Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 9219176, 22 pages
https://doi.org/10.1155/2021/9219176

Research Article

Modeling Intercity Travel Mode Choice with Data Balance Changes: A Comparative Analysis of Bayesian Logit Model and Artificial Neural Networks

Xiaowei Li[1], **Yuting Wang**[1], **Yao Wu**[2], **Jun Chen**[1], and **Jibiao Zhou**[3]

[1]*School of Civil Engineering, Xi'an University of Architecture & Technology, Xi'an 710055, China*
[2]*School of Modern Posts and Institute of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[3]*College of Transportation Engineering, Tongji University, Shanghai 201804, China*

Correspondence should be addressed to Yao Wu; wuyao@njupt.edu.cn

Received 1 July 2021; Revised 26 August 2021; Accepted 1 September 2021; Published 14 September 2021

Academic Editor: Xiangyang Guan

This study conducts a comprehensive comparative analysis of regression-based multinomial models and artificial neural network models in intercity travel mode choices. The four intercity travel modes of airplane, high-speed rail (HSR), train, and express bus were used for analysis. Passengers' activity data over the process of intercity travel were collected to develop the models. The standard multinomial logit (MNL) regression and Bayesian multinomial logit (BMNL) regression were compared with the radial basis function (RBF) and multilayer perceptron (MLP). The results show that MLP performs best in terms of predictive accuracy, followed by BMNL and MNL, and RBF is the least accurate. The performances of all models were examined against changes in data balance, and it was found that rebalancing can improve fitting performance while slightly reducing the predictive performance. This comparative study and its parameter estimation shed new light on the comparison of traditional and emerging models in travel behavior studies, and the findings can be used as heuristic guidance for all stakeholders.

1. Introduction

To model passengers' travel behaviors is of value to better understand mobility modes in the complex travel environment [1]. Policies and managerial strategies rely on the accurate estimation of travel mode choices of passengers. In 2020, COVID-19 has profoundly influenced passengers' travel behaviors, causing a dramatic shift in intracity and intercity mobility modes, inevitably affecting society, production, and the global economy. Scholars have investigated contextual factors that influence travel modes, aiming to better understand passengers' choices and develop suitable models.

Previous studies have shown that travel mode choice can be affected by social and demographic factors, including gender [2–4], age [4–6], occupation [3], income [2, 4, 7, 8], and car ownership [4]. Miskeen et al. [4] found that males were more likely to use public transportation than cars, while females were less likely to shift to public transportation. Cheng et al. [6] indicated that age was the most significant individual-related attribute. Tourists were more likely to choose a plane or train than a coach [3]. Forinash et al. [7] found that high- and low-income groups preferred air travel and bus, respectively. It was similarly reported that an increase in passengers' incomes decreased their use of buses [4]. Lower-income individuals were found to be more sensitive to cost and less sensitive to out-of-vehicle time than middle- and high-income individuals [8]. Related attributes, such as travel demand, service quality of transport modes, and accessibility of transportation hubs, have been found to influence travel mode choices [1, 3, 9, 10].

The most widely used modeling techniques in travel mode choice are discrete choice models, such as the binomial logit (BL) [11], multinomial logit (MNL) [4, 12], multinomial probit (MNP) [3], nested logit (NL) [13, 14], and mixed logit (ML) models [15–17], which have high interpretability of estimation results on input variables, as well as high transferability and validity. Regression-based models form maximum likelihood estimates of parameters [4, 5, 11, 12, 17, 18]. Apart from the popular logit model, Bayesian parameter estimation methods have shown good accuracy and performance [19–22]. For example, Wong and Farooq [23] developed an algorithm based on the restricted Boltzmann machine, which has multiple discrete-continuous layers and can be expressed as a variational Bayesian inference optimization problem.

Emerging machine learning techniques have been studied for travel mode choice [24–34]. Lindner et al. [33] found an artificial neural network (ANN) and classification tree (CT) to outperform binary logit regression in motorized travel mode choice. Cheng et al. [6] found the random forest (RF) to have significantly better prediction accuracy than support vector machine (SVM), adaptive boosting (AdaBoost), and MNL in modeling travel mode choice. Zhao et al. [24] compared the model development, evaluation, and behavior interpretation of MNL and ML with that of the naive Bayes, CT, AdaBoost, bag fruit tree, RF, SVM, and ANN machine learning classifiers. Among machine learning approaches, the multilayer perceptron (MLP) and radial basis function (RBF) have been widely applied due to their better classification accuracy compared to naive Bayes, K-nearest neighbors, and backpropagation neural networks [35–37]. Hence, they have potential use in the study of travel mode choice.

The influence of data balance on the accuracy of multimode choice models has not been widely reported. Imbalanced sample data can influence the accuracy of estimation in multiclass discrete choice prediction [38, 39], and methods such as oversampling and undersampling have been proposed to address this issue [40, 41]. However, there is no commonly agreed best method to resolve this issue in multiclass classification. This is a well-known issue in travel mode choice, and the effectiveness of rebalancing methods when using different regression-based and neural network models in empirical studies of modeling travel mode choice requires study.

This study has three objectives: (1) to investigate the predictive performance of modeling techniques including Bayesian multinomial logit (BMNL), MNL, MLP, and RBF for intercity travel mode choice; (2) to assess the predictive performance of the above techniques after data balancing; and (3) to evaluate the factors affecting intercity travel mode choice and their relative importance using a comprehensive dataset. Passengers' activity data over the whole process of intercity travel were collected. The travel modes of airplane, HSR, train, and express bus were investigated. The BMNL, MNL, MLP, and RBF models were developed and validated. A receiver operating characteristic (ROC) curve and confusion matrix were employed to evaluate the models' predictive performance.

The remainder of this paper is organized as follows. Section 2 introduces the methodological background of the selected models, followed by a description of the dataset in Section 3. Section 4 presents the results and findings. We summarize our conclusions and propose future work in Section 5.

2. Methodology

2.1. Bayesian Multinomial Logit Model. MNL regression generalizes logistic regression into multiclass problems that consist of more than two possible discrete groups [1, 19]. It can be expressed as [19]

$$P\left(Z_j = i\right) = \frac{\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}{\sum_{i=1}^{I}\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}, \tag{1}$$

where $\mathbf{X} = [x_{j1}, x_{j2}, \ldots, x_{jk}]$ is a vector of independent variables x_{jk}, $\beta = [\beta_0^i, \beta_1^i, \beta_2^i, \ldots, \beta_k^i]^T$ is the corresponding coefficient vector, and $Z_j = i$ is the choice of travel mode i for the j^{th} observation.

The likelihood function can be expressed as

$$\begin{aligned} f(\mathbf{Z} \mid \boldsymbol{\beta}) &= \prod_{j=1}^{N}\prod_{i=1}^{I}\left[\varepsilon_{ji} \times P\left(Z_j = i\right)\right] \\ &= \prod_{j=1}^{N}\prod_{i=1}^{I}\left[\varepsilon_{ji} \times \frac{\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}{\sum_{m=1}^{I}\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}\right], \end{aligned} \tag{2}$$

where N is the number of samples, I is the number of outcomes, and ε_{ij} equals 1 when the discrete outcome of sample j is i and is 0 otherwise.

The Bayesian approach using Markov chain Monte Carlo (MCMC) was utilized for model estimation. Based on Bayesian inference, the posterior joint distribution of parameters β conditional on dataset $\mathbf{Z}$ can be estimated as [19]

$$f(\boldsymbol{\beta} \mid \mathbf{Z}) = \frac{f(\mathbf{Z}, \boldsymbol{\beta})}{f(\mathbf{Z})} = \frac{f(\mathbf{Z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})}{f(\mathbf{Z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})d(\boldsymbol{\beta})} \propto f(\mathbf{Z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}), \tag{3}$$

where $f(\mathbf{Z}, \beta)$ is the joint probability distribution of $\mathbf{Z}$ and β, $f(\mathbf{Z} \mid \beta)$ is the likelihood of the conditional function based on β, and $\pi(\beta)$ is the prior distribution of β. Due to lack of information on the random parameters, we used the non-informative prior distributions [1]:

$$\boldsymbol{\beta} \sim N\left(0_k, 10^6 M_k\right), \tag{4}$$

where 0_k is a vector of zeros and M_k is the $k \times k$ identity matrix.

The posterior joint distribution can be derived as [42]

$$f(\boldsymbol{\beta} \mid \mathbf{Z}) \propto f(\mathbf{Z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}) = \prod_{j=1}^{N}\prod_{i=1}^{I}\left[\varepsilon_{ji} \times \frac{\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}{\sum_{m=1}^{I}\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}\right]$$
$$\times \prod_{j=1}^{N}\prod_{i=1}^{I}\left[\frac{1}{\sqrt{2\pi}\,10^3}\exp\left(-\frac{1}{2}\frac{\left(\beta_k^i\right)^2}{10^6}\right)\right] \propto \exp\left\{\sum_{j=1}^{N}\sum_{i=1}^{I}\left[\varepsilon_{ij} \times \frac{\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}{\sum_{m=1}^{I}\exp\left(\beta_0^i + \beta_1^i x_{j1} + \beta_2^i x_{j2} + \cdots + \beta_k^i x_{jk}\right)}\right] - \sum_{j=1}^{N}\sum_{i=1}^{I}\left[\frac{1}{2}\frac{\left(\beta_k^i\right)^2}{10^6}\right]\right\}. \tag{5}$$

2.2. Radial Basis Function (RBF) Neural Network. An RBF neural network is a typical three-layer neural network model with input, hidden, and output layers, as shown in Figure 1, where k is the number of input variables, H is the number of hidden neurons, I is the number of output neurons (travel modes), $\mathbf{X} = [x_1, x_2, \ldots, x_k]^T$ is the input, $\mathbf{Y} = [y_1, y_2, \ldots, y_I]^T$ is the output, and w_{hi} is the connection weight of the h^{th} hidden layer neuron to the i^{th} output layer neuron.

A Gaussian function is generally used as the hidden layer excitation function. The output of the h^{th} hidden layer neuron is

$$G_h(x) = e^{\left(-X - c_h^2/2\sigma_h^2\right)}, \quad h = 1, 2, \ldots, H, \tag{6}$$

and the linear mapping relationship between $G_h(x)$ and the i^{th} output layer neuron is

$$G_h(x) = y_i = \sum_{h=1}^{H} w_{hi} G_h(x), \quad i = 1, 2, \ldots, I, \tag{7}$$

where c_h and σ_h are, respectively, the center vector of the Gaussian function and the base width of the h^{th} hidden neuron.

RBF has been criticized as a “black box” that lacks interpretability [43]. Various tools have been developed to address this issue, the most common being variable importance analysis [44–46], which measures the relative importance of each independent variable in predicting dependent variables.

2.3. Multilayer Perceptron (MLP) Neural Network. MLP is a commonly used supervised ANN model that can be used for both pattern recognition and function approximation. Compared to RBF, MLP can have multiple hidden layers (shown in Figure 2) [47]. The hyperbolic tangent function is selected as the activation function of MLP hidden neurons. The output from a hidden neuron is

$$y = \frac{e^u - e^{-u}}{e^u + e^{-u}}, \tag{8}$$

and the connection weight is the output of the net function,

$$u = b + \sum_{p=1}^{k} w_p x_p, \tag{9}$$

where k is the number of inputs, x_p is the input, w_p is the weight of the corresponding input (w_p; $1 \le p \le k$), b is the bias weight, and the Levenberg–Marquardt training algorithm is selected [28].

2.4. Model Comparison and Validation. The multi-classification confusion matrix (see Table 1) is used to calculate the accuracy of each model [48], where s_{im} is the number of samples in which mode i is predicted as mode m. The recall and precision of mode i are

$$\text{Recall}_i = \frac{S_{ii}}{\sum_{m=1}^{I} s_{im}}, \qquad \text{Precision}_i = \frac{S_{ii}}{\sum_{m=1}^{I} S_{mi}}, \tag{10}$$

and the accuracy of the model can be calculated as [1]

$$\text{accuracy} = \frac{\sum_{i=1}^{I} S_{ii}}{N}. \tag{11}$$

The ROC curve and area under the curve (AUC) were also used to measure the predictive ability. A higher AUC value indicates better predictive accuracy [42, 49].

3. Data Collection

Data from Li et al.’s work [42] were used in this study. A total of 985 random samples collected in Xi’an from March 1–10, 2018, were used for analysis, where 161 samples reported the choice of airplane, accounting for 16.3% of intercity travel records, and 369 (37.5%) were reported as HSR, 299 (30.4%) as train, and 156 (15.8%) as express bus. Among them, 80% were randomly selected for training, and the remaining 20% were used for prediction. In addition to the original information included in the database, the travel distance was calculated by Baidu Maps using the real route between the cities of origin and destination. The intercity travel time was obtained according to the identification number of the carrier, transportation schedule, and origin and destination cities.

Undersampling and oversampling are the most frequently used techniques to balance data for machine learning and pattern recognition [38–41]. Undersampling

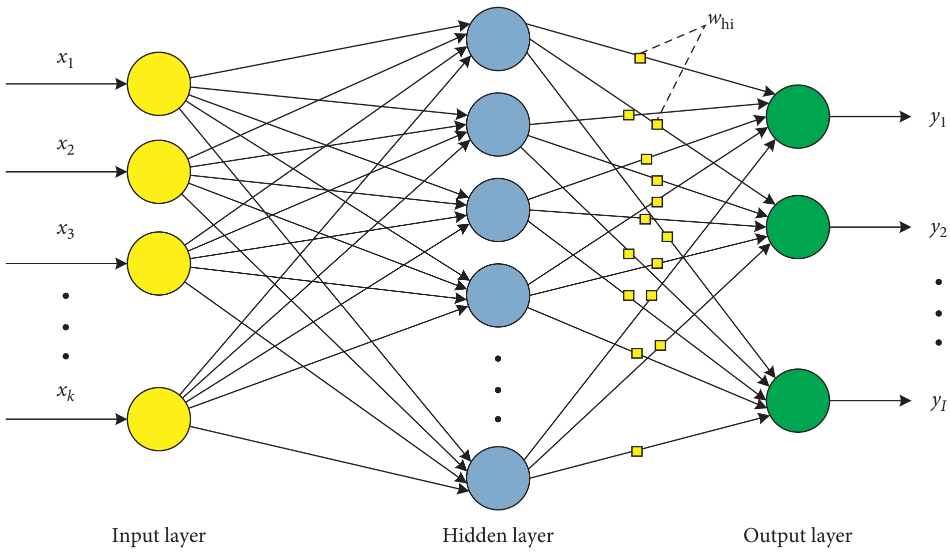

Figure 1: RBF network.

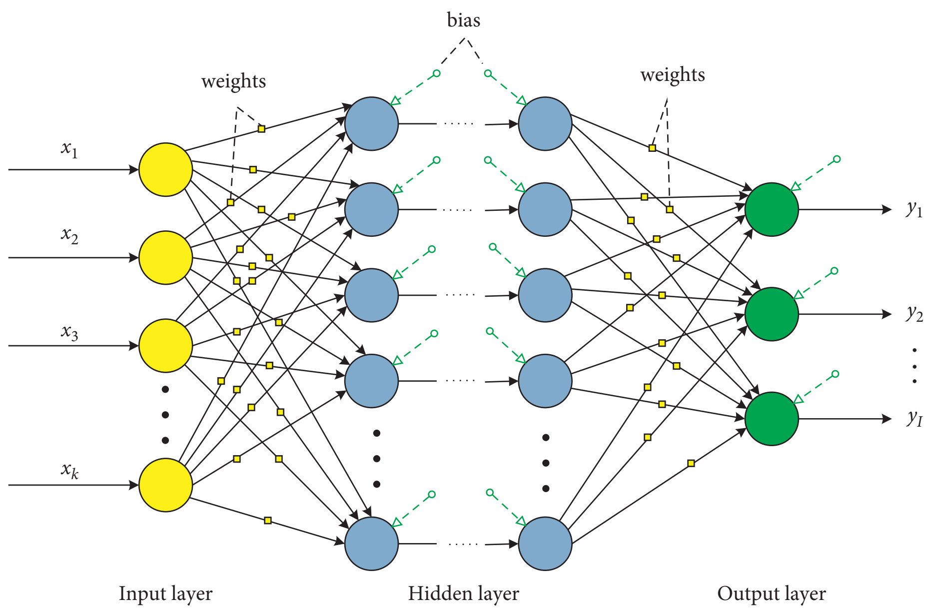

Figure 2: General topology of MLP.

Table 1: Multiclassification confusion matrix.

		Predictive class					
	Mode	1	2	3	...	I	Recall
Actual class	1	s_{11}	s_{12}	s_{13}	...	s_{1I}	Recall_1
	2	s_{21}	s_{22}	s_{23}	...	s_{2I}	Recall_2
	3	s_{31}	s_{32}	s_{33}	...	s_{3I}	Recall_3
	...	...	...	...	...	...	
	I	s_{I1}	s_{I2}	s_{I3}		s_{II}	Recall_I
	Precision	Precision_1	Precision_2	Precision_3		Precision_I	Accuracy

achieves relative equilibrium among classes by reducing the number of samples of classes with more samples. Using this method, the number of samples of each travel mode was 156, with 80% randomly selected for training, and the remaining 20% selected for prediction. Oversampling is to add samples of classes with fewer samples to equal the number of samples in a class with more samples. Through oversampling, the sample size of each transportation mode became 369; 80% of samples were randomly selected for training, and the remaining 20% were selected for prediction.

Tables 2 and 3 describe the categories and continuous variables for imbalanced and rebalanced data, respectively.

4. Results

Stata 15.0 software was used for parameter estimation of the BMNL and MNL models, confusion matrix, and ROCs. SPSS 25.0 was used for relative importance analysis of variables by the RBF and MLP models.

4.1. Model Results. Table 4 presents the estimated means of variables from the BMNL model, and Tables 5–7 show their parameter estimates. The frequently used train was considered as the reference in the model. The typical variables including gender, age, occupation, travel purpose, monthly income, intercity travel distance, intercity travel cost, intercity travel time, safety, comfort, punctuality, access time, and departure time were selected for modeling after collinearity testing. The MCMC simulation-based full Bayesian approach was employed to estimate the posterior distributions of parameters. Variables with confidence intervals not including zero were regarded as significant [19]. As shown in Table 4, we found that the parameter estimates of certain variables differed slightly between the imbalanced and balanced data. For example, the intercity travel distance is significantly related to the choice of express bus when using balanced data, but not when using imbalanced data. The signs of variables were found to be consistent between balanced and imbalanced data.

Table 8 shows the estimated coefficients of variables from the MNL model using the same variables. Parameter estimates of variables are shown in Tables 9–11. Similar to the BMNL model, the parameter estimates differ to some extent between imbalanced and balanced data, and the signs of significant variables are consistent. The symbols of significant variables in the MNL model were consistent with those in the BMNL model. However, the significant variables in MNL were not completely consistent with those in the BMNL model. For example, the travel purpose is significant in the BMNL model but not in the MNL model. Gender was significantly related to the choice of HSR in the BMNL model, but not in the MNL model.

Figures 3 and 4 show the relative importance of the factors obtained by RBF and MLP, respectively. There is a slight difference in the order of relative importance of factors. For example, using imbalanced data, intercity travel cost is most important in the RBF model, but second in importance in the MLP model, after intercity travel time. Slight differences exist in the relative importance of factors between imbalanced and balanced data in the same model. For example, in the MLP model, gender is the least important using imbalanced data, and travel purpose is the least important using balanced data. Overall, intercity travel cost, intercity travel time, intercity travel distance, comfort, safety, and punctuality are the most important factors in the intercity travel mode choice, followed by the monthly income, age, and occupation. Access and departure times, which reflect the accessibility of a transport hub, show moderate importance. Travel purpose and gender are the least important.

4.2. Model Comparison and Validation

4.2.1. Model Performance for Imbalanced Data. AUCs and confusion matrices were employed to compare the fitting and predictive performance of the MNL, BMNL, MLP, and RBF models. The confusion matrix of the four models using imbalanced data is shown in Table 12. Through the analysis of the accuracy, it can be found that MLP has the best fitting performance (80.70%), and RBF is the worst (67.30%). BMNL (76.36%) and MNL (76.10%) have similar fitting performance. For the predictive set, the predictive performance of MLP (78.70%) is the best, followed by MNL (75.76%), BMNL (75.25%), and RBF (65.50%).

The ROC curves of the four models are shown in Figures 5 and 6. For the training set, the AUC of the MLP for the airplane is 0.9857, which indicates that its fitting performance is better than that of BMNL (0.9732), MNL (0.9731), and RBF (0.9443). The MLP model is almost perfect, as its ROC curve rises rapidly toward the upper-left corner of the graph. Similarly, the AUCs of MLP for HSR and train are the largest, followed by BMNL, MNL, and RBF. These findings confirm that the MLP model outperforms BMNL and MNL, followed by RBF.

For the predictive set, the AUC of MLP for airplane is 0.9905, which is better than RBF (0.9823), BMNL (0.9784), and MNL (0.9767). Similarly, MLP is almost perfect, as its ROC curve rises rapidly toward the upper-left corner of the graph. The AUC of MLP for HSR is also the largest, followed by BMNL, MNL, and RBF. The AUC of MLP for train is 0.9054, which indicates that its predictive performance is significantly better than that of MNL (0.8637), BMNL (0.8624), and RBF (0.8280). However, the AUC of BMNL for express bus is the largest, followed by MNL, MLP, and RBF.

4.2.2. Model Performance for Rebalanced Data. MNL, BMNL, MLP, and RBF were used to train and verify the balanced data with the same variables used for the imbalanced data. The confusion matrices of each model for undersampled and oversampled balanced data are shown in Tables 13 and 14. The ROC curves for the undersampled training and predictive set are presented in Figures 7 and 8. The ROC curves for the oversampled training and predictive sets are presented in Figures 9 and 10.

The results show that MLP provides the best fitting for both oversampled and undersampled data, followed by BMNL and MNL, and RBF has the poorest fitting performance. The results are consistent with those of the four models using imbalanced data. Hence, whether the data are balanced will not affect the relative fitting performance of the models.

Table 2: Description of categorical variables.

Variable	Description	Value	Imbalanced data				Oversampling balanced data				Undersampling balanced data			
			Training set		Predictive set		Training set		Predictive set		Training set		Predictive set	
			Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)
Dependent														
Travel modes	Airplane	1	129	16.39	32	16.16	295	25	74	25	125	25	31	25
	HSR	2	295	37.48	74	37.37	295	25	74	25	125	25	31	25
	Train	3	239	30.37	60	30.31	295	25	74	25	125	25	31	25
	Express bus	4	124	15.76	32	16.16	295	25	74	25	125	25	31	25
Independent														
Gender	Female	0	335	42.57	81	40.91	475	40.25	125	42.23	213	42.6	58	46.77
	Male	1	452	57.43	117	59.09	705	59.75	171	57.77	287	57.4	66	53.23
Age	<19	1	18	2.29	5	2.53	26	2.2	4	1.35	9	1.8	1	0.81
	20–29	2	324	41.17	83	41.92	501	42.46	131	44.26	211	42.2	54	43.55
	30–39	3	261	33.16	58	29.29	377	31.95	83	28.04	157	31.4	37	29.84
	40–49	4	116	14.74	38	19.19	191	16.19	43	14.53	83	16.6	23	18.55
	50–59	5	54	6.86	11	5.56	63	5.34	28	9.46	30	6	7	5.65
	60 and above	6	14	1.78	3	1.51	22	1.86	7	2.36	10	2	2	1.60
Occupation	Enterprise unit	1	168	21.35	45	22.73	245	20.76	66	22.3	102	20.4	29	23.39
	Personnel of institutions	2	134	17.03	40	20.20	197	16.69	48	16.22	94	18.8	19	15.32
	Student	3	225	28.59	53	26.77	366	31.02	92	31.08	152	30.4	33	26.61
	Farmer	4	50	6.35	7	3.54	65	5.51	18	6.08	28	5.6	10	8.06
	Self-employed	5	113	14.36	30	15.15	171	14.49	37	12.5	67	13.4	21	16.94
	Other	6	97	12.33	23	11.61	136	11.53	35	11.82	57	11.4	12	9.68
Monthly income	<3K yuan	1	249	31.64	59	29.80	393	33.31	99	33.45	166	33.2	37	29.84
	3-4K yuan	2	147	18.68	41	20.71	198	16.78	53	17.91	98	19.6	20	16.13
	4-5K yuan	3	200	25.41	54	27.27	301	25.51	64	21.62	123	24.6	37	29.84
	5-6K yuan	4	112	14.23	29	14.65	177	15	50	16.89	68	13.6	21	16.94
	6-7K yuan	5	29	3.68	4	2.02	38	3.22	6	2.03	11	2.2	2	1.61
	>7K yuan	6	50	6.35	11	5.55	73	6.19	24	8.11	34	6.8	7	5.65
Travel purpose	Mandatory	1	380	48.28	94	47.47	584	49.49	159	53.72	251	50.2	66	53.23
	Leisure	0	407	51.72	104	52.53	596	50.51	137	46.28	249	49.8	58	46.77
Access mode	Public transit	1	548	69.63	129	65.15	808	68.47	200	67.57	358	71.6	74	59.68
	Private car or taxi	0	239	30.37	69	34.85	372	31.53	96	32.43	142	28.4	50	40.32
Access time	0–30 min	1	252	32.02	66	33.33	383	32.46	104	35.14	166	33.2	44	35.48
	30–60 min	2	283	35.96	80	40.41	414	35.08	97	32.77	165	33	42	33.87
	60–90 min	3	252	32.02	52	26.26	383	32.46	95	32.09	169	33.8	38	30.65
Safety	Very unsafe	1	0	0	0	0	0	0	0	0	0	0	0	0
	Unsafe	2	8	1.02	2	1.01	17	1.44	3	1.01	6	1.2	2	1.61
	General	3	145	18.42	30	15.15	216	18.31	67	22.64	92	18.4	31	25
	Safe	4	390	49.56	105	53.03	594	50.34	144	48.65	251	50.2	55	44.35
	Very safe	5	244	31	61	30.81	353	29.91	82	27.7	151	30.2	36	29.03

Table 2: Continued.

Variable	Description	Value	Imbalanced data				Oversampling balanced data				Undersampling balanced data			
			Training set		Predictive set		Training set		Predictive set		Training set		Predictive set	
			Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)
Comfort	Very uncomfortable	1	5	0.64	3	1.52	15	1.27	2	0.68	6	1.2	2	1.61
	Uncomfortable	2	40	5.08	12	6.06	70	5.93	27	9.12	32	6.4	10	8.06
	General	3	198	25.16	47	23.74	301	25.51	74	25	119	23.8	32	25.81
	Comfortable	4	411	52.22	105	53.03	609	51.61	140	47.3	262	52.4	59	47.58
	Very comfortable	5	133	16.9	31	15.65	185	15.68	53	17.91	81	16.2	21	16.94
Punctuality	Very unpunctual	1	0	0	0	0	0	0	0	0	0	0	0	0
	Unpunctual	2	30	3.81	9	4.54	46	3.9	22	7.43	24	4.8	5	4.03
	General	3	204	25.92	47	23.74	340	28.81	76	25.68	141	28.2	37	29.84
	Punctual	4	414	52.6	111	56.06	624	52.88	160	54.05	248	49.6	70	56.45
	Very punctual	5	139	17.66	31	15.66	170	14.41	38	12.84	87	17.4	12	9.68
Departure mode	Public transit	1	571	72.55	142	71.72	816	69.15	220	74.32	356	71.2	90	72.58
	Private car or taxi	0	216	27.45	56	28.28	364	30.85	76	25.68	144	28.8	34	27.42
Departure time	0–30 min	1	387	49.24	102	51.52	594	50.42	134	45.27	260	52.1	45	36.29
	30–60 min	2	254	32.31	68	34.34	369	31.32	100	33.78	151	30.26	51	41.13
	60–90 min	3	145	18.45	28	14.14	215	18.25	62	20.95	88	17.64	28	22.58

Table 3: Description of continuous variables.

Data	Variable	Unit	Training set Min	Max	Mean	SD	Predictive set Min	Max	Mean	SD
Imbalanced data	Intercity travel distance	km	16.00	2831.00	797.54	579.86	16.00	2540.00	792.66	632.00
	Intercity travel cost	Yuan	7.00	2600.00	310.46	319.44	7.00	1400.00	283.07	288.70
	Intercity travel time	Hour	0.22	52.00	6.48	6.49	0.25	33.00	6.00	5.87
	Access cost	Yuan	1.00	300.00	11.59	22.89	1.00	100.00	10.19	16.36
	Departure cost	Yuan	1.00	150.00	13.52	20.72	1.00	150.00	13.23	21.89
Oversampling balanced data	Intercity travel distance	Km	16.00	2831.00	807.54	621.35	16.00	2801.00	808.45	617.20
	Intercity travel cost	Yuan	7.00	2600.00	321.23	343.35	7.00	2600.00	324.57	344.48
	Intercity travel time	Hour	0.22	52.00	5.92	6.33	0.25	37.00	5.85	6.50
	Access cost	Yuan	1.00	300.00	13.03	25.51	1.00	150.00	11.57	20.38
	Departure cost	Yuan	1.00	150.00	15.19	23.36	1.00	150.00	14.24	22.69
Undersampling balanced data	Intercity travel distance	Km	17.00	2831.00	823.55	616.21	27.00	2500.00	731.52	597.32
	Intercity travel cost	Yuan	7.00	2600.00	330.79	360.54	13.50	1200.00	277.20	275.22
	Intercity travel time	Hour	0.22	37.00	5.92	6.14	0.25	28.00	5.22	5.90
	Access cost	Yuan	1.00	300.00	12.77	25.43	1.00	120.00	12.49	19.88
	Departure cost	Yuan	1.00	150.00	14.68	23.11	1.00	120.00	14.51	22.79

Table 4: Parameter estimation in BMNL.

Variable	Imbalanced data			Oversampling balanced data			Undersampling balanced data		
	Airplane Mean	HSR Mean	Express bus Mean	Airplane Mean	HSR Mean	Express bus Mean	Airplane Mean	HSR Mean	Express bus Mean
Gender									
Male vs. female	0.555	0.186	0.368	0.683	0.197	0.713	1.231	0.384	0.733
Age	0.282	0.350	0.247	0.416	0.357	0.346	0.396	—	0.380
Occupation									
Personnel of institutions vs. enterprise unit	0.710	0.408	0.651	0.344	0.143	0.313	—	—	—
Student vs. enterprise unit	—	−0.300	1.348	0.316	−0.349	1.930	—	—	1.454
Farmer vs. enterprise unit	−0.473	—	—	−0.898	−0.561	−0.333	−2.097	−0.584	−0.574
Self-employed vs. enterprise unit	—	—	0.686	−1.157	−0.384	0.558	−2.540	−1.114	—
Others vs. enterprise unit	−0.844	−0.193	0.679	−1.480	−0.475	0.616	−2.160	−0.704	0.473
Monthly income	—	−0.273	—	—	−0.221	—	−0.163	−0.222	—
Travel purpose									
Mandatory travel vs. leisure travel	−0.477	−0.284	−0.425	−0.341	−0.388	—	—	−0.230	0.138
Intercity travel distance	0.004	0.001	—	0.002	—	−0.002	0.003	—	−0.002
Intercity travel cost	0.018	0.015	—	0.027	0.022	0.001	0.026	0.023	—
Intercity travel time	−1.239	−0.397	—	−1.265	−0.417	0.002	−1.255	−0.482	0.036
Safety	0.566	0.609	—	0.818	0.685	—	−0.098	0.621	—
Comfort	—	0.314	−0.433	0.195	0.369	−0.463	0.639	0.312	−0.480
Punctuality	−0.335	0.319	−0.574	−0.377	0.317	−0.525	−0.496	0.323	−0.817
Access time	0.390	—	−0.268	0.455	—	−0.284	0.351	−0.220	—
Departure time	0.694	—	—	1.026	0.264	—	0.796	—	−0.427
Constant	−7.703	−5.803	3.089	−9.503	−6.798	2.328	−5.985	−5.589	3.884

Note: parameters that were significant at the 95% confidence level are shown in the table.

For the predictive performance of the models, we found that MLP performs best regardless of oversampling or undersampling balanced data. More importantly, BMNL and MNL show the same predictive performance when using oversampling balanced data, and RBF models have the worst predictive performance. Similarly, BMNL and MNL have the same predictive performance using undersampled data, but their performance is lower than that of RBF.

The fitting performance of models based on balanced data is a slight improvement over using imbalanced data. For example, the fitting performance of MLP model is 80.70% using imbalanced data, and 81.80% and 83.10%, respectively, with undersampled and oversampled data. However, except for the RBF model, the predictive performance of these models based on balanced data is slightly lower than that when using imbalanced data.

The ROC curve was also used to intuitively judge the predictive performance of each model, and AUCs were used to quantitatively compare their predictive accuracy under different modeling techniques. We found that the results

Table 5: Parameter estimation in BMNL using imbalanced data.

Variable	Airplane				HSR				Express bus			
	Mean	SD	Credit interval 2.50%	Credit interval 97.50%	Mean	SD	Credit interval 2.50%	Credit interval 97.50%	Mean	SD	Credit interval 2.50%	Credit interval 97.50%
Gender												
Male vs. female	0.555	0.143	0.275	0.833	0.186	0.057	0.067	0.295	0.368	0.167	0.040	0.677
Age	0.282	0.089	0.114	0.454	0.350	0.091	0.180	0.538	0.247	0.084	0.072	0.407
Occupation												
Personnel of institution vs. enterprise unit	0.710	0.173	0.377	1.051	0.408	0.059	0.295	0.527	0.651	0.129	0.403	0.913
Student vs. enterprise unit	0.114	0.171	−0.217	0.435	−0.300	0.068	−0.440	−0.167	1.348	0.181	1.002	1.727
Farmer vs. enterprise unit	−0.473	0.226	−0.930	−0.027	−0.078	0.063	−0.205	0.423	−0.031	0.158	−0.321	0.288
Self-employed vs. enterprise unit	−0.147	0.187	−0.522	0.228	0.074	0.082	−0.083	0.235	0.686	0.139	0.409	0.967
Others vs. enterprise unit	−0.844	0.161	−1.173	−0.526	−0.193	0.066	−0.316	−0.059	0.679	0.144	0.396	0.953
Monthly income	−0.098	0.090	−0.270	0.089	−0.273	0.062	−0.394	−0.154	−0.121	0.070	−0.251	0.016
Travel purpose												
Mandatory travel vs. leisure travel	−0.477	0.196	−0.871	−0.096	−0.284	0.053	−0.384	−0.185	−0.425	0.130	−0.699	−0.192
Intercity travel distance	0.004	0.001	0.003	0.005	0.001	0.001	0.001	0.002	−0.001	0.001	−0.002	0.001
Intercity travel cost	0.018	0.002	0.014	0.022	0.015	0.002	0.011	0.186	0.001	0.002	−0.005	0.004
Intercity travel time	−1.239	0.096	−1.433	−1.059	−0.397	0.044	−0.485	−0.311	−0.009	0.029	−0.064	0.048
Safety	0.566	0.149	0.268	0.847	0.609	0.086	0.440	0.782	0.039	0.065	−0.079	0.173
Comfort	0.245	0.124	−0.017	0.474	0.314	0.053	0.211	0.427	−0.433	0.099	−0.634	−0.254
Punctuality	−0.335	0.13	−0.586	−0.089	0.319	0.066	0.189	0.440	−0.574	0.04	−0.655	−0.498
Access time	0.390	0.151	0.074	0.677	−0.08	0.056	−0.189	0.034	−0.268	0.106	−0.472	−0.066
Departure time	0.694	0.142	0.428	0.981	0.029	0.060	−0.080	0.151	−0.161	0.132	−0.419	0.094
Constant	−7.703	0.081	−7.863	−7.537	−5.803	0.098	−5.997	−5.62	3.089	0.124	2.857	3.342

Table 6: Parameter estimation in BMNL using oversampling of balanced data.

Variable	Airplane				HSR				Express bus			
	Mean	SD	Credit interval 2.50%	Credit interval 97.50%	Mean	SD	Credit interval 2.50%	Credit interval 97.50%	Mean	SD	Credit interval 2.50%	Credit interval 97.50%
Gender												
Male vs. female	0.683	0.118	0.474	0.947	0.197	0.075	0.052	0.339	0.713	0.085	0.556	0.872
Age	0.416	0.106	0.190	0.615	0.357	0.028	0.298	0.410	0.346	0.054	0.241	0.452
Occupation												
Personnel of institution vs. enterprise unit	0.344	0.076	0.218	0.510	0.143	0.054	0.041	0.256	0.313	0.107	0.050	0.491
Student vs. enterprise unit	0.316	0.078	0.126	0.436	−0.349	0.050	−0.444	−0.256	1.930	0.060	1.817	2.041
Farmer vs. enterprise unit	−0.898	0.058	−1.014	−0.787	−0.561	0.077	−0.712	−0.417	−0.333	0.054	−0.438	−0.229
Self-employed vs. enterprise unit	−1.157	0.120	−1.398	−0.928	−0.384	0.075	−0.531	−0.232	0.558	0.142	0.269	0.818
Others vs. enterprise unit	−1.480	0.072	−1.643	−1.348	−0.475	0.116	−0.695	−0.249	0.616	0.101	0.423	0.794
Monthly income	−0.044	0.061	−0.192	0.058	−0.221	0.047	−0.315	−0.129	0.015	0.052	−0.079	0.128
Travel purpose												
Mandatory travel vs. leisure travel	−0.341	0.033	−0.396	−0.264	−0.388	0.086	−0.556	−0.216	−0.127	0.064	−0.243	0.004
Intercity travel distance	0.002	0.001	0.001	0.004	−0.001	0.001	−0.002	0.001	−0.002	0.001	−0.003	−0.001
Intercity travel cost	0.027	0.002	0.023	0.031	0.022	0.002	0.018	0.026	0.001	0.002	−0.002	0.005
Intercity travel time	−1.265	0.075	−1.426	−1.128	−0.417	0.022	−0.459	−0.373	0.002	0.024	−0.047	0.045
Safety	0.818	0.069	0.671	0.947	0.685	0.047	0.594	0.780	0.155	0.104	−0.042	0.359
Comfort	0.195	0.120	0.007	0.490	0.369	0.033	0.303	0.434	−0.463	0.094	−0.644	−0.284
Punctuality	−0.377	0.116	−0.602	−0.164	0.317	0.041	0.242	0.399	−0.525	0.080	−0.680	−0.370
Access time	0.455	0.081	0.255	0.587	−0.136	0.110	−0.357	0.066	−0.284	0.058	−0.409	−0.175
Departure time	1.026	0.171	0.717	1.356	0.264	0.033	0.201	0.329	−0.202	0.111	−0.418	0.012
Constant	−9.503	0.080	−9.678	−9.362	−6.798	0.092	−6.976	−6.605	2.328	0.067	2.164	2.433

Table 7: Parameter estimation in BMNL using undersampling of balanced data.

Variable	Airplane				HSR				Express bus			
	Mean	SD	Credit interval 2.50%	Credit interval 97.50%	Mean	SD	Credit interval 2.50%	Credit interval 97.50%	Mean	SD	Credit interval 2.50%	Credit interval 97.50%
Gender												
Male vs. female	1.231	0.266	0.719	1.764	0.384	0.114	0.167	0.620	0.733	0.165	0.435	1.075
Age	0.396	0.102	0.206	0.595	0.232	0.143	−0.080	0.506	0.380	0.083	0.220	0.541
Occupation												
Personnel of institution vs. enterprise unit	−0.162	0.130	−0.403	0.081	−0.176	0.106	−0.393	0.035	0.139	0.171	−0.235	0.444
Student vs. enterprise unit	0.041	0.170	−0.274	0.366	−0.121	0.210	−0.529	0.288	1.454	0.131	1.198	1.718
Farmer vs. enterprise unit	−2.097	0.191	−2.450	−1.722	−0.584	0.061	−0.709	−0.458	−0.574	0.223	−1.017	−0.128
Self-employed vs. enterprise unit	−2.540	0.232	−3.011	−2.105	−1.114	0.081	−1.280	−0.954	0.021	0.231	−0.449	0.485
Others vs. enterprise unit	−2.160	0.211	−2.563	−1.735	−0.704	0.085	−0.906	−0.574	0.473	0.134	0.217	0.740
Monthly income	−0.163	0.067	−0.305	−0.031	−0.222	0.101	−0.418	−0.027	−0.128	0.075	−0.284	0.020
Travel purpose												
Mandatory travel vs. leisure travel	−0.038	0.215	−0.470	0.352	−0.230	0.113	−0.471	−0.013	0.138	0.205	−0.253	0.554
Intercity travel distance	0.003	0.001	0.001	0.005	0.000	0.001	−0.002	0.002	−0.002	0.001	−0.004	−0.001
Intercity travel cost	0.026	0.004	0.020	0.035	0.023	0.003	0.016	0.031	0.000	0.003	−0.006	0.006
Intercity travel time	−1.255	0.062	−1.378	−1.146	−0.482	0.056	−0.598	−0.388	0.036	0.043	−0.049	0.126
Safety	−0.098	0.133	−0.357	0.151	0.621	0.099	0.432	0.824	0.077	0.096	−0.113	0.274
Comfort	0.639	0.035	0.574	0.706	0.312	0.120	0.056	0.526	−0.480	0.134	−0.753	−0.223
Punctuality	−0.496	0.158	−0.800	−0.190	0.323	0.018	0.290	0.363	−0.817	0.118	−1.065	−0.582
Access time	0.351	0.073	0.216	0.490	−0.220	0.030	−0.278	−0.159	0.042	0.070	−0.098	0.178
Departure time	0.796	0.109	0.584	1.029	−0.050	0.063	−0.178	0.064	−0.427	0.115	−0.641	−0.200
Constant	−5.985	0.135	−6.242	−5.731	−5.589	0.207	−6.001	−5.200	3.884	0.173	3.540	4.203

Table 8: Parameter estimation in MNL.

Variable	Imbalanced data			Oversampling balanced data			Undersampling balanced data		
	Airplane	HSR	Express bus	Airplane	HSR	Express bus	Airplane	HSR	Express bus
	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
Gender									
Male vs. female	—	—	—	—	—	0.681	—	—	0.723
Age	—	0.349	—	—	0.366	0.360	—	—	—
Occupation									
Personnel of institutions vs. enterprise unit	—	—	—	—	—	—	—	—	—
Student vs. enterprise unit	—	—	1.298	—	−0.352	1.856	—	—	1.642
Farmer vs. enterprise unit	—	—	—	—	—	—	—	—	—
Self-employed vs. enterprise unit	—	—	—	—	—	—	—	—	—
Others vs. enterprise unit	—	—	—	—	—	—	—	—	—
Monthly income	−0.128	−0.289	—	—	−0.226	—	—	—	—
Travel purpose									
Mandatory travel vs. leisure travel	—	—	—	—	—	—	—	—	—
Intercity travel distance	0.004	0.001	—	0.002	—	−0.002	0.003	—	—
Intercity travel cost	0.017	0.014	—	0.026	0.022	—	0.031	0.027	—
Intercity travel time	−1.211	−0.380	—	−1.242	−0.421	—	−1.395	−0.638	—
Safety	—	0.585	—	0.801	0.678	—	—	—	—
Comfort	—	—	−0.458	—	—	−0.461	—	—	−0.418
Punctuality	—	0.299	−0.575	—	—	−0.513	—	—	−0.617
Access time	—	—	—	—	—	—	—	—	—
Departure time	0.731	—	—	1.030	—	−0.185	—	—	—
Constant	−7.673	−5.821	3.058	−9.522	−6.798	2.303	—	−5.297	3.420

Note: parameters that were significant at the 95% confidence level are shown in the table.

Table 9: Parameter estimation in MNL using imbalanced data.

Variable	Airplane				HSR				Express bus			
	Coefficient	$P > z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P > z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P > z$	Credit interval 2.50%	Credit interval 97.50%
Gender												
Male vs. female	0.474	0.321	−0.462	1.410	0.160	0.543	−0.355	0.674	0.409	0.121	−0.107	0.926
Age	0.326	0.269	−0.252	0.904	0.349	0.035	0.024	0.675	0.253	0.128	−0.073	0.579
Occupation												
Personnel of institution vs. enterprise unit	0.725	0.290	−0.618	2.069	0.410	0.316	−0.391	1.210	0.642	0.191	−0.320	1.604
Student vs. enterprise unit	0.116	0.885	−1.453	1.685	−0.362	0.390	−1.189	0.464	1.298	0.002	0.464	2.132
Farmer vs. enterprise unit	−0.425	0.658	−2.307	1.458	−0.080	0.892	−1.234	1.074	−0.059	0.929	−1.359	1.241
Self-employed vs. enterprise unit	−0.207	0.789	−1.720	1.306	0.081	0.849	−0.754	0.916	0.611	0.195	−0.314	1.537
Others vs. enterprise unit	−0.824	0.366	−2.610	0.962	−0.173	0.732	−1.161	0.816	0.705	0.178	−0.32	1.730
Monthly income	−0.128	0.565	−0.563	0.307	−0.289	0.013	−0.516	−0.062	−0.106	0.329	−0.318	0.106
Travel purpose												
Mandatory travel vs. leisure travel	−0.443	0.340	−1.352	0.466	−0.302	0.256	−0.823	0.219	−0.412	0.112	−0.921	0.096
Intercity travel distance	0.004	0.001	0.002	0.005	0.001	0.278	−0.001	0.002	−0.001	0.009	−0.003	0.001
Intercity travel cost	0.017	0.001	0.013	0.021	0.014	0.001	0.010	0.018	0.001	0.828	−0.005	0.004
Intercity travel time	−1.211	0.001	−1.444	−0.978	−0.380	0.001	−0.483	−0.276	−0.003	0.917	−0.061	0.055
Safety	0.491	0.137	−0.156	1.138	0.585	0.004	0.188	0.982	0.059	0.760	−0.319	0.437
Comfort	0.251	0.410	−0.345	0.846	0.358	0.056	−0.01	0.725	−0.458	0.011	−0.811	−0.105
Punctuality	−0.278	0.356	−0.866	0.311	0.299	0.104	−0.061	0.660	−0.575	0.002	−0.936	−0.213
Access time	0.404	0.180	−0.187	0.994	−0.035	0.841	−0.374	0.304	−0.221	0.172	−0.539	0.096
Departure time	0.731	0.017	0.131	1.331	0.047	0.791	−0.304	0.398	−0.210	0.269	−0.584	0.163
Constant	−7.673	0.002	−12.518	−2.828	−5.821	0.001	−8.592	−3.051	3.058	0.012	0.667	5.450

Table 10: Parameter estimation in MNL using oversampling of balanced data.

Variable	Airplane				HSR				Express bus			
	Coefficient	$P > z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P > z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P > z$	Credit interval 2.50%	Credit interval 97.50%
Gender												
Male vs. female	0.661	0.118	−0.167	1.490	0.196	0.443	−0.305	0.697	0.681	0.001	0.275	1.087
Age	0.386	0.126	−0.108	0.880	0.366	0.020	0.058	0.673	0.360	0.007	0.097	0.624
Occupation												
Personnel of institution vs. enterprise unit	0.343	0.564	−0.824	1.510	0.138	0.723	−0.627	0.903	0.318	0.401	−0.424	1.061
Student vs. enterprise unit	0.298	0.672	−1.082	1.679	−0.352	0.399	−1.171	0.466	1.856	0.001	1.193	2.519
Farmer vs. enterprise unit	−0.898	0.309	−2.627	0.831	−0.546	0.376	−1.757	0.664	−0.338	0.488	−1.291	0.616
Self-employed vs. enterprise unit	−1.107	0.126	−2.525	0.312	−0.387	0.376	−1.243	0.469	0.554	0.119	−0.142	1.250
Others vs. enterprise unit	−1.518	0.055	−3.067	0.030	−0.479	0.305	−1.394	0.437	0.545	0.179	−0.250	1.340
Monthly income	−0.040	0.836	−0.416	0.336	−0.226	0.043	−0.445	−0.007	−0.017	0.840	−0.182	0.148
Travel purpose												
Mandatory travel vs. leisure travel	−0.343	0.413	−1.166	0.479	−0.384	0.140	−0.895	0.126	−0.134	0.506	−0.529	0.261
Intercity travel distance	0.002	0.001	0.001	0.004	−0.001	0.264	−0.002	0.001	−0.002	0.001	−0.003	−0.001
Intercity travel cost	0.026	0.001	0.021	0.031	0.022	0.001	0.017	0.026	0.001	0.580	−0.003	0.005

Table 10: Continued.

Variable	Airplane				HSR				Express bus			
	Coefficient	$P>z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P>z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P>z$	Credit interval 2.50%	Credit interval 97.50%
Intercity travel time	−1.242	0.001	−1.432	−1.053	−0.421	0.001	−0.524	−0.318	0.001	0.964	−0.050	0.052
Safety	0.801	0.006	0.225	1.378	0.678	0.001	0.279	1.077	0.146	0.333	−0.149	0.440
Comfort	0.236	0.404	−0.318	0.791	0.371	0.051	−0.001	0.743	−0.461	0.001	−0.737	−0.185
Punctuality	−0.353	0.194	−0.886	0.179	0.321	0.082	−0.040	0.682	−0.513	0.001	−0.808	−0.219
Access time	0.456	0.076	−0.048	0.959	−0.122	0.462	−0.447	0.203	−0.244	0.060	−0.499	0.010
Departure time	1.030	0.001	0.506	1.554	0.271	0.130	−0.080	0.621	−0.185	0.232	−0.487	0.118
Constant	−9.522	0.001	−13.873	−5.170	−6.798	0.001	−9.530	−4.066	2.303	0.020	0.356	4.250

Table 11: Parameter estimation in MNL using undersampling of balanced data.

Variable	Airplane				HSR				Express bus			
	Coefficient	$P>z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P>z$	Credit interval 2.50%	Credit interval 97.50%	Coefficient	$P>z$	Credit interval 2.50%	Credit interval 97.50%
Gender												
Male vs. female	0.847	0.137	−0.268	1.962	0.175	0.637	−0.552	0.901	0.723	0.010	0.176	1.271
Age	0.439	0.213	−0.252	1.130	0.305	0.183	−0.144	0.754	0.259	0.135	−0.080	0.598
Occupation												
Personnel of institution vs. enterprise unit	0.252	0.748	−1.284	1.787	−0.057	0.916	−1.113	0.999	0.220	0.657	−0.751	1.191
Student vs. enterprise unit	0.317	0.742	−1.565	2.199	−0.128	0.831	−1.304	1.048	1.642	0.001	0.774	2.510
Farmer vs. enterprise unit	−1.600	0.179	−3.931	0.732	−0.704	0.399	−2.339	0.931	−0.518	0.420	−1.777	0.741
Self-employed vs. enterprise unit	−1.939	0.059	−3.953	0.076	−0.828	0.200	−2.096	0.439	0.646	0.176	−0.291	1.584
Others vs. enterprise unit	−1.860	0.093	−4.028	0.307	−0.782	0.289	−2.229	0.664	0.632	0.256	−0.458	1.721
Monthly income	−0.139	0.612	−0.676	0.398	−0.223	0.180	−0.550	0.103	−0.048	0.671	−0.271	0.175
Travel purpose												
Mandatory travel vs. leisure travel	−0.337	0.549	−1.439	0.764	−0.313	0.408	−1.055	0.428	−0.283	0.300	−0.819	0.252
Intercity travel distance	0.003	0.006	0.001	0.005	0.001	0.960	−0.002	0.002	−0.002	0.008	−0.003	0.001
Intercity travel cost	0.031	0.001	0.023	0.040	0.027	0.001	0.019	0.036	0.001	0.875	−0.006	0.005
Intercity travel time	−1.395	0.001	−1.686	−1.105	−0.638	0.001	−0.837	−0.439	−0.013	0.763	−0.095	0.070
Safety	0.084	0.838	−0.727	0.896	0.514	0.074	−0.050	1.077	0.014	0.945	−0.392	0.421
Comfort	0.301	0.424	−0.437	1.039	0.466	0.074	−0.044	0.976	−0.418	0.023	−0.780	−0.057
Punctuality	−0.389	0.256	−1.061	0.283	0.189	0.448	−0.298	0.675	−0.617	0.002	−1.003	−0.230
Access time	0.422	0.237	−0.277	1.120	−0.194	0.430	−0.676	0.288	−0.135	0.434	−0.475	0.204
Departure time	0.655	0.073	−0.061	1.371	−0.034	0.892	−0.530	0.461	−0.226	0.259	−0.617	0.166
Constant	−5.696	0.050	−11.396	0.004	−5.297	0.006	−9.039	−1.555	3.420	0.008	0.912	5.928

from AUCs are consistent with those from the confusion matrices for each model.

4.3. Model Interpretation. We use the results of the BMNL with better statistical performance and MLP models with better predictive performance to explain the effects of factors on intercity travel mode choice.

From Table 4 and Figure 4, it is found that gender was positively correlated with the choice of HSR and express bus, indicating that men were prone to traveling by HSR or express bus, and women by train. This finding is consistent with a previous study [2], which revealed that women preferred using train more than men. The models show that personnel of government-sponsored institutions were more likely than enterprise personnel to choose an airplane. Farmers and the self-employed were less likely than enterprise personnel to travel by airplane. Similarly, students and farmers were not prone to choosing HSR, and farmers were prone to using an express bus [2, 8]. These results are supported by a previous study [1, 3] that found that passengers working in the state sector are likely to choose airplane over coach. Monthly income was found to be positively associated with airplane choice, and negatively

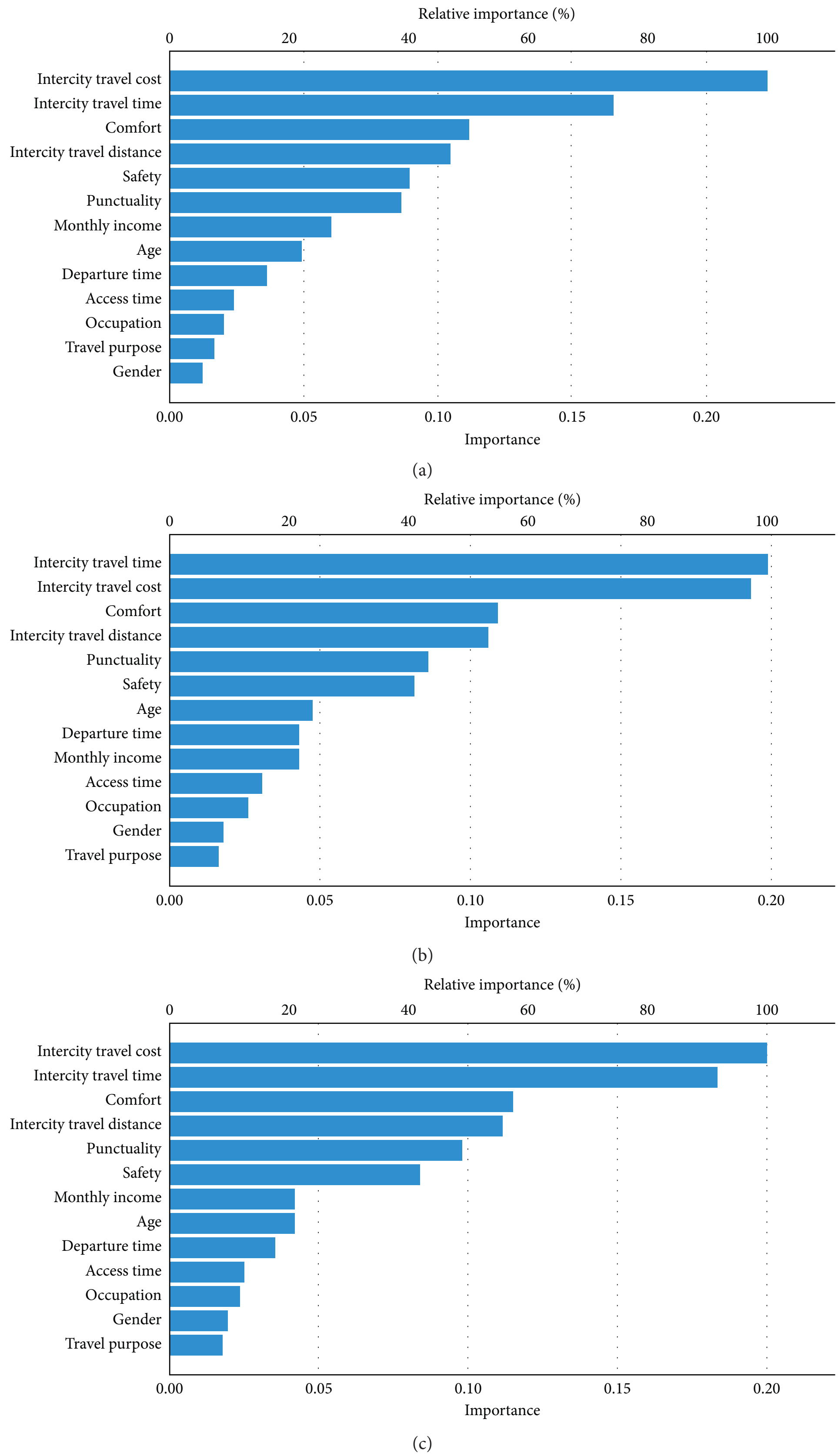

FIGURE 3: Relative importance of each variable using RBF. (a) Imbalanced data. (b) Oversampling balanced data. (c) Undersampling balanced data.

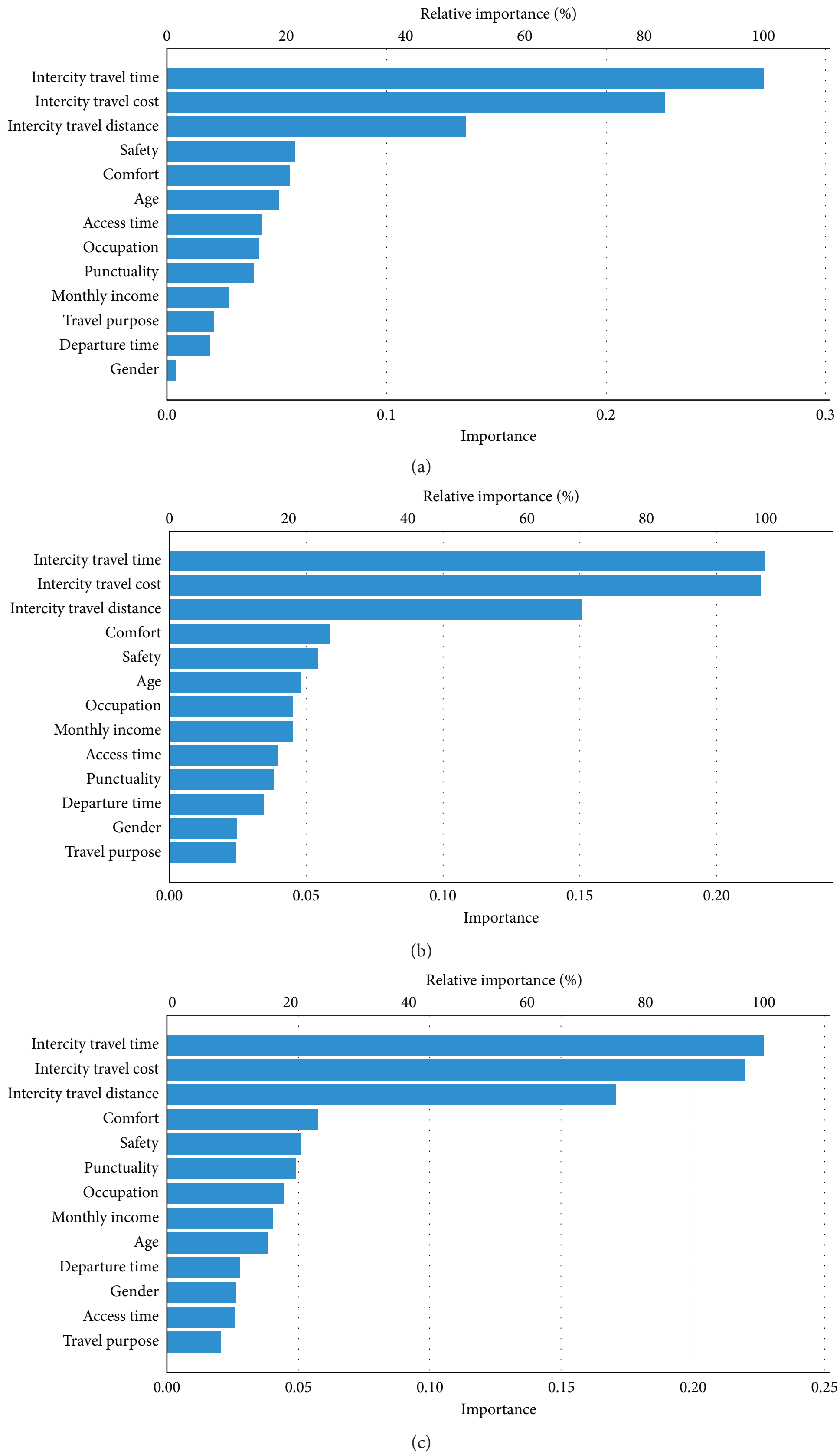

Figure 4: Relative importance of each variable using MLP. (a) Imbalanced data. (b) Oversampling balanced data. (c) Undersampling balanced data.

Table 12: Confusion matrix and recall, precision, and accuracy of each model for imbalanced data.

Model			Training set					Predictive set				
			Predictive class					Predictive class				
		Mode	Airplane	HSR	Train	Express bus	Recall	Airplane	HSR	Train	Express bus	Recall
MNL	Actual class	Airplane	109	18	2	0	84.50%	30	2	0	0	93.80%
		HSR	16	246	27	6	83.40%	3	57	11	3	77.00%
		Train	3	25	181	28	76.40%	3	5	44	8	73.30%
		Express bus	0	8	54	59	48.80%	0	1	12	19	59.40%
		Precision	85.16%	82.83%	68.56%	63.44%	**76.10%**	83.33%	87.69%	65.67%	63.33%	**75.76%**
BMNL	Actual class	Airplane	113	14	2	0	87.60%	30	2	0	0	93.80%
		HSR	18	243	28	6	82.40%	3	55	13	3	74.30%
		Train	5	24	184	26	77.00%	3	5	45	7	75.00%
		Express bus	3	7	53	61	49.20%	0	0	13	19	59.40%
		Precision	81.29%	84.38%	68.91%	65.59%	**76.36%**	83.33%	88.71%	63.38%	65.52%	**75.25%**
MLP	Actual class	Airplane	118	9	2	0	91.50%	29	3	0	0	90.60%
		HSR	12	268	13	2	90.80%	3	64	6	1	86.50%
		Train	1	30	183	23	77.20%	0	7	45	7	76.30%
		Express bus	0	5	54	62	51.20%	0	2	13	17	53.10%
		Precision	90.08%	85.90%	72.62%	71.26%	**80.70%**	90.63%	84.21%	70.31%	68.00%	**78.70%**
RBF	Actual class	Airplane	80	41	6	2	62.00%	21	11	0	0	65.60%
		HSR	17	216	45	17	73.20%	4	53	12	5	71.60%
		Train	0	39	181	17	76.40%	1	9	38	11	64.40%
		Express bus	0	18	54	49	40.50%	0	1	14	17	53.10%
		Precision	82.47%	68.79%	63.29%	57.65%	**67.30%**	80.77%	71.62%	59.38%	51.52%	**65.50%**

The bold values represent the accuracy of models.

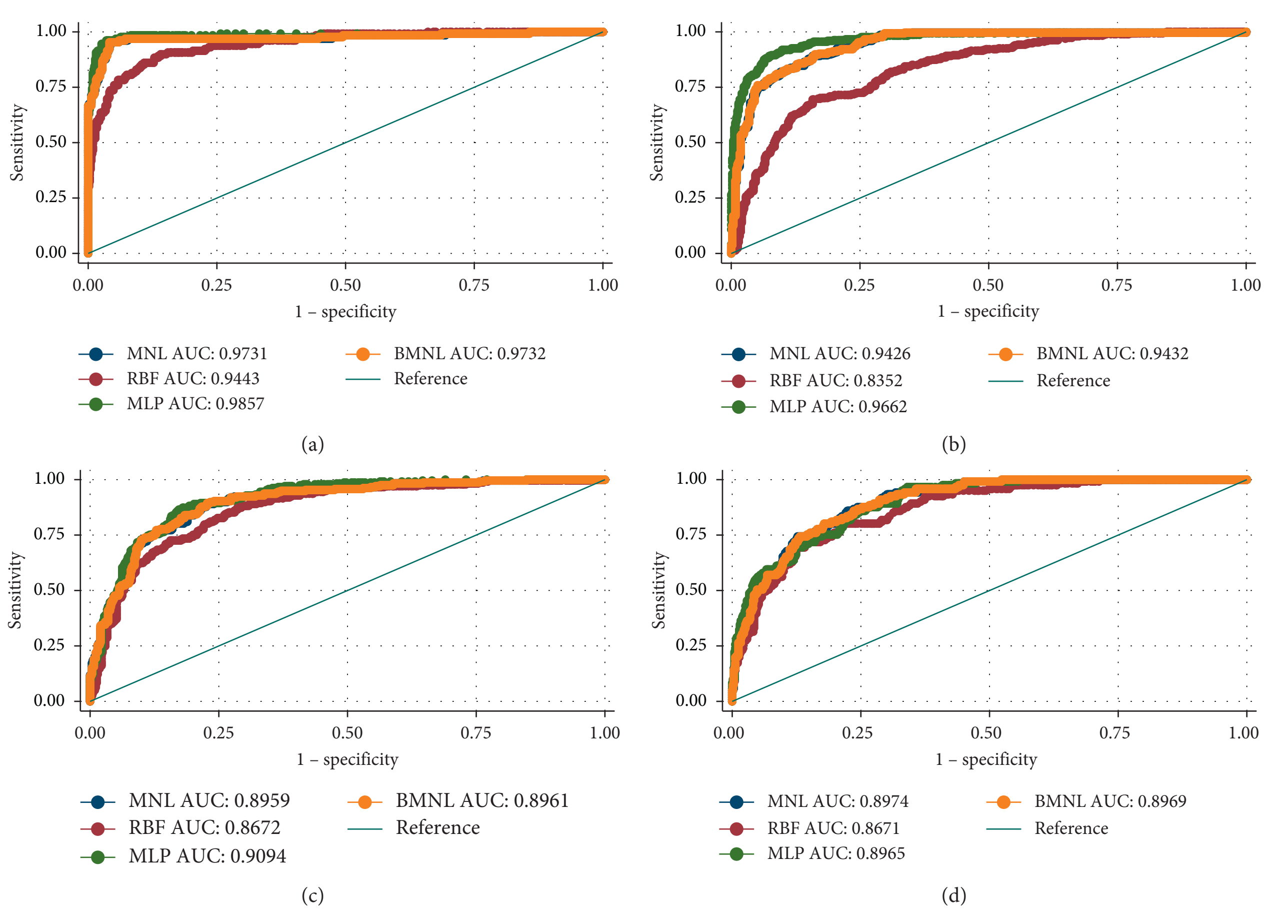

Figure 5: ROC curves of models for imbalanced data training set. (a) Airplane. (b) HSR. (c) Train. (d) Express bus.

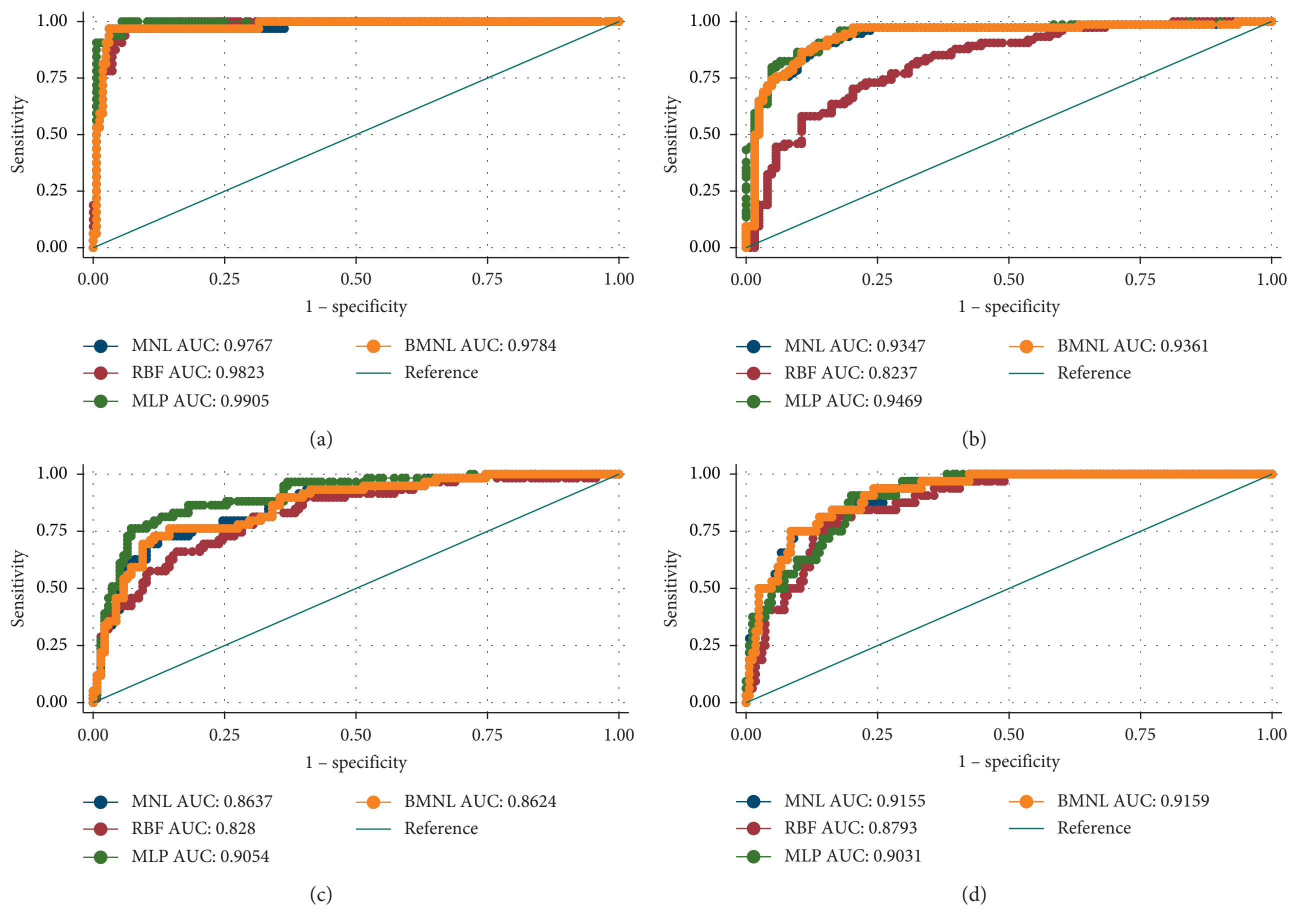

Figure 6: ROC curves of models for imbalanced data predictive set. (a) Airplane. (b) HSR. (c) Train. (d) Express bus.

Table 13: Confusion matrix and recall, precision, and accuracy of each model for undersampling of balanced data.

Model			Training set					Predictive set				
			Predictive class					Predictive class				
		Mode	Airplane	HSR	Train	Express bus	Recall	Airplane	HSR	Train	Express bus	Recall
MNL	Actual class	Airplane	115	8	1	1	92.00%	27	4	0	0	87.10%
		HSR	8	102	7	8	81.60%	5	21	1	4	67.74%
		Train	3	15	79	28	63.20%	0	4	15	12	48.39%
		Express bus	3	3	25	94	75.20%	0	1	8	22	70.97%
		Precision	89.15%	79.69%	70.54%	71.76%	**78.00%**	84.38%	70.00%	62.50%	57.89%	**68.55%**
BMNL	Actual class	Airplane	116	7	1	1	92.80%	28	3	0	0	90.32%
		HSR	12	88	14	11	70.40%	5	18	3	5	58.06%
		Train	3	6	81	35	64.80%	0	2	17	12	54.84%
		Express bus	3	0	28	94	75.20%	0	1	8	22	70.97%
		Precision	86.57%	87.13%	65.32%	66.67%	**75.80%**	84.85%	75.00%	60.71%	56.41%	**68.55%**
MLP	Actual class	Airplane	121	2	1	1	96.80%	28	3	0	0	90.30%
		HSR	6	107	7	5	85.60%	2	25	1	3	80.60%
		Train	1	13	92	17	74.80%	0	4	19	8	61.30%
		Express bus	0	3	30	89	73.00%	0	2	10	19	61.30%
		Precision	94.53%	85.60%	70.77%	79.46%	**81.80%**	93.33%	73.53%	63.33%	63.33%	**73.40%**
RBF	Actual class	Airplane	105	12	3	5	84.00%	25	5	0	1	80.60%
		HSR	20	79	15	11	63.20%	3	20	3	5	64.50%
		Train	1	12	83	27	67.50%	0	2	22	7	71.00%
		Express bus	1	9	28	84	68.90%	0	2	8	21	67.70%
		Precision	82.68%	70.54%	64.34%	66.14%	**70.20%**	89.29%	68.97%	66.67%	61.76%	**71.00%**

The bold values represent the accuracy of models.

Table 14: Confusion matrix of each model for oversampling of balanced data.

Model		Mode	Training set Predictive class Airplane	HSR	Train	Express bus	Recall	Predictive set Predictive class Airplane	HSR	Train	Express bus	Recall
MNL	Actual class	Airplane	275	18	1	1	93.22%	71	1	1	1	95.95%
		HSR	18	235	23	19	79.66%	9	55	5	5	74.32%
		Train	6	23	190	76	64.41%	1	10	40	23	54.05%
		Express bus	7	8	63	217	73.56%	0	1	27	46	62.16%
		Precision	89.87%	82.75%	68.59%	69.33%	**77.71%**	87.65%	82.09%	54.79%	61.33%	**71.62%**
BMNL	Actual class	Airplane	276	17	1	1	93.56%	71	1	1	1	95.95%
		HSR	18	235	22	20	79.66%	9	54	6	5	72.97%
		Train	6	23	194	72	65.76%	1	10	41	22	55.41%
		Express bus	7	5	73	210	71.19%	0	0	28	46	62.16%
		Precision	89.90%	83.93%	66.90%	69.31%	**77.54%**	87.65%	83.08%	53.95%	62.16%	**71.62%**
MLP	Actual class	Airplane	277	16	1	1	93.90%	72	0	1	1	97.30%
		HSR	13	241	21	20	81.70%	6	55	5	8	74.30%
		Train	3	21	209	58	71.80%	1	13	42	18	56.80%
		Express bus	0	13	31	244	84.70%	0	2	15	57	77.00%
		Precision	94.54%	82.82%	79.77%	75.54%	**83.10%**	91.14%	78.57%	66.67%	67.86%	**76.40%**
RBF	Actual class	Airplane	244	33	6	12	82.70%	58	10	3	3	78.40%
		HSR	51	176	37	31	59.70%	14	43	8	9	58.10%
		Train	2	37	188	64	64.60%	2	15	41	16	55.40%
		Express bus	2	18	72	196	68.10%	0	3	18	53	71.60%
		Precision	81.61%	66.67%	62.05%	64.69%	**68.14%**	78.38%	60.56%	58.57%	65.43%	**65.90%**

The bold values represent the accuracy of models.

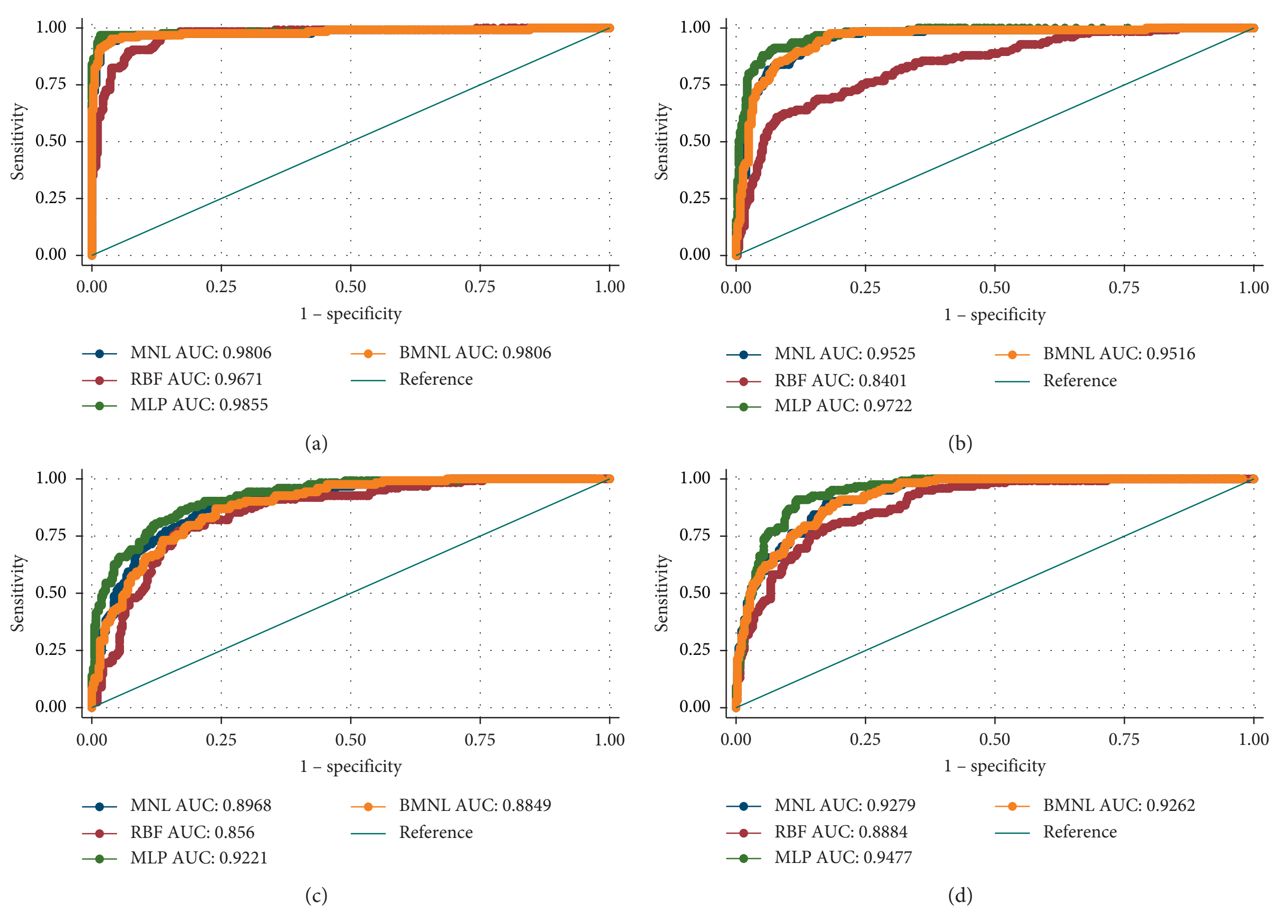

Figure 7: ROC curves for undersampled balanced data training set. (a) Airplane. (b) HSR. (c) Train. (d) Express bus.

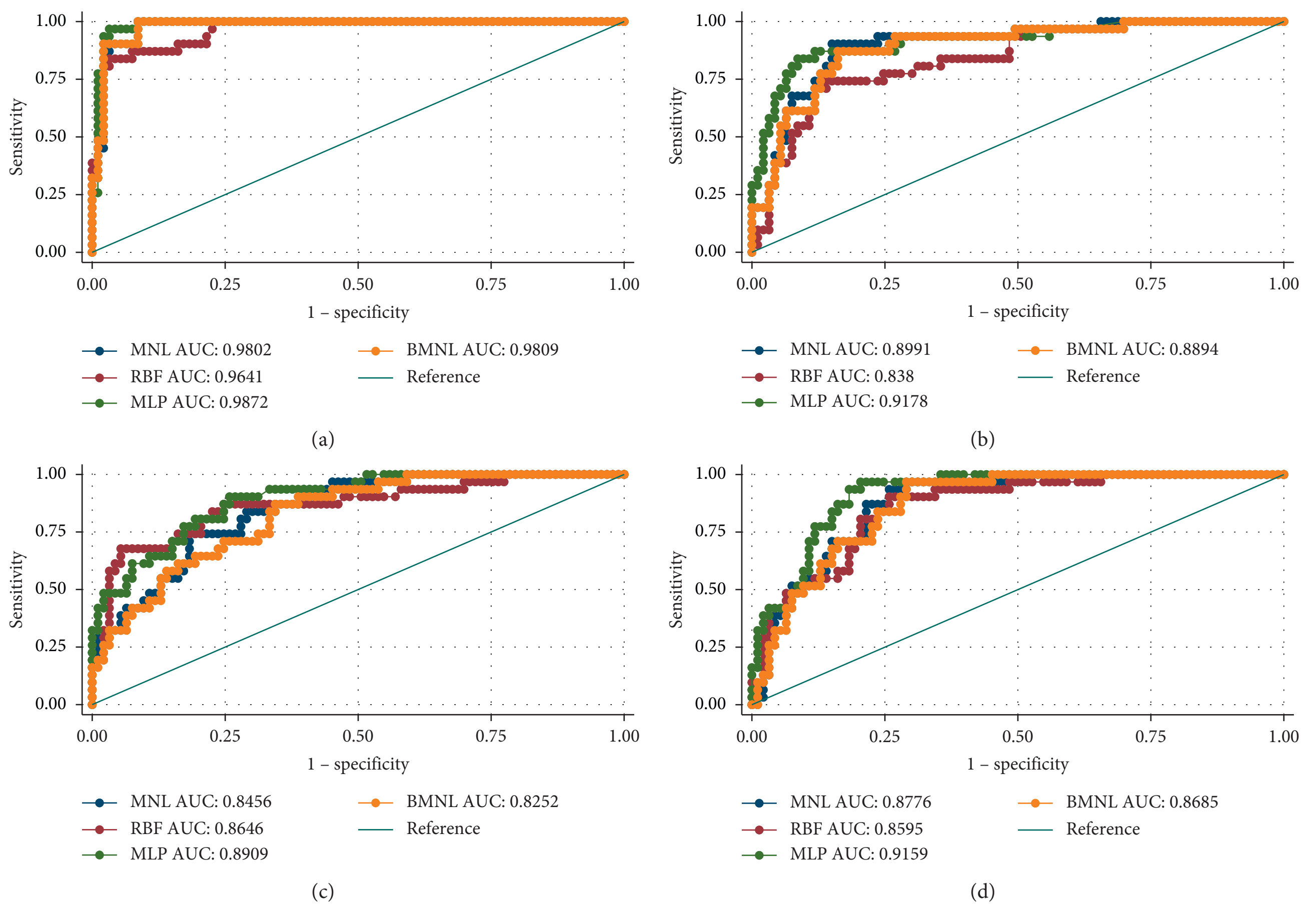

Figure 8: ROC curves for undersampled balanced data predictive set. (a) Airplane. (b) HSR. (c) Train. (d) Express bus.

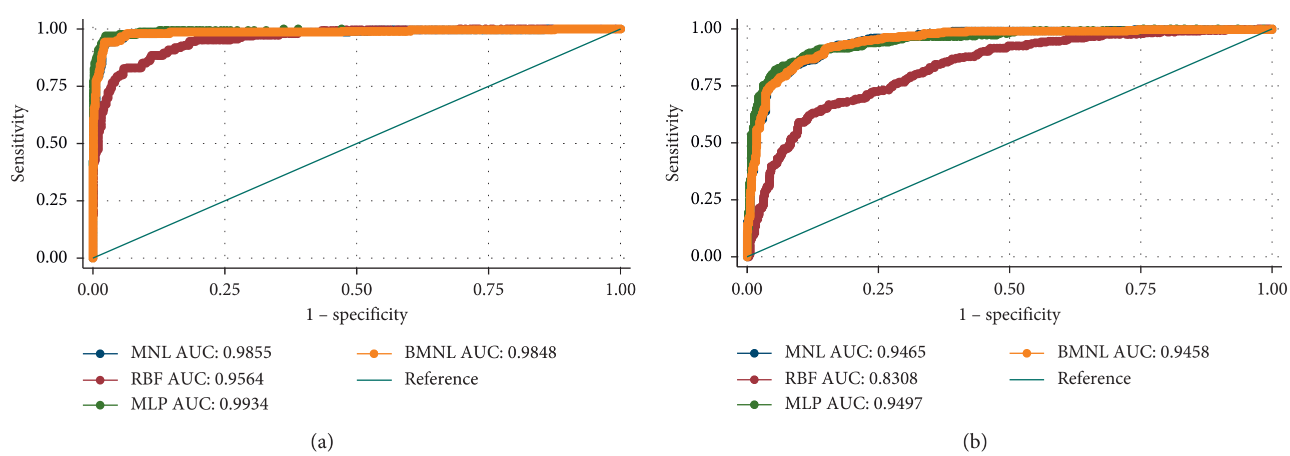

Figure 9: Continued.

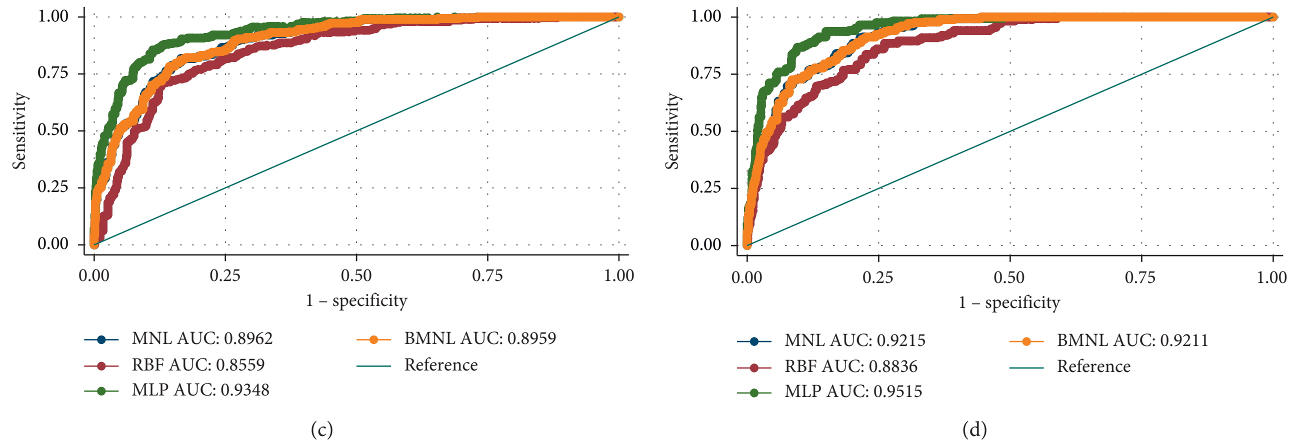

Figure 9: ROC curves for oversampled balanced data training set. (a) Airplane. (b) HSR. (c) Train. (d) Express bus.

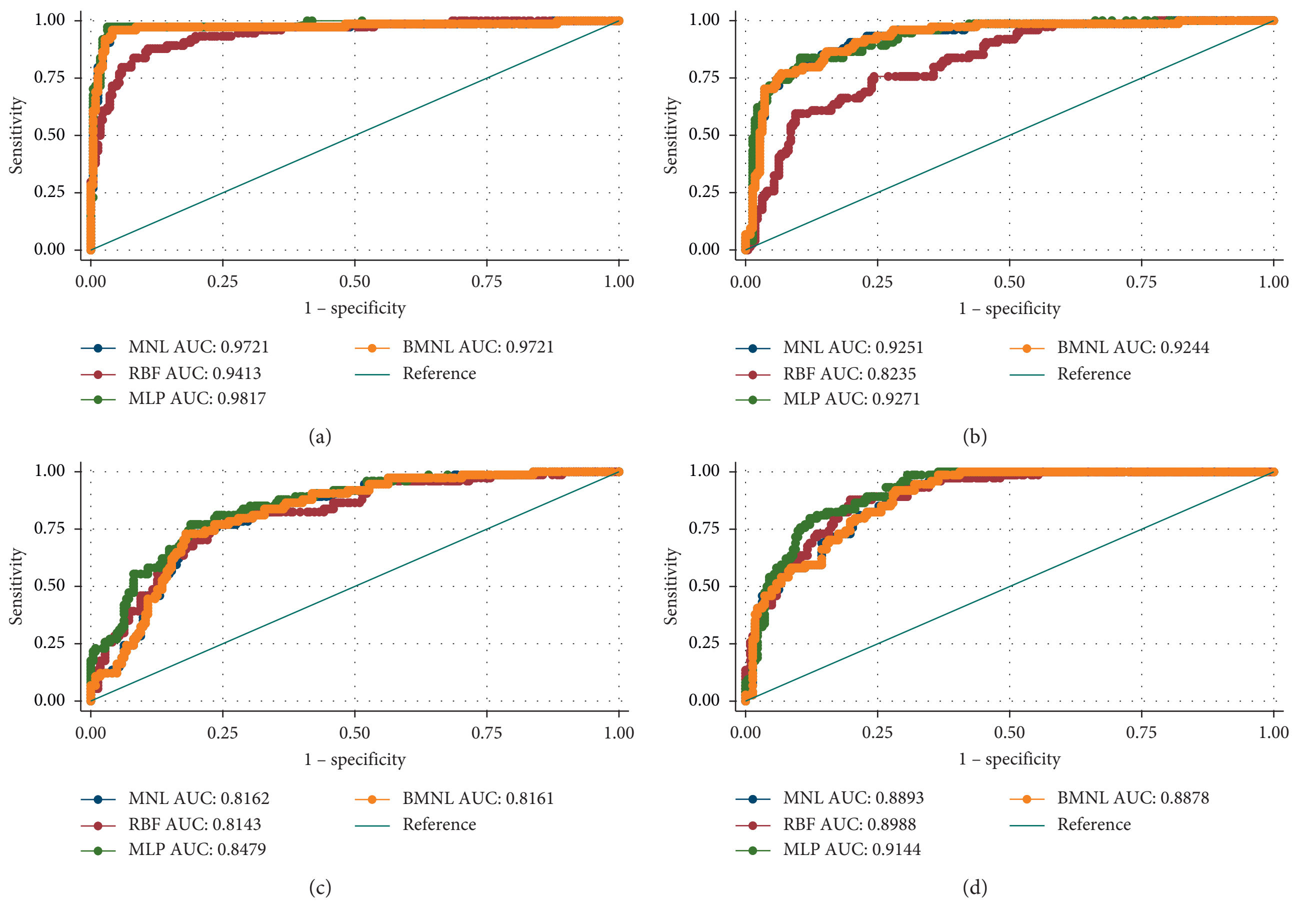

Figure 10: ROC curves for oversampled balanced data training set. (a) Airplane. (b) HSR. (c) Train. (d) Express bus.

with express bus choice. This result agrees with the result of a study [4, 7] that found that higher- and lower-income individuals favor air and bus travel, respectively.

The variable of travel purpose had a significant positive effect on HSR choice and was ranked 11th in the relative importance of all variables. This finding is similar to that of a past study [1] and reveals that, compared to the train, leisure passengers prefer HSRs or airplanes more than passengers for mandatory travel. It is possible that leisure passengers can afford the higher travel cost and are more willing to travel in a comfortable mode. The modeling results also show the significant impact of travel distance on intercity travel mode choice, which is third most important. This result implies that, compared to the train, the longer travel distance favors airplane and HSR, and the shorter distance favors express bus, which is consistent with previous studies [3, 10]. Intercity travel cost is the most important variable and is a positive sign for the airplane or HSR choice, indicating that passengers incurring higher travel costs are more likely to travel by airplane or HSR. Intercity travel time is the second most important factor, showing a negative association with the choice of airplane and HSR, indicating that passengers spending less travel time are more likely to select airplane or HSR. This finding is intuitive because airplanes and HSRs are faster than trains.

Safety ranked fourth, and this variable affects the choice of airplane and HSR, showing that passengers with a higher safety demand are more likely to travel by airplane or HSR. Comfort is the fifth most important factor; it positively influences the choice of airplane and HSR and is negatively associated with express bus. This result is expected, as airplanes and HSRs have better service facilities and environments than trains [1]. Punctuality ranked ninth and is positively related to HSR choice and negatively associated with airplane and express bus. This shows that a higher punctuality demand favors HSR and does not favor airplane and express bus. This result is expected, as external conditions such as bad weather can easily affect the operation of airplanes and express buses, but its impact on HSRs and trains is relatively small [1].

Access time ranked seventh in relative importance and is found to have a positive effect on airplane choice and a negative effect on express bus compared to train, indicating that passengers spending longer access time prefer traveling by airplane and are less likely to travel by express bus. The finding is straightforward because the airport is generally farther than the railway station from the city center, and the highway passenger station is closer [10]. A similar result was found for the effect of departure time.

5. Discussion and Conclusions

We investigated modeling techniques BMNL, MNL, MLP, and RBF for passengers' intercity travel mode choices. Data from a large individual-level survey in the city of Xi'an were used to develop the model. More comprehensive factors such as socioeconomics, travel demand, service quality, and accessibility of transport hub were incorporated in the models.

The comparison results show that MLP has the best predictive performance, BMNL and MNL have approximately equal predictive accuracy, and RBF has the poorest performance using imbalanced data. It was found that the fitting performance of the four models with balanced data was slightly higher than those with imbalanced data. However, it was surprising that the predictive performance of these models with balanced data was slightly lower than those with imbalanced data. A potential reason could be that the degree of imbalance for the original data is very small. These findings suggest that the MLP and BMNL modeling approaches are recommended for the analysis of passengers' intercity travel mode choice. Significant variables in the BMNL model include gender, age, occupation, travel purpose, intercity travel distance, intercity travel cost, intercity travel time, safety, punctuality, access time, and departure time, which is not completely consistent with those in the MNL model. However, the signs of significant variables in the BMNL model were in line with those in the MNL model. Regarding the MLP modeling results, the travel cost was found to be the most important factor in intercity mode choice, followed by travel time and travel distance. Comfort, safety, and punctuality were relatively important factors for passenger travel mode choices. The influence of individual characteristics on intercity travel mode choices was relatively low, and monthly income was the most important factor among individual characteristics.

These findings can provide a reference for traffic management departments to formulate traffic demand management strategies and provide technical support for data analysts and high-tech enterprises to develop intelligent decision-making systems for the choice of passenger intercity travel modes. Through our research conclusion, we can find that intercity travel time, intercity travel cost, intercity travel distance, and the service quality of a transportation mode are important factors affecting intercity travel mode choices. Traffic transportation management departments can accordingly develop a green transportation development strategy by optimizing ticket prices, increasing vehicle speeds, and improving the quality of service, so as to push travelers from transportation with high energy consumption to that with low energy consumption. Our findings show that the predictive performance of models does not significantly improve when using balanced data instead of imbalanced data. This can provide a basis for data analysts to fully understand the impact of data structures on the predictive performance of models.

There are some limitations to this study. The results may only apply to the selected dataset and therefore must be verified using datasets from more cities. The degree of data imbalance and proportion between the training set and the prediction set may also affect the fitting and predictive performance of the models, and it is necessary to explore the fitting and predictive performance of models using extremely unbalanced data and other proportions in the future. In addition, although no significant multicollinearity was found in the independent variables for the models, intercity travel time and intercity travel cost varied with travel distance. It is necessary to generate the fare rate and intercity travel time per kilomile by standardizing the intercity travel time and intercity travel cost and incorporate the transformed variables into the models to eliminate the potential impact of travel distance. Moreover, more variables that might be associated with intercity travel mode choices, such as the characteristics of the destination city, weather, and coronavirus disease, should be investigated. Advanced modeling techniques, such as the Bayesian random parameter model capturing more unobserved heterogeneity, the Probit model with endogenous variables, and the XGBoost model, should be applied in future studies.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 52002282), Shaanxi Natural Science Foundation Youth Project (Grant no. 2017JQ5086), Shaanxi Education Department Special Science and Technology Project Science (Grant no. 19JK0477), and Philosophy and Social Science Foundation of Zhejiang Province (Grant no. 21NDJC163YB).

References

[1] X. Li, J. Tang, X. Hu, and W. Wang, "Assessing intercity multimodal choice behavior in a touristy city: a factor analysis," *Journal of Transport Geography*, vol. 86, Article ID 102776, 2020.

[2] C. R. Bhat, "Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 7, pp. 495–507, 1998.

[3] V. V. Can, "Estimation of travel mode choice for domestic tourists to Nha Trang using the multinomial probit model," *Transportation Research Part A: Policy and Practice*, vol. 49, pp. 149–159, 2013.

[4] M. A. A. B. Miskeen, A. M. Alhodairi, and R. A. A. B. O. Rahmat, "Modeling a multinomial logit model of intercity travel mode choice behavior for all trips in Libya," *International Journal of Civil & Environmental Engineering*, vol. 7, no. 9, pp. 636–645, 2013.

[5] S. Hess, G. Spitz, M. Bradley, and M. Coogan, "Analysis of mode choice for intercity travel: application of a hybrid choice model to two distinct US corridors," *Transportation Research Part A: Policy and Practice*, vol. 116, pp. 547–567, 2018.

[6] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour and Society*, vol. 14, pp. 1–10, 2019.

[7] C. V. Forinash and F. S. Koppelman, "Application and interpretation of nested logit models of intercity mode choice," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1413, no. 1413, pp. 98–106, 1993.

[8] C. R. Bhat, "An endogenous segmentation mode choice model with an application to intercity travel," *Transportation Science*, vol. 31, no. 1, pp. 34–48, 1997.

[9] Z.-C. Li and D. Sheng, "Forecasting passenger travel demand for air and high-speed rail integration service: a case study of Beijing-Guangzhou corridor, China," *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 397–410, 2016.

[10] C. Román, J. C. Martín, R. Espino et al., "Valuation of travel time savings for intercity travel: the Madrid-Barcelona corridor," *Transport Policy*, vol. 36, pp. 105–117, 2014.

[11] Y. Guo, Z. Li, Y. Wu, and C. Xu, "Evaluating factors affecting electric bike users' registration of license plate in China using Bayesian approach," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 59, pp. 212–221, 2018.

[12] C. R. Bhat, "A heteroscedastic extreme value model of intercity travel mode choice," *Transportation Research Part B: Methodological*, vol. 29, no. 6, pp. 471–483, 1995.

[13] Y. Wang, L. Li, L. Wang, A. Moore, S. Staley, and Z. Li, "Modeling traveler mode choice behavior of a new high-speed rail corridor in China," *Transportation Planning and Technology*, vol. 37, no. 5, pp. 466–483, 2014.

[14] S. Ashiabor, H. Baik, and A. Trani, "Logit models for forecasting nationwide intercity travel demand in the United States," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2007, no. 1, pp. 1–12, 2007.

[15] J.-H. Lee, K.-S. Chon, and C. Park, "Accommodating heterogeneity and heteroscedasticity in intercity travel mode choice model: formulation and application to HoNam, South Korea, high-speed rail demand analysis," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1898, no. 1, pp. 69–78, 2004.

[16] Y. Guo, P. Liu, Y. Wu, and J. Chen, "Evaluating how right-turn treatments affect right-turn-on-red conflicts at signalized intersections," *Journal of Transportation Safety & Security*, vol. 12, no. 3, pp. 419–440, 2020.

[17] S. Srinivasan, C. R. Bhat, and J. Holguin-Veras, "Empirical analysis of the impact of security perception on intercity mode choice," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1942, no. 1, pp. 9–15, 2006.

[18] X. Li, W. Wang, C. Xu, Z. Li, and B. Wang, "Multi-objective optimization of urban bus network using cumulative prospect theory," *Journal of Systems Science and Complexity*, vol. 28, no. 3, pp. 661–678, 2015.

[19] Y. Guo, Z. Li, Y. Wu, and C. Xu, "Exploring unobserved heterogeneity in bicyclists' red-light running behaviors at different crossing facilities," *Accident Analysis & Prevention*, vol. 115, pp. 118–127, 2018.

[20] C.-X. Zhang, S. Xu, and J.-S. Zhang, "A novel variational Bayesian method for variable selection in logistic regression models," *Computational Statistics & Data Analysis*, vol. 133, pp. 1–19, 2019.

[21] A. P. Afghari, M. M. Haque, S. Washington, and T. Smyth, "Effects of globally obtained informative priors on Bayesian safety performance functions developed for Australian crash data," *Accident Analysis & Prevention*, vol. 129, pp. 55–65, 2019.

[22] X. Zhou, M. Wang, and D. Li, "Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning," *Journal of Transport Geography*, vol. 79, 2019.

[23] M. Wong and B. Farooq, "A bi-partite generative model framework for analyzing and simulating large scale multiple discrete-continuous travel behaviour data," *Transportation Research Part C: Emerging Technologies*, vol. 110, pp. 247–268, 2020.

[24] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models," *Travel Behaviour and Society*, vol. 20, pp. 22–35, 2020.

[25] C. Xie, J. Lu, and E. Parkany, "Work travel mode choice modeling with data mining: decision trees and neural networks," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1854, no. 1, pp. 50–61, 2003.

[26] Y. Zhang and Y. Xie, "Travel mode choice modeling with support vector machines," *Transportation Research Record:*

Journal of the Transportation Research Board, vol. 2076, no. 1, pp. 141–150, 2008.
[27] Y. Wang, S. Peng, and M. Xu, "Emergency logistics network design based on space-time resource configuration," *Knowledge-Based Systems*, vol. 223, Article ID 107041, 2021.
[28] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, vol. 78, pp. 273–282, 2017.
[29] Y. Wang, Y. Yuan, X. Guan et al., "Collaborative two-echelon multicenter vehicle routing optimization based on state-space-time network representation," *Journal of Cleaner Production*, vol. 258, Article ID 120590, 2020.
[30] F. Wang and C. L. Ross, "Machine learning travel mode choices: comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 47, 2018.
[31] M. Wong, B. Farooq, and G.-A. Bilodeau, "Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling," *Journal of Choice Modelling*, vol. 29, pp. 152–168, 2018.
[32] Y. Wang, S. Peng, X. Zhou, M. Mahmoudi, and L. Zhen, "Green logistics location-routing problem with eco-packages," *Transportation Research Part E: Logistics and Transportation Review*, vol. 143, Article ID 102118, 2020.
[33] A. Lindner, C. S. Pitombo, and A. L. Cunha, "Estimating motorized travel mode choice using classifiers: an application for high-dimensional multicollinear data," *Travel Behaviour and Society*, vol. 6, pp. 100–109, 2017.
[34] S. Rasouli and H. J. P. Timmermans, "Using ensembles of decision trees to predict transport mode choice decisions: effects on predictive success and uncertainty estimates," *European Journal of Transport and Infrastructure Research*, vol. 14, pp. 412–424, 2014.
[35] H. B. Celikoglu, "Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling," *Mathematical and Computer Modelling*, vol. 44, no. 7-8, pp. 640–658, 2006.
[36] L. Wang and Z. Zuo, "Travel mode recognition using RBF neural network," *CICTP 2014: Safe, Smart, and Sustainable Multimodal Transportation Systems*, pp. 711–721, 2014.
[37] J. Sun, J. Sun, and P. Chen, "Use of support vector machine models for real-time prediction of crash risk on urban expressways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2432, no. 2432, pp. 91–98, 2014.
[38] J. Xie and Z. Qiu, "The effect of imbalanced data sets on LDA: a theoretical and empirical analysis," *Pattern Recognition*, vol. 40, no. 2, pp. 557–562, 2007.
[39] A. D'Addabbo and R. Maglietta, "Parallel selective sampling method for imbalance and large data classification," *Pattern Recognition Letters*, vol. 62, pp. 61–67, 2015.
[40] W. A. Rivera and P. Xanthopoulos, "A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets," *Expert Systems with Applications*, vol. 66, pp. 124–135, 2016.
[41] V. García, J. S. Sánchez, A. I. Marqués, R. Florencia, and G. Rivera, "Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data," *Expert Systems with Applications*, vol. 158, Article ID 113026, 2020.
[42] X. Li, R. Ma, Y. Guo, W. Wang, B. Yan, and J. Chen, "Investigation of factors and their dynamic effects on intercity travel modes competition," *Travel Behaviour and Society*, vol. 23, pp. 166–176, 2021.
[43] H. A. Klaiber and R. H. Von Haefen, "Do random coefficients and alternative specific constants improve policy analysis? An empirical investigation of model fit and prediction," *Environmental and Resource Economics*, vol. 73, pp. 75–91, 2011.
[44] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[45] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational & Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.
[46] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Lulu, Morrisville, NC, USA, 2018.
[47] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," *Proceedings of the International Conference on Neural Networks*, vol. 3, pp. 11–13, 1987.
[48] S. V. Stehman, "Estimating standard errors of accuracy assessment statistics under cluster sampling," *Remote Sensing of Environment*, vol. 60, no. 3, pp. 258–269, 1997.
[49] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a real-time crash risk prediction model incorporating the various crash mechanisms across different traffic states," *Traffic Injury Prevention*, vol. 16, no. 1, pp. 28–35, 2015.