Prediction of Resident Travel Modes Based on XGBoost and Analysis of Behavioral Interaction Characteristics

Xia Yang  
Faculty of Transportation Engineering, Kunming University of Science and Technology

Jin Zhang (✉ 55715389@qq.com)  
Faculty of Transportation Engineering, Kunming University of Science and Technology

Kai Wang  
Faculty of Transportation Engineering, Kunming University of Science and Technology

Chao-qin Feng  
Faculty of Transportation Engineering, Kunming University of Science and Technology

Jie-mei Li  
Faculty of Transportation Engineering, Kunming University of Science and Technology

Research Article

Keywords: Urban traffic, Travel mode prediction, Integrated learning, XGBoost, TreeSHAP

Posted Date: October 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3434939/v1

License: ☺️ ☝️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Prediction of Resident Travel Modes Based on XGBoost and Analysis of Behavioral Interaction Characteristics

Xia Yang¹, Jin Zhang², Kai Wang³, Chao-qin Feng⁴, Jie-mei Li⁵

Abstract
Figuring out the characteristics of urban residents' travel mode choices is the key to the forecasting of residents' travel demand as well as an important basis for transportation system management and planning. The integrated learning model based on the Boosting framework has high prediction accuracy and strong feature selection and combination ability and has become the preferred algorithm for building travel demand prediction models. In this article, the authors use the resident travel survey data of Kunming City, choose four integrated learning classifiers, XGBoost, LightGBM, CatBoost, and GBDT, to predict the travel mode of the residents, select the best parameters of the model by using grid search and five-fold cross-validation, analyze the importance of the features of the prediction model by using TreeSHAP, and finally explore the selection of travel modes under the interaction of important feature variables. The results of the study show that (1) the XGBoost model performs better than the other models, and the accuracy, precision, recall, and F1 value of the XGBoost model reach 90%, respectively, and the prediction accuracy of the four modes of travel, namely walking, two-wheeled electric motorcycle, public transportation, and car, reaches 94%, 90%, 85%, and 90%, respectively, and the corresponding AUC values reach 0.99, 0.97, 0.96, and 0.98, respectively. (2) Compared with household size and annual income, the actual distance of travel paths, ownership of cars and 2-wheeled electric motorcycles, age and gender of travelers, and the built environment are more important factors influencing the prediction of residents' travel choices. (3) The characteristics of travel mode choice under the interaction of several factors are obvious; except for the group over 55 years old, the ownership of travel means of transportation in the family significantly affects the choice of travel mode of residents; men between 20 and 55 years old have more medium-distance and long-distance trips, and they are the main group of people who use cars; when the travel distance is less than 15km, the 2-wheeled electric motorcycle and cars have a certain mutual substitution effect. In order to comprehensively promote the high-quality development of transportation, it is necessary to focus on the travel needs of women and the elderly while controlling the number of motor vehicles in the household, introducing policies to encourage the use of two-wheeled electric motorcycles, and improving the city's public transportation and commercial support facilities.

Keywords Urban traffic • Travel mode prediction • Integrated learning • XGBoost • TreeSHAP
Introduction

With the rapid economic and social development and continuous expansion of the city, people's production and living space are gradually expanding, the dependence on motorized transportation is increasing, and the problem of traffic congestion and environmental pollution is gradually becoming more and more prominent. In order to alleviate the problem of urban diseases caused by motorized travel, researchers have tried to predict and analyze the travel modes of urban residents, which can be used as a basis for policy formulation and urban planning and construction (Liu et al. 2022). In the 21st century, people's lifestyles are complex and varied, and travel mode choices are flexible and diverse. How to accurately model and predict residents' travel modes is an important topic in modern transportation planning research.

For a long time, scholars at home and abroad have used traditional linear regression models and discrete choice models to predict residents' travel modes and analyze the influencing factors such as socio-economic attributes, spatial environment, and activity travel characteristics of residents' travel mode choices (Huang et al. 2020; Yang et al. 2023). For example, Wookjae and Guang et al. used travel data from 29 areas in the U.S. to study the relationship between land use characteristics and travel mode choice and found that neighborhoods with higher activity densities, higher intersection densities, better accessibility to public transit, and a balance of residential employment are more likely to have non-motorized travel choices (Wookjae et al. 2023). Nurul proposed RI-MNL and RI-NL econometric models to predict the commuting patterns of residents in the Greater Toronto and Hamilton area using household travel survey data and found that inducing RI in discrete choice models (MNL and NL) can greatly improve the fit of the models (Nurul 2023). Linear regression models are highly interpretable but perform poorly when dealing with nonlinear relationships or non-normally distributed data. The discrete choice model based on random utility maximization assumes the rational behavior of individual utility maximization, which can better explain individual choice behavior, but it is easy to ignore individual preference and behavioral heterogeneity, mutual compensation between travel modes, and cannot accurately capture the nonlinear relationship between variables.

With the development of information technology, machine learning has been widely used in the field of transportation, and more and more alternatives to random utility models have been proposed, such as decision trees, support vector machines, neural networks, etc. (Koushik et al. 2020; Hagenauer and Helbich 2017), which are not based on strict a priori assumptions, are able to automatically learn patterns and regularities from the data, efficiently deal with high-dimensional and complex data structures, and are capable of modeling nonlinear relationships with strong robustness and good generalization ability. For example, Xia and Chen et al. demonstrated that Random Effect Bayesian Neural Network (RE-BNN) have better prediction performance and the ability to reveal regional heterogeneity of preferences than MNL using national travel data from the UK (Xia and Chen 2023). In addition, there are also integrated learning models such as random forest and GBDT (Yin et al. 2023; Sekhar et al. 2016; Hillel et al. 2020), which combine different underlying learners to model the data from different perspectives, which can improve the prediction accuracy, generalization ability, and noise immunity and improve the stability of the model, which makes it show better performance than a single machine learning model in most of the practical applications. Jin and Wang et al. used
a gradient-boosting decision tree (GBDT) model to explore the relationship between the built environment and e-scooter sharing (ESS) route flow. This study found a significant nonlinear relationship between both distance to the city center and gradient with e-scooter sharing (ESS) line flow. More bicycle facilities and commercial corridor supply can be more attractive for urban residents to choose e-scooter sharing (ESS) trips (Jin et al. 2023).

The integrated learning based on the Boosting framework adjusts the weights of the samples according to the errors of the previous base model, pays more attention to misclassified samples, improves the accuracy of the model, and at the same time reduces the influence of individual samples on the model to improve its robustness. For this reason, this paper is based on the resident travel survey data in Kunming, using the mainstream integrated learning model of the Boosting framework to predict and analyze the comparison of travel modes, using TreeSHAP as a post-hoc explanatory model, intuitively reflecting the relative importance of the features on the prediction results through the SHAP summary graph, and finally analyzing the influence of each factor on the choice of travel modes in detail. The XGBoost model used in this paper provides a more accurate means of travel mode prediction research as well as a theoretical basis for urban transportation planning and residents' travel decisions.

**Research data**

**Sources of research data**

In this study, we selected Kunming, the capital city of Yunnan, China, as the study area. At the same time, the data we used are mainly residential travel survey data and built environment data. This data is obtained by geographically whole cluster sampling, and the data includes attributes of 4427 members and more than 11,000 travel data. It should be noted that our survey data contains personal information, household information and travel information, where travel modes include walking, bicycles, two-wheeled electric motorcycles, assisted vehicles, buses, subways, cars, and passenger shuttles. Since Kunming is in the early stage of subway construction, the subway travel sample accounts for less than 1% of the overall travel ratio, and to ensure a balanced sample, only 98 subway travel samples from the total sample are excluded. In addition, 158 samples with unclear travel modes and other abnormal samples are excluded, and finally 10,700 valid samples are retained. The OD point distribution of the study travel samples is shown in Figure 1.
Selection of variables

It has been shown (Zhang and Chai 2018; Wang et al. 2019; Kamargianni et al. 2015; Mahdi and Sakth 2023) that the choice of travel mode of residents is influenced by a variety of factors, such as personal attributes, family attributes, travel attributes, and the built environment. The built environment in our study specifically refers to the spatial environment facilities around the residence, which includes the distance of the residence from the city center, the population density within 500 meters of the residence, the total POI density, the density of bus stops, and the density of the road network. Among them, POI types include shopping malls, office buildings, neighborhoods, government, schools, hospitals, banks, parks, factories, food markets, restaurants, supermarkets, hotels, gas stations, and postal services. In addition, the road network with class IV and above and the length of the single line are selected for the road network density calculation. Based on previous studies, the variables we selected are shown in Table 1. In order to better predict travel mode choices, we classify travel modes into four categories: the first category is "walking", which includes walking (92%) and bicycling (8%); the second category is "electric motorcycle", which includes two-wheeled electric motorcycles or shared mobility scooters; the third category is "public transportation" the fourth category is "mini-motorcycle" and the third category is "public transportation". The third category is public transportation; the fourth category is "car", which includes cars (97%) and passenger shuttles (3%), and the statistics are shown in Table 1.

Table 1 Selection and description of variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Symbol</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traveling attributes</td>
<td></td>
<td>1: walking, bicycling (30.5%); 2: 2-wheeled electric motorcycle, shared mobility scooter (26.9%); 3: public transportation (21.4%); 4: cars, passenger shuttles (21.2%)</td>
</tr>
</tbody>
</table>
Purpose of travel $X_1$:
1: commuting (51.2%); 2: commuting to school (8.9%); 3: transportation, medical care, and living (14.3%); 4: shopping, leisure, visiting friends and relatives (25.6%); The actual path distance crawled from the Gaode Map API interface according to the longitude and latitude of the traveling OD point and the corresponding travel mode (5.0km);

Route distance $X_2$:

Personal attributes

Gender $X_3$:
0: female (50.8%); 1: male (49.2%);

Age $X_4$:
Age of the traveler (40.0 years old);

Educational attainment $X_5$:
1: junior high school and below (34.0%); 2: high school (33.1%); 3: university and above (32.9%);
1: Students, retired, and unemployed (36.8%); 2: Agriculture, forestry, animal husbandry, and fishery workers; business services and community workers; private and self-employed owners (38.5%); 3: Employees of production enterprises, scientific research and design, information services, civil servants, clerical, educational, and healthcare employees, doctors, nurses, and other employees (24.7%);

Occupation $X_6$:

Household attributes

Income $X_7$:
Annual household income .1: less than RMB 50,000 (33.1%); 2: RMB 50~100,000 (47.7%); 3: more than RMB 100,000 (19.2%);

Car $X_8$:
Number of private cars owned in the household (0.6);

Electric motorcycle $X_9$:
Number of 2-wheeled electric motorcycles owned in the household (0.8);

Household size $X_{10}$:
Number of people residing in the household (3.0);

Employed people $X_{11}$:
Number of employed people in the household (1.6);

Less than 6 years old $X_{12}$:
Number of people under 6 years of age in the household (0.2);

Built environment

Distance to city center $X_{13}$:
Straight-line distance between the residence and the city center (4.3km);

Population density $X_{14}$:
The density of the population in the area of 1km around the settlement (28446.8 / km$^2$);

Total POI density $X_{15}$:
Density of total POIs in the area of 0.5km around the settlement (191.8 / km$^2$);

Bus stop density $X_{16}$:
Density of bus stops in the area of 0.5km around the settlement (8.1 / km$^2$);

Road network density $X_{17}$:
Density of road network in the area of 1km around the settlement (2.4km/km$^2$);

Note: Regarding the contents of "()", if the corresponding variable is of subtype, it is the percentage; if the corresponding variable is of continuous numeric type, it is the average value.

Models and Methodologies

Extreme gradient boosting

Extreme Gradient Boosting (XGBoost) is a machine learning model based on the Gradient Boosting Tree algorithm (Chen and Guestrin 2016), which integrated learning by iteratively training multiple weak learners (decision trees) to efficiently model complex nonlinear relationships and feature interactions. Characterized by parallelization, flexibility, and scalability, the model excels in prediction accuracy and performance and is widely used in machine learning tasks such as classification and regression.

The learning task of the new tree in the XGBoost model is to de-fit the residuals of the original tree, which makes the prediction of the generated tree closer to the true value by
continuously adding decision trees. At the same time, regular terms such as leaf node weights and tree depth are added to the loss function to control the complexity and stability of the model.

Given a dataset \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) containing \( n \) samples with \( m \) features, the predicted output of the XGBoost integrated model is represented as:

\[
\hat{y}_i^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\]

Equation (1) represents that, given an input \( x_i \), the output value is the sum of the predicted values obtained from \( K \) decision trees. \( K \) is the number of decision trees; \( f_k \) denotes the decision trees; \( f_t(x_i) \) is the predicted value of the decision tree that needs to be newly added; and \( \hat{y}_i^{(t)} \) is the predicted value of the \( i \)th sample in terms of the \( t \)th tree (the \( t \)-th iteration).

The objective function The object of XGBoost consists of two parts: the loss function and the regular term.

\[
Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{T} \Omega(f_i) \tag{2}
\]

\[
\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{3}
\]

where \( l \) denotes the loss function, which is used to measure the difference between the model prediction result \( \hat{y}_i \) corresponding to \( x_i \) and the true value \( y_i \); \( \Omega \) is the penalty term used to control the complexity of the decision tree. \( \gamma \) denotes the penalty coefficient of the leaf nodes of the whole tree, which is used to control the number of leaf nodes; \( \lambda \) denotes the penalty coefficient of the leaf node dimension; \( T \) denotes the number of leaf nodes of the given tree, and the fewer leaf nodes, the simpler the model is; \( w_j \) denotes the score of the \( j \)-th leaf node leaf.

The forward stepwise algorithm is often used in objective function optimization, and the specific steps are as follows: First, the second-order Taylor expansion of \( l \left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \) is carried out, and in the \( t \)-th iteration, the leaf nodes of the tree before \( t \) and the weights as well as the loss function have been determined, i.e., they are constants, and the following results are obtained:

\[
Obj^{(t)} \approx \sum_{i=1}^{n} \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + c \tag{4}
\]

\[
g_i = \partial_{\hat{y}_i^{(t-1)}} l \left(y_i, \hat{y}_i^{(t-1)}\right) \tag{5}
\]

\[
h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l \left(y_i, \hat{y}_i^{(t-1)}\right) \tag{6}
\]

Second, instead of traversing the samples, traverse the leaf nodes, so that \( I_j = \{i | q(x_i) = j\} \) is the set of samples of the \( j \)-th leaf node.

\[
Obj^{(t)} = \sum_{j=1}^{T} \left[ G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \tag{7}
\]

\[
G_j = \sum_{i \in I_j} g_i \quad H_j = \sum_{i \in I_j} h_i \tag{8}
\]

where \( G_j \) and \( H_j \) are constants that have been determined from the previous \( t - 1 \) steps, and the first-order derivative of the objective function with respect to \( w_j \) is obtained to obtain the leaf node weights \( w_j^{*} \) corresponding to the smallest time of the objective function.
\[ w_j^* = -\frac{g_j}{H_j + \lambda} \quad (9) \]

\[ Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{g_j^2}{H_j + \lambda} + \gamma T \quad (10) \]

Again, assume that the decision tree completes the feature split at a certain node. \( I_L \) and \( I_R \) are the sample sets of the left and right nodes, respectively, after the split, and \( I = I_L \cup I_R \), and the post-split payoff for the objective function is as follows:

\[ Gain = \frac{1}{2} \left[ \frac{g_L^2}{H_L + \lambda} + \frac{g_R^2}{H_R + \lambda} - \frac{(g_L + g_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (11) \]

Finally, to determine the cut \( Gain \), the larger the gain value, indicating that the split can make the objective function value reduce, the more the better. The reason is that there is the introduction of new leaves of the penalty term. If the split brought about by the gain is less than the threshold value of the time, you can cut off this split.

**Shapley additive explanations**

The Shapley Additive exPlanations (SHAP) explains machine learning based on the SHAP value of the cooperative game theory approach (Ji et al. 2022), while TreeSHAP is a variant of SHAP (Campbell et al. 2022). Unlike SHAP, which traverses all combinations of features for a decision tree, TreeSHAP simply traverses all possibilities for the path of the tree to reduce the computational complexity, increase the speed and accuracy of computation, and correctly estimate the SHAP value when there is correlation between features. The SHAP value is correctly estimated when there is a correlation between the features. For an input sample \( x \), the model calculates the SHAP value \( \phi_i \) for each feature \( i \) through the structure of the decision tree, where \( i \) denotes the index of the feature, as follows:

\[ \phi_i = \sum_{S \subseteq D} \frac{|S|!(|D|-|S|-1)!}{|D|!} \left[ f(x_{S \cup \{i\}}) - f(x_S) \right] \quad (12) \]

where \( f(x_S) \) denotes the prediction result of input sample \( x \) on the feature subset \( S \) that does not contain feature \( i \); \( f(x_{S \cup \{i\}}) \) denotes the prediction result of input sample \( x \) on the feature subset \( S \cup \{i\} \) that contains feature \( i \); \( D \) is the set of features; \( S \) is a subset of \( D \); \( |S| \) denotes the size of \( S \), and \( |D| \) denotes the size of \( D \). The average absolute SHAP values of features are computed by feature dimension aggregation and visualized to obtain a SHAP summary map for global feature importance analysis.

**Technical routes**

We preprocess the raw data and then import it into the model. The specific technical route is shown in Figure 2. Firstly, the feature columns \([X]\) and target variables \([Y]\) are stored, and the training set and test set are divided according to the ratio of 8:2. Secondly, the XGBoost classifier is created, and the specified parameter intervals are traversed by GridSearchCV, and five-fold cross-validation is performed on each parameter combination to select the parameter combination that performs the best in the training set, to obtain the optimal hyper-parameters, and to perform a prediction evaluation on the test set and evaluate the model performance. The prediction evaluation is performed on the test set, and the model performance is evaluated.
Finally, to further understand the degree of contribution of different features to the model prediction results, TreeSHAP is used to interpret the prediction results of the XGBoost model, SHAP values are calculated, and summary plots are drawn to show the importance of features.

![Technical route](image)

**Grid search and 5-fold cross-validation**

Grid search traversal is the key to debugging the XGBoost model, and the specific parameter descriptions and optimal values are shown in Table 2.

**Table 2** Description of XGBoost's parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Range of adjustment parameters</th>
<th>Optimum value</th>
<th>Parameters Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_estimators</td>
<td>(50, 500)</td>
<td>285</td>
<td>Number of decision trees, i.e., number of base learners in integrated learning; The maximum depth of the decision tree, i.e. the number of nodes in the longest path of the decision tree;</td>
</tr>
<tr>
<td>Max_depth</td>
<td>(1, 30)</td>
<td>7</td>
<td>Learning rate, controlling the magnitude of weight adjustment at each training step; Minimum value of sample weights on child nodes;</td>
</tr>
<tr>
<td>Learning_rate</td>
<td>(0.1, 1.1)</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Min_child_weight</td>
<td>(1, 10)</td>
<td>1</td>
<td>Minimum value of sample weights on child nodes;</td>
</tr>
<tr>
<td>Subsample</td>
<td>(0.1, 1.01)</td>
<td>0.9</td>
<td>Proportion of samples sampled in each decision tree;</td>
</tr>
<tr>
<td>Colsample_bytree</td>
<td>(0.1, 1.1)</td>
<td>0.8</td>
<td>Proportion of features sampled in each decision tree;</td>
</tr>
<tr>
<td>Reg_alpha</td>
<td>(0, 1.01)</td>
<td>0.5</td>
<td>Penalty coefficients for regularization terms to control model complexity(L1);</td>
</tr>
<tr>
<td>Reg_lambda</td>
<td>(0, 1.01)</td>
<td>0</td>
<td>Penalty coefficients for regularization terms to control model complexity(L2);</td>
</tr>
</tbody>
</table>
Model comparison

To confirm the advantages of the XGBoost model, we simultaneously selected three mainstream integrated learning models based on the Boosting framework, LightGBM (Suchismita and Debapratim 2023), CatBoost (Dorogush et al. 2022), and GBDT (Laviolette et al. 2022) (Table 3), to model the residents' travel mode choices, and compared the results with the modeling results of the XGBoost model.

Table 3  Advantages and disadvantages of integrated learning models based on boosting framework

<table>
<thead>
<tr>
<th>Model</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>It builds on GBDT by adding regularization terms and approximate learning algorithms to improve performance and optimize the model training process, with excellent performance on high-dimensional sparse data and large datasets;</td>
<td>Easily overfitted and memory-intensive;</td>
</tr>
<tr>
<td>LightGBM</td>
<td>It introduces histogram-based decision tree segmentation and mutually exclusive feature bundling on the basis of GBDT, etc., which can efficiently handle large-scale datasets, classification features, and fast training speed;</td>
<td>Poor performance on small datasets, sensitive to noise and outliers;</td>
</tr>
<tr>
<td>CatBoost</td>
<td>Based on GBDT, it introduces classification features, automatic processing of missing values, multi-threaded training, etc., which effectively handle large-scale datasets and are robust to outliers;</td>
<td>Small datasets perform poorly and training is relatively slow;</td>
</tr>
<tr>
<td>GBDT</td>
<td>It combines the prediction results of multiple decision trees to fit nonlinear relationships based on the Boosting algorithm, which can handle multiple data types at the same time and performs well with small data sets;</td>
<td>Poor performance on high-dimensional sparse datasets, noise-sensitive;</td>
</tr>
</tbody>
</table>

Model results

We edited the model code using Python on the Jupyter Notebook (Anaconda3) platform. In the model, we use residents' travel mode as the target variable and personal attributes, household attributes, as well as travel purpose and path distance as the feature vectors, and train, parameterize, and predict the test set with four models: XGBoost, LightGBM, CatBoost, and GBDT, respectively.

Model performance evaluation

According to the characteristics of the classification model, accuracy, precision, recall (also known as true case rate), F1 value (F-Score), ROC curve, and AUC value are selected as the comprehensive evaluation indexes of model performance and precision, and the calculation method of each index is shown in Table 4.
Table 4 Calculation of evaluation indicators for models

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>False positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FP</td>
<td>(\frac{TP}{TP + TN})</td>
<td>(\frac{TP}{TP + FP})</td>
<td>(\frac{TP}{TP + FN})</td>
<td>(\frac{Precision \times Recall}{Precision + Recall})</td>
<td>(\frac{FP}{FP + TN})</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
<td>TN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: \(TP\) denotes the number of samples where the prediction was positive and the truth was also positive; \(TN\) denotes the number of samples where the prediction was negative and the truth was also negative; \(FP\) denotes the number of samples where the prediction was positive but the truth is negative; \(FN\) denotes the number of samples where the prediction was negative but the truth is positive.

**General indicators**

Accuracy rate indicates the ratio of the number of samples predicted by the model to be classified correctly to the total number of samples, measuring the overall correctness of the model; precision rate indicates the proportion of samples predicted by the model to be positive cases that are truly positive cases, measuring the accuracy of the model in predicting positive cases; recall rate indicates the proportion of samples that are truly positive cases that are predicted by the model to be positive, measuring the model's rate of checking for positive cases; The value of F1 is the harmonic average of the precision rate and the recall rate. The F1 value is the average of precision rate and recall rate; the F1 value is the combined value of precision rate and recall rate; the closer these values are to 1, the better the performance of the model is. The accuracy, precision, recall, and F1 values of the four models based on the Boosting framework are calculated separately (Table 5), and the results show that the accuracy, precision, recall, and F1 values of the XGBoost model reach 90%, which is significantly better than the other three models.

Table 5 Predictive accuracy of models

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td>LightGBM</td>
<td>88%</td>
<td>88%</td>
<td>87%</td>
<td>87%</td>
</tr>
<tr>
<td>CatBoost</td>
<td>80%</td>
<td>80%</td>
<td>79%</td>
<td>79%</td>
</tr>
<tr>
<td>GBDT</td>
<td>76%</td>
<td>76%</td>
<td>75%</td>
<td>75%</td>
</tr>
</tbody>
</table>

**Confusion Matrix**

The Confusion Matrix (CM) displays the correspondence between the model predictions and the actual labels in a tabular matrix for visualizing and assessing the model's prediction accuracy across multiple categories. Each row represents the actual category, each column represents the predicted category, and each element in the matrix represents the proportion of total samples for which the model predicted the actual category i table as the category j table. The model confusion matrix shows that the XGBoost model predicts accuracy better than the other three models overall and has the highest prediction accuracy of 94% for bicycling and walking; 90% for electric motorcycles; 90% for cars and passenger shuttles; and 85% for public transit (Figure 3).
Fig. 3 Confusion matrix for models

Fig. 4 ROC curve of models
ROC curves and AUC values

The ROC (Receiver Operating Characteristic Curve) curve is generated by plotting the relationship between the false positive rate as the horizontal axis and the true case rate as the vertical axis for different thresholds. The true case rate (recall rate) measures the model's detection rate of positive cases, while the false positive rate indicates the proportion of true negative cases that are predicted to be positive by the model and measures the model's misclassification rate of negative cases. The closer the curve is to the upper left corner, the better the performance of the model. The results show that XGBoost has a high detection rate and a low false positive rate under different thresholds, so it outperforms the other three models (Figure 4).

AUC (Area Under the ROC Curve) is the area under the ROC curve. A positive sample and a negative sample are randomly selected, and the probability that the model predicts a positive sample with a higher score than a negative sample is the AUC value. When the AUC is equal to 0.5, it means that the model cannot distinguish between positive and negative cases, and the performance is equivalent to random guessing, and the closer it is to 1 means that the model performs better and is able to distinguish between positive and negative cases more accurately, and the AUC value provides a global metric for evaluating classification models. The results show that XGBoost has an AUC value of up to 0.99 for bicycles and walking, 0.97 for electric motorcycles, and 0.98 for small cars, and although LightGBM has a slightly higher AUC value (0.97) for public transit than XGBoost, overall, XGBoost is superior (Figure 4).

TreeSHAP interpretation

The XGBoost model has a better overall prediction effect, so the XGBoost model is selected for global feature importance analysis, and a summary graph is drawn according to the TreeSHAP interpretation model (Figure 5). Each bar position in the graph corresponds to a specific feature, and the length of the bar indicates the importance of the corresponding feature; the longer the bar, the greater the contribution, i.e., the greater the impact of the feature on the prediction results.
The results show that the variables have different strengths and weaknesses in influencing the prediction of the travel mode choice of urban residents. Specifically, the influence of the variables on the prediction of travel mode choice, in descending order of strength, is: actual path distance in travel attributes; car ownership and electric motorcycle ownership in household attributes; age, gender, and occupational category in personal attributes; distance from the city center of the place of residence in the built-up environment; and the density of the road network.

In terms of specific factors, there are large differences in the sensitivity of the impact of each variable on the prediction of residential travel mode choice. Among these, path distance has the most sensitive impact on walking travel choice; car ownership has the most sensitive impact on car travel choice, primarily because car ownership promotes individuals to choose car travel; electric motorcycle ownership has the most sensitive impact on electric motorcycle travel choice, further illustrating the importance of transportation ownership on travel choice; gender has the most sensitive impact on car travel choice, likely due to the fact that male and female travelers have large differences in their preferences for use of small cars; distance from the city center has the most sensitive impact on walking mode choice, suggesting that built environment amenities are a key variable for walking trips.

**Analysis of variable interaction characteristics**

The above TreeSHAP results indicate that the path distance of a trip is the most important characteristic influencing the prediction of travel mode choice. In order to analyze the multifactorial effects of travel mode choice in detail, we further explore the travel mode choice
characteristics of the interaction of path distance, car ownership, electric motorcycle ownership, age, gender, distance of residence from the city center, and road network density.

From the distribution of the interaction of car ownership, age, and path distance, there are significant differences in the mode choice characteristics of the intertwined influence of various factors (Figure 6). Taken together, the critical path distance characteristics of various travel mode choices are obvious, and the critical path distance values of walking, electric motorcycles, and public transportation are 3km, 15km, and 20km, respectively. Car ownership is a necessary condition for residents to choose car travel. The proportion of individuals between 20 and 55 years of age who use cars for medium- and long-distance trips is higher than that of the other age groups, and the carless group tends to choose electric motorcycles and public transportation more and travel shorter distances, which may be caused by employment trips. Public transit and travel shorter distances, which can be attributed to employment trips. It is interesting to note that for 20- to 55-year-olds without a car, there is a particularly high concentration of people using electric motorcycles for medium- and long-distance trips, which may be a manifestation of the role of electric motorcycles as a substitute for automobiles. 55-year-olds and older no longer rely on automobiles for their trips, even if they own them, which may be due to the weakening of rigid time constraints and the reduced need for motorized travel in this age group. Residents who do not own cars are more likely to choose transit for medium- and long-distance trips after age 55, while those who do own cars use transit for shorter distances, a difference that may be attributable to pre-55 transit use habits.

Fig. 6 Distribution of travel modes under the interaction of car ownership

In terms of the distribution of the interaction of electric motorcycle ownership, age, and path distance, there were significant differences in the mode choice characteristics of the
multifactorial cross-talk (Figure 7). From the cumulative distribution, the critical characteristics of path distance for various travel mode choices remained the same regardless of electric motorcycle ownership. In the group with electric motorcycle, the electric motorcycle travel volume is large and more medium and long-distance travel, and the proportion of car and bus travel is less, indicating that the presence or absence of electric motorcycle directly affects the choice of electric motorcycle, car, and public transportation travel, and electric motorcycle has a certain substitution role for the other two in the travel. From the distribution of individual travel mode choice in different age groups, owning an electric motorcycle is a necessary condition for the choice of electric motorcycle mode for residents of all ages, and residents between the ages of 20 and 55 years old are the main group of electric motorcycle users. 55 years old and older, individuals travel more by walking and public transportation, probably because their activities are mainly recreational and they choose to travel in a way that is low-cost and beneficial to their health. More specifically, residents without scooters make a large number of medium- and long-distance trips on public transit, while only a relatively small number of residents with scooters also choose scooters for their trips, possibly because individuals reserve scooters for younger members of the family in their overall decision-making.

![Fig. 7](image)

**Fig. 7** Distribution of travel modes under the interaction of electric motorcycle ownership

In terms of the distribution of the interaction of gender, age, and path distance, the multivariate intersection of mode choice is characterized by distinct differences (Figure 8). Taken together, the nonlinear relationship between travel mode choice and the age of individuals of different genders is characterized by significant differences in the age characteristics of various travel mode choices, with residents choosing electric motorcycle trips
being the youngest overall, those choosing car trips being the next oldest, and those walking and public transit trips being the oldest overall on average. The detailed characteristics of the mode choice predictions show that electric motorcycle trips are comparable between males and females in the 20 to 55 year old group; males between the ages of 20 and 55 are the dominant group of small car users, possibly due to the fact that males in this even-age stage tend to be employed for long-distance commuting, with longer distances in their daily paths of travel and a higher need to motorize their trips. Females between the ages of 20 and 55 choose small cars for long-distance trips only and for short and medium distances, indicating that women prefer active transportation when spatial constraints are not particularly strong. Women after age 50 travel almost exclusively by transit and walking, while men after age 50 still choose cars and motorized vehicles, mainly because women in this age group have a lower proportion of employment and their daily trips are dominated by close-to-home recreational living activities, while men have a higher proportion of employment and a larger space for activities. The main reason is that women in this age group have a lower employment rate, and their daily travel is mainly for leisure activities near home, while men have a higher employment rate and more space for activities.

In terms of the distribution of the interaction of distance of residence from the city center, road network density, and path distance, the characteristics of mode choice varied considerably across built-environment contexts (Figure 9). Overall, there is a non-linear relationship between built environment and mode choice; e.g., if not any range, the further the residence is from the city center, the more likely residents are to choose to travel by car. Residents living between 2km and 9km from the city center make the majority of urban motorized trips, possibly
due to the fact that residents in this area are more dispersed, have more lateral commutes, and travel longer distances. When living too far or too close to the city center, the travel distance is shorter, and the mode choice distinction is not obvious, probably because the high facility density supply in the city center and the inner-circle living circle in the suburbs promote the short-distance travel of residents. When the road network density is lower than 4 km/km², residents have more long-distance motorized trips, while when it is higher than 4 km/km², residents concentrate on short-distance active trips, probably due to the better public transport infrastructure in high road network density areas, the higher mix of land uses, and the closer proximity of residents to their homes and workplaces, which results in shorter travel distances.

Fig. 9  Distribution of travel modes under the interaction of built environment

Conclusion and Discussion

Considering the limitations of traditional regression frameworks as well as conventional machine learning models, this paper utilizes the 2016 Kunming City residents' travel survey data to model residents' travel mode choices using XGBoost, LightGBM, CatBoost, and GBDT, evaluates the performance of the models, and predicts residents' travel modes using comprehensive indicators. Secondly, considering the high accuracy of XGBoost, TreeSHAP is used to visualize and analyze the prediction results of the model. Finally, the characteristics of travel mode choice under the interaction of multiple variables are analyzed based on the importance of the influential factors in the degree. The main research conclusions are as follows:

(1) the XGBoost model performs better than the other models, and the accuracy, precision, recall, and F1 value of the XGBoost model reach 90%, respectively, and the prediction accuracy of the four modes of travel, namely walking, two-wheeled electric motorcycle, public
transportation, and car, reaches 94%, 90%, 85%, and 90%, respectively, and the corresponding AUC values reach 0.99, 0.97, 0.96, and 0.98, respectively.

(2) Interpretation results from the TreeSHAP model show that the actual path distance traveled, car and 2-wheeled electric motorcycle ownership, age, and gender are more important factors influencing the prediction of residents' travel mode choices than household size and annual income, while the distance of residence from the city center and the density of the road network are also key factors in the prediction of an individual's travel mode choices.

(3) Characteristics of travel mode choice vary significantly across the interaction of multiple factors. Specifically, owning cars and 2-wheeled electric motorcycles in the household significantly affects the travel mode choice of the majority of residents, but excludes the group of people over 55 years old who often travel by public transportation or on foot; the group of males between the ages of 20 and 55 years old makes more trips over medium and long distances and is the main group of people who use cars; for trips with path distances of less than 15km for both males and females, 2-wheeled electric motorcycles and cars have a certain degree of substitution for each other.

Considering the above conclusions, controlling the number of motorized vehicles in the family and, at the same time, introducing relevant incentive policies for 2-wheeled electric motorcycle travel and improving the corresponding supporting facilities can effectively promote the transformation of residents' travel to non-motorized. In addition, in order to comprehensively promote the high-quality development of transportation, it is also necessary to focus on the travel needs of female groups and the elderly and improve the bus and pedestrian facilities in the areas where such groups are concentrated. Although the research in this paper has certain practical significance, there are still many shortcomings: First, this paper uses data from Kunming City in 2016 to predict the travel modes of urban residents and subsequently needs to update the data to further analyze the differences in the travel characteristics of residents in different attribute categories; Second, this paper focuses on the characteristics of travel mode choice under the interaction of multiple factors, and there is insufficient discussion of the nonlinear relationship between variables and travel mode choice, especially under the role of the built environment, which needs to be explored in depth.

Authors' contributions  The authors confirm their contribution to the paper as follows: study conception and design: Xia Yang, Chao-qin Feng, and Jin Zhang; data collection: Xia Yang and Kai Wang; analysis and interpretation of results: Xia Yang, Jin Zhang, Kai Wang, and Jie-mei Li; draft manuscript preparation: Xia Yang and Kai Wang. All authors reviewed the results and approved the final version of the manuscript.

Funding  This work was supported by National Natural Science Foundation of China (72263022)

Declarations  
Conflicts of interest  On behalf of all authors, the corresponding author states that there is no conflict of interest.
Reference


Xia Y T, Chen H F, Zimmermann R. "A Random Effect Bayesian Neural Network (RE-BNN) for travel mode choice analysis across multiple regions", Travel Behaviour and Society, 2023, 30: 118-134.


Suchismita N, Debapratim P. A joint and simultaneous prediction framework of weekday and weekend daily-activity


Xia Yang is a third-year master's student at the Faculty of Transportation Engineering at Kunming University of Science and Technology, with research interests in transportation planning and management. One paper titled "Analysis of urban household job-housing patterns and the factors influencing their commuting carbon emissions" has been accepted by the 2024 TRB Annual Meeting.

Jin Zhang is an associate professor and master's tutor at the Faculty of Transportation Engineering at Kunming University of Science and Technology. Her research direction is passenger transportation planning and management, and she has presided over more than 10 scientific research projects of various kinds, edited and participated in the publication of three books, won the first prize of the Science and Technology Advancement Award of the Yunnan Highway Society, and was granted four patents by the state.

Kai Wang is a master's degree graduate from the Faculty of Transportation Engineering at Kunming University of Science and Technology, with research interests in transportation planning and management. A paper entitled "Jobs-housing space types and commuting efficiency of urban dual-earner households: a case study of Kunming" was accepted by the Journal of Kunming University of Science and Technology(Natural Science). A paper entitled "Influence of jobs-housing space on commuting efficiency from the perspective of time geography: a case study of dual-job family in Kunming" was accepted by the Urban Transport of China.

Chao-qin Feng is a third-year master's student at the Faculty of Transportation Engineering at Kunming University of Science and Technology, with research interests in transportation big data. One paper entitled "Forecasting for urban traffic grid clusters based on Bi-LSTM" was accepted by the Journal of Chongqing University.

Jie-mei Li is a professor at the Faculty of Transportation Engineering at Kunming University of Science and Technology, Ph.D. She has presided over two National Natural Science Foundations of China, published more than 30 academic papers in core journals at home and abroad, and edited and participated in the publication of four books.

Authors and Affiliations

Xia Yang¹, Jin Zhang², Kai Wang³, Chao-qin Feng⁴, Jie-mei Li⁵

Xia Yang
Tel: +86-18314423735
Email: 330029176@qq.com

Jin Zhang, Corresponding Author
Tel: +86-15987199303
Email: 55715389@qq.com

¹, ², ³, ⁴, ⁵ Kunming University of Science and Technology, China
Kai Wang  
Email: 1652353513@qq.com

Chao-qin Feng  
Email: 1279351025@qq.com

Jie-mei Li  
Email: 1263703867@qq.com

1/2/3/4/5  Faculty of Transportation Engineering, Kunming University of Science and Technology  
727 Jingming South Road, Kunming, Yunnan Province, China 650500