Modeling Car-Free Household Prevalence in U.S. Census Block Groups Using Socioeconomic and Spatial Indicators from the Smart Location Database

Lukas Becker

Universität Leipzig

Probabilistic Machine Learning (SoSe 2025)

Dr. Alvaro Diaz-Ruelas

August 15, 2025

Modeling Car-Free Household Prevalence in U.S. Census Block Groups Using Socioeconomic and Spatial Indicators from the Smart Location Database

Introduction

In recent years, transportation equity and sustainability have become central concerns in urban planning and policy. Among the key indicators of social and infrastructural access is household vehicle ownership, particularly the percentage of households without access or need to a private vehicle. These zero-vehicle households are often more dependent on public transit and walkable environments, and their distribution is shaped by a complex mix of socio-economic, spatial, and infrastructural variables.

This research aims to predict the percentage of car-free households in new and changed US Census Block Groups (CBGs). It is using the US EPA's Smart Location Database (SLD), a rich source of standardized spatial and socio-economic indicators. The focus is on identifying how evolving social, economic, or political changes, such as new transit infrastructure, economic shifts, or population dynamics, can influence car ownership trends.

To model these dynamics, this study employs a range of machine learning techniques, from linear models such as Ordinary Least Squares (OLS), Lasso, Ridge, Elastic Net, and Bayesian Ridge regression to more sophisticated approaches including k-Nearest Neighbors, Random Forest, and XGBoost. The goal is to determine which models best capture the nuanced, multidimensional patterns influencing car ownership at the CBG level.

The analysis will be conducted using a carefully selected subset of variables from the SLD dataset, chosen for their theoretical relevance and practical predictive potential. These features span demographic composition, employment density, urban design, and transit access. All of which are thought to influence household decisions regarding car ownership.

Data Exploration and Preprocessing

Table 1

Variables used in the prediction model with descriptions and relevance.

Column Name	Description	Relevance
CBSA_POP	Total population in the CBSA	Indicates scale of urbanization, which affects transit and car dependency
CBSA_EMP	Total employment in CBSA	Economic size often correlates with public transportation investment and job accessibility
CBSA_WRK	Workers living in the CBSA	Reflects commuting patterns and urban form
TOTAL_POPULATION	Total population of the CBG	More people can imply denser neighborhoods with better transit
HOUSEHOLDS	Number of occupied households	Base unit for computing car ownership and household ratios
P_WORKING_AGE	Percent of population aged 18 to 64	Working-age population drives demand for work-related travel
P_ZERO_CARS_HOUSEHOLDS	**Percent of households with zero cars**	**Target variable**
WORKERS	Count of workers living in CBG	High numbers of workers may increase car ownership unless transit is strong
P_LOW_WAGE_WORKERS	Share of workers earning $\leq$\$1,250/month	Low-income households are more likely to be car-free
TOTAL_EMPLOYMENT	Number of jobs located in the CBG	More nearby jobs may reduce need for car commuting
P_LOW_WAGE_EMPLOYMENT	Share of local jobs that are low wage	Indicates local affordability and types of employment accessible without a car
HOUSEHOLD_P_ACRE	Residential density per acre	Denser housing often supports transit and walking
POPULATION_P_ACRE	Population density per acre	Key factor in predicting transit viability and car ownership
JOBS_P_ACRE	Job density per acre	Dense employment centers reduce reliance on personal vehicles
JOBS_P_HOUSEHOLD	Ratio of jobs to households	Indicates job-housing balance, affecting travel needs
ROAD_NETWORK_MILES	Miles roads per square mile	Roads are mostly car-oriented infrastructure and often correlate with higher car ownership
METERS_NEXT_TRANSIT_STOP	Distance to nearest transit stop (meters)	Key factor-longer distances reduce the viability of car-free living
TIMES_P_HOUR_TRANSIT_SERVICE	Frequency of transit service near CBG	More frequent service supports households not owning cars

Null Values

Missing CSA and CBSA Values

The CSA and CBSA columns contain many null values and provide only supplementary location information. Therefore they will not be used in the model. Although `CBSA_POP`, `CBSA_EMP`, and `CBSA_WRK` could serve as features, their strong correlation with null values in the CSA and CBSA columns renders them non-informative, and they will consequently be excluded from model training.

Missing Households Values

The values correspond to the total counts for the Northern Mariana Islands and Guam. Since these territories are not part of the continental United States and their data is incomplete, they will be excluded from the analysis. Additionally, other special territories that are not U.S. states will also be removed.

Threshold Values for Transit Features

The variable `METERS_NEXT_TRANSIT_STOP` uses a placeholder value of -99999 for all CBGs whose population-weighted centroids are located more than three-quarters of a mile (1.2 km) from the nearest transit stop. This negative placeholder is inappropriate, and instead, a value of 2000 meters is used to better represent greater distance. Similarly, `TIMES_P_HOUR_TRANSIT_SERVICE` assigns -99999 to CBGs without transit service, which is misleading; this is replaced with zero to accurately indicate no transit frequency.

Distributions

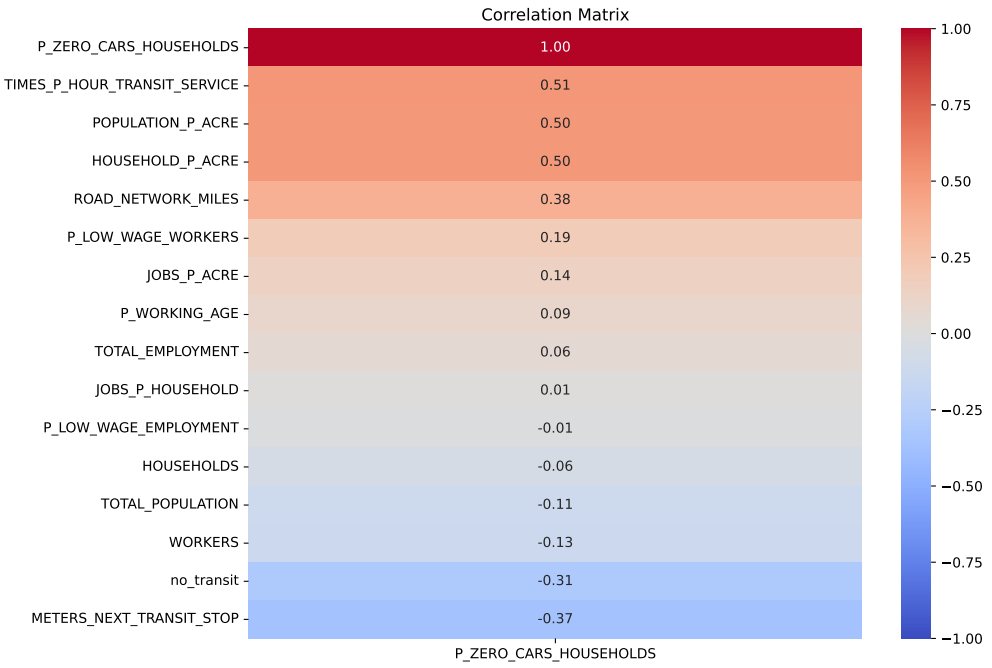
Several variables exhibit distinctive distributional patterns. The proportion of zero-car households is strongly left-skewed, with most areas showing very low values. Population density (population per acre) is generally low but displays a pronounced right tail, reflecting a small number of highly dense areas. Transit-related measures highlight a sharp divide: many CBGs have no transit service, with distances to the nearest stop capped at 2000 m and service frequency recorded as zero, while others have high accessibility and frequent service.

Correlation Analysis

Several variables exhibit substantial correlation with target feature.

Figure 1

Correlation matrix



Probabilistic Modeling Approach

Linear Models

Ordinary Least Squares (OLS)

OLS minimizes the residual sum of squares:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2$$

Suitability: Serves as a baseline due to interpretability and closed-form solution. May overfit with multicollinearity or noisy features.

Ridge Regression (L2 Regularization)

Introduces an L2 penalty on the coefficients:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

Suitability: Controls overfitting by shrinking coefficients; effective with correlated predictors.

Lasso Regression (L1 Regularization)

Introduces an L1 penalty:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1$$

Suitability: Encourages sparsity and performs automatic feature selection.

Elastic Net

Combines L1 and L2 regularization:

$$\hat{\beta}_{\text{EN}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$$

Suitability: Balances sparsity and regularization; useful for correlated features.

Bayesian Ridge Regression

Probabilistic version of Ridge regression with priors:

$$p(\beta) = \mathcal{N}(0, \alpha^{-1}I), \quad p(y \mid X, \beta) = \mathcal{N}(X\beta, \eta^{-1}I)$$

Posterior distribution:

$$p(\beta \mid D) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{S}_N = (\alpha I + \eta X^T X)^{-1}, \quad \mathbf{m}_N = \eta \mathbf{S}_N X^T y$$

Suitability: Captures uncertainty in weights and includes automatic regularization.

Nonlinear Models

To address potential nonlinear relationships in the data, the following models are considered.

k-Nearest Neighbors (k-NN) Regression

Predicts by averaging target values of the k nearest neighbors:

$$\hat{y}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i$$

Suitability: Non-parametric, data-driven. Best for low-dimensional, smooth problems.

Decision Tree Regression

Recursively partitions the feature space to minimize within-node variance:

$$\text{MSE}(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y}_S)^2$$

Suitability: Captures nonlinear relationships and interactions; interpretable but prone to overfitting without regularization.

Random Forest Regression

An ensemble of decision trees trained on bootstrapped samples:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

Suitability: Reduces overfitting compared to a single tree; robust and effective for tabular data.

XGBoost Regression

Implements gradient boosting by sequentially adding trees to correct previous errors, with regularization to control complexity and prevent overfitting.:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta f_t(x), \quad f_t \in \mathcal{F}$$

Suitability: Highly efficient and accurate; handles complex nonlinear relationships and large datasets.

Table 2

Summary of Suitability for Different Models

Model	**Strengths**	**When to Use**
OLS	Simple, interpretable	Linear relationships
Ridge	Handles multicollinearity	Correlated predictors
Lasso	Feature selection	Sparse models
Elastic Net	Hybrid regularization	Correlated + sparse features
Bayesian Ridge	Uncertainty-aware, regularized	Noisy/small data
k-NN	Local, non-parametric	Small, low-dim, smooth data
Decision Tree	Nonlinear, interpretable	Simple, explainable models
Random Forest	Nonlinear, robust	Tabular data, interactions
XGBoost	High accuracy, efficient	Large, complex datasets

Model Evaluation, Results and Discussion

Table 3

Model performance comparison using RMSE and R^2

Model	**RMSE**	R^2
OLS	0.098	0.478
Ridge	0.098	0.478
Lasso	0.098	0.475
Elastic Net	0.098	0.475
Bayesian Ridge	0.098	0.478
k-NN	0.091	0.546
Random Forest	0.083	0.628
XGBoost	0.082	0.633

The linear models show similar performance with modest accuracy. Nonlinear models perform noticeably better, capturing more complex patterns in the data. Ensemble methods, in particular, achieve the highest predictive accuracy.

Regularization Analysis

Hyperparameter tuning for Lasso (L1), Ridge (L2), and Elastic Net regularization yielded identical optimal mean squared errors (MSE ≈ 0.010) at $\alpha = 0.001$ for all models. Comparison of Ridge performance on training and test sets ($\text{RMSE}_{\text{train}} = 0.0996$, $R^2_{\text{train}} = 0.462$; $\text{RMSE}_{\text{test}} = 0.0978$, $R^2_{\text{test}} = 0.478$) indicates no signs of overfitting. These results suggest that regularization provides no tangible benefit in this case, as the unregularized model is already stable and the predictors do not induce coefficient instability. Therefore, only OLS, Bayesian Ridge, and nonlinear models are considered in the following comparisons.

Boundaries of Predictions

In linear regression, some predictions fall below 0 or exceed 1, which is inappropriate given that the target variable is a proportion bounded between 0 and 1. This occurs because linear models extrapolate without constraints on the output range. In contrast, KNN and Random Forest implicitly learn the observed bounds from the training data and thus do not produce predictions outside the feasible range. XGBoost generally stays close to the observed range but can still produce slight out-of-bound values.

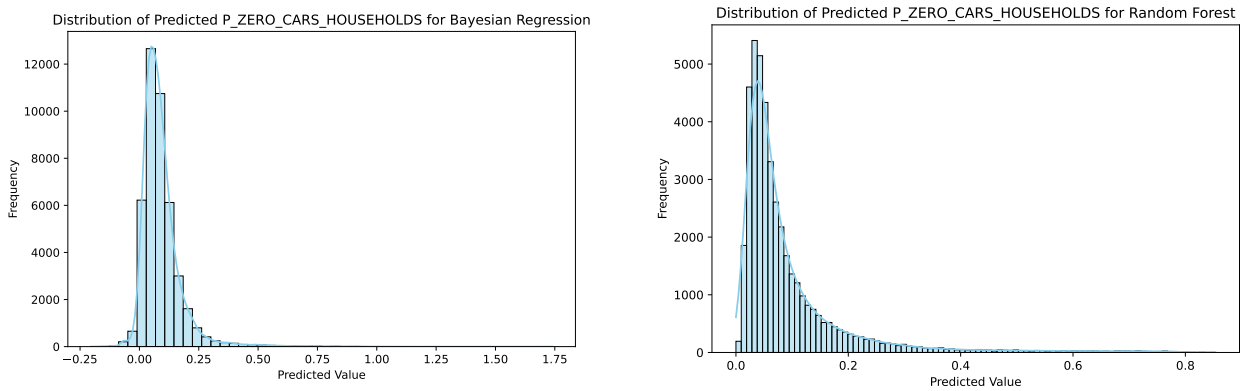


Figure 2

Comparison of two prediction boundaries: Bayesian vs. Random Forest-

Skewed Distributions of Target Variable

In the test data, the target distribution is notably skewed, with a pronounced spike at zero corresponding to rural areas without transit service. In contrast, the Random Forest predictions exhibit a more normalized distribution, smoothing over the large concentration of zero values. This behavior indicates that while the model captures overall trends, it tends to underrepresent extreme sparsity in the data, likely due to its averaging across decision trees

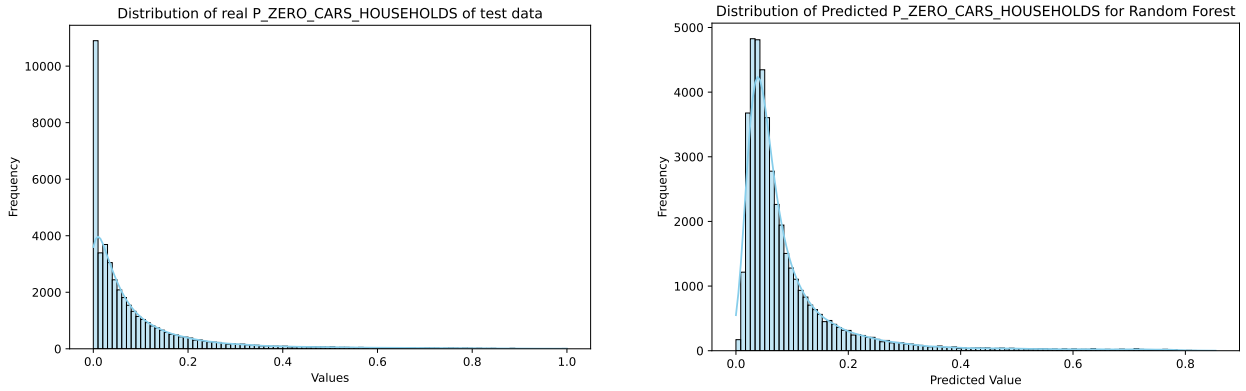


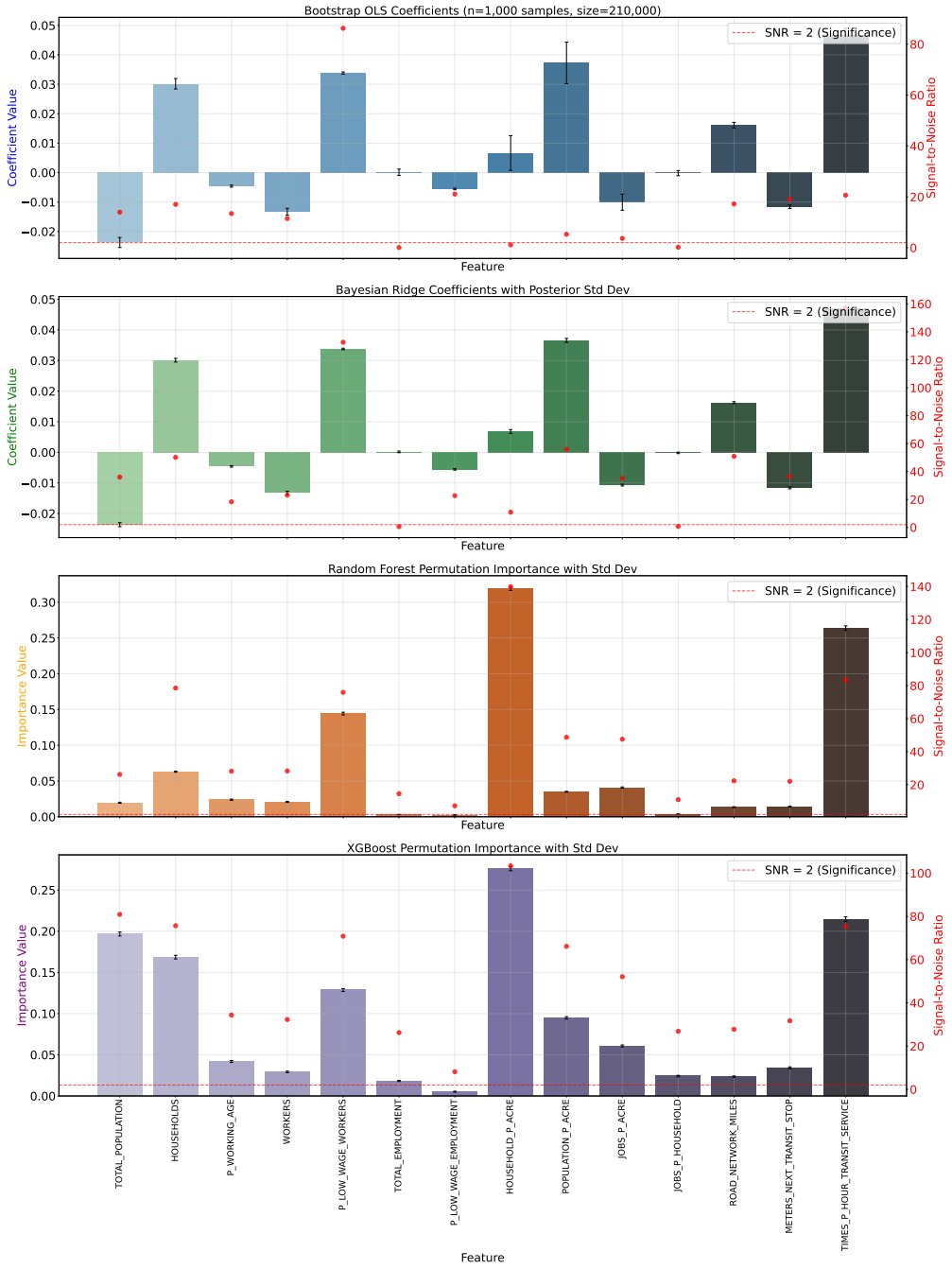
Figure 3

Comparison of two distributions: test data vs. Random Forest predictions.

Feature Strengths and Uncertainty

Figure 4

Feature Analysis



Linear Models - Magnitude and Direction: Both Bayesian Ridge and OLS yield similar coefficient patterns but differ in uncertainty. OLS bootstrapping reveals high standard errors for many predictors, reflecting the difficulty of fitting linear models to inherently non-linear relationships. For

example, TIMES_P_HOUR_TRANSIT_SERVICE exhibits the strongest positive linear association across both models. P_LOW_WAGE_WORKERS also shows consistently large, precise effects in both OLS and Bayesian Ridge. Linear models favor POPULATION_P_ACRE over HOUSEHOLD_P_ACRE (despite their high correlation), with moderate positive effects. Bayesian Ridge produces notably low uncertainties, likely due to imposing linear assumptions on non-linear data, resulting in overconfident but potentially biased estimates.

Tree Based Models - Magnitude Only: Both tree-based models capture non-linearities and interactions, though their importance rankings differ. Random Forest assigns its highest importance to HOUSEHOLD_P_ACRE, followed closely by TIMES_P_HOUR_TRANSIT_SERVICE, and attributes far greater importance to density than do linear models. XGBoost, while also ranking HOUSEHOLD_P_ACRE and TIMES_P_HOUR_TRANSIT_SERVICE near the top, places relatively greater emphasis on TOTAL_POPULATION and HOUSEHOLDS, with substantial weight also given to P_LOW_WAGE_WORKERS and POPULATION_P_ACRE. This suggests XGBoost may capture smoother, additive patterns, while Random Forest detects sharper threshold effects.

Consistency: TIMES_P_HOUR_TRANSIT_SERVICE and P_LOW_WAGE_WORKERS emerge as important across Random Forest, XGBoost, and linear models (albeit with different rankings), strengthening confidence in their substantive influence.

Discrepancies: The density measures show the strongest divergence. Linear models prefer POPULATION_P_ACRE, whereas both Random Forest and XGBoost favor HOUSEHOLD_P_ACRE dramatically so for Random Forest, with very high importance and Signal-to-Noise Ratio (SNR), indicative of sharp non-linear effects. XGBoost shares this preference but moderates its magnitude, while assigning higher weight to broad demographic totals such as TOTAL_POPULATION and HOUSEHOLDS. These differences reflect XGBoost's tendency toward smoother decision boundaries and Random Forest's strength in detecting abrupt shifts.

Uncertainty: OLS bootstrapping identifies high uncertainty when fitting linear models to non-linear data. Bayesian Ridge's lower uncertainties likely arise from model misspecification. Both Random Forest and XGBoost achieve high SNRs. Random Forest generally exhibits slightly higher SNRs for its top-ranked features, suggesting greater confidence in its primary variable set, while XGBoost's SNR values are more evenly distributed across predictors.

Model Uncertainty

Both linear models (OLS and Bayesian Ridge) exhibit near-ideal 95% coverage in their theoretical uncertainty estimates, but their prediction intervals are extremely wide ($\sim 40\%$ of the 0–1 target range), making them largely uninformative for practical decision-making. Random Forest provides narrower intervals ($\sim 23\%$) with slightly lower coverage ($\sim 92.5\%$), yielding more actionable uncertainty estimates.

XGBoost was not included in the formal interval calculation because standard uncertainty estimates are not readily available, and bootstrap approximations proved unreliable. However, a practical evaluation within a ±3% absolute error range shows XGBoost achieves the highest accuracy (44.1%), followed by Random Forest (42.8%) and the linear models (35.1%), indicating that XGBoost most effectively captures the complex, non-linear relationships in the data.

Prediction accuracy within ±3% absolute error:

OLS: 0.351, Bayesian Ridge: 0.351, Random Forest: 0.428, XGBoost: 0.441

Conclusion

Nonlinear and ensemble models clearly outperform linear methods in both predictive accuracy and practical reliability. Random Forest provides strong performance with narrow prediction intervals and well-constrained outputs, while XGBoost achieves the highest accuracy within a ±3% absolute error range. However, XGBoost occasionally produces slight out-of-bound predictions, which is a minor disadvantage compared to Random Forest. Despite these differences, both ensemble models represent optimal choices given the current feature set and data complexity. Including additional relevant features could further improve predictions and help better capture the high number of zero values observed in the target distribution.

References

[1] Python libraries and tools used in this study: `pandas`, `numpy`, `seaborn`, `matplotlib.pyplot`, `scikit-learn` (`LinearRegression`, `Ridge`, `Lasso`, `ElasticNet`, `BayesianRidge`, `KNeighborsRegressor`, `RandomForestRegressor`, `RandomizedSearchCV`, `permutation_importance`, `train_test_split`, `StandardScaler`, `mean_squared_error`, `mean_absolute_error`, `r2_score`, `resample`), `xgboost` (`XGBRegressor`), `statsmodels.api`, `tqdm`.

[2] U.S. Environmental Protection Agency, *Smart Location Database*, 2023. `https://www.epa.gov/smartgrowth/smart-location-mapping#SLD`

[3] Data.gov, *Smart Location Database*, 2023. `https://catalog.data.gov/dataset/smart-location-database8`