Supplementary File 1 Description of analysis methods.

Binary logistic regression

The binary logistic regression model is mainly used to examine the probability that an event will occur, which is represented in dichotomic form as: Y=1 (the occurrence of an event) and Y=0 (the occurrence of the non-event). The relevant formulaes used are as follows[94,95]:

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots + \beta_k X_{ki} \quad (1)$$

Where:

α is the constant;

β_j $(j = 1, 2, \ldots, k)$ are the estimated parameters for each explanatory variable;

X_j are the explanatory variables;

$i = 1, 2, \ldots, n$ and n is the sample size;

$p_i = P(Y_i = 1)$ is the probability of the occurrence of the event of interest for each observation.

Therefore, the formula for calculating the probability of the occurrence of the event is as follows[94]:

$$p_i = \frac{1}{1+e^{-Z_i}} \quad (2)$$

Where $Z_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots + \beta_k X_{ki}$.

Similarly, the formula for calculating the probability that the event will not occur[94]:

$$1 - p_i = \frac{1}{1+e^{Z_i}} \quad (3)$$

Multinomial logistic regression

When the dependent variable Y generates more than two outcomes, multinomial logistic regression should be used to calculate the occurrence probabilities for each category. Generically speaking, if the dependent variable Y produces outcomes for M answer categories, there will be (M−1) logits. The formulae for the multinomial logistic regression model are similar to those used for the binary logistic regression model, as shown below[94]:

$$Z_{i_m} = \alpha_m + \beta_{1m} X_{1i} + \beta_{2m} X_{2i} \ldots + \beta_{km} X_{ki} \quad (4)$$

$$p_{i_m} = \frac{e^{Z_{i_m}}}{\sum_{m=0}^{M-1} e^{Z_{i_m}}} \quad (5)$$

where $Z_{i_0} = 0$; thus, $e^{Z_{i_0}} = 1$. And $m = 0, 1, \ldots, M-1$.

Naive Bayes classifier

Bayes theorem can be written as the following formula - applied to a set of individual variables to form Naive Bayes[75]:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)} \quad (6)$$

In this study, “Y” is the travel time and “X” is the set of dependent variables $X = X_1, X_2, ---, X_n$, i.e., six socio-demographic factors (gender, age, annual income, living area, car ownership and whether respondents have children) and five travel behaviour factors (walking time to bus stops, average waiting time for buses, whether travelling in the peak period, average number of trips per day, and attitudes towards Yangzhou’s transport system)

Since Naive classifier assumes the condition of independence among the feature variables, therefore the Eq. 6 for bayes rule is applied to the set of independent variables to get the output probabilities of particular class given X as:

(Probability of outcome/evidence) = (Probability of likelihood of evidence * Prior) /(Probability of evidence)

In mathematical way:

$$P(Y|X_1, X_2, ---, X_n) = \frac{P(X_1|Y)P(X_2|Y) --- P(X_n|Y) * P(Y)}{P(X_1)P(X_2) --- P(X_n)} \quad (7)$$

Posterior probabilities for each class is calculated and class with maximum value is the final predicted class (result) of the model.