IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 7, 2020, accepted October 20, 2020, date of publication October 28, 2020, date of current version November 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034355

Demand Forecasting of Online Car-Hailing With Stacking Ensemble Learning Approach and Large-Scale Datasets

YUMING JIN[1], XIAOFEI YE[2], QIMING YE[2], TAO WANG[3], JUN CHENG[4], AND XINGCHEN YAN[5]

[1]School of Transportation Engineering, Chang'an University, Xi'an 710064, China
[2]Ningbo Collaborative Innovation Center for Port Trade Cooperation and Development, School of Maritime and Transportation, Ningbo University, Ningbo 315211, China
[3]School of Architecture and Transportation, Guilin University of Electronic Science and Technology, Guilin 541004, China
[4]School of Transportation, Southeast University, Nanjing 211189, China
[5]College of Automobile and Traffic Engineering, Nanjing Forestry University, Nanjing 210037, China

Corresponding author: Xiaofei Ye (yexiaofei@nbu.edu.cn)

This work was supported in part by the Projects of the Natural Science Foundation of Zhejiang Province, China, under Grant LY20E080011; in part by the National Natural Science Foundation of China under Grant 71971059, Grant 71701108, and Grant 71861006; in part by the Guangxi Natural Science Foundation under Grant 2020GXNSFAA159153; in part by the Guangxi Science and Technology Base and Talent Special Project under Grant AD20159035; in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK20180775; in part by the National Key Research and Development Program of China-Traffic Modeling, Surveillance, and Control With Connected and Automated Vehicles under Grant 2017YFE9134700; and in part by Data Source: Didi Chuxing GAIA Initiative.

ABSTRACT With the rapid development and convenient service of online car-hailing, it has gradually become the preferred choice for people to travel. Accurate forecasting of car-hailing trip demand not only enables the drivers and companies to dispatch the vehicles and increase the mileage utilization, but also reduces the passengers' waiting-time. The rebalance of spatiotemporal demand and supply could mitigate traffic congestion, reduce traffic emission, and guide people's travel patterns. This study aimed to develop a short-term demand forecasting model for car-hailing services using stacking ensemble learning approach. The spatial-temporal characteristics of online car-hailing demand were analyzed and extracted through data analysis. The region-level spatial characteristics, time features, and weather conditions were added into the forecasting model. Then the stacking ensemble learning model was developed to predict the car-hailing demand at region-level for different time intervals, including 10 min, 15 min, and 30 min. The validation results suggested that the proposed stacking ensemble learning model has reasonable good prediction accuracy for different time intervals. The comparison results show that the short-term demand forecasting model based on stacking ensemble learning is better than single LSTM, SVR, lightGBM and Random Forest models. MAE and RMSE increased by 6.0% and 5.2% respectively at 30 min time interval, which further verifies the effectiveness and feasibility of the proposed model.

INDEX TERMS Stacking ensemble learning approach, online car-hailing, feature selection.

I. INTRODUCTION

Taxi is an important part of passenger transport. Its purpose is to provide people with fast and convenient travel services. Because there are the gaps between the taxi quantities and demand distributions in many cities, the imbalance of supply and demand are particularly serious. In recent years, online car-hailing services such as Didi and Uber have been widely used in numerous cities worldwide [1]. The trips of online car-hailing have increased rapidly in China and other countries. For example, the trips of online car-hailing completed about 20 billion person times in China. Compared with traditional taxis, online car-hailing is more flexible and convenient, and its on-demand mobility services also provide new ideas for improving the balance of travel supply and demand. Due to the unbalanced spatiotemporal distribution

The associate editor coordinating the review of this manuscript and approving it for publication was Razi Iqbal.

of passengers and online car-hailing services, it is difficult to reveal the complicated dependencies among different regions and temporal period [2]. Actually, the demand of an area is usually influenced by its surrounding neighbors and at the same time correlated with various historical observations, e.g. half an hour ago, a day ago or even a month ago. The fluctuations of spatial and temporal demand also lead to inefficient vehicle dispatching and high operating costs in the on-demand mobility services. The dynamic demand forecasting is critical important for rebalancing the car-hailing services in the different regions and temporal observations. Accurate forecasting of car-hailing trip demand not only enables the drivers and companies to dispatch the vehicles and increase the mileage utilization, but also reduce the passengers' waiting-time. The rebalance of spatiotemporal demand and supply could mitigate traffic congestion, reduce traffic emission, and guide people's travel patterns [3].

This study investigated the demand pattern of car-hailing based on the large amount of historical data from online car-hailing platform. The data contained the spatiotemporal characteristics of the demand for online car-hailing services, so as to be used for forecasting the demand of various regions in the city. The dynamic demand forecasting models were developed to predict the trip demand in short time. The existing researches only consider the temporal or spatial characteristics of online car-hailing demand separately. In reality, the demand for online car-hailing is often related to time and space at the same time. In addition, a single model like long short-term memory (LSTM) network and convolutional neural network (CNN) was often used for establishing the forecasting model, which is well in capturing the time-series or spatial characteristics from historical data. However, to the best of our knowledge, the combination of each model based stacking integrated learning approach has not been conducted to predict the trip demand of the car-hailing services.

Therefore, the purpose of this paper is to develop a short-term demand forecasting model for online car-hailing services by applying the stacking integrated learning approach with large-scale travel data. The temporal, spatial and weather characteristics as influential factors were also extracted to input into the model. The main contributions of this paper are as follows: (a) we reveal the temporal and spatial characteristics of regions and identify the correlations among regions. Then we further apply the stacking ensemble learning approach to explicitly model these correlations. (b) This paper proposes a demand forecasting model of online car-hailing services based on stacking ensemble learning approach. Compared with the single model, the prediction accuracy is improved significantly. The results of this study have the potential to provide useful information for online car-hailing rebalance and to improve the operational efficiency of the car-hailing system.

II. LITERATURE REVIEW

Demand for online car-hailing service is easily disturbed by random factors, and has strong uncertainty and variability [4]. Therefore, how to accurately forecast the demand of online car-hailing is a hot issue in recent years. Especially, some ride-hailing platforms like Didi and Uber shared their open data of trip order and vehicle trajectory with transportation researchers. They have considerable interest in forecasting the demand of taxi and car-hailing trips. Demand forecasting methods can fall into three categories including the traditional statistical model, machine learning and deep learning approach [5]. The first kind of algorithm only considers historical data and growth rate, which could not accurately reflect the uncertainty of online car-hailing demand. For example, clustering algorithms were applied to mine the historical data and forecast the demand distributions of traditional taxi [6]. Time-series techniques were often used to predict the demand of taxi and car-hailing [7]. The second model with machine learning approach uses the potential flaw of data to reach the prediction of target value. For example, X. Jia used WAVE-SVM coupling model to predict residents' travel demand [8]. Y. Zhong established a XGBoost model to predict short-term traffic flow on the basis of considering the spatial-temporal characteristics of traffic flow [9]. I. Saadi *et al.* used single decision tree, bootstrap aggregation (bagged) decision tree, random forest, enhanced decision tree and artificial neural network to predict the short-term demand of online car-hailing. The third type of model is based on deep learning prediction model [10]. R. Pedro *et al.* On the basis of historical data, ARIMA and ANN are used to establish demand forecasting models to forecast the demand of taxis [11]. J. Xu proposed a taxi demand model based on long-term memory neural network (LSTM) and New York City taxi data [12]. G. Xu gridded the experimental area and extracted the geographical relationship of each grid. Then multi-graph convolution neural network and cyclic neural network were applied to construct the prediction model of ride-hailing demand [13]. C. Wang proposed a convolutional neural network (CNN)-based deep learning model for multi-step ride-hailing demand prediction using the trip data of Didi [14].

The findings from the previous studies provide valuable insights and referential experience in forecasting the travel demand of car-hailing services. Most of them focused on historical data mining from the temporal and spatial features respectively. The most recently developed deep learning methods, such as Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) and recurrent neural network (RNN) have shown better performance in other prediction applications, but the ensemble learning approach has not been explored in car-hailing demand forecasting yet. On the other hand, although a small amount of studies combine the spatiotemporal correlations of features of car-hailing services into their demand prediction models, car-hailing demand predictions that also carries spatiotemporal features is overlooked yet. Actually, car-hailing demand can exhibit certain spatiotemporal characteristics related to measures of travel choice behavior, traffic flow and weather conditions. Therefore, it is essential to model the spatial-temporal correlations among all the multi-hierarchies data in order to most effectively predict parking usage.

III. METHODOLOGY

A. STACKING ENSEMBLE LEARNING

This paper proposes an ensemble learning model to forecast the online car-hailing demand with multi-hierarchical datasets considering both temporal and spatial features. The basic idea of stacking ensemble learning is to integrate several different types of classifiers to generate a strong classifier to improve the generalization ability of strong classifiers [15]. The training process is shown in Fig.1. Firstly, the total training set is divided into five parts by cross-validation, including four training subsets and one verification subset. Then, the training subset is trained by using the model of the first layer, and the verification subset is predicted to generate five prediction sets. Then, the five predicted values are combined into generate a new input feature of the training set. At the same time, the testing set is predicted and averaged to generate a new test of input feature. Finally, the new features generated by the first level model are input into the second level model to obtain the final prediction results. The effect of Stacking ensemble learning depends on two aspects: one is the effect of the first layer base classifier. The better the effect of base classifier, the better the effect of ensemble learning model; the second is the combination mode of model. There should be certain differences between base classifiers, so that each base classifier can take its advantages, the effect of ensemble learning model is better.

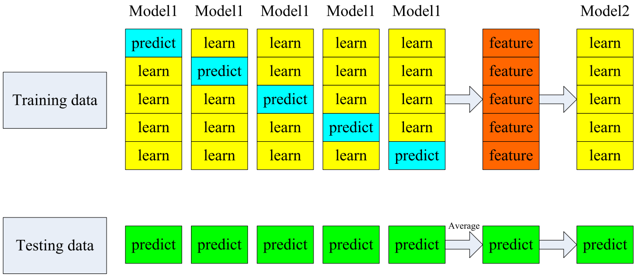


FIGURE 1. Stacking ensemble learning method.

B. RANDOM FOREST

Random Forest (RT) was proposed by Breiman in 2001 [16]. It is an algorithm based on bagging idea. Random Forest takes CART as a basic learner. Its basic idea is to select several sample sets randomly from the sample set and model each sample set for decision tree. At the same time, random sampling of features is also introduced. Finally, the final prediction result is obtained by combining the prediction results of multiple decision trees. Generally, the simple average method is used for regression problems. The process of Random Forest is as follows:

(1) The training set is randomly sampled to generate N different training subsets, and then each training subset is trained to construct n decision trees.

(2) In the process of building decision tree, m ($m < M$) feature subsets are randomly selected from each node.

(3) Each decision tree is generated completely without pruning process.

(4) For the regression problem, the mean value of the predicted value of each tree is in the final result.

C. LONG AND SHORT TERM MEMORY

Long and short term memory (LSTM) is an improvement of recurrent neural network (RNN). The hidden layer information of RNN at this time t only comes from the current input t and the previous time t-1. As shown in Fig. 2, different neurons have no direct connection in the RNN model network structure, and cannot obtain far-reaching information [17]. Therefore, RNN can handle certain short-term dependence, but cannot deal with long-term dependence.

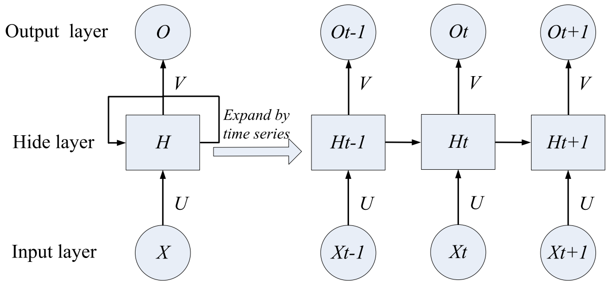


FIGURE 2. Illustration of the recurrent neural network (RNN) structure.

In order to solve the long-term dependence of RNN, that is, the problem of gradient disappearance, LSTM introduces cell state and uses three gate structures: input gate, forgetting gate and output gate, which not only avoids gradient disappearance and gradient dispersion, but also enables LSTM to process longer sequence data [18]. The structure of LSTM model is shown in Fig.3.

(1) Forget gate decides which information should be discarded from the previous cell state. Forget gate's input h_{t-1} X_{t}, and f_t are obtained by an activation function sigmoid. The value range of f_t is (0, 1),0 means forget all, 1 means keep all. The formula of forget gate is as follows:

$$f_t = \sigma(W_f{}^*[h_{t-1}, X_t] + b_f) \tag{1}$$

wherer h_{t-1} represents the output of the previous cell, X_{t} is the input of this cell, σ is the sigmoid function, and W_f is the forget matrix.

(2) Input gate determines the information stored in the cell state. The input gate consists of two parts: the first part is the sigmoid function, which determines the value for updating. The second part uses the tanh activation function to create a new candidate vector C_t, and then combines the two vector constructions to create the updated value C_{t}. The input gate formula is as follows:

$$i_t = \sigma(W_i{}^*[h_{t-1}, X_t] + b_i) \tag{2}$$

$$C_t = \tanh(W_c{}^*[h_{t-1}, X_t] + b_c) \tag{3}$$

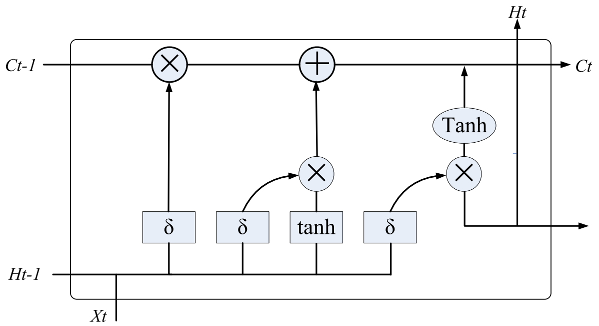


FIGURE 3. Network structure diagram of LSTM model.

where W_i and W_C are the weight matrices; b_c and b_i are the bias vectors; and tanh denotes the he activation function given by:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4}$$

$$C_t = \sigma(f_t{}^* C_{t-1} + i_t{}^* C_t') \tag{5}$$

where the current cell states C_t are the combinations of long term memory C_{t-1} and the current memory C_t. The forget gate f_t controls the amount of information from long term memory C_{t-1} stored in the current cell states. And the input gate determines the amount of information from the current memory C_t stored in the current cell states.

(3) Output gate: determines the output content. output gate O_t determines which information of cell state is used to generate output h_t [19]. The output gate formula is as follows:

$$O_t = \sigma(W_o{}^*[h_{t-1}, X_t] + b_o) \tag{6}$$

$$h_t = O_t{}^* \tanh(C_t) \tag{7}$$

where W_o are the weight matrices; b_o are the bias vectors; and h_t is the output vector of the LSTM layer.

D. LightGBM

GBDT is widely used because of its effectiveness, accuracy and interpretability. LightGBM model is a variant of GBDT model proposed by Microsoft Asia Research Institute in 2016 [20]. GBDT and XGBoost adopt level wise strategy. Each tree is divided into leaf nodes layer by layer, which wastes a lot of computing resources. However, LightGBM adopts the optimal leaf wise strategy, and selects the point with the largest splitting income to split [21]. Compared with GBDT and XGBoost, LightGBM model has faster processing speed and higher accuracy.

E. SVR

Support vector machine (SVM) is a kind of supervised learning linear classifier, which aims at binary classification problem. Support vector regression (SVR) is an important application branch of SVM, which is used to solve the regression prediction problem [22]. The basic principles are as follows:

Given data set $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots, (x_n, y_n)\}$, the result $f(x)_i$ is obtained by inputting x_i into the model. The prediction is correct if the maximum deviation of ε can be tolerated between the output $f(x)_i$ and the correct value y_i The SVR model is defined as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(f(x)_i - y_i) \tag{8}$$

where C is the regularization constant.

By introducing the relaxation variables ξ and $\widehat{\xi}$, equation (9) can be rewritten as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i - \hat{\xi}_i) \tag{9}$$

$$s.t. \begin{cases} f(x)_i - y_i \le \xi_i + \varepsilon \\ y_i - f(x)_i \le \hat{\xi}_{i\iota} + \varepsilon \\ \xi_i, \hat{\xi}_i \ge 0 \end{cases} \tag{10}$$

The optimal model can be obtained as follows:

$$f(x) = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)x_i^T x + b \tag{11}$$

where $\hat{\alpha}$, α is the Lagrange multiplier.

IV. DATA ANALYSIS AND RESULTS

A. TIME FEATURES

1) TIME INTERVAL FEATURE

As shown in Fig.4, there are obvious differences in the demand for online car-hailing in different periods of the day. The demand for online car-hailing is less at 0:00-5:00, reaches the bottom at 4:00-5:00, and then begins to increase. The demand starts to increase sharply at 6:00 a.m. and reaches the first peak at 7:00 a.m., which is due to the demand for commuting from home to work or school. In the evening peak, the number of orders reaches the peak due to home-based commuting, and then begins to decline. Therefore, the time period is considered as an important feature when training the model.

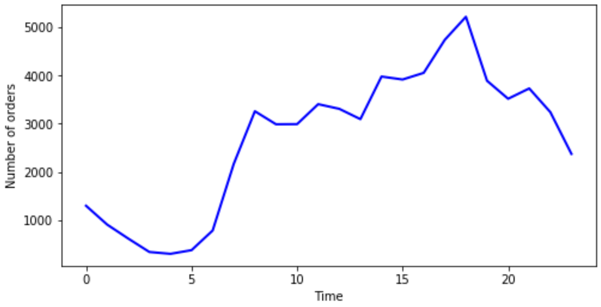


FIGURE 4. Time distribution of online car-hailing travel orders.

2) WEEK FEATURE

Fig.5 shows the trend of online car-hailing services in the current urban area changing with the date. The time range was from May 1, 2017 to May 28, 2017, in which may 6, 7, 13,

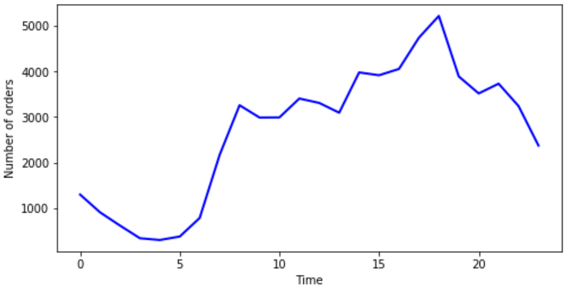


FIGURE 5. Week characteristic of online car-hailing demand.

14, 20, 21, 27 and 28 were weekends. It can be seen that the demand for online car-hailing generally fluctuates in a week. In addition, May 1 is the Labor Day holiday; the demand for online car-hailing is large. May 20 is Chinese Valentine's day with the largest demand for online car-hailing. Therefore, this paper would put forward the time characteristics: 0-6 represents Monday to Sunday; and 7 represents Labor Day on May 1; and 8 represents Valentine's Day on May 20.

3) TIME SERIES FEATURE

The demand of online car-hailing in urban area is continuous. The demand of online car-hailing in the current time period t is affected by the demand in the past time period. Therefore, the demand of online car-hailing in the next period can be predicted by the demand of previous several time periods. Fig.4 and 5 also illustrate the temporal distributions of online car-hailing demand are quite different across different time period. In addition to the commuting peak periods of 7:00 am to 9:00 am, and 17:00 pm to 19:00 pm, considerable number of orders occurred in morning and evening peak hours. In this paper, the demand of the first five time periods (t-1, t-2, t-3, t-4, t-5) is intercepted to construct the time series characteristics.

B. SPATIAL FEATURES

1) SIMILARITY

K-means is a clustering algorithm based on distance. The main idea is: for a given sample set, the sample set is divided into K clusters according to the distance between samples. Let the points in the cluster as close together as possible, and make the distance between clusters as large as possible [23]. In this paper, K-means clustering algorithm is used to analyze the time series of demand in each region, and the input model of regional demand time series similar to the predicted region is obtained by clustering analysis based on the similarity with the time series of predicted regional demand. In this paper, the contour coefficient is used as the evaluation index to determine the K value as 3. As shown in Table 1, these results are named as category 1-3, of which there are 6 areas with the same type as the prediction region 2887, accounting for 28%.

2) NEIGHBORHOOD

The demand of online car-hailing has certain correlations between the predicted region and its adjacent regions.

TABLE 1. Clustering result graph.

Category	Number of similar areas	Proportion
1	15	60%
2	7	28%
3	3	12%

The area is divided into sub-regions. The area of each region is about 1 km2 on average. As shown in Fig.6, the region-level distributions of online car-hailing demand are quite different across the 28 regions. The demand in the central region No.2887 has the similar orders with its 8 adjacent regions and significant differences from distant regions. In this paper, the demand time series of 8 adjacent regions with the predicted region 2887 is used as the feature input model.

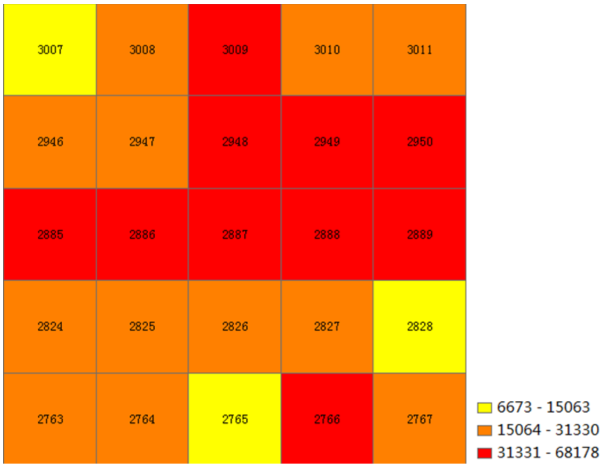


FIGURE 6. Regional distribution of online car-hailing demand.

C. OTHER FEATURES

Weather conditions also affect people's choice of travel mode, and different weather has a certain impact on the demand for online car-hailing, as shown in Fig.7. According to the weather data, the weather on May 4, 2017 (Thursday) was moderate rain, May 9 (Tuesday) and May 11, 2017 (Thursday) was cloudy, and May 16, 2019 (Tuesday) was heavy rain. When the weather is cloudy and rainy, the demand for online car-hailing is greater than that of other weather conditions. It can be seen that the weather conditions have a certain impact on the demand of online car-hailing. Therefore, weather related information can be extracted when constructing model features. The available weather data were collected at each 1 hour, which is the commonly used time aggregation interval for weather data. The weather variables for each observation at each region were extracted based on the time of each observation. The observations during one hour were assigned with the same values of weather variables.

D. ALGORITHM FLOW

The structure of online car-hailing demand prediction model based on Stacking ensemble learning is shown in Fig.8. Random Forest, LightGBM and LSTM are selected as the first

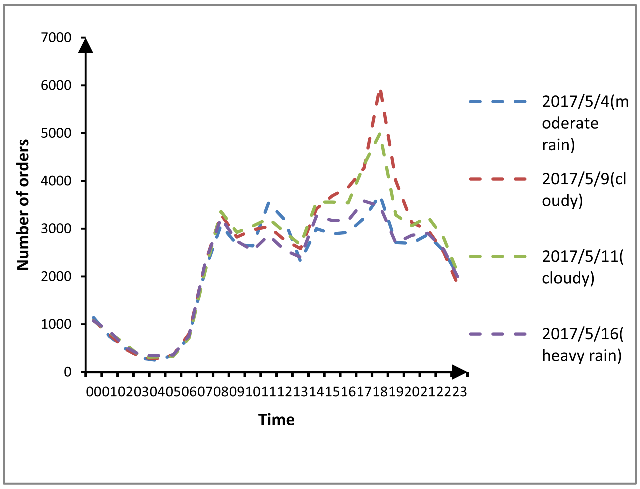


FIGURE 7. Distribution of online car-hailing trip order for different weather conditions.

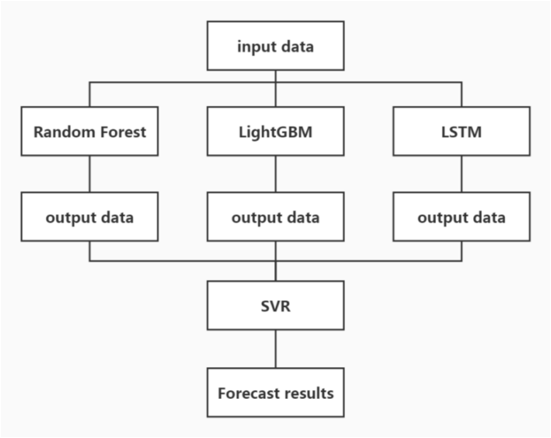


FIGURE 8. Network online car-hailing demand forecast model structure.

layer of base learners, and the output of the first layer of base learners is input into the SVR model of the second layer, and then the final prediction results are obtained. The algorithm steps are as follows:

(1) Select data to preprocess the data, construct training set and testing set, and divide the training set into 5 fold cross validation method.

(2) The training set is input into the Random Forest, LightGBM and LSTM of the first layer base learner respectively for training, and obtain the predicted results.

(3) The predicted results of the first layer base learners are spliced into the SVR of the second layer learners, and trained again to obtain the final prediction results.

E. ERROR ANALYSIS

In order to better analyze the prediction effect of stacking model. This paper introduces two evaluation indexes: Mean Absolute Error $MAE = \frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i|$and Root Mean Square Error $RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2}$.

V. CASE ANALYSIS

A. DATA SOURCES

The data set used in this paper is the two-week online car-hailing order data of Haikou City, China. The dataset mainly includes: trip start time, trip end time, pick-up longitude and latitude, drop-off longitude and latitude, passenger carrying time, the number of passengers, order ID and trip distance. The order time, order ID and GPS position were applied to extract trip data from the raw geo-location information sample. The origin, destination, start time and end time associated with each trip were extracted by scanning whether the GPS position of one vehicle changed between two continuous samples. The dataset was from May 1, 2017 to May 28, 2017, with a total of more than 800000 trips. ArcGIS was used to extract the order data from a 1km * 1km grid, and 25 regions in the city center were taken as experimental regions. ArcGIS was then applied to match the order data with each region and time to generate the demand data. Then 10 minutes, 15 minutes and 30 minutes were taken as different time intervals to count the order quantity for each region. Time period feature and week feature were extracted, and weather feature and time series feature were added to the sample dataset. Finally, all the features were statistically analyzed, including weather characteristics, time period characteristics, week characteristics and regional characteristics. The classification variables were processed by one-hot coding. The m categories of classification variables were converted into m variables with the value of 0 or 1. Although the transformation increases the number of features and improves the complexity of the model, it also improves the accuracy of prediction. The training data was from May 1 to May 21, 2017. The testing data is from May 22, 2017 to May 28, 2017.

B. MODEL PARAMETERS

As shown in Table 2, 3, 4 and 5, the optimal parameters of the model for the three time intervals are obtained by grid search method. The main parameters of Random Forest are the number of trees and the maximum depth of each tree. The main parameters of LightGBM are learning rate, the number of trees and the maximum depth of each tree. The main parameters of SVR model are penalty coefficient and kernel function parameters. The activation function selected by LSTM is SGD function. SGD function training speed is fast, which is a practical and feasible training algorithm. In addition, in order to prevent overfitting problem, dropout rate is introduced into the model to reduce the interaction between feature detectors (hidden layer nodes). The specific parameters of LSTM model are shown in Table 5.

TABLE 2. Optimal parameters of random forest.

Time interval	Parameters	
	*N*_estimators	Max_depth
10-minutes	175	10
15-minutes	200	8
30-minutes	200	10

TABLE 3. Optimal parameters of SVR.

Time interval	Parameters	
	C (Penalty coefficient)	Gamma(RBF kernel function parameter)
10-minutes	100	0.001
15-minutes	37	0.00001
30-minutes	100	0.0001

TABLE 4. Optimal parameters of LightGBM.

Time interval	Parameters		
	*N*_estimators	Max_depth	Learning_rate
10-minutes	150	5	0.03
15-minutes	150	5	0.03
30-minutes	200	4	0.02

TABLE 5. Optimal parameters of LSTM.

Parameters	10-minutes	15-minutes	30-minutes
Input layer	1	1	1
Output layer	1	1	1
Hidden layer	1	1	1
Number of neurons	32	32	75
Time step	1	1	1
Activation function	Tanh	Tanh	Tanh
Optimization function	SGD	SGD	SGD
Batch_size	200	220	50
Dropout rate	0.4	0.6	0.5

In order to verify the effectiveness of the proposed forecasting model of car-hailing demand based on Stacking ensemble learning, the prediction results of Random Forest, LightGBM, SVR, LSMT and stacking ensemble learning model were compared and analyzed. The models are implemented by Python 2.7 sklearn module and keras module.

All the models were trained by using the same samples and exogenous variables to stacking ensemble learning model training. Table 6 and Fig.9 compare the predictive performance of stacking ensemble learning approach with those of four methods on the validation data sample. The MAE and RMSE were as the evaluation index to measure the predictive performance. As shown in Fig.10, 11 and 12, the predicted values of the five models are close to the observed values. The fluctuation of the results of the online car-hailing demand forecasting model based on stacking ensemble learning is closer to the observed value. Compared with the other four models, the error of stacking ensemble learning models is lower. MAE index is 5.163, 3.337 and 2.994, and RMSE index is 6.746, 4.387 and 3.929, for 10 min, 15 min, and 30 min time intervals respectively. Therefore, the stacking ensemble learning model produces significantly higher prediction accuracy than other four models. Comparing the prediction accuracy of stacking ensemble learning model for different time intervals, it can be found that the predictive performance increases with an increase in the aggregation time interval. The reason is that the car-hailing order data aggregated at the 10-min interval have greater data noises and more useless fluctuation information than the longer time intervals. The data aggregated at shorter time intervals are more difficult to be predicted.

TABLE 6. Prediction performances of different models on the validation sample.

Model	30 min		15 min		10 min	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
SVR	5.473	7.227	3.578	4.586	3.040	3.978
LightGBM	5.532	7.119	3.384	4.399	3.020	3.936
LSTM	6.077	8.130	3.490	4.434	3.047	4.063
Random Forest	6.247	8.303	3.725	4.902	3.405	4.704
Stacking ensemble learning	5.163	6.746	3.337	4.387	2.994	3.929

For illustrative purposes, Fig.12 (e) descripts the prediction accuracy of the stacking ensemble learning model over the regional demand at the 30 min time interval. The gap between the predicted values and the observed values for 30 min is the narrowest in the different time intervals. The results of MAE and RSME in Table 6 further illustrate that the stacking ensemble learning model provides reasonably accurate forecasts of car-hailing demand. As expected, the predicted demand by stacking ensemble learning model is very close to the real observations, indicating that the predictive performances are satisfactory.

VI. THE APPLICATION OF ONLINE CAR-HAILING DEMAND FORECAST

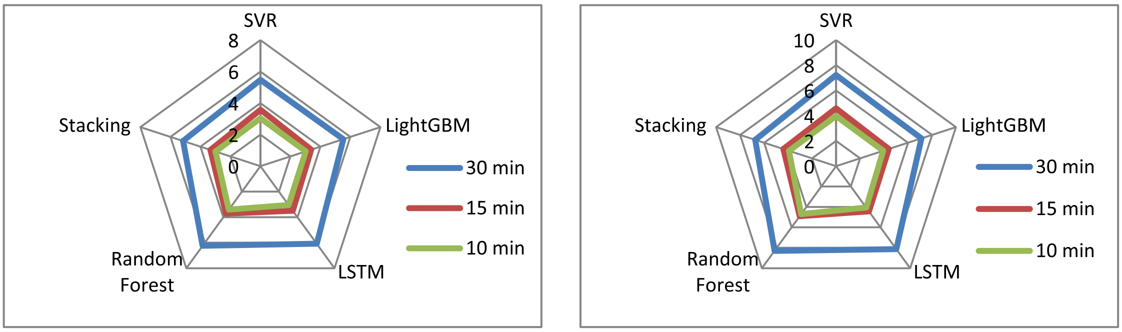
It is of great significance to predict the short-term demand of the regional network in advance for the scheduling and dynamic pricing of the network.

A. EARLY SCHEDULING

An accurate prediction of the short-term demand of a network in a certain area can predict passengers' travel demand in a short time in the future, and use this as guidance information to deploy vehicles ahead of time, to a certain extent, to solve the problem of supply and demand between the passenger's demand for taxi and the supply of vehicles on the net. On the one hand, reasonable scheduling of online car-hailing can optimize the relationship between supply and demand, reduce empty mileage and energy consumption, reduce air pollution, and improve the operation efficiency and driver income of online car-hailing. On the other hand, it can save passengers' waiting time on the roadside and improve their satisfaction with the service.

B. DYNAMIC PRICING

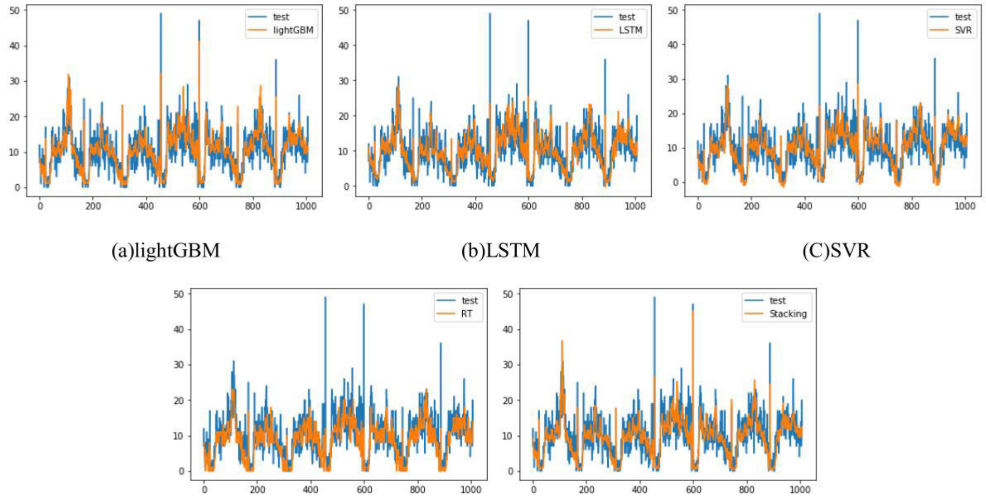
The online car-hailing operation platform can adjust the online car-hailing price by predicting the future demand of online car-hailing. The cost of online car-hailing is one of the main factors of passengers' travel choice. Therefore, considering the supply-demand relationship of online car-hailing platform, determining the optimal ride price, adjusting the



(a) MAE comparison of different models

(b) RMSE comparison of different models

FIGURE 9. Comparison of prediction errors of different models.



(a)lightGBM

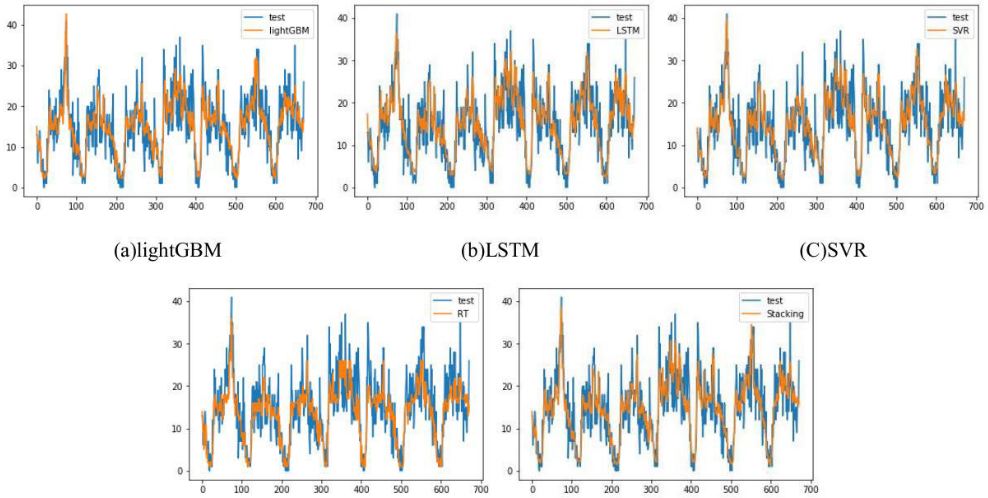
(b)LSTM

(C)SVR

(d)Random Forest

(e)Stacking

FIGURE 10. Comparison of predicted values in 10-minutes.



(a)lightGBM

(b)LSTM

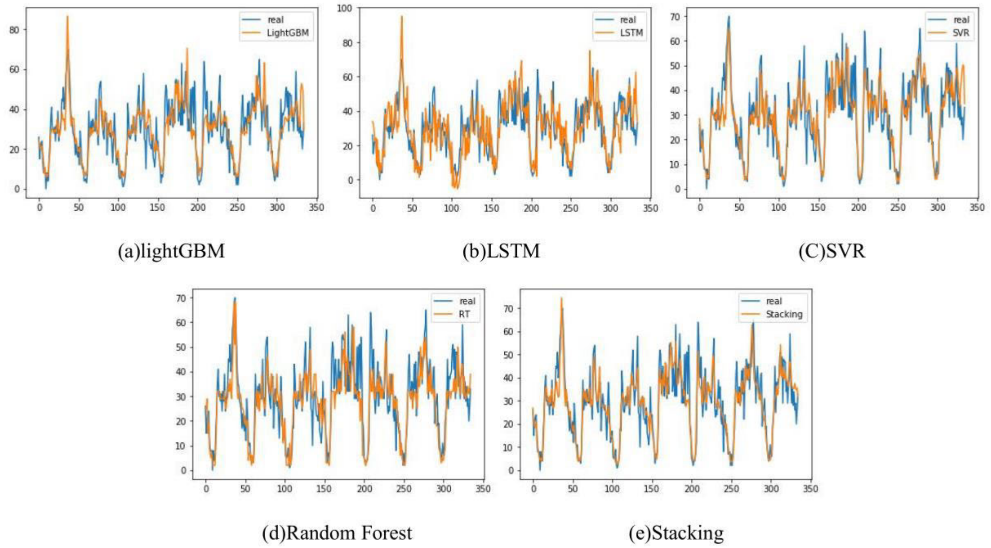
(C)SVR

(d)Random Forest

(e)Stacking

FIGURE 11. Comparison of predicted values in 15-minutes.

supply capacity of online car-hailing, reducing order delay and idle travel resources, and meeting the demand of online car-hailing platform are of great significance for the operation management and optimization of online car-hailing platform.



(a)lightGBM (b)LSTM (C)SVR

(d)Random Forest (e)Stacking

FIGURE 12. Comparison of predicted values in 30-minutes.

VII. CONCLUSION

Accurate demand prediction for a certain region can provide support for early scheduling and dynamic pricing of car-hailing services. This paper studied the short-term demand forecasting of online car-hailing based on Didi's data in Haikou City. The purpose of this paper is to develop a short-term demand forecasting model for online car-hailing services by applying the stacking integrated learning approach with large-scale travel data. The temporal, spatial and weather characteristics as influential factors were also extracted to input into the model. Finally, taking area 2887 as an example, the demand for 10 min, 15 min and 30 min time interval were predicted respectively, and compared with the predictions of other four single models. The results show that when the time interval is 30 minutes, the short-term forecasting model based on stacking ensemble learning performs best. MAE and RMSE increased by 6.0% and 5.2% respectively. Therefore, the proposed model n this paper has higher prediction accuracy and applicability.

Due to the limited urban area covered by the data set, it is not possible to predict the short-term demand of online car-hailing in other areas of the city. In addition, although many factors are considered in this paper, there are still many characteristics that affect the short-term demand of online car-hailing. In order to improve the performance of the model, we will further mine the spatiotemporal correlation of data and the influence of external factors (such as traffic congestion index) on the model in the future work.

ACKNOWLEDGMENT

The authors thank his mentor, Xiaofei Ye of the Ningbo University, who gave instruction on writing this paper. They also thank the interviewers for their assistance in the survey.

REFERENCES

[1] X. F. Ye, M. Li, Z. Yang, X. C. Yan, and J. Chen, "A dynamic adjustment model of cruising taxicab fleet size combined the operating and flied survey data," *Sustainability*, vol. 12, no. 2276, pp. 1–18, Apr. 2020, doi: 10.3390/su12072776.

[2] X. Feng, "Research on the scale of taxi development under the balance of supply and demand," M.S. thesis, Dept. Transp. Eng., Southwest Jiaotong Univ., Chengdu, China, 2010.

[3] X. Guo, "Prediction of taxi demand based on gradient ascending regression tree," in *Proc. WTC*, Beijing, China, 2018, pp. 310–320.

[4] K. Niu, C. Wang, X. Zhou, and T. Zhou, "Predicting ride-hailing service demand via RPA-LSTM," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4213–4222, May 2019, doi: 10.1109/TVT.2019.2901284.

[5] P. Lin and N. Zhou, "Short term traffic flow prediction of toll stations based on multi feature gbdt model," *J. Guangxi Univ.*, vol. 43, no. 3, pp. 1192–1199, Mar. 2018.

[6] H. Chang, Y. Tai, and J. Y. Hsu, "Context-aware taxi demand hotspots prediction," *Int. J. Bus. Intell. Data Mining*, vol. 5, no. 1, pp. 12–19, Jan. 2010.

[7] L. Moreira-Matias, J. Gama, M. Ferreira, and L. Damas, "A predictive model for the passenger demand on a taxi network," in *Proc. 15th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 1014–1019, doi: 10.1109/ITSC.2012.6338680.

[8] X. Jia, "Analysis and forecast of residents' travel demand based on online car Hailing data," *Transp. Eng.*, vol. 18, no. 05, pp. 39–45, May 2018.

[9] Y. Zhong, Y. Shao, W. Wu, and G. Hu, "Short term traffic flow prediction model based on XGBoost," *Sci. Technol. Eng.*, vol. 19, no. 30, pp. 337–342, Jul. 2019.

[10] I. Saadi, M. Wong, B. Farooq, J. Teller, and M. Cools, "An investigation into machine learning approaches for forecasting spatio-temporal demand in ride-hailing service," 2017, *arXiv:1703.02433*. [Online]. Available: http://arxiv.org/abs/1703.02433

[11] P. Rodrigues, A. Martins, S. Kalakou, and F. Moura, "Spatiotemporal variation of taxi demand," *Transp. Res. Procedia*, vol. 47, pp. 664–671, Dec. 2020, doi: 10.1016/j.trpro.2020.03.145.

[12] J. Xu, R. Rahmatizadeh, L. Boloni, and D. Turgut, "A sequence learning model with recurrent neural networks for taxi demand prediction," in *Proc. IEEE 42nd Conf. Local Comput. Netw. (LCN)*, Oct. 2017, pp. 261–268, doi: 10.1109/LCN.2017.31.

[13] G. Xu, Y. Li, L. Wang, L. Zhang, Q. Yang, J. Ye, and Y. Liu, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proc. AAAI Conf. Artif. Intell.*, Worcester, MA, USA, 2019, vol. 33, no. 1, pp. 3656–3663, doi: 10.1609/aaai.v33i01.33013656.

[14] C. Wang, Y. Hou, and M. Barth, "Data-driven multi-step demand prediction for ride-hailing services using convolutional neural network," in *Proc. Adv. Comput. Vis.*, Las Vegas, NV, USA, 2019, pp. 11–22, doi: 10.1007/978-3-030-17798-0_2.

[15] L. C. Lei, S. Gao, and E.-Y. Zeng, "Regulation strategies of ride-hailing market in China: An evolutionary game theoretic perspective," *Electron. Commerce Res.*, vol. 1, pp. 1–29, Apr. 2020.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[17] Y. Luo, X. Jia, S. Fu, and M. Xu, "PRide: Privacy-preserving ride matching over road networks for online ride-hailing service," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1791–1802, Jul. 2019.

[18] A. Graves, A. Rahman Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, vol. 13, May 2013, pp. 6645–6649.

[19] C. Xu, J. Ji, and P. Liu, "The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets," *Transp. Res. C, Emerg. Technol.*, vol. 95, pp. 47–60, Oct. 2018, doi: 10.1016/j.trc.2018.07.013.

[20] Y. Guo, Y. Zhang, J. Yu, and X. Shen, "A spatiotemporal thermo guidance based real-time online ride-hailing dispatch framework," *IEEE Access*, vol. 8, pp. 115063–115077, 2020, doi: 10.1109/ACCESS.2020.3003942.

[21] K. Xu, L. Sun, J. Liu, and H. Wang, "An empirical investigation of taxi driver response behavior to ride-hailing requests: A spatio-temporal perspective," *PLoS ONE*, vol. 13, no. 6, Jun. 2018, Art. no. e0198605.

[22] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Taxi-Passenger-Demand modeling based on big data from a roving sensor network," *IEEE Trans. Big Data*, vol. 3, no. 3, pp. 362–374, Sep. 2017.

[23] D. Pu, F. Xie, and G. Yuan, "Active supervision strategies of online ride-hailing based on the tripartite evolutionary game model," *IEEE Access*, vol. 8, pp. 149052–149064, Jul. 2020, doi: 10.1109/ACCESS.2020.3012584.

YUMING JIN was born in Yiwu, China, in 1982. He received the degree in business management (traffic and transportation planning) from the School of Economics and Management, Chang'an University, in 2007, where he is currently pursuing the Ph.D. degree in traffic and transportation planning and management. His research interests include the research on the demand of network car-hailing and cruising taxi, and large data technology.

XIAOFEI YE was born in Chengde, China, in 1984. He received the B.S. degree in traffic engineering from the Heilongjiang Institute of Technology in 2007, the M.S. degree in transportation planning and management from the Harbin Institute of Technology in 2009, and the Ph.D. degree in transportation engineering from Southeast University in 2013.

From 2013 to 2017, he was a Lecturer with Ningbo University, where he has been an Assistant Professor with the Department of Logistics Management, School of Maritime and Transportation, since 2018. He has authored two books and more than 30 articles, and holds eight patents. His research interests include parking planning and management, non-motorized vehicles and pedestrian traffic flow theory, traffic reliability, and advanced traffic information systems.

Dr. Ye was a recipient of science and technology awards from the China General Chamber of Commerce in 2016.

QIMING YE was born in Ningbo, China, in 1996. He received the bachelor's degree in traffic engineering from the Ningbo Institute of Engineering in 2019. He is currently pursuing the master's degree with Ningbo University. His research interests include the research on the demand of network car hailing and large data technology.

TAO WANG received the bachelor's degree in traffic engineering from the Guilin University of Electronic Science and Technology in 2007, the master's degree in traffic planning and management from Southeast University in 2010, and the Ph.D. degree in traffic engineering from Southeast University in 2017.

From 2010 to 2019, he served as a Teacher with the School of Architecture and Transportation Engineering, Guilin University of Electronic Science and Technology. His research interests include traffic behavior and safety, urban traffic planning and design, traffic data analysis, and intelligent transportation.

Dr. Wang's awards and honors include the Guangxi Science and Technology Progress Award and the 1000 Young and Middle-Aged Backbone Teachers in Guangxi Colleges and Universities.

JUN CHENG received the bachelor's degree in electronics and mechanics from the Xi'an University of Electronic Science and Technology in 1995, and the master's degree in automatic control theory and applications and the Ph.D. degree in transportation planning and management from Southeast University, in 1998 and 2000, respectively.

From 2000 to 2019, he worked as a Teacher at the School of Communications, Southeast University. His research interests include urban comprehensive transportation planning and management, urban parking facilities planning and management, urban public transport planning, and intelligent transportation management and control.

Dr. Cheng's awards and honors include the Ministry of Education's Supporting Plan for Excellent Talents in the New Century, the Ministry of Transportation's "Young Transportation Science and Technology Talents" Title, and the Jiangsu's "333 Talents Project" Supporting Plan.

XINGCHEN YAN received the bachelor's degree in transportation from Nanjing Forestry University in 2008, the master's degree in transport planning and management from Southeast University in 2009, and the Ph.D. degree in forestry economic management from Nanjing Forestry University in 2012.

From 2004 to 2019, he served as a Teacher at the College of Mechanical and Electrical Engineering and the College of Automobile and Transportation Engineering, Nanjing Forestry University. His research interests include transportation planning, intelligent transportation, and complex network modeling.