Modeling of Residents' Travel Mode Choice Decisions during Peak Commuting Hours via RL-SVM Method

Tian Xie[1,2,a,*], Yan Fang[2]

[1]China Design Group Co., Ltd., Nanjing, China
[2]College of Transport & Communications, Shanghai Maritime University, Shanghai, China
[a]sdodingnan123@gmail.com
**Corresponding author*

Abstract: *In response to worsening traffic congestion, cities worldwide have prioritized the development of public transportation systems, striving to become 'public transport cities.' To address this issue, conducting an in-depth study of residents' travel choice behavior during peak hours is of vital importance. The paper aims to consider the key factors affecting residents' travel decisions by utilizing revealed preference (RP) samples and proposes a novel RL-SVM model based on the random parameter logit (RL) and support vector machine (SVM) theories for travel mode prediction. Comparative experimental analysis shows that our model outperforms traditional prediction methods regarding classification sensitivity. Therefore, we can further evaluate the anticipated impact of implementing public transport priority strategies to obtain the internal shift patterns of commuters' travel modes from private cars to public transport. It holds significant implications for promoting healthy and sustainable urban development.*

Keywords: *Commuting travel; Public transport; Travel mode choice; Random parameter logit; Support vector machine*

1. Introduction

The completeness of public transport is an essential characteristic for measuring the urban modernization level, as it supports the orderly operation of the urban economy and social activities. To effectively implement public transport priority strategies, it is crucial to establish the goal of 'public transport leading urban development' and concentrate efforts on creating a 'public transport city.' It will help accomplish an urban development pattern where the speed of public transport development exceeds that of urbanization, and the growth rate of public transport capacity exceeds car ownership. The percentage of public transport utilization should be the primary factor considered when evaluating the reliability of a city's transportation system.

Scholars have done much research on residents' travel mode choice decisions using disaggregation models and machine learning algorithms. The disaggregation model, such as logit, assumes a random utility term and boasts a higher data utilization rate. Xianyu[1] used the multinomial logit (MNL) and co-evolutionary method to determine the relationship between residents' travel mode choices and trip-chaining behavior. Ermagun et al.[2] introduced a two-level cross-nested logit (CNL) model to quantify the active travel component of the public transport mode for journeys to school. Liu et al.[3] analyzed the residents' motorized travel decisions by incorporating the psychological factors into a random coefficient logit model. In recent years, there has been mounting attention on data mining, leading to a growing interest in developing supervised learning algorithms. Unlike statistical models, such algorithms can intuitively capture nonlinear relationships among multiple variables. Omrani[4] applied various machine learning algorithms to predict individuals' travel modes in Luxembourg City. The cross-validation results using the average probability of correct assessment score (APCA) proved that the ANN network had a more advanced model performance. Chen et al.[5] presented an SVM model based on planned behavior while considering the latent variables of low-carbon travel to improve the travel mode prediction performance; Xu et al.[6] discussed the critical factors that affect residents' travel satisfaction based on the MNL model. Furthermore, they used the SVM to evaluate residents' travel satisfaction. Presently, several researchers have drawn specific conclusions regarding residents' travel behavior. However, the reliability of the traditional model is often hindered by factors such as insufficient information on RP samples and limitations of the algorithms used. In that case, In this study, we propose a novel RL-SVM approach to predict the travel mode choices of residents. The contribution of this paper is in the following ways:

(1) Our ensemble model combines the strengths of the stochastic utility maximization theory and the support vector classifier thoughts. It has performed better for multi-classification tasks, particularly for small sample sizes.

(2) Previous studies rarely focused on analyzing the behavior of residents changing transport modes during peak commuting hours. This research can motivate commuters to opt for low-carbon travel modes like subway or bus rapid transit.

2. RL-SVM theory

2.1 Utility maximization theory

The random parameter logit model has developed with its distinctive advantages of overcoming IIA[7] and preference for randomness. Unlike the MNL model, the condition constraint of the RL model is relatively flexible. The primary thought of RL is the utility maximization theory[7]. The total utility *U* comprises a deterministic term *V,* and a stochastic term ξ. ξ represents the error portion that cannot be directly estimated. In general, the more concise linear utility function expression is:

$$V_{ik} = \sum_{l=1} \beta_l X_{ikl} \tag{1}$$

In the above equation, X_{ikl} represents the *i*-th variable of the *i*-th sample for the category *k*. β_l is the undetermined coefficient. The probability selection expression of the RL model is:

$$P_{ik} = \int (e^{V_{ik}} / \sum_{k=1} e^{V_{ik}}) f(\beta | \theta) d\beta \tag{2}$$

P_{nk} is the probability weight; $f(\beta | \theta)$ represents a specific distribution density function, such as log-normal, normal, uniform distribution, etc. θ is the unknown parameter of the $f(\beta)$, such as the mean (*M*) and standard deviation (*S*) of the normal distribution.

2.2 Support vector machine

Support vector classifier is widely used because of its powerful iterative and nonlinear fitting ability. The basic theory of this technique is to seek the optimal hyperplane by mapping the sample points *x* in the low-dimensional space to the high-dimensional feature space through a specific nonlinear method. The standard expression of SVM based on the structured risk minimization (SRM) idea[5] is defined as:

$$\begin{aligned} &\min_{\omega,b} Q = \frac{1}{2} ||\omega||^2 \\ &s.t. \quad 1 - y_i(\omega^* x_i + b) \leq 0 \end{aligned} \tag{3}$$

Transforming into the dual problem using the Lagrangian function:

$$\begin{aligned} &\max_{\alpha} Q = \sum\nolimits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &s.t. \quad \sum\nolimits_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{4}$$

Therefore, the final solution is obtained by iterative SMO algorithm:

$$\omega^* = \sum\nolimits_{i=1}^{n} \alpha_i^* y_i x_i \quad b^* = y_i - \sum\nolimits_{i=1}^{n} \alpha_i y_i x_i^T x_i \tag{5}$$

In the above equation, ω, b are the weight vector and bias, respectively. α is the Lagrangian multiplier variable. $x_i, x_i, y_j, y_j (i, j = 1, 2, .., n)$ represents the different feature matrices and class labels.

2.3 RL-SVM model

we introduce an RL-SVM approach to address the limitations of traditional SVM classifiers in achieving the SRM objective. The approach is inspired by a binary classification algorithm that combines the logistic and SVM models proposed by Hua Z et al.[8]. It takes advantage of the RL model's ability to fit any random utility function and effectively captures the heterogeneity between individuals with different feature dimensions. The fundamental thought of the RL-SVM model is to seek the best classification hyperplane and utilizes the probability weight estimated by the RL model to adjust the SVM model's classification outcomes. The SVM structure can take on binary or multi-classification forms, such as one vs. one (OVO) or one vs. rest (OVR) structures. Assuming that both training set *a* (with *k* categories) and test set *b* are known, the steps to implement an RL-SVM algorithm for classification purposes are as follows:

Step 1: we establish an RL model M_R and use it as a feature extractor. This model is utilized to estimate the coefficient β for each variable and calculate the probability weights P_{ik} of each sample *i* using different utility functions V_{ik}. The resulting vector P_{ik} is a latent representation that M_R extracts.

Step 2: Concatenating the original feature matrix and the latent vectors as the overall input for training the SVM classifier M_S. Moreover, the fitted model outputs the class C_i^* with the highest prediction probability.

Step 3: Ensemble model optimization. Parameter tuning is conducted to obtain the best models M_R^*, M_S^*, with the average accuracy and micro-f1 score selected as the classification evaluation metrics.

3. Residents' travel mode prediction model

3.1 Residents' travel characteristic attributes system

Table 1: Residents' travel characteristic variables

Category	Variable	Parameter	Variable Definition
Individual attributes	Driver's license (*x*1)	B_1	0—No; 1—Yes
	Age (*x*2)	B_2	0—1-17; 1—18-30; 2—31-45; 3—46-60; 4—61-100
	Gender (*x*3)	B_3	0—Female; 1—Male
	Education (*x*4)	B_4	0—High school or below; 1—College and bachelor; 2—Postgraduate or above
	Migrant population (*x*5)	B_5	0—No; 1—Yes
	Occupation (*x*6)	B_6	0—Private; 1—Freelance; 3—Others 2—Officeworkers or school-goers
Family attributes	Number of cars (*x*7)	B_7	Continuous variable, unit—vehicle
	Number of bicycles (*x*8)	B_8	Continuous variable, unit—vehicle
	Family members (*x*9)	B_9	Continuous variable
	Consumption (*x*10)	B_{10}	0—Low level; 1—Medium level;2—High level
	Residential location (*x*11)	B_{11}	0—Central area; 1—Suburban area; 2—General residential area
	Annual household income (*x*12)	B_{12}	0—0-69,999; 1—70,000-119,999; 2—120,000-199,999; 3—200,000 or above, unit—Yuan
Individual travel mode preference	Public transport (*y*1)	A_1	0
	Cycling (*y*2)	A_2	1
	Taxi (*y*3)	A_3	2
	Private cars (*y*4)	A_4	3
	Other modes (*y*5)	A_5	4
Traffic accessibility	In-vehicle time (*x*13)	B_{13}	Continuous variable, unit—minutes
	Out-of-vehicle time (*x*14)	B_{14}	Continuous variable, unit—minutes
	Driving costs (*x*15)	B_{15}	Continuous variable, unit—Yuan
	Public transport fare (*x*16)	B_{16}	Continuous variable, unit—Yuan
	Parking Fee (*x*17)	B_{17}	Continuous variable, unit—Yuan
	Travel comfort and safety (*x*18)	B_{18}	0—Low; 1—Medium; 2—High; 3—Very concerned
Travel characteristic	Weather conditions (*x*19)	B_{19}	0—Sunny; 1—Foggy; 2—Rainy; 3—Snowy; 4—Others
	Average daily commuting distance (*x*20)	B_{20}	0—0-5km; 1—6-10km; 2—11-15km; 3—16km or above
	Transfer times (*x*21)	B_{21}	Continuous variable

This paper uses the travel survey data of commuters in Dongcheng District, Beijing, in 2021 for analysis, with 2863 survey questionnaires received. We eliminated the RP samples with incomplete information to establish the residents' travel characteristic attributes system.

Specifically, this research focuses on four transport modes: public transport, taxi, cycling, and private cars. The system's variables involve personal socioeconomic attributes, traffic accessibility, and travel characteristic variables. Among them, continuous variables use actual survey values, and categorical variables are considered dummy variables. Table 1 shows the final residents' travel characteristic attribute system. The *in-vehicle time* refers to the duration individual spends traveling in a particular transport mode during peak commuting hours. In contrast, *out-of-vehicle time* is the duration individuals spend transferring, walking, and waiting for public transport. The *taxi* transport mode includes online car-hailing services and licensed taxis.

3.2 RL parameter calibration

In our study, we introduce ridge regression[9] and information gain[10] methods to identify the most significant variables. Depending on the feature analysis, five variables ($x1$, $x3$, $x7$, $x8$, $x10$) were excluded due to strong correlations. The remaining 16 variables were used to construct the RL using the Stata software. After conducting several pre-modeling experiments, the results indicate that the model exhibits excellent convergence only when the variables ($x12$, $x13$, $x14$, $x17$, $x21$) follow a normal distribution while the rest are fixed values. Most variables have a significance level of 90% or higher.

Table 2: RL model results

Variables	Random Parameter Logit Estimated Coefficient(ρ^2 =0.229)				
	Public transport	Cycling	Taxi	Private cars	Other modes
Constant	-9.348***	0.379**	-1.639**	-2.504*	5.293***
B_2_M	1.409**	0.612*	1.323*	0.821*	0.155*
B_4	-1.593**	1.025**	/	/	3.376**
B_5	1.516*	2.874**	/	/	/
B_6	1.247***	0.384***	-0.678***	1.792***	-.0.914*
B_9	/	-1.185**	/	/	0.286**
B_{11}	2.176*	-0.504**	1.548*	1.193*	-2.055***
B_{12}_M	1.513*	-0.706*	-1.153**	-0.863**	/
B_{12}_S	0.016***	0.054***	0.012*	0.027**	/
B_{13}_M	2.562*	-0.724*	2.991***	2.702*	-0.261*
B_{13}_S	0.055*	0.036**	0.018*	0.138***	0.029*
B_{14}_M	-2.029***	/	0.992**	/	0.547***
B_{14}_S	0.117**	/	0.007*	/	0.062*
B_{15}	0.258*	0.083***	-2.152***	-1.89**	-.618***
B_{16}	0.525*	/	/	/	/
B_{17}_M	/	/	-1.084*	-2.143**	0.572**
B_{17}_S	/	/	0.129***	0.152***	0.005**
B_{18}	2.071**	1.508***	1.098***	1.255*	0.276*
B_{19}	1.905***	-2.601**	-0.977**	-1.337**	-0.529***
B_{20}	0.831*	-2.092***	1.249*	0.513***	-0.224***
B_{21}_M	-0.749***	-0.217*	/	1.255*	/
B_{21}_S	0.024*	0.066*	/	0.074***	/

Note: "/" represents p>0.1; "*" represents 0.05<p<=0.1; "**" represents 0.01<p<=0.05; "***" represents p<=0.01. p is the level of statistical significance.

According to Table 2, the fitness ratio ρ^2 of the RL model falls within an acceptable range of [0.2,0.4], indicating an excellent model-fitting effect. Additionally, the potential relationship between travel time and an individual's travel mode choice follows the normal distribution, with the variance reflecting individual preference differences consistent with reality. As a result, we can determine the final utility function expression of the model for different travel modes. Using the public transport mode as an illustration:

$$V_{PT} = -9.348 + 1.409x_2 - 1.593x_4 + 1.516x_5 + 1.247x_6 + 2.176x_{11} + N(1.513, 0.016)x_{12} + N(2.562, 0.055)x_{13}$$

$$+N(\text{-}2.029, 0.117)x_{14} + 0.258x_{15} + 0.525x_{16} + 2.071x_{18} + 1.905x_{19} + 0.831x_{20} + N(-0.749, 0.024)x_{21}$$

V_{PT} represents the tendency of residents to choose public transport as their preferred transport mode during peak commuting hours.

3.3 RL-SVM model establishment

After completing parameter estimation for all variables in section 3.2, we can easily calculate each sample's feature vectors P_{ik} of different travel modes. It will be an additional portion of the input utilized for the subsequent training of the SVM classifier. Based on the thoughts of multiclass SVM, we choose the OVR-SVM structure to predict each travel mode employing a partial binary tree approach[11]. Thus the modeling flow of OVR-SVM is shown in Figure 1:

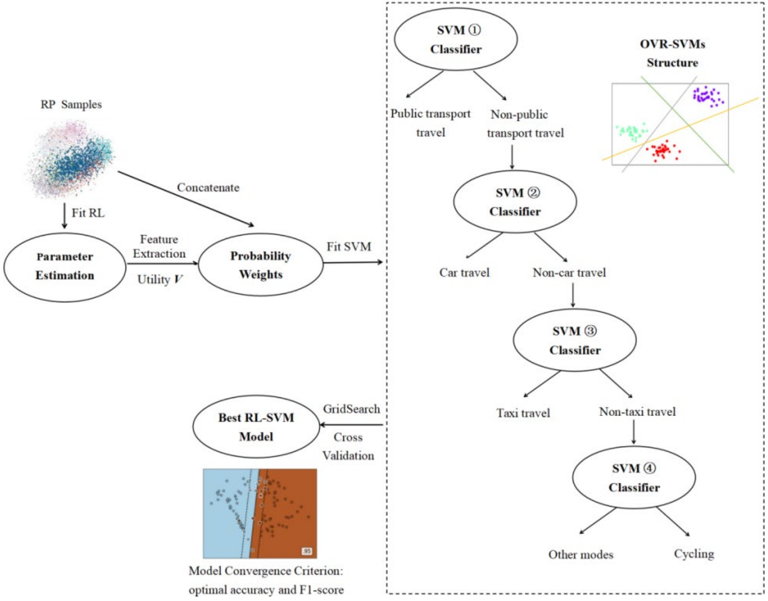
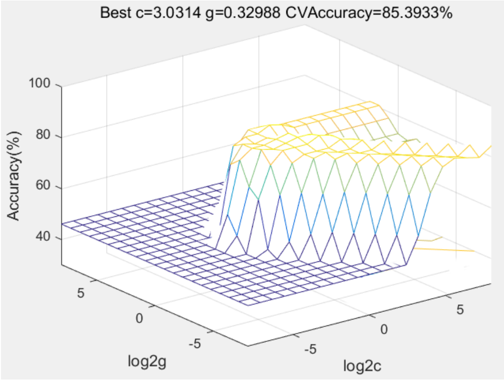


Figure 1: OVR-SVM Modeling Flow

The SVM model was implemented using the Python Scikit-learn library and adopted an RBF kernel form. The optimal model parameters were obtained using the improved GridSearch[12]. The best value of penalty coefficient *c* was 3.0314, and kernel function coefficient *g* was 0.3298, yielding an average accuracy of 85.393% for each travel mode, as shown in Figures 2 and 3.



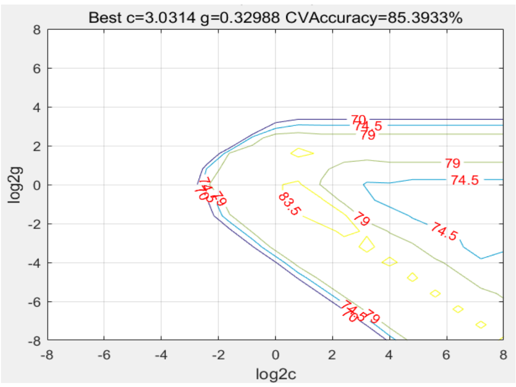


Figure 2: SVM GridSearch 3D view

Figure 3: SVM parameter selection contour plot

To further determine the effectiveness of our proposed model in multi-classification tasks, we consider comparing it with the general algorithms used by previous scholars in their research of residents' travel behavior. The baseline models include the nested-logit[2] (NL), LightGBM[13], and the least squares support vector classifier[11] (LS-SVM) model. As shown in Figure 4, our model achieves the highest classification ability, far exceeding other models. Among them, the NL and LS-SVM models perform the worst, with sensitivity and F1 scores of only about 80%. Meanwhile, using an ensemble of RL and SVM models is a more practical approach for making robust classification predictions than a single model. The F1 score of the RL-SVM model has improved by 1.1% and 1.3%, respectively, compared to LightGBM and RL models. Moreover, our model's high sensitivity indicates its ability to accurately identify positive samples, reducing the possibility of underreporting.

Figure 5 shows the multiclass ROC curve, with the test set classification AUC value is 0.84. The results reflect that the RL-SVM model exhibits a remarkable capacity to differentiate between various travel modes. Notably, the AUC for the category of other transport modes is merely near 0.75, while private cars and public transport modes both obtain high scores. It can be attributed to the collected samples' relatively smaller proportion of walking, shuttle bus travel, etc. Thus the findings reveal two essential outcomes. Firstly, the RL's probability weights depict the heterogeneity among RP samples concerning distinct feature dimensions. Employing these weights as an additional part of the SVM classifier's input can enhance the model's interpretability and robustness. Secondly, the ensemble model leverages SVM's exceptional classification competence, particularly for small datasets, resulting in a reliable classifier with superior generalization abilities.

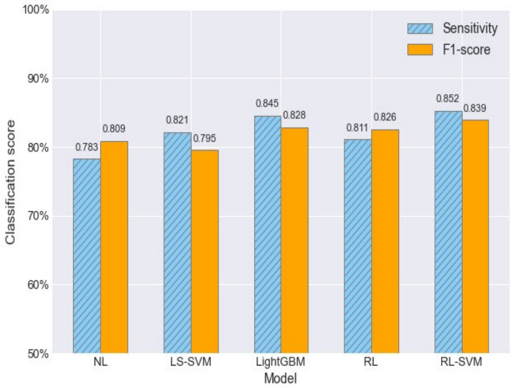


Figure 4: Comparative results of each model

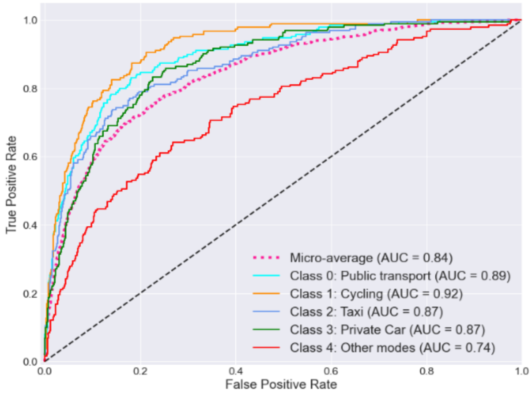


Figure 5: RL-SVM Multi-classification ROC curve

3.4 Analysis of residents' public transport travel intentions

In order to reduce the ever-increasing number of private cars, it is imperative to conduct a thorough analysis of commuters' willingness to shift from private cars to public transport during peak hours. As such, we assume that the social and economic conditions of residents listed in Table 1 remain constant and reclassify the RP samples into three categories: public transport, private cars, and other modes. Additionally, we consider using the RL-SVM model to calculate the impact of different traffic accessibility factors on residents' travel decisions.

Table 3: The prediction change of the RL-SVM model

Variable assignment condition		The probability of shifting from cars to public transport increases(%)
Travel Time ($x13$, $x14$)	Both reduce by one level	5.27%
	Both reduce by two levels	7.08%
Travel Cost ($x15$,$x16$,x17)	Fuel cost increases by 1 yuan per liter	6.57%
	Fuel cost increases by 2 yuan per liter	10.02%
	5 yuan traffic congestion fees	4.61%
	10 yuan traffic congestion fees	6.29%
	Parking fee increases by 3 yuan	6.83%
	Parking fee increases by 6 yuan	10.58%
	Public transport fare reduces by 10%	1.21%
	Public transport fare reduces by 20%	1.65%
$x18$ increases by one level		3.94%
$x18$ increases by two levels		5.81%

As shown in Table 3, there is a low correlation between residents' travel preferences and public transport fares. The travel time variables improve the model's ability to distinguish between private cars and public transport modes. As the travel time values gradually decrease, the willingness of car commuters to take public transport increases strongly. This result reflects that residents are concerned about the value of travel time. The intention of residents to choose public transport mode can be affected by factors such as long walking or transfer distances. Besides, the fuel price and parking fees significantly impact residents' travel mode choice decision-making than congestion charges. Increasing the travel costs may effectively reduce car usage, shifting towards public transport travel.

Additionally, individuals have high demands for traveling safety and comfort. The increase of two levels in these factors can raise the proportion of public transport usage by 5.81%. Residents have developed a deep subjective awareness of various travel modes, such as choosing a conservative public transport mode during stormy weather or traffic congestion conditions.

4. Conclusion

Optimizing urban passenger transport corridors and traffic travel structure poses a formidable challenge due to the distinct impact of factors such as commuting distance, transportation connectivity, and environmental conditions on residents' travel mode choices during peak hours. Therefore, This study introduces an RL-SVM multi-classification model that predicts residents' transport mode choices by combining the concepts of stochastic mixed utility maximization and SRM. Experimental results demonstrate that our ensemble model has a higher multiclass prediction sensitivity than the baseline model, with an F1 score of up to 0.839. It can efficiently capture the potential nonlinear crucial features and the multidimensional diversity across samples. Furthermore, we investigate the impact of different traffic accessibility variables on private car owners' travel preferences. The findings suggest that fuel prices and parking fees significantly influence residents' travel decisions more than travel time costs. The probability of shifting to public transport can increase to 10.02% and 10.58%, respectively. These results provide a theoretical basis for improving urban transportation management and enhancing the effectiveness of the public transport system.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 71871137.

References

[1] Xianyu J. An exploration of the interdependencies between trip chaining behavior and travel mode choice[J]. Procedia-Social and Behavioral Sciences, 2013, 96: 1967-1975.

[2] Ermagun A, Levinson D. Public transit, active travel, and the journey to school: a cross-nested logit analysis[J]. Transportmetrica A: Transport Science, 2017, 13(1): 24-37.

[3] Liu J, Hao X. Travel mode choice in city based on random parameters logit model[J]. Journal of Transportation Systems Engineering and Information Technology, 2019, 19(5): 6.

[4] Omrani H. Predicting travel mode of individuals by machine learning[J]. Transportation research procedia, 2015, 10: 840-849.

[5] Chen Y, Chen L, Zha Q, et al. Forecasting model of travel mode based on latent variable SVM[J]. Journal of Southeast University, 2016, 46(6): 1314-1317.

[6] Xu Z, Shao C, Wang S, et al. Analysis and Prediction Model of Resident Travel Satisfaction[J]. Sustainability, 2020, 12(18):7522.

[7] Hensher D A, Greene W H. The mixed logit model: the state of practice and warnings for the unwary[M]. General Information, 2002, 30(2): 133-176.

[8] Hua Z, Yu W, Xu X, et al. Predicting corporate financial distress based on integration of support vector machine and logistic regression[J]. Expert Systems with Applications, 2007, 33(2):434-440.

[9] McDonald Gary C. Ridge regression[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2009, 1(1): 93-100.

[10] Omuya E O, Okeyo G O, Kimwele M W. Feature selection for classification using principal component analysis and information gain[J]. Expert Systems with Applications, 2021, 174: 114765.

[11] Yu Q, Liu R. Least Squares Twin SVM Based on Partial Binary Tree Algorithm[C]//2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2018: 1-4.

[12] Syarif I, Prugel-Bennett A, Wills G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance [J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2016, 14(4): 1502-1509.

[13] Zhang Y, Zhu C, Wang Q. LightGBM-based model for metro passenger volume forecasting[J]. IET Intelligent Transport Systems, 2020, 14(13): 1815-1823.