Proceedings of the 2018 International Conference on Machine Learning and Cybernetics, Chengdu, China, 15-18 July, 2018

PREDICTING FUTURE VISITORS OF RESTAURANTS USING BIG DATA

XU MA[1], YANSHAN TIAN[1], CHU LUO[2], YUEHUI ZHANG[3]

[1]School of Mathematics and Computer Science, Ningxia Normal University, Guyuan, China
[2]School of Computing and Information Systems, University of Melbourne, Australia
[3]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China
E-MAIL: maxu20046@hotmail.com, albberk@163.com, chu.luo@unimelb.edu.au, zyh@sjtu.edu.cn

Abstract:

For efficient and economical operation, restaurant owners need to accurately estimate the number of future customers. In this paper, we propose an approach to predict how many future visitors will go to a restaurant using big data and supervised learning. The included big data involves restaurant information, historical visits and historical reservations. With features constructed from the big data, our approach generates predictions by performing regression using a mix of K-Nearest-Neighbour, Random Forests and XGBoost. We evaluate our approach using large-scale real-world datasets from two restaurant booking websites. The evaluation results show the effectiveness of our approach, as well as useful insights for future work.

Keywords:

Machine learning; Big data; Business intelligence; Random forests; XGBoost

1. Introduction

In service industry, an increasing number of business owners improve the quality of service and reduce cost using big data techniques. Based on historical data of business, predictive models of machine learning can estimate future moves of customers [1]. Especially, forecasting the number of future visitors can help restaurant owners [2] make the best operations to maximise the revenue. With an accurate visitor forecasting model, restaurant owners can prepare suitable amount of ingredients that exactly satisfy future visitors. In addition, restaurant owners can also schedule a suitable number of staff that can serve these visitors.

Despite various visitor forecasting techniques for other purposes such as national tourism and hotel demand (e.g., [3], [4] [5] [6] [7]) in the literature, little is known for restaurant owners to estimate the number of future visitors using big data. Previous work such as [6] focused only on visitors' revisit. Note that new customers may outnumber old customers in some restaurants of tourist destinations. Hence, it is necessary to develop a new method that can predict the total number of future visitors to a restaurant on a specific day.

To bridge this gap, in this paper we propose an approach to forecast how many future visitors will go to a restaurant using big data and supervised learning. Our approach collects big data involving restaurant information, historical visits and historical reservations. Regular restaurants can easily collect such big data on their own without any complex computing infrastructure (e.g., third-party cloud computing services). From such big data together with time information, we constructed four groups of features correspondingly. With these features, our approach generates predictions by performing regression using a mix of K-Nearest-Neighbour, Random Forests and XGBoost. Compared with techniques such as deep learning, these algorithms have relatively low computational cost so that restaurants can deploy them on common PCs. Rather than using only one regressor, using a mix of different regressors can reduce the inductive bias from one algorithm. We evaluate our approach using large-scale real-world datasets from two restaurant booking websites. The evaluation results show the effectiveness of our approach. To understand the usefulness of different factors for the prediction, we quantified the importance of each feature using a function of XGBoost. We found that time-related features (such as week of year) and historical visitor records (such as mean visitors on a day) are the strongest indicators of the future visitor number to a restaurant. From the results, we also offer several useful insights for future work.

In summary, our paper makes the following contributions:

- We propose a novel approach to predict how many future visitors will go to a restaurant using big data and supervised learning. This approach relies on low-computation

978-1-5386-5214-5/18/$31.00 ©2018 IEEE

algorithms so that restaurants can easily deploy them on common PCs.

- We show the effectiveness of our approach by an evaluation using large-scale real-world datasets from two restaurant booking websites.
- We quantify the usefulness of different factors for the prediction. We found that time-related features and historical visitor records are the strongest indicators.

2. Related work

2.1 Predicting future visitors

Forecasting the number of future visitors is a meaningful task in service industry. Researchers have proposed various techniques to predicting the volume of future visitors using big data in different areas.

For instance, search engine data can be used to forecast the number of future tourists to a certain popular destination in China [3]. A strong correlation between the search queries and visitor numbers was found. Similarly, web traffic data in an area can be used to predict local hotel demand [4]. With the popularity of devices connected to the Internet, this prediction relies on the fact that more travellers generate more traffic data.

However, obtaining web search data and web traffic data is not a trivial task. Especially, many tourist destinations do not have the network connection, e.g., high mountains and small islands. Hence, many techniques predict the number of future visitors using historical visitor numbers only, e.g., predicting visitors to a city using past tourist traffic [5]. Also, the characteristics of a location can be used for the prediction, e.g., predicting customers' revisit to a restaurant using the attributes of the restaurant [6]. Moreover, the geographical influence is important factor in predicting future visitors going to Points-of-Interests (POIs) [7].

Although various techniques are proposed for many areas to predicting the number of future visitors in service industry, little is known for restaurant owners to forecast the number of future visitors using big data. [6] focused only on visitors' revisit. Note that new customers may outnumber old customers in certain restaurants of tourist destinations.

2.2 Machine learning algorithms for regression

Based on big data, many machine learning algorithms for regression can find correlations between factors and consequences to predict future numbers.

Support Vector Machine (SVM) [8] is a widely used method for regression (also called SVR). It can also work for classification. It builds a hyperplane to represent the patterns in the training data. Users can change the kernel functions of SVM to complete learning tasks on non-linear distributions of training data. However, the quality of features heavily affect the performance of SVM. If training data has irrelevant features, SVM may have very low accuracy.

Random Forests (RF) [9] is a decision-tree-based method, also for both regression and classification. As decision trees are non-linear by nature, RF can work for either linear or non-linear data without any prior knowledge about linearity. Relying on a number of decision trees, it uses the mean prediction to generate the final prediction. Since each tree contains a subset of all features, it is robust against irrelevant features.

Deep Neural Networks [10] is also a powerful method. Since it can use multiple layers of networks from input to output, it can model highly non-linear relations in training data. However, it requires huge amount of computation, significantly increasing the hardware cost and computation time for users to complete a machine learning task.

Recently, a new decision-tree-based method is proposed, called XGBoost [11]. It is based on the idea of Gradient Boosting. XGBoost aims to provide high performance without long time of computation. It had the best performance in many real-world datasets, outperforming all other algorithms.

3. Methodology

3.1 Problem statement

Let R be the set of all restaurants and r_i be the number i restaurant. Let $v_i(t)$ be the number of visitors to the number i restaurant at moment t. Given a future time t and a restaurant r_i, our task is to construct a function $f(D, i, t) = v$ to predict the number of visitors v using big data D about R (Note that the big data involves not only r_i but also any other restaurants and their visitor number history).

3.2 Performance measure

According to the convention, performance of predictions is measured on the root mean squared logarithmic error (RMSLE), i.e.: (1).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln(p_i+1)-\ln(a_i+1))^2} \tag{1}$$

where n is the number of predictions in the test set; p_i the number i prediction value; a_i the number i actual value of visitors.

3.3 Raw data and features

In our approach, the raw data (i.e., big data D) involves four data sources: time t, restaurants' attributes, restaurant visitor history, and reservation history from restaurant booking websites. With these data sources, we constructed four groups of features correspondingly.

The first group of features are related to time. Using time information, we constructed the following features: year, month, week of year, day of week, and whether the data is a holiday.

The second group of features are from restaurant attributes. To compare different restaurants, we constructed several features: their unique ID, latitude, longitude, genre, and location area. Since some features are categorical, we use one-hot encoding for preprocessing, so that distance-based algorithms can process them. Hence, the training data has a significantly large number of columns.

The third group of features are from restaurant visitor history. We constructed several features: mean, median, minimum, maximum of visitors, and the total number of visitors before a day (note that we count the repeated visits of a visitor).

The last group of features are from reservation history. We constructed several features: mean reserved seats, the total number of reserved seats before a day, mean time of the reservation before expected visit, and whether the reservation is within 48 hours.

3.4 Regression

Considering that restaurant owners may not have access to high-performance computers for their large-scale datasets, we selected three efficient algorithms: K-Nearest-Neighbour, Random Forests and XGBoost. Although SVM and DNN are also popular regressors, they require significant memory space and computation time for large-scale datasets, increasing the cost of restaurants.

To avoid the bias brought from a single regressor, we employed a weighted regressor which uses the mean of all predictions for one input instance from the three regressors as the final prediction, i.e.: (2).

$$f(D,i,t) = \sum_{j=1}^{3} \frac{1}{3} f_j(D,i,t) \tag{2}$$

K-Nearest-Neighbour regressor has very low computational cost. It has a hyperparameter K determining how many neighbours the input refers to. As stated in [12], we set $K = 10$ so that the regressor can have high robustness against noise without heavy computation.

Similarly, to achieve high robustness with relatively low computation, we chose 500 to be the number of trees for Random Forests and XGBoost. To achieve higher robustness, for XGBoost, we also set the learning rate of XGBoost as 0.2; the subsample ratio of training data as 0.8; and the subsample ratio of columns used for constructing each tree as 0.8. Rather than using 1 for the subsample ratios, using 0.8 can effectively avoid overfitting of the model on a small amount of noise in the training data.

4. Evaluation

4.1 Implementation

We implemented our approach using Python 3.6 with scikit-learn [13] and XGBoost package. The computer to run the program is a regular HP laptop with Intel Core i5-5300 CPU. The computer has 8GB physical memory. The program does not require GPU for computation.

4.2 Dataset

We selected two open large-scale real-world datasets from two restaurant booking websites: HOTPEPPER GOURMET [14] and AirREGI [15]. The two datasets contain the visitor and reservation history of 150 restaurants from year 2016 to the first season of 2017. We combined the two datasets to form a training dataset for our evaluation (2350992 instances in total). The test set, containing 2019 instances, records how many visitors go to these restaurants in the two following months in year 2017.

4.3 Regression performance

Figure 1 shows the training error of each single regressor in RMSLE. RF obtained a significantly lower error (0.194), compared to KNNR and XGBoost (both above 0.470).

Figure 2 shows the test error of each single regressor and our approach in RMSLE. Although RF's training error is the lowest, its test error is the highest (0.522). Combining the 3 regressors, our approach achieved the lowest test error (0.490) which is better than that of any individual regressor. XGBoost is the best individual regressor, obtaining a test error of 0.502.

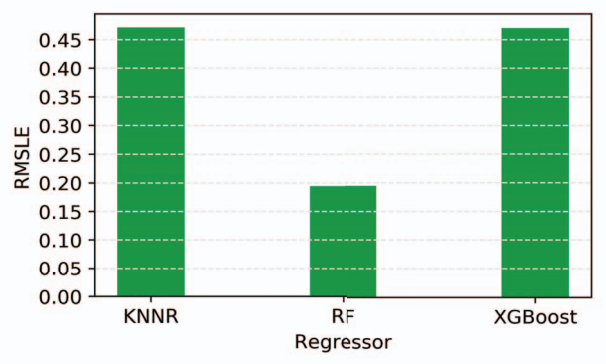

FIGURE 1. Training error of each single regressor (lower is not necessarily better)

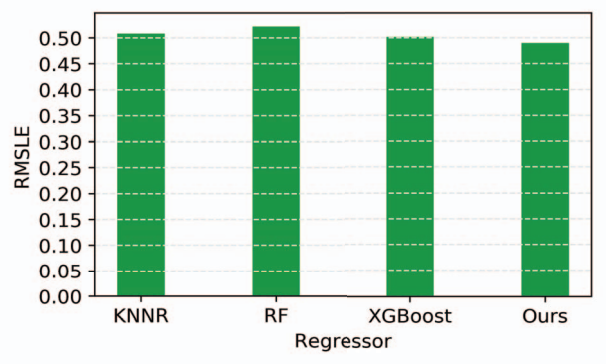

FIGURE 2. Test error of each single regressor and our approach (lower is better)

4.4 Feature importance

Beyond the regression performance, we investigated what factors are the strongest indicators of future visitor numbers to restaurant. We measured the importance of each feature using a function of XGBoost. This function automatically evaluates each feature by computing its usage of all the trees during training. The details of this function are illustrated in [16].

Figure 3 lists the features with 10 highest importance. The most important feature is "week of year" (0.158), followed by mean visitors (0.079), restaurant unique ID (0.076) and so on.

4.5 Computation time

On our computer, we quantified the computation time of each phase in the analysis with 2350992 training instances and 2019 test instances.

Figure 4 shows the computation time of each phase in training. Preprocessing required only 20s. In the process of training the three regressors, our computer can complete KNNR (326.57s) and XGBoost (334.64s) in 6 minutes, however, RF (1527.94s) in about 25 minutes. Hence, our approach used 2209s to build a predictive model based on the training data.

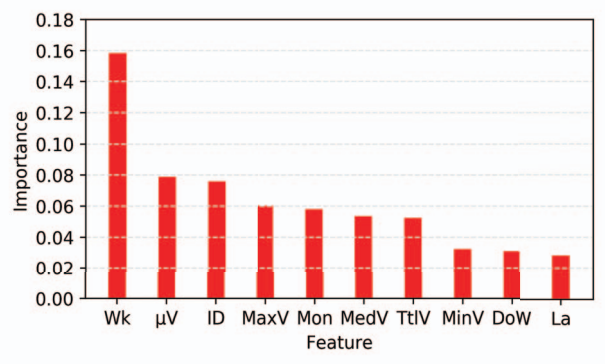

FIGURE 3. 10 most important features and their importance measured by XGBoost (higher is more important). Wk = week of year; μV = mean visitors, ID = restaurant unique ID, MaxV = maximal visitors, Mon = month of year, MedV = median visitors, TtlV = total visitors, MinV = minimal visitors, DoW = day of week, La = latitude

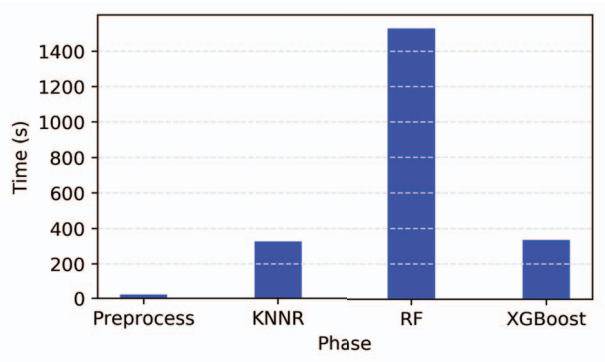

FIGURE 4. Computation time of each phase for training

Figure 5 shows the computation time of each phase in test. Compared to training, all regressors worked significantly faster in test: KNNR (31s) , RF (133.03s) and XGBoost (9.73s). This means that, based on the training model, our approach required 173.76s to generate predictions for all 2019 test instances.

5. Discussion

5.1 Implications

We show that it is feasible to predict how many future visitors will go to a restaurant using big data and supervised learn-

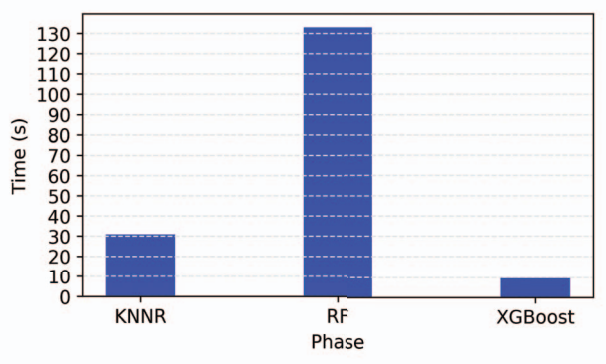

FIGURE 5. Computation time of each phase for test with a training model

ing. To drive our approach, users only need to collect big data, including restaurant information, historical visits and historical reservations.

Considering that restaurant owners may not have access to high-performance computers, we show that the prediction can be achieved by a mix of low-computation regressors: K-Nearest-Neighbour, Random Forests and XGBoost, rather than relying on computation-intensive methods such as Deep Learning (requiring long time of training and mostly requiring high-performance GPU) and SVM (requiring huge amount of memory).

In evaluation, we found that a low training error cannot ensure a low test error for a regressor. Hence, users should not minimise training error by all means during training. A relatively high training error often reflects the robustness of a model.

Regarding the strongest indicators of future visitors, we found that time-related features (such as week of year, month of year and day of week) and historical visitor records (such as mean, median, maximal and minimal visitors on a day) are most useful for the prediction. In addition, the unique ID and location of restaurant are also important features. This indicates that each individual restaurant follows a unique pattern, even if other factors are similar. By contrast, historical reservations can hardly help the prediction.

We quantified the computation time on our computer in evaluation. We show that the training time and the time for a model to generate a prediction are acceptable for regular restaurant owners with limited computational resources. Hence, with simple deployment, our approach is suitable for most usage scenarios.

5.2 Limitation and future work

Beyond features considered by our approach, many more factors can facilitate accurate prediction of future visitors to a restaurant. For instance, bad weather of the restaurant location may reduce visitors. Also, if a new restaurant is opened next to an existing restaurant, the number of future visitors going to this existing restaurant may drop. Moreover, social events can bring more visitors to restaurants of the related venues.

Hence, in future work, it is necessary to include more information to the predictive model, such as weather, competitors and social events.

For some kinds of restaurants, it is challenging to accurately count how many visitors have come on a day. For example, one customer may buy food for all its friends at fast food restaurants, such as McDonald's (especially when customers in a car buy food via "Drive-Thru"). Future work can also explore new customer counting methods for restaurants based on new technology (such as combining video camera and the Internet of Things [17], or new software/hardware sensing infrastructure [18, 19, 20]).

6. Conclusion

In this paper, we present an approach to estimate how many future visitors will go to a restaurant using big data and supervised learning. The included big data involves restaurant information, historical visits and historical reservations. With features constructed from the big data, our approach generates predictions by performing regression using a mix of K-Nearest-Neighbour, Random Forests and XGBoost. We evaluate our approach using large-scale real-world datasets from two restaurant booking websites. The evaluation results show the effectiveness of our approach, as well as useful insights for future work.

Acknowledgements

This work is supported by the NSFC grant 11671258 and 11771280, the National Science Foundation of Shanghai Municipal (17ZR1415400), the National Natural Science Foundation of China under grant no. 11361046, Ningxia High Education research fund with grant no. NGY2017180, Undergraduate teaching project of Ningxia High Education with grant no.NXJG2016060, and the fund of Ningxia High Education Construction of First-Class Disciplines (Education) with grant No.NXYLXK2017B11.

References

[1] Amir Gandomi, and Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, Vol 35, no. 2, pp. 137-144, 2015.

[2] Daryl Ansel, and Chris Dyer, "A framework for restaurant information technology", Cornell Hotel and Restaurant Administration Quarterly, Vol 40, no. 3, pp. 74-84, 1999.

[3] Xin Yang, Bing Pan, James A. Evans, and Benfu Lv, "Forecasting Chinese tourist volume with search engine data", Tourism Management, Vol 46, pp. 386-397, 2015.

[4] Yang Yang, Bing Pan, and Haiyan Song, "Predicting hotel demand using destination marketing organizations web traffic data", Journal of Travel Research, Vol 53, no. 4, pp. 433-447, 2014.

[5] Vincent Cho, "A comparison of three different approaches to tourist arrival forecasting", Tourism management, Vol 24, no. 3, pp. 323-330, 2003.

[6] Usep Suhud, and Arifin Wibowo, "Predicting Customers Intention to Revisit A Vintage-Concept Restaurant", Journal of Consumer Sciences, Vol 1, no. 2 (2016).

[7] Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee, "POI2Vec: Geographical Latent Representation for Predicting Future Visitors", In AAAI, pp. 102-108, 2017.

[8] Corinna Cortes, and Vladimir Vapnik, "Support-vector networks", Machine learning, Vol 20, no. 3, pp. 273-297, 1995.

[9] Leo Breiman, "Random forests", Machine learning, Vol 45, no. 1, pp. 5-32, 2001.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning", nature, Vol 521, no. 7553, pp. 436, 2015.

[11] Tianqi Chen, and Carlos Guestrin, "Xgboost: A scalable tree boosting system", In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. ACM, 2016.

[12] Cheng Zhang, Juyuan Yang, Yi Zhang, Jing Fan, Xin Zhang, Jianjun Zhao, and Peizhao Ou, "Automatic parameter recommendation for practical API usage", In Proceedings of the 34th International Conference on Software Engineering, pp. 826-836. IEEE Press, 2012.

[13] scikit-learn. http://scikit-learn.org/stable/index.html

[14] HOTPEPPER GOURMET. https://www.hotpepper-gourmet.com/en/

[15] AirREGI. https://air-regi.com/

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. New York: Springer series in statistics, 2001.

[17] Chu Luo, "Video Summarization for Object Tracking in the Internet of Things", Next Generation Mobile Apps, Services and Technologies (NGMAST), 8th International Conference on. IEEE, pp. 288-293, September 2014.

[18] Huber Flores, Denzil Ferreira, Chu Luo, Vassilis Kostakos, Pan Hui, Rajesh Sharma, Sasu Tarkoma and Yong Li, "Social-aware Device-to-Device Communication: A Contribution for Edge and Fog Computing?", Adjunct Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp16), pp. 1466-1471, September 2016.

[19] Chu Luo, Angelos Fylakis, Juha Partala, Simon Klakegg, Jorge Goncalves, Kaitai Liang, Tapio Sepp?nen, and Vassilis Kostakos, "A data hiding approach for sensitive smartphone data", In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 557-568. ACM, September 2016.

[20] Chu Luo, Henri Koski, Mikko Korhonen, Jorge Goncalves, Theodoros Anagnostopoulos, Shin'Ichi Konomi, Simon Klakegg, and Vassilis Kostakos, "Rapid clock synchronisation for ubiquitous sensing services involving multiple smartphones", In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, pp. 476-481. ACM, 2017.