Big Data Applications in Guangzhou Restaurants Analysis
Vicent Chang,1,*, Ziyang Ji,2 and Qianwen Ariel Xu1,2

Abstract
With the development of modern information and communication technologies, such as the internet of things and big data analytics, businesses and users have become more adaptable to rapid changes. Both consumers and merchants have obtained great convenience. Meanwhile, a huge amount of data is generated. However, many businesses lack the ability to process these data, which contain critical business values. Therefore, this article uses data from the Dianping website to show how to use big data analytics techniques to exploit the valuable information from these raw data. First, descriptive analysis is conducted by using kernel density estimation. Then, multilinear regression analysis, Naive Bayes, and J48 are used to predict the level of restaurants. We found that flavor, environment, and service score are essential factors to the restaurant level. Moreover, J48 performs best among the three models with an accuracy of 88.89%.

Keywords: internet of things; big data analytics; catering industry; machine learning

Introduction
In the internet economy era, many industries have been affected, and the catering industry is no exception. The internet economy has brought tremendous impact to the catering industry, which innovate operation mode and broaden the source of customers.1 With the development of big data applications, merchants can rely on data analysis to target better business strategies. Similarly, customers can also find restaurants that meet their tastes. Dianping website or its application is the first third-party review website in China, where merchants can provide consumers with various intelligent services and users can fully interact and share the experience of the restaurant.

By integrating the emerging information technologies, such as the internet of things (IoT), this kind of application brings both merchants and consumers great convenience. For example, when consumers plan to go to the nearest Chinese restaurant for dinner, through Dianping and GPS, they can quickly find a qualified Chinese restaurant and learn about the queuing situation of the restaurant. If the customers need to queue, they can queue online, and the queue information will be automatically uploaded to the restaurant’s terminal. Whenever a customer enters the restaurant, the waiter only needs to confirm by clicking on the merchant terminal, and the updated queue status will be sent to other customers in the queue. Therefore, customers do not need to queue up at the scene but just reserve time to go to the restaurant, saving time and improving efficiency.

If the customer does not know the route to the restaurant, Dianping will use GPS to navigate the customer. After arriving at the restaurant, customers can also scan the QR code on the table through the scan function of the application to order, pay, and write reviews.2 There is plenty of restaurant information submitted by merchants and consumers. Thus, the website offers an effective way to conclude the characteristics of the spatial distribution of food consumption in a particular region by collecting and analyzing the data from the Dianping website.

With the development of the internet, most consumers have begun to post comments on various online information, which rapidly disseminates the positive or negative report to the consumer groups. The word-of-

1Artificial Intelligence and Information Systems Research Group, School of Computing and Digital Technologies, Teesside University, Middlesbrough, United Kingdom.
2IBSS, Xi’an Jiaotong-Liverpool University, Suzhou, China.
iORCID ID (https://orcid.org/0000-0002-8012-5852).

*Address correspondence to: Victor Chang, Artificial Intelligence and Information Systems Research Group, School of computing and Digital Technologies, Teesside University, G1.20, Greig Building, Middlesbrough TS1 3BX, United Kingdom, E-mail: victorchang.research111@gmail.com
mouth effect of online commentary is becoming more significant, which has a substantial impact on consumers’ decision-making behavior. However, many businesses lack the ability to process these data, which contain critical business values.

This article uses data from the Dianping website as an example to show how merchants and consumers can use big data analytics techniques to exploit the valuable information from these raw data. Dianping website is a reliable platform for consumers to post comments and rate restaurants based on their experience, which boosts the impacts of word of mouth. This article uses a data set provided by Si, including consumer reviews for 3124 restaurants in Guangzhou. This article aims to explore the valuable information and compare the accuracy of the restaurant rating under three algorithms.

The rest of our article is organized as follows. The Literature Review section introduces the literature of online word of mouth and the applications of data analytics in the catering industry, as well as the potential issues. The Research Methodology section describes the data and methodology employed in our research. In the Descriptive Analysis section, the descriptive analysis is conducted to provide a brief understanding of the catering industry of Guangzhou. Then, the level prediction is performed by using the multiple stepwise regression method, Naïve Bayes, and J48 machine learning methods in the Level Prediction section. The results could help both merchants improve their operation methods and consumers to explore a satisfactory restaurant to experience. The Conclusion section concludes this article and provides suggestions for merchants.

**Literature Review**

**Online word of mouth**

Online word of mouth attracts numerous researchers. Li believed that consumers use the internet platform to disseminate positive or negative comments about products or services. Zhang and Dong found that internet word of mouth is the evaluation of products and services published by consumers online. Yan et al. pointed out that internet word of mouth is a new form of word-of-mouth communication, exchanging information between consumers on the products or services through online media. Kon believed that internet word of mouth is an informal noncommercial communication, but this kind of post-purchase experience or knowledge sharing takes place on the network platform.

In recent years, when scholars study the relationship between online word of mouth and consumers’ willingness to purchase, the quality of online word of mouth is one of the crucial variables of research. Zuo et al. introduced the social capital theory in their study, constructed a model of online word of mouth and consumer purchase intention from the perspective of social capital, and allowed using the apps such as Mushroom Street and Beauty to investigate the object. The quality of online word of mouth will positively influence consumers’ willingness to purchase.

Besides, Song et al. mainly used film as a background to explore the influence of various factors on consumer attitudes, but the film has its particularity. This conclusion may not be universally applicable in other product categories. Research should focus on the type of product. In their study, Schindler and Bickart divided consumers into information about goods from online forums and product information from corporate home pages. The results showed that consumers’ interest in assets mainly depends on obtaining product information from online forums. In other words, it indicates that online forums have a more significant impact on consumer trust.

The number of internet word of mouth is usually an indispensable variable in the study of internet word-of-mouth communication. Chen et al. used the number of internet word of mouth as one of the variables influencing consumers’ willingness to purchase in the process of internet word-of-mouth communication. Constructing a theoretical model verified that the number of internet word of mouth in the social network has a significant impact. Shan and Zhao studied the interaction mechanism between consumers’ contradictory attitudes and internet word of mouth. Through two-stage experimental research, it was found that consumers’ conflicting opinions steadily regulate consumers’ handling of external information.

**Data analytics and its implications in the catering industry**

With the rapid development of the internet and mobile applications, our life becomes smarter and smarter. The integration of big data and IoT technologies has been applied in many areas, including smart environment applications for monitoring, protecting, and improving natural resources and smart tourism for monitoring tourist flow. The catering industry is no exception. Researchers can have instant access to online catering data, for example, locations, reviews,
evaluation scores, and transaction data. The characteristics of catering data, such as the spatial distribution of restaurants and commercial activities, are able to provide consumers with a large number of references for making decisions and provide necessary guidance for catering industry design. Therefore, the application of big data analytics in this field is essential.

Previous literature include statistical response patterns in restaurant data sets, sales revenue prediction, and location analysis of restaurants. In the study by Laksono et al., they conducted a sentiment analysis to analyze customer satisfaction based on the restaurant customer reviews. They compared the performance of the Naive Bayes method and TextBlob and found that Naive Bayes had higher prediction accuracy. Additionally, restaurant data can also help analyze or predict socioeconomic characteristics. By using restaurant data from Facebook, Lin et al. showed how to identify promising business locations.

There are also a number of issues about the application of data analytics. First, Mariani stated that while the techniques of generating, analyzing, reporting, and visualizing data are becoming mature, the interpretation of data still needs improvement. Second, the focus of existing research is on running big data analytics and ignoring the difficulty of facilitating fog data analytics. To solve challenges such as accessibility, scalability, fog node communication, and so on, Kumari et al. discussed the unique nature and complexity of FDA. Moreover, they tried to deal with the latency issue during the data analytics in smart grid by investigating the smart grid architecture in the context of fog computing. Finally, ethical, privacy, and security issues on emerging technologies have received great attention.

**Research Methodology**

**Data collection and management**

This article uses data from the Dianping website as it is a reliable platform and the first third-party review website in China, where merchants can provide consumers with various intelligent services and users can fully interact and share the experience of the restaurant.

There is plenty of restaurant information submitted by merchants and consumers. Thus, the website offers an effective way to conclude the characteristics of the spatial distribution of food consumption in a particular region by collecting and analyzing the data from the Dianping website.

The data set contains 3124 restaurant data, which are classified as business items. It includes 722 Cantonese cuisine restaurants, 572 porridge noodles restaurants, 566 Sichuan cuisine restaurants, 595 Japanese cuisine restaurants, and 669 western food restaurants.

To conduct descriptive analysis and multiple stepwise regression method, the data are organized into 15 attributes: Level, ReviewNum (RN), FlavorScore (FS), EnvironmentScore (ES), ServiceScore (SS), StarReviewNum (SRN), HighqualityMerchant (HM), PopularArea (PA), PictureNum (PN), ParkingInfo (PI), GroupPurchase (GP), Promotion (Promotion), TakeOut (TO), PerConsumption (PC), and BusinessDay (BD). In addition, the data set is further divided into a training group and a test group to conduct the level prediction using machine learning algorithms. The training data include 2944 objects with 21 attributes, whereas the testing data have 179 objects with the same qualities.

**Research methods**

First, the descriptive analysis part consists of the analysis of the number of restaurants in different regions, high score restaurants distribution, the popularity of each cuisine, distribution of score and price, and the history of these restaurants. Kernel density estimation is calculated by ArcGIS to present a more accurate result of the delivery. Additionally, a multiple stepwise regression method is needed to determine the impact of different variables on the level of restaurants. Correlation analysis has been taken into this part. Fifteen variables in the data set are taken into consideration. “Level” is recognized as the dependent variable. It measures the level of restaurants, and the higher level indicates the restaurant is more satisfactory. The independent variables are RN, FS, ES, SS, SRN, HM, PA, PN, PI, GP, Promotion, TO, PC, and BD. Their meanings are discussed in detail in the Level Prediction section.

After the stepwise regression, the level prediction is conducted based on machine learning algorithms. Naive Bayes and J48 are two classical machine learning methods that could classify the data set in a comparably accurate way. Naive Bayes is one of the supervised learning models that is simple to implement, has no iteration, and has a solid mathematical theory (i.e., Bayes’ theorem) as support. J48 is a decision tree algorithm based on the C4.5 algorithm. In this supervising learning method, each of these tuples can be described by a set of attribute values. The level of a restaurant will be predicted under those three algorithms, and their prediction accuracy will be evaluated.
Descriptive Analysis

Number of restaurants in each region

From the bar chart (Fig. 1), we could observe that most restaurants are in Tianhe District, up to 901, and Yuexiu District is in the second place with 653 restaurants. The analysis is in line with the actual situation. The above two districts are the two most economically developed regions in Guangzhou.

To observe the distribution of these samples, we will use the histogram method to visually display them. This method is simple and easy to calculate, but it needs us to determine the exact interval when plotting the histogram. If the ranges are different, the final result will be very different. Additionally, the histogram shows a distribution curve that is not smooth. We can increase the number of intervals so that each sample has a probability of its own. There is also a disadvantage to this; the likelihood that the value does not appear in the example is 0, so the probability density function is not continuous.

To solve this problem, we take a nuclear density estimation method. Kernel density estimation is a density function used to estimate unknowns in probability theory. It is one of the nonparametric test methods. It was proposed by Rosenblatt and Parzen, also known as Parzen window. It uses a kernel to fit the observed data points to simulate an actual probability distribution curve. We can get a more accurate data distribution through the kernel density estimation function in ArcGIS.

Related information is also displayed in the heat map (Fig. 2). It helps us to understand the restaurant distribution more intuitively.

Among the food consumption spaces in the downtown area of Guangzhou, the areas with the highest kernel density values are the Tianhe Sports Center area in Tianhe District, the Beijing Road Cultural Tourism Area in Yuexiu District, and the vicinity of Zhongshan University in Haizhu District. Among them, the Tianhe Sports Center has the maximum value of kernel density estimation and is the core consumption center. The area involves several business districts, with large shopping malls in the south. These include the following:

1. Tianhe City Department Store, Zhengjia Plaza, and so on;
2. West Victoria Plaza, Tianhe Building, and Baixi Shangdu in the west; Fortune Plaza, Economic and Trade Building in the east, and numerous commercial buildings;
3. Neighboring is Guangzhou’s highest point—CITIC Plaza, Mayor Building, and Guangdong East Railway Station; and
4. Beijing Road Cultural Tourism Zone is the starting point of Guangzhou City. It has been the political,
FIG. 2. Heat map of restaurant distribution.
cultural, and commercial center of Guangzhou for more than 2000 years. It has several characteristic commercial streets, such as Beijing Road Pedestrian Street, Laozihao Street, Huifu Food Street, and Wende Road Cultural Street. The area near Sun Yat-sen University in Haizhu District mainly involves the Jiangnan West Business Circle, including several commercial districts such as Guangbaixin City, Runhui Building, Mingdian City, and Jiangnan Xindi. There are a large number of teachers and students here, providing the power of consumption. We could then filter the restaurants with low-level values, with the threshold of 4.

Opening date
A majority of restaurants start operating in 2015, and those account for more than half of the whole share. The most significant number of restaurants opened in 2016 was 689. As shown in Figure 3, only 16 restaurants have been in business for more than 10 years, indicating that the catering industry owns high liquidity.

Data analysis related to food
The data set contains only four different cuisine styles, and the number of restaurants from diverse styles is almost the same. Due to the influence of regional culture, Cantonese cuisine accounts for the largest share, as shown in Figure 4.

The statistics in Figure 5 illustrate three kinds of scores, including flavor score, service score, and environment score. All of them have the same upper score of 9.3, whereas there are differences among the bottom values. The median value of them is diverse as well. The distribution of the flavor scores shows that consumers generally think that these restaurants have tasty food.

According to Figure 6, bar chart illustrates that restaurants providing food with a price below 100 Yuan...
FIG. 5. Distribution of three scores.

FIG. 6. Bar of the price distribution.
FIG. 7. Dot plots of price versus taste.
account for the vast majority of all restaurants. In contrast, only 36 restaurants ask consumers to spend more than 300 CNY for dinner.

As shown in Figure 7, we can divide these restaurants into different categories according to two dimensions of price and taste. Surprisingly, there are a large number of restaurants that have delicious food at a low price.

**Level Prediction**  
Level prediction based on multiple linear regression

This part is considered to illustrate the impact of selected factors on the level of restaurants, and 15 variables in the data set are taken into consideration. The list of variables is shown in Table 1. "Level" is recognized as the dependent variable. It measures the level of restaurants, and the higher level indicates that the restaurant is more satisfactory. The correlation between the variables is presented in Figure 8.

For achieving the purpose, following the model has been displayed.

\[
Y_{\text{level}} = \beta_1 \text{RN} + \beta_2 \text{FS} + \beta_3 \text{ES} + \beta_4 \text{SS} + \beta_5 \text{SRN} + \beta_6 \text{HM} \\
+ \beta_7 \text{PA} + \beta_8 \text{PN} + \beta_9 \text{Park} + \beta_{10} \text{GP} \\
+ \beta_{11} \text{Promotion} + \beta_{12} \text{AR} + \beta_{13} \text{TO} + \beta_{14} \text{PerC} \\
+ \beta_{15} \text{BD} + \varepsilon
\]

(Model 1)

In Model (1), \( \beta_i \) represents the multilinear regression coefficient between the \( Y_{\text{level}} \) and RN, FS, ES, SS, SRN, HM, PA, PN, Park, GP, Promotion, AR, TO, PerC, and BD variables, respectively. There are also assumptions about the error term (\( \varepsilon \)). The errors are assumed to be mean zero and would be normally distributed random variables.

The model used multiple stepwise regression methods for analysis. The stepwise regression method is simple to implement, and the obtained regression equation has fewer variables and retains the most significant variables. In practice, this method has been proven to be more productive and the prediction accuracy is higher. Variables may have multiple collinearities, but stepwise regression can correct multicollinearity to some extent.

The results of the stepwise regression method are shown in Table 2. From the above analysis results, it can be observed that a total of 3124 samples participated in the analysis. The \( F \)-value (7, 3116) of the model is 2444.44 and the \( p \)-value is 0.0000, which indicate that the model is very significant overall. The R-squared of the model is 0.8459 and the adjusted R-squared is 0.8456, which suggest that the model has excellent explanatory power (Table 3).

The model gets the final result after eight rejections of the variable (Equation below). In the final regression model, we get seven valid variables (Fig. 3), all of which have \( p \)-values < 0.05, and each coefficient is significant.

\[
Y_{\text{level}} = 0.000025 \text{RN} + 0.457396 \text{FS} + 0.047503 \text{ES} \\
+ 0.187566 \text{SS} + (-0.0000403) \text{SRN} + 0.016421 \text{HM} \\
+ 0.000027 \text{BD} + (-1.57768)
\]

From the multilinear regression model, we could conclude that the environment score is the most critical variable. One unit increase in environment score will result in an \( \sim 0.46 \) increase at the restaurant level. It

### Table 1. List of variables

| RN       | How many reviews the restaurant received, and the comments are submitted by consumers voluntarily. The types of study could be divided into three methods, and consumers could select “good,” “bad,” or “so so” according to their feelings about the food when they upload their reviews. |
| FS       | A significant factor, which represents how consumers consider the food of the restaurant, and this score results from the sum of every users’ comment. |
| ES       | How consumers feel about the environment of the restaurant, and the higher score means a more satisfactory dining environment. |
| SS       | The impression of service, and it may result from the speed of serving, the courtesy of the waiter, etc. |
| SRN      | Counts how many comments are submitted by 5-star consumers. These users are considered as a higher class and their opinions are more valuable. |
| HM       | Whether the restaurant earns a reputation from the official management department of the Dianping website. |
| PA       | A dummy that if the restaurant is located in the business district. It may have an impact on the possibility of attracting consumers. |
| PN       | How many pictures posted on the website, whether uploaded by merchants or consumers. |
| PI       | How many parking spaces are owned by the restaurant? It is more convenient for users if there are a large number of parking spaces. |
| GP       | A method that consumers could buy a coupon of the restaurant online, which effectively attracts individuals. |
| Promotion| A method that consumers could buy a coupon of the restaurant online, which attracts individuals in productive ways. |
| TO       | A dummy variable that indicates if the restaurant provides a takeout or takeaway service. |
| PC       | A significant variable that the average consumption of the food. |
| BD       | How long the restaurant has been opening. |
needs to be noticed that the level range is from 0 to 5, which means that a 0.46 increase is a significant upgrade. Moreover, the favor score is another critical variable, so the restaurant needs to pay more attention to the taste. The R square of the model is 0.8456.

Level prediction based on machine learning
Classification is one of the commonly used methods in data mining. Whether it is a practical application

**Table 2. Result of the stepwise regression method**

<table>
<thead>
<tr>
<th>No. of obs</th>
<th>F(7, 3116)</th>
<th>Prob &gt; F</th>
<th>R-squared</th>
<th>Adj</th>
<th>R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>3124</td>
<td>2444.44</td>
<td>0</td>
<td>0.8459</td>
<td></td>
<td>0.8456</td>
</tr>
</tbody>
</table>

**Table 3. Significant variables**

<table>
<thead>
<tr>
<th>Level</th>
<th>Coef.</th>
<th>p &gt; T</th>
</tr>
</thead>
<tbody>
<tr>
<td>RN</td>
<td>0.000025</td>
<td>0.001</td>
</tr>
<tr>
<td>FS</td>
<td>0.457396</td>
<td>0</td>
</tr>
<tr>
<td>ES</td>
<td>0.047503</td>
<td>0</td>
</tr>
<tr>
<td>SS</td>
<td>0.018766</td>
<td>0</td>
</tr>
<tr>
<td>SRN</td>
<td>-0.0000403</td>
<td>0.005</td>
</tr>
<tr>
<td>HM</td>
<td>0.016421</td>
<td>0.032</td>
</tr>
<tr>
<td>BD</td>
<td>0.0000274</td>
<td>0</td>
</tr>
<tr>
<td>_cons</td>
<td>-1.57768</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 4. The result of Naive Bayes algorithm**

| Correctly classified instances | 2689 | 91.34% |
| Incorrectly classified instances | 255 | 8.66% |
| Kappa statistic | 0.7212 |
| Total no. of instances | 2944 |
of scientific research, they are inseparable from them. For classification problems, we usually get the data set representing the actual object or event. We know the category of each piece of data in the data set. These categories divide the pieces of data into different classes.

The goal of data processing here is to train the classification model based on existing data, which can then be used to classify new samples.

In this part, two algorithms are considered to classify the data set and then predict the level of restaurants based on the best classification method. The data set is split into two versions, one for data training and another one for data testing. The training data include 2944 objects with 21 attributes, whereas the testing data have 179 objects with the same qualities.

**Naive Bayes.** Naive means condition independence and Bayes means that it is based on Bayes’ theorem. It is one of the supervised learning models that is simple to implement, has no iteration, and has a solid mathematical theory (i.e., Bayes’ theorem) as support. It will perform well under a large number of samples. Weka could be utilized to complete the classification and the result is shown below (Table 4).

From the statistics mentioned above, we could observe that the correctly classified instances are only 2689 units, which is the lowest among the above-three classification models. Kappa compares an observed accuracy with an expected accuracy (random chance) and measures the inter-rater reliability or agreement. According to Cohen, values ≤0.2 indicate poor agreement, 0.21–0.40 mean fair agreement, 0.41–0.60 indicate moderate, 0.61–0.80 mean substantial, and 0.81–1.00 indicate that the agreement is quite good. In this model, the kappa statistic is 0.7212, and it is a not bad result but still lower than the other two models.

**J48.** J48 is a decision tree algorithm based on the C4.5 algorithm. C4.5 is a series of algorithms for classification in machine learning and data mining. It is one of the supervising learning methods: given a data set, each of these tuples can be described by a set of attribute values.

Implement process can be explained as follows:

1. We are constructing a decision tree. Naturally, each layer of the tree represents the value of an attribute, and the last leaf node points to the class.

2. The fundamental problem is how to choose an appropriate node in each layer to construct the tree and make the structure as optimal as possible. In other words, the search path is as short as possible.

3. The most critical problem is how to find the most suitable splitting node in each layer from the remaining nodes that have not yet been assigned.

4. The C4.5 algorithm uses the attribute with the highest information gain rate. The advantage of this is that the tree can be avoided.

5. After the tree is constructed, some pruning operations are required.

The decision tree is displayed in Figure 9 as well. The number of leaves is 32 and the size of the tree is 57. According to Table 5, the number of correctly classified instances is 2876, and the accuracy is 97.6902% that is the highest accuracy among all three algorithms. Kappa statistic is 0.9108, which is a considerably high value, showing that the accuracy of this classification model is almost perfect.

**Level prediction by using the most accurate method**

Compared with the result of the above three models, the J48 algorithm results in the best result, so this model is considered to be taken into the prediction part. Weka could be utilized to complete the classification and the result is shown below.

The prediction based on the training set runs well that we could classify a new sample to the extent of 88.89% accuracy (Table 6). Kappa statistic is 0.7578, showing that it is substantially accurate. For the misclassified samples (Fig. 10), seven samples that should be classified into the normal class are assigned to the perfect class. Then, 13 units should be divided into the perfect class but to the normal class.

### Table 5. The result of the J48 algorithm

<table>
<thead>
<tr>
<th>Correctly classified instances</th>
<th>2876</th>
<th>97.69%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrectly classified instances</td>
<td>68</td>
<td>2.31%</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.9108</td>
<td></td>
</tr>
<tr>
<td>Total no. of instances</td>
<td>2944</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6. The result of the prediction

<table>
<thead>
<tr>
<th>Correctly classified instances</th>
<th>160</th>
<th>88.89%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrectly classified instances</td>
<td>20</td>
<td>11.11%</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.7578</td>
<td></td>
</tr>
<tr>
<td>Total no. of instances</td>
<td>180</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

With the development of modern information and communication technologies, such as the IoT and big data analytics, our cities are becoming more and more friendly and convenient. The IoT allows every physical thing to connect to each other and a large amount of data is then produced. It has been applied in many different areas, for example, the internet of nano-things and the internet of vehicles. Meanwhile, big data analytics techniques can extract valuable information from these raw data and support the decision-making for the businesses. These emerging information technologies not only make our life intelligent but also the business. In this article, the research is focused on the application of IoT and big data analytics in the catering industry.

E-commerce applications with these techniques, such as Dianping, enable the merchants to provide consumers with various intelligence services and also allow consumers to find the merchants satisfied their requirements efficiently. Because of the great convenience obtained from them, the frequency of using such applications has become higher and higher. It has gradually become an indispensable part of our lives, and a large amount of data has been generated. However, many businesses lack the ability to process these data, which contain insightful business values and useful decision-making information. This article uses Dianping’s data as an example to show how businesses and consumers can use big data analytics to exploit and valuable information from these raw data.

In this article, the data set about restaurants opened in Guangzhou City could be analyzed from three aspects.

The first part was a descriptive analysis. As shown in the above statistics, the number of restaurants in the Tianhe District is the largest, and high-level restaurants are in this place. Because of the influence of regional culture, Cantonese cuisine wins the largest share among the whole five cuisine styles. The flavor, environment, and service score of the restaurants are at a very high level. Cost-effective food is typical to find in Guangzhou City, which offers good news for the gourmet. Most of the restaurants have been opened in recent years, and there are only 16 restaurants with a history of more than 10 years.

The second part is to predict the level of restaurants. Multilinear regression analysis, Naive Bayes, and J48 are taken into consideration. Initially, we create a formulation with 15 variables, but several variables are deleted after checking the relationships among those variables and related p-value and t-value. Finally, a formulation $Y_{level} = 0.000025RN + 0.457396FS + 0.047503ES + 0.187566SS + (-0.0000403)SRN + 0.016421HM + 0.0000274BD + (-1.57768)$ is more explanatory for analysis, and this model is significantly meaningful. The level prediction based on Naive Bayes and J48 is the knowledge from machine learning. The performance of J48 is the best among those three models. Then, this model is applied to prediction with accuracy of 88.89% that is sufficiently high.
Using big data analytics, merchants can timely understand industry trends, competitors’ advantages and their own shortcomings, as well as consumer preferences. Therefore, with the support of this valuable information, they can formulate business strategies better, marketing plans and customers can also efficiently identify the merchants that best meet their needs.

Several issues should be investigated for future work. The data set contains limited samples, which requires more data to have a more effective analysis. The multilinear regression formulation results are worse than expected, and 11 variables are not significantly meaningful in the first regression model. Most of them are dummy variables that need more attention to figure out the cause of problems. Additionally, there are numerous classification models, and only three are taken into this article. In our future work, more patterns can be used to improve the accuracy of classification.

Author Disclosure Statement
No competing financial interests exist.

Funding Information
This work is partly supported by VC Research (VCR 0000099).

References
5. Li J. Research on the impact of Internet word of mouth on consumer trust—Taking Taobao as an example. Shandong University. 2009 (last accessed August 14, 2020).
10. Song XB, Cong Z, Dong DH. Research on the influence mechanism of Internet word of mouth on consumer product attitude. Chin J Manage. 2011;8:559–566.


Abbreviations Used
BD = BusinessDay
ES = EnvironmentScore
FDA = fog data analytics
FS = FlavorScore
GP = GroupPurchase
HM = HighqualityMerchant
IoT = internet of things
PA = PopularArea
PC = PerConsumption
PI = ParkingInfo
PN = PictureNum
RN = ReviewNum
SRN = StarReviewNum
SS = ServiceScore
TO = TakeOut