PAPER • OPEN ACCESS

Research on hotel online sales forecast model based on improved WaveNet

View the article online for updates and enhancements.

Research on hotel online sales forecast model based on improved WaveNet

Dongmei Duan[1*]

Institute of Information, Beijing University of Technology, No100 Pingleyuan,

Chaoyang District, Beijing, China

ddm9227@126.com

Abstract. In recent years, with the rapid development of the Internet in China, online transactions have grown greatly. For example, OTAs with a large number of hotels have accumulated a large amount of hotel data and user consumption data. And the online sales of hotels is the basis and core of revenue management. Time series prediction has always been one of the main application fields of machine learning algorithm. From the classical traditional time series prediction methods to long-term and short-term memory networks and closed-loop neural networks, the prediction ability is constantly improving. With the development of deep neural networks, convolution neural networks show superior performance in the prediction of time series. This paper proposes a new prediction model based on the improved WaveNet using not only the parameters of historical sales and hotel property, but also the parameters of holiday time and time position in the prediction range, which are processed by serialization. Simulation results are presented in details in this paper, where these results indicate the effectiveness of the proposed forecasting tool as an accurate technique.

1. Introduction

In recent years, China's tourism market has maintained a continuous development trend, and the Internet has constantly penetrated into all aspects of national life. People tend to book hotels through the Internet, and OTA (online travel agency) with a large number of hotels has accumulated a large number of hotel data and user consumption data. Accurate prediction of hotel online sales plays an important role in improving the efficiency of the whole hotel industry. However, online sales of hotels are affected by holidays, hotel environment, regional development and other factors. Although each factor has different effect, some of them must be more important than the others, or there are more important parameters that have not been mined out. The problem of hotel online sales forecast is essentially a time series prediction problem. Time series refers to a group of statistical data arranged in chronological order observed or recorded for the same phenomenon [1]. The core of time series prediction problem is to mine the trend of this group of data changing with time and then use it to estimate the future data. One of the main issues in sales forecasting is determining the most effective parameters and their importance in hotel online sales. The second is to apply an effective online hotel sales forecasting method.

Many scholars have done a lot of research on time series prediction algorithm and application in various fields. There are two main methods of time series prediction, one is based on statistics, the other is based on machine learning. There are two main prediction methods of time series, one is based on statistics, and the other is based on machine learning. The prediction methods based on probability statistics can be divided into linear and nonlinear modeling methods. In 1927, the autoregressive model (AR model) proposed by the British statistician Yule in the study of sunspots laid the foundation for the

discipline of time series prediction. Inspired by AR model, Walker, another mathematician, established MA model and ARMA model in 1931. Box and Jenkins formally puts forward the time series analysis method in the book *Time Series Analysis: Forecasting and Control* [2] which points out that it is theoretically applicable to time series analysis in various fields, and systematically expounds the ARIMA model for the first time. Now, the ARIMA model is also known as box Jenkins model, which first transforms the non-stationary time series into a stable one that is easy to handle, and then uses the ACF and PACF pattern recognition model according to three basic models (AR, Ma and ARMA), and then take the identification, parameter estimation, model test and control. This has been the most common method in time series modeling.

Due to the disadvantages of traditional time series prediction methods, researchers have invented many time series prediction algorithms based on machine learning. In the past decades, people have proved that machine learning algorithm can effectively deal with a large number of high-dimensional nonlinear data. In machine learning algorithm, the classical models that can be used for regression prediction are Logistic Regression [3], GBRT (MART) [4], XGBoost [5], Neural Network model [6], Support Vector Machine (SVM) [7]. The traditional boost algorithm is to assign the same weight 1/N to each sample at the beginning of calculation, and gets a weak learner every time of the training. There will be differences when using this weak learner to estimate the samples. After n weak learners are obtained by N iterations, the learners with high accuracy are given high weight, and then the N weak learners are combined by weighting to get a final model. The basic idea of gradient lifting regression tree [8] is to calculate a series of simple regression trees. Each tree is built to predict the residual of the previous tree, which is the difference between the predicted value and the real value. WaveNet is a deep learning based speech generation model [9] launched by Deepmind company of Google in 2016, which improves the efficiency of convolution by using the dilated convolution network. Shaojie Bai et al. [10] had proved that a simple convolution structure is superior to the standard recursive network, and they concluded that the common association between sequence modeling and recurrent networks should be reconsidered, and convolutional networks should be regarded as a natural starting point for sequence modeling tasks. The architecture of WaveNet model enables it to take advantage of the efficiency of the convolution layer, while alleviating the challenge of learning long-term dependencies across a large number of time steps (1000 +). In face of high-dimensional time series, sequence to sequence prediction model [11] plays a great role. At the cost of constructing and adjusting the complexity of the model, one model can capture the whole prediction problem in all sequences. In this paper, the sequence to sequence framework and WaveNet framework are combined to construct a hotel online sales forecast model based on improved WaveNet. After experimental verification, when more relevant parameters are input into the model, the proposed forecast model reflects higher accuracy.

2. Factors influencing hotel online sales forecast

In the free economic market, the fluctuation of hotel online sales is a common phenomenon, which has been influenced by both individual development factors and the development of the environment, some of which have a greater impact on hotel online sales[12]. The following categories of influencing factors are used in this paper.

2.1. Time Features

The hotel online sales are different from the sales of a single commodity. It refers to the room rentals of a hotel in a period of time. At present, the general measurement unit of hotel sales is the consumption room nights, that is, the room nights = the number of rooms to be occupied * the number of days to be occupied. This paper forecasts the hotel online sales in a day dimension, that is, the actual consumption room nights of a hotel a day in a period of time in the future on the OTA platform. It can be seen from Figure.1 that the consumed room nights of the hotel is relatively stable as a whole, showing a periodicity in weeks, with peak occupancy on Friday and Saturday. At the same time, due to the influence of holidays, during the typical holidays, there will be a huge explosion in the hotel online sales. For example, during the Spring Festival, the hotel online sales will be significantly lower than that on

weekdays, while on the International Labor Day and National Day, the hotel online will increase greatly, and at the same time, it will also have an impact on the sales before and after the holidays. Holiday time is an important feature that affects the hotel online sale to encode holidays. At the same time, this paper will forecast the future hotel online sales on a weekly basis. The longer the forecast period is, the more time will be available for the OTA platform and the hotel to manage the revenue. However, the accuracy of the forecast will also decline. It is a reasonable time range to determine the forecast period as 2 weeks.
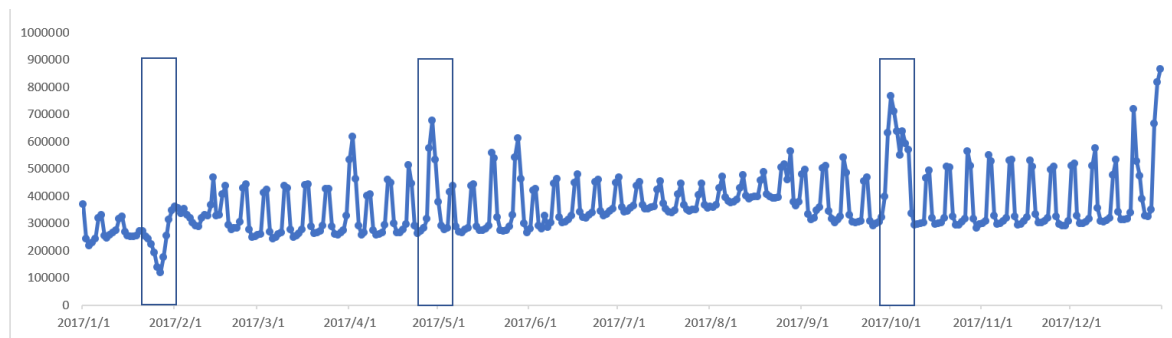

Figure.1 the consumed room nights in 2017

2.2. Basic features of hotel

The development of hotel industry is related to GDP, consumption level, consumption concept, infrastructure, etc. There are differences in hotel sales levels among different regions. The overall trend of hotel sales in each province is consistent with the overall trend of all hotels. However, we can see in Figure.2 and Figure.3 there are obvious differences in the sales of hotels in different provinces, and there are also differences in the hotel sales in different cities of the same province. Therefore, features such as provinces and cities are also characteristic labels of hotel sales.
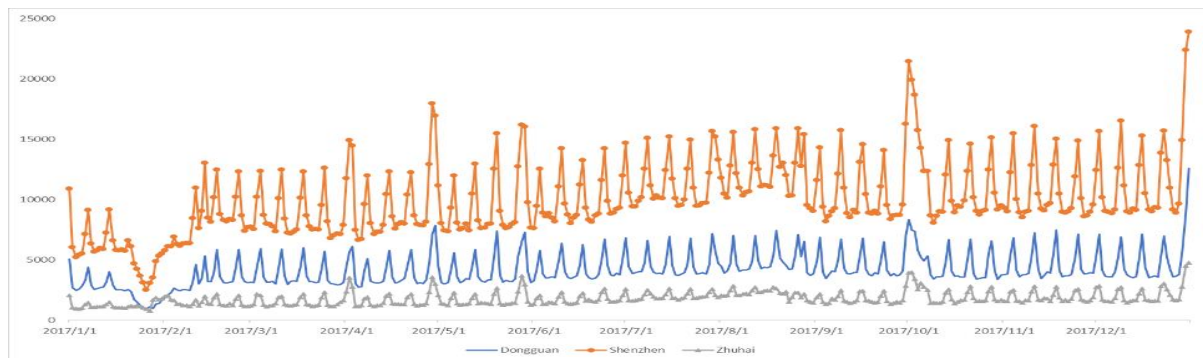

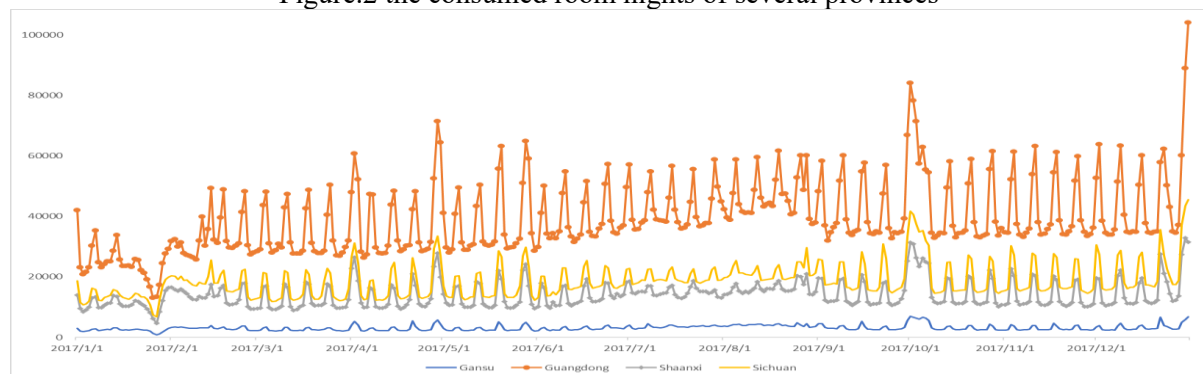Figure.2 the consumed room nights of several provinces

Figure.3 the consumed room nights of several cities in the same province

2.3. The features of sales

Sales show significant differences in different seasons, and in terms of annual cycles, there is similar volatility, sales itself can also be an important feature.

3. Improved WaveNet

In the hotel online sales forecasting, the main goal is to get the future sales according to the effective features. Convolution network is an effective method to solve this problem, because the relationship between sales and other features is nonlinear. It is a complex and time-consuming work to build a forecast model which is suitable for large data volume, high dimension and non-linear. The architecture of WaveNet model enables it to take advantage of the efficiency of volume accumulation layer, and at the same time relieves the challenge of learning long-term dependency across a large number of time steps. Similar architectures were used for predicting Uber demand in NYC[13]. Facing high-dimensional time series, sequence to sequence prediction model plays a great role. It uses one model to capture the whole prediction problem in all series at the cost of building and adjusting the complexity of the model. The effective combination of the two can solve the above problems and get good forecast results.

3.1. WaveNet

The main component of WaveNet is the dilated causal convolution network as Figure.4 showed, which can correctly handle the time sequence and long-term dependence without causing the model complexity to surge. The system is a generative model: it can generate the sequences of real-valued data starting from some conditional inputs. The behavior is mainly due to the dilated causal convolutions. A big number of layers and large filters are used to increase the receptive field within the causal convolutions。 It only allows the input to be connected to the future time step by causality.

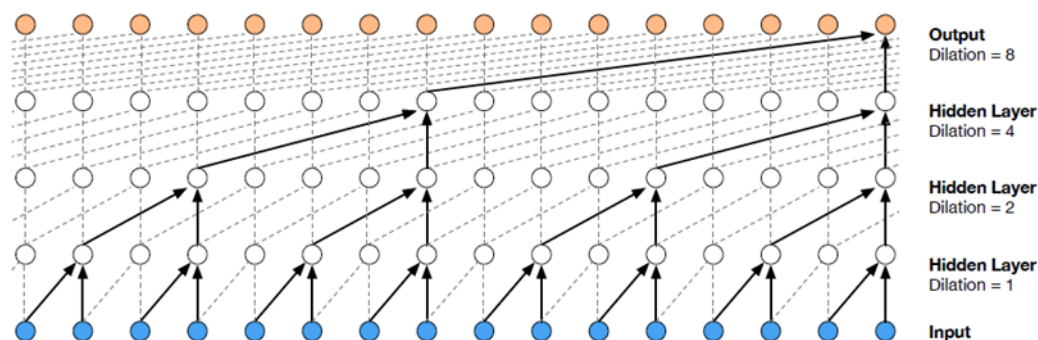Figure.4 Visualization of a stack of dilated causal convolutional layers[9]

In fact, by offsetting the traditional convolution output by multiple time steps, this one-dimensional structure of causality can be easily realized. The dilated convolution allows the receptive field to grow exponentially according to the depth of the convolution layer. In the dilated convolution layer, the filter is not applied to the input in a simple sequential manner, but instead skip a constant dilation rate inputs in between each of the inputs they process. Through in each layer (e.g. 1, 2, 4, 8...), the exponential relationship between the required depth of the layer and the size of the receptive field can be achieved by increasing the expansion rate. In Figure.4, we can see that we can now connect all 16 input sequence values to the highlighted output (for example, the 17th time step value) in only four layers. By extension, when dealing with daily time series, only nine dilated convolution layers in this form can capture more than one year's history. The dilated convolution output is divided into two branches, and then recombined by element multiplication. This is the application of the door control activation unit. At the same time, the skip connection is used to retain the earlier feature layer output when the network passes forward signals for final prediction processing. Residual connections facilitate the use of deeper networks by allowing more direct gradient flows in back propagation. Due to the length of the back

propagation chain, it is usually difficult to train the early layers of the deep network effectively, but residual connection and skip connection will create a more easy information highway. Given an additional input h, WaveNet can model the conditional distribution of this given input:

$$p(x|h) = \prod_{t=1}^{T} p(x_t | x_1, \ \cdots, \ x_{t-1}, \ h) \tag{1}$$

$$x = \{x_1, \ \cdots, \ x_T\} \tag{2}$$

3.2. Sequence to sequence learning

In the traditional time series prediction, we usually consider the series one by one, and then fit the prediction model with the specific parameters of the series. Unfortunately, they can't extend well to the problem that the number of sequences to be predicted increases to thousands or even hundreds of thousands. Multi-step time series prediction can be expressed as the prediction problem of sequence to sequence supervision, which is suitable for the framework of modern neural network model. At the cost of constructing and adjusting the complexity of the model, one model can capture the whole prediction problem in all series. Because neural network is a natural feature learner, it can also use the simple method to carry out feature engineering when preparing the model. This model uses "encoder decoder" framework to map any length of input sequence to any length of output sequence with intermediate coding state.

WaveNet is trained using the next step prediction, so in the absence of condition information, errors may accumulate due to the growth sequence generated by the model. To solve this problem, we trained the model to minimize the loss in step 14. We adopt sequence to sequence method, in which the encoder and decoder do not share the parameters. This allows the decoder to deal with accumulated noise when generating long sequences. The flow chart of the whole prediction model is shown in Figure 5.
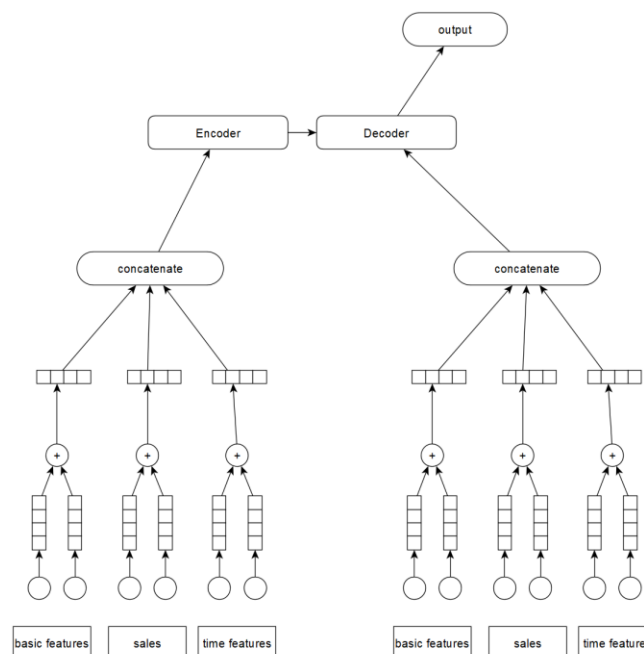

Figure.5 Flow chart of the hotel online sales forecast model based on improved WaveNet

4. Experiments

In order to prove the effectiveness of the proposed method, some real data that have not been used in the training process are used to test the prediction model. In order to measure the accuracy of the

proposed method, RMSE and MAPE are used to evaluate the prediction results of the model, RMSE is defined as (3):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{3}$$

The mean absolute percentage error (MAPE) is defined as (4) to determine the mean error of all the test results:

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\frac{|y_i - \hat{y}_i|}{y_i} \tag{4}$$

Where, M is the size of the data set, $\hat{y}_i$ is the prediction value of the model for the consumption of sample I, and y_i is the real consumption of sample I.

In the previous literature of hotel online sales forecast, the feature that each predicted date in the time position of the whole forecast range is not used. However, due to the particularity of hotel online sales, not only the sales will increase sharply during the holidays, but also the sales before and after the holidays will be affected. The following two cases confirm the proposed method is proved to be an accurate forecast method. The two experiments take the forecast of online hotel sales from December 18, 2018 to December 31, 2018 as an example.

Case A: In this case, no holiday and predicted date time location features are added, and the experimental results are shown in the Figure.6 and Table.1.

Case B: In this case, all the features mentioned above are added to the sales forecast model. The prediction results are shown in the Figure.7 and Table.1.
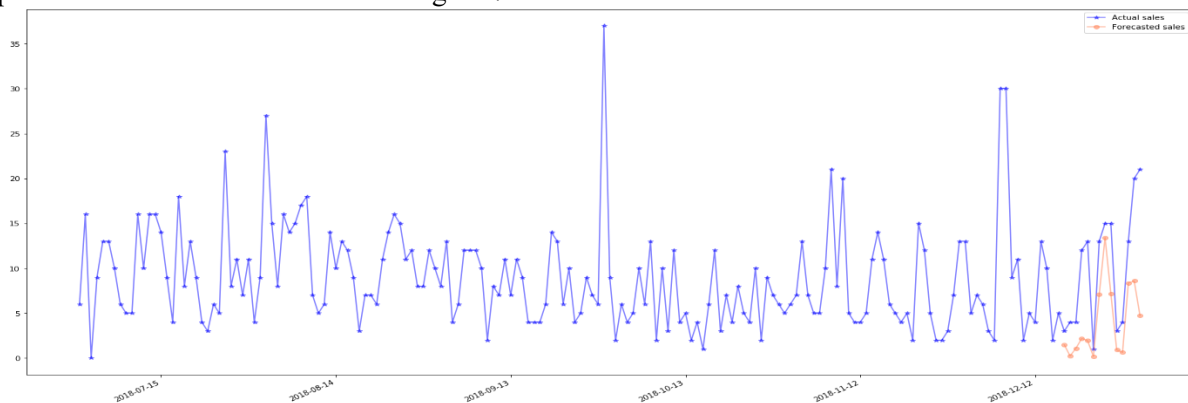

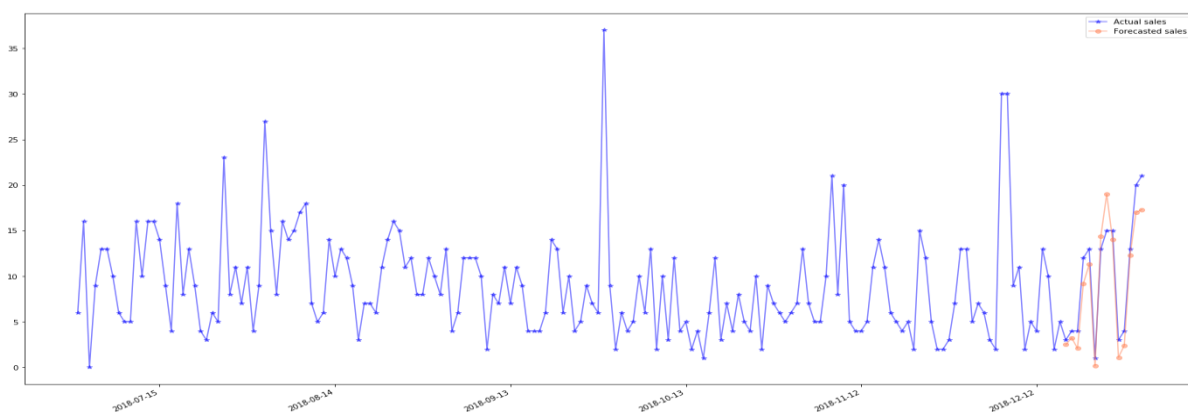Figure.6 Forecasted online sales for 2 weeks in Case A

Figure.7 Forecasted online sales for 2 weeks in Case B

Table.1 RMSE and MAPE of the Cases

Case	RMSE	MAPE
A	7.4412	0.6407
B	2.1429	0.2791

5. Conclusion

Based on the above analysis, this paper proposes an improved WaveNet based hotel online sales forecasting model, and describes the principle, design and implementation process of the model. Through the verification of actual data, compared with other traditional statistical prediction methods and machine learning models, the model proposed in this paper has obvious advantages in processing high-dimensional and non-linear data, and the prediction error is less than 30%, which is acceptable for OTA enterprises. The whole experimental process reflects the efficiency and accuracy of the prediction model.

References

[1]   Shi Lee. Application statisticsM]. Beijing: Tsinghua University Press, Beijing Jiaotong University Press, 2005: 115-120.

[2]   Box, G. E. P, Jenkins, G. M. Time series analysis: Forecasting and control[M]. San Francisco: Holden-Day, 1970.

[3]   Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (p. 217–244). American Psychological Association.

[4]   Li, Chenhui, Liu, Yuhua, Huang, Yifei, et al. GBRTVis: online analysis of gradient boosting regression tree[J]. Journal of visualization, 2019, 22(1): 125-140.

[5]   KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust 2016 Pages 785–794https: //doi. org/10. 1145/2939672. 2939785.

[6]   Bao Liu, Daiping Hu, Some applications of neural network in prediction of neural networks [J]. System Engineering News, 1999, 14(4): 338-344. DOI: 10. 3969/j. issn. 1000-5781. 1999. 04. 008.

[7]   Shifei Ding, Bingjuan Qi, Hongyan Tan. A survey of support vector machine theory and algorithm[J]. Journal of University of Electronic Science and technology, 2011, 40(01): 2-10.

[8]   Li, Chenhui, Liu, Yuhua, Huang, Yifei, et al. GBRTVis: online analysis of gradient boosting regression tree[J]. Journal of visualization, 2019, 22(1): 125-140.

[9]   Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio[J]. 2016, 2–5.

[10]   Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neuralnetworks. In Advances in Neural Information Processing Systems, pages 3104–3112, 2014.

[11]   Bai, S., Zico Kolter, J. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. Preprint at arXiv https://arxiv.org/abs/1803.01271 (2018).

[12]   Xiaofei Yu. Study on the influencing factors of hotel passenger online ordering [D]. South China University of Technology, 2014.

[13]   Chen, L. , Ampountolas, K., Thakuriah, P. Predicting Uber Demand in NYC withWavenet. In Proceedings of the Fourth International Conference on Universal Accessibility in the Internet of Things and Smart Environments, Athens, Greece, 24–28 February 2019; p. 1.