consistent and statistically robust, they reflect only a snapshot of post-COVID-19 conditions. Longitudinal surveys, repeated measures, or quasi-experimental designs would provide stronger evidence of causal mechanisms and behavioral change over time. Third, the models did not incorporate attitudinal or perceptual factors such as environmental concern, risk perception, or lifestyle preferences. These are increasingly recognized as central to mobility choices and should be included in future research through hybrid choice models, latent class approaches, or psychometric surveys. Finally, fairness auditing was not conducted, leaving open the risk that algorithmic outputs could reinforce existing social inequities. Incorporating fairness metrics and bias detection frameworks will be essential for ensuring equitable mobility outcomes.

Beyond these limitations, the findings advance a broader agenda for policy-relevant, explainable, and equity-sensitive analytics in transportation. Anticipating household vehicle ownership is critical for forecasting parking demand, emissions, and infrastructure needs, as well as for shaping congestion management and accessibility strategies. As urban systems adapt to hybrid work, demographic change, and decarbonization mandates, models that combine behavioral realism with interpretive transparency will be indispensable.

In conclusion, this study contributes a behaviorally grounded and computationally robust framework for analyzing household vehicle ownership in the post-pandemic era. It demonstrates how machine learning can be used not only to predict but also to explain and justify outcomes, thereby supporting sustainable, inclusive, and context-responsive transport policy.

Author Contributions: Conceptualization, M.H.; methodology, M.H.; software, S.S.S.; validation, M.H., M.B.A., F.A. and Z.N.; formal analysis, S.S.S.; investigation, F.A.; resources, Z.N.; data curation, S.S.S.; writing—original draft preparation, M.H.; writing—review and editing, M.B.A., S.S.S.; F.A., Z.N.; visualization, M.H. and S.S.S.; supervision, Z.N.; project administration, Z.N.; funding acquisition, M.B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study didn't receive any external funds.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were obtained from the 2022 NextGen National Household Travel Survey Core Data, conducted by the Federal Highway Administration (FHWA), U.S. Department of Transportation, Washington, DC. The dataset is publicly available at http://nhts.ornl.gov (accessed on 24 December 2024) and is distributed under the U.S. Government's Public Domain license, permitting free use, distribution, and reproduction for research and educational purposes. The official citation is as follows: *Federal Highway Administration. (2022). 2022*

NextGen National Household Travel Survey Core Data, U.S. Department of Transportation, Washington, DC. Available online: http://nhts.ornl.gov.

Acknowledgments: This research was supported by the "University of Debrecen Program for Scientific Publication".

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Moody, J.; Farr, E.; Papagelis, M.; Keith, D.R. The value of car ownership and use in the United States. *Nat. Sustain.* **2021**, *4*, 769–774. [CrossRef]
2. Taghvaee, V.M.; Arani, A.A.; Nodehi, M.; Shirazi, J.K.; Agheli, L.; Ghojogh, H.M.N.; Salehnia, N.; Mirzaee, A.; Taheri, S.; Saber, R.M.; et al. Sustainable development goals: Transportation, health and public policy. *Rev. Econ. Politi. Sci.* **2021**, *8*, 134–161. [CrossRef]
3. Fried, T.; Tun, T.H.; Klopp, J.M.; Welle, B. Measuring the Sustainable Development Goal (SDG) transport target and accessibility of Nairobi's matatus. *Transp. Res. Rec. J. Transp. Res. Board* **2020**, *2674*, 196–207. [CrossRef]
4. Hermelin, B.; Henriksson, M. Transport and mobility planning for sustainable development. *Plan. Pract. Res.* **2022**, *37*, 527–531. [CrossRef]
5. Fang, H.A. A discrete–continuous model of households' vehicle choice and usage, with an application to the effects of residential density. *Transp. Res. Part B Methodol.* **2008**, *42*, 736–758. [CrossRef]
6. Anowar, S.; Eluru, N.; Miranda-Moreno, L.F. Alternative modeling approaches used for examining automobile ownership: A comprehensive review. *Transp. Rev.* **2014**, *34*, 441–473. [CrossRef]
7. Wang, X.; Shaw, F.A.; Mokhtarian, P.L.; Watkins, K.E. Response willingness in consecutive travel surveys: An investigation based on the National Household Travel Survey using a sample selection model. *Transportation* **2022**, *50*, 2339–2373. [CrossRef]
8. Cresswell, T. Valuing mobility in a post COVID-19 world. *Mobilities* **2020**, *16*, 51–65. [CrossRef]
9. Cirillo, C.; Liu, Y. Vehicle ownership modeling framework for the state of Maryland: Analysis and trends from 2001 and 2009 NHTS data. *J. Urban Plan. Dev.* **2013**, *139*, 1–11. [CrossRef]
10. Paul, J. Temporary versus Permanent Pandemic Transit Leavers: Findings from the 2022 US National Household Travel Survey. *Findings* **2024**. [CrossRef]
11. Nithila, A.N.; Mitra, S. Rural-Urban Differences in Older Adults' Travel Behavior in the Us: Evidence from the 2022 National Household Travel Survey. *SSRN* **2025**, 5272765. [CrossRef]
12. Buehler, R.; Pucher, J. Socioeconomic variations in walking rates in the United States: Recent evidence from the 2022 National Household Travel Survey. *J. Transp. Health* **2024**, *38*, 101875. [CrossRef]
13. Cairns, D.; França, T.; Calvo, D.M.; de Azevedo, L. An immobility turn? The Covid-19 pandemic, mobility capital and international students in Portugal. *Mobilities* **2021**, *16*, 874–887. [CrossRef]
14. Bricka, S.; Reuscher, T.; Schroeder, P.; Fisher, M.; Beard, J.; Sun, L. Summary of Travel Trends: 2022 National Household Travel Survey. 2024. Available online: https://rosap.ntl.bts.gov/view/dot/73764 (accessed on 17 September 2025).
15. Hassan, M.; Al Nafees, A.; Shraban, S.S.; Paul, A.; Mahin, H.D. Application of machine learning in intelligent transport systems: A comprehensive review and bibliometric analysis. *Discov. Civ. Eng.* **2025**, *2*, 98. [CrossRef]
16. Hagenauer, J.; Helbich, M. A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Syst. Appl.* **2017**, *78*, 273–282. [CrossRef]
17. Xu, F. Data-Driven Modeling and Optimization of Active Travel Behavior Under Digital Distraction. *Intell. Transp. Smart Cities* **2025**, *70*, 232–240. [CrossRef]
18. Ali, N.F.M.; Sadullah, A.F.M.; Majeed, A.P.A.; Razman, M.A.M.; Zakaria, M.A.; Nasir, A.F.A. Travel Mode Choice Modeling: Predictive Efficacy between Machine Learning Models and Discrete Choice Model. *Open Transp. J.* **2021**, *15*, 241–255. [CrossRef]
19. Wang, J.; Lian, Z.; Feng, T.; Tang, L.; Liu, K. A review and outlook of machine learning-based travel choice behavior research. *Prog. Geogr.* **2024**, *43*, 1649–1665. [CrossRef]
20. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2021**, *55*, 3503–3568. [CrossRef]
21. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
22. Eiksa, K.; Vatne, J.E.; Lekkas, A.M. Explaining Deep Reinforcement Learning Policies with SHAP, Decision Trees, and Prototypes. In Proceedings of the 2024 32nd Mediterranean Conference on Control and Automation, MED 2024, Crete, Greece, 11–14 June 2024; pp. 700–705. [CrossRef]
23. Li, M.; Sun, H.; Huang, Y.; Chen, H. Shapley value: From cooperative game to explainable artificial intelligence. *Auton. Intell. Syst.* **2024**, *4*, 2. [CrossRef]

24. Mersha, M.; Lam, K.; Wood, J.; AlShami, A.K.; Kalita, J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* **2024**, *599*, 128111. [CrossRef]
25. Malwade, S.S.; Budhavale, S.J. Exploring Explainable AI: Current Trends, Challenges, Techniques and its Applications. In Proceedings of the ICIMMI '23: Proceedings of the 5th International Conference on Information Management & Machine Intelligence, Jaipur, India, 23–25 November 2023; Association for Computing Machinery: New York, NY, USA, 2023. [CrossRef]
26. Berri, A. A cross-country comparison of household, car ownership: A Cohort Analysis. *IATSS Res.* **2009**, *33*, 21–38. [CrossRef]
27. Sillaparcharn, P. Modeling of vehicle ownership: Case study of Thailand. *Transp. Res. Rec.* **2007**, *2038*, 98–104. [CrossRef]
28. Bhat, C.R.; Eluru, N. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transp. Res. Part B Methodol.* **2009**, *43*, 749–765. [CrossRef]
29. Ha, T.V.; Asada, T.; Arimura, M. Determination of the influence factors on household vehicle ownership patterns in Phnom Penh using statistical and machine learning methods. *J. Transp. Geogr.* **2019**, *78*, 70–86. [CrossRef]
30. Bas, J.; Cirillo, C.; Cherchi, E. Classification of potential electric vehicle purchasers: A machine learning approach. *Technol. Forecast. Soc. Change* **2021**, *168*, 120759. [CrossRef]
31. Zambang, M.A.M.; Jiang, H.; Wahab, L. Modeling vehicle ownership with machine learning techniques in the Greater Tamale Area, Ghana. *PLoS ONE* **2021**, *16*, e0246044. [CrossRef]
32. Tangirala, S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 612–619. [CrossRef]
33. Naseri, H.; Waygood, E.; Wang, B.; Patterson, Z. Interpretable machine learning approach to predicting electric vehicle buying decisions. *Transp. Res. Rec. J. Transp. Res. Board* **2023**, *2677*, 704–717. [CrossRef]
34. Ma, M.; Pinsky, E. Using machine learning to identify primary features in choosing electric vehicles based on income levels. *J. Inf. Technol. Data Manag.* **2024**, *7*, 1–6. [CrossRef]
35. Dixon, J.; Koukoura, S.; Brand, C.; Morgan, M.; Bell, K. Spatially disaggregated car ownership prediction using deep neural networks. *Futur. Transp.* **2021**, *1*, 113–133. [CrossRef]
36. Ali, A.; Kalatian, A.; Choudhury, C.F. Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions. *Transp. Res. Part A Policy Pract.* **2023**, *173*, 103727. [CrossRef]
37. Xu, Z.; Aghaabbasi, M.; Ali, M.; Macioszek, E. Targeting sustainable transportation development: The support vector machine and the Bayesian optimization algorithm for classifying household vehicle ownership. *Sustainability* **2022**, *14*, 11094. [CrossRef]
38. Shah, S.; Rajiv, R.M.; Lokre, A. Moving Toward Gender-Equitable Transportation in Post-COVID-19 Urban South Asia. *Transp. Res. Rec. J. Transp. Res. Board* **2022**, *2677*, 865–879. [CrossRef]
39. Soltani, A.; Azmoodeh, M.; Doostvandi, M.; Ahmadi, A.S.; Rahimi, M. Post-COVID-19 campus commuting patterns and influential factors: Evidence from a developing country. *Transp. Plan. Technol.* **2024**, *47*, 566–597. [CrossRef]
40. Hsieh, H.S. Understanding post-COVID-19 hierarchy of public transit needs: Exploring relationship between service attributes, satisfaction, and loyalty. *J. Transp. Health* **2023**, *32*, 101656. [CrossRef]
41. Feng, F.; Anastasopoulos, P.C.; Guo, Y.; Wang, W.; Peeta, S.; Li, X. Willingness to use ridesplitting services for home-to-work morning commute in the post-COVID-19 era. *Ransportation* **2024**, 1–34. [CrossRef]
42. Mahin, H.D.; Hassan, M.; An, H.K.; Sameer, M. Socioeconomic Mobility and Urban Travel Patterns in Post-Covid US: 2022 Nhts. *SSRN* **2025**, 5192486.
43. Navarro, D. Learning Statistics with R. 2013. Available online: https://books.google.com/books?hl=en&lr=&id=rV5pCQAAQBAJ&oi=fnd&pg=PA3&dq=R+statistics&ots=4JSv96zvOI&sig=l5f7nDKClAnYVtkOQklBpKscPvM (accessed on 31 July 2025).
44. Stopher, P.R.; Greaves, S.P. Guidelines for samplers: Measuring a change in behaviour from before and after surveys. *Transportation* **2006**, *34*, 1–16. [CrossRef]
45. Larson, M.G. Descriptive statistics and graphical displays. *Circulation* **2006**, *114*, 76–81. [CrossRef]
46. Cabello-Solorzano, K.; de Araujo, I.O.; Peña, M.; Correia, L.; Tallón-Ballesteros, A.J. The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In *18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023)*; Springer: Berlin/Heidelberg, Germany, 2023; Volume 750, pp. 344–353. [CrossRef]
47. Byun, J.; Lin, X.; Ward, J.; Cheng, G. Risk In Context: Benchmarking Privacy Leakage of Foundation Models in Synthetic Tabular Data Generation. *arXiv* **2025**, arXiv:2507.17066. [CrossRef]
48. Anable, J.; Brown, L.; Docherty, I.; Marsden, G. *Less is More: Changing Travel in a Post-Pandemic Society*; Centre for Research into Energy Demand Solutions: Oxford, UK, 2022. Available online: https://www.creds.ac.uk/wp-content/uploads/CREDS-Less-is-more-web.pdf (accessed on 1 August 2025).
49. Shivam, K. *Modelling Activity Generation and Scheduling Decisions, and Exploring the Interplay with Telecommuting*; University of British Columbia: Vancouver, BC, Canada, 2024. Available online: https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/24/items/1.0443912 (accessed on 1 August 2025).

50. Gao, Y.; Zhao, P. Tracing long-term commute mode choice shifts in Beijing: Four years after the COVID-19 pandemic. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1566. [CrossRef]
51. Javadinasr, M.; Rahimi, E.; Mohammadian, A.K. Travel behavior and extreme events: Cases of the Covid-19 pandemic and transit disruption and their impacts on activity patterns. In *Handbook of Travel Behaviour*; Edward Elgar Publishing: Cheltenham, UK, 2024. Available online: https://www.elgaronline.com/edcollchap/book/9781839105746/book-part-9781839105746-32.xml (accessed on 1 August 2025).
52. Obeid, H. Causal Inference, Transportation, and Travel Demand: A Conceptual Review with Applications Using Observational and Experimental Data. 2022. Available online: https://search.proquest.com/openview/6e9de81d24865c08698140800b77ce2d/1?pq-origsite=gscholar&cbl=18750&diss=y (accessed on 1 August 2025).
53. Thompson, C.G.; Kim, R.S.; Aloe, A.M.; Becker, B.J. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic Appl. Soc. Psychol.* **2017**, *39*, 81–90. [CrossRef]
54. Cheng, J.; Sun, J.; Yao, K.; Xu, M.; Cao, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2021**, *268*, 120652. [CrossRef] [PubMed]
55. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
56. Naidu, G.; Zuva, T.; Sibanda, E.M. A review of evaluation metrics in machine learning algorithms. In *Artificial Intelligence Application in Networks and Systems*; Springer: Berlin/Heidelberg, Germany, 2023; Volume 724, pp. 15–25. [CrossRef]
57. Liu, Y.; Wang, Y.; Zhang, J. New machine learning algorithm: Random forest. In *Information Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7473, pp. 246–252. [CrossRef]
58. Demir, S.; Şahin, E.K. Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter. *Environ. Earth Sci.* **2022**, *81*, 459. [CrossRef]
59. Salcedo-Sanz, S.; Rojo-Álvarez, J.L.; Martínez-Ramón, M.; Camps-Valls, G. Support vector machines in engineering: An overview. *WIREs Data Min. Knowl. Discov.* **2014**, *4*, 234–267. [CrossRef]
60. Wickramasinghe, I.; Kalutarage, H. Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Comput.* **2020**, *25*, 2277–2293. [CrossRef]
61. Makhtar, M.; Neagu, D.C.; Ridley, M.J. Binary classification models comparison: On the similarity of datasets and confusion matrix for predictive toxicology applications. In *International Conference on Information Technology in Bio- and Medical Informatics*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6865, pp. 108–122. [CrossRef]
62. Noferesti, V.; Mirzahossein, H. Leveraging Machine Learning to Predict Residential Location Choice: A Comparative Analysis. *Results Eng.* **2025**, *25*, 104214. [CrossRef]
63. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 2020. Available online: https://arxiv.org/pdf/1811.12808 (accessed on 1 August 2025).
64. Branco, P.; Torgo, L.; Ribeiro, R.P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* **2016**, *49*, 31. [CrossRef]
65. Markatou, M.; Tian, H.; Biswas, S.; Hripcsak, G.M. Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.* **2005**, *6*, 1127–1168.
66. Bhatnagar, S.; Agrawal, R. Understanding explainable artificial intelligence techniques: A comparative analysis for practical application. *Bull. Electr. Eng. Inform.* **2024**, *13*, 4451–4455. [CrossRef]
67. Hassan, M.; Mahin, H.D.; Ahmed, F.; Rahaman, A.; Abdullah, M. Assessing Public Transit Network Efficiency and Accessibility in Johor Bahru and Penang, Malaysia: A Data-Driven Approach. *Results Eng.* **2025**, *27*, 106126. [CrossRef]
68. Dugarova, E.; Gülasan, N. *Challenges and Opportunities in the Implementation of the Sustainable Development Goals*; One United Nations Plaza: New York, NY, USA, 2017; Available online: https://cdn.unrisd.org/assets/library/books/pdf-files/global-trends_undp-and-unrisd_final.pdf (accessed on 17 September 2025).