A Machine Learning Approach to Suggest Ideal Geographical Location for New Restaurant Establishment

Ibne Farabi Shihab
Computer Science and Engineering
BRAC Univeristy
Dhaka, Bangladesh
ibnfarabishihab@gmail.com

Maliha Moonwara Oishi
Computer Science and Engineering
BRAC Univeristy
Dhaka, Bangladesh
oishimaliha1994@gmail.com

Samiul Islam
Computer Science and Engineering
BRAC Univeristy
Dhaka, Bangladesh
sami141215@gmail.com

Kalyan Banik
Reasearch and Development
Codemen Solution Inc
Dhaka, Bangladesh
kalyan.b.aninda@gmail.com

Hossain Arif
Computer Science and Engineering
BRAC Univeristy
Dhaka, Bangladesh
hossain.arif@bracu.ac.bd

***Abstract*— Restaurant business is a prospective and profitable business nowadays. However, ensuring quality food, good stuff, inner-environment etc. is a big concern and most importantly before facing all these, the trickiest part is to choose a perfect place where a restaurant business will flourish. Without doing a perfect research on this area, setting up a restaurant may lead to an immediate downfall. In recent time, for choosing a preferred restaurant location and calculating the estimated risk, people are now hiring professionals to do ground check and here the data scientists are coming into play as a bigshot. This research is focused on suggesting a suitable place for setting up a restaurant business based on the existing data from Yelp where 75 features have been extracted for supervised machine learning. Our model will also calculate the expected rating that a restaurant will get depending on the features the restaurant possesses. Several machine learning algorithms (Support Vector Machine, Decision Tree, Logistic Regression and Decision Tree with presort) have been used and juxtaposed to nurture out the suitable one. As yelp's review is authentic and it is maintained regularly, we have considered the rating of a business as the point of suggestion. We have also looked at the comparative analysis of these algorithms and searched for an algorithm that gives us the best result.**

Keywords*— *Support vector machine (SVM), Decision tree, Linear regression

I. Introduction

In every field, expert's suggestions do play a significant role in taking an appropriate decision. In the old days, we used to go to an expert for the valuable suggestions in the context of our need. In this era, things are changing dramatically. The ever-growing amount of available digital information and the number of visitors to the internet have created a potential amount of information which promotes the data-driven approach. The improvement of computational power is fueling this data-driven approach more. Due to this improvement, the machine learning models, which once were slow and impractical, are being energized. As a result, in 2012 and 2013, AlexNet acquired almost human level's accuracy in case of image classification [1]. The whole concept that humans are better than a machine, falls apart when Alpha Go defeated the Go world champion and Google's inception network crossed the human level's accuracy in the context of classifying an image [2]. This revolution has proved that given a sufficient amount of data, the machine can surpass human-level accuracy. Later, people started to try machine-learning approach in a different field and following this path, we have come up with the idea of suggesting a suitable place for setting up a business and suggesting the expected rating of a business from the features that a restaurant possesses. In this field, works have been done mostly on personalizing the food suggestion for users using fuzzy logic [4], [5], [6], [7], [8]. Almost in all the cases, user satisfaction, helping users etc. were the main concerns. From the point of view of the entrepreneurs, risk assessment before setting up a business is an important factor [6]. To the best of our knowledge, there is no existing system which will help the entrepreneurs in this. Thus, our concern was to aid the entrepreneurs who will invest in the restaurant business to take an optimal decision. The lack of work in giving emphasis on the entrepreneurs inspired us to build up a model which suggest them a suitable place and the expected rating from the features of the restaurants which will help them in assessing the risk associated with. Blending the necessity of entrepreneurs and extensive computational power, we build up a supervised machine learning model to serve this purpose.

In our work, we are trying to do the rating prediction in the context of the restaurant business. If anyone wants to set up a business, our system would tell them the suitable places to set up that business depending on the average rating of the users from the data using machine learning. Add to that, it will also tell the expected average rating that a restaurant will get depending on the features of the restaurants which will pave the way for accessing the suitable features.

The rest of the paper is categorized as follows where section II represents the related works done in this field, section III describes data preprocessing, section IV presents working procedure, result, and analysis where V concludes the article by drawing future plans.

II. Background Study

In a research article [9], Lunkad predicted the rating of a restaurant using data from Yelp. His concentration was on the two sub-datasets (business and review) of yelp dataset [13]. In his work, he tried to predict the rating of the restaurant using the review subset. The business sub-dataset contains state, city, name, average stars, business id etc. and review file contains user id, business id, and review. Six major attributes were considered from these two sub-datasets. Later, he used a support vector machine, linear regression, and naïve bias model to have a look at how it works. Among these three

978-1-5386-5051-6/18/$31.00 ©2018 IEEE

algorithms, linear regression and support vector machine works well though linear regression (53.13%) was slightly better than support vector machine (52.35%).

Another work by Wang [10] et al. have aimed to predict new restaurant success as well as rating. Moreover, their goal was to recognize the restaurant features that control the success of the restaurant business. They wanted to predict the range in which a restaurant business can be successful. They have used yelp dataset for the prediction of restaurant success and rating as this dataset is very standard and easy to use. For classifying restaurant features, that contains most weight, they have run the Chi-squared test as well as stochastic gradient descent. For this purpose, they have used a variety of binary and multi-class classification algorithms such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and Multilayer Neural Networks. Their aim was to guess a restaurant's success and they have rounded ratings to the nearest star. After performing these algorithms, they have found that two algorithms among the four, performed really well. These two algorithms are Random Forest and Multilayer Neural Networks. The accuracy of these two algorithms is 56% for multi-class classification and 60% for binary classification. Later, they have performed sentiment analysis on restaurant reviews and after undergoing several processes the accuracy increased to 85% using clustering algorithms combined with the mentioned work.

In another paper, Yu [11] et al. have used yelp dataset for predicting business success and rating, as yelp dataset provides the connection of people in the local business. In their paper, the authors have mainly concentrated on the reviews for the restaurants. Moreover, they have also noticed that the sentiment features are very useful for rating prediction. Star rating is the most useful option where users can make their choices among all the businesses that are available. The higher the rating is, the higher the chance to be liked by the users because the users know that higher star rating ensures the high quality of service. So, in yelp dataset, the star rating encourages the users to judge specific business as they know that people will only give a higher rating if they are actually satisfied with the business. Therefore, the star rating is a way to evaluate a business success. For the entire investigation, they have used the yelp dataset. In their paper, the authors aimed to predict the star rating for the review of the restaurant. For this purpose, they have used three machine learning algorithms: Linear Regression, Random Forest Tree and Latent Factor Model, which were then combined with sentiment analysis. They have evaluated each individual model to check which algorithm gives the best accuracy. After the evaluation, they have found that the random forest tree algorithm gives the maximum accuracy of around 85%.

In another paper [12], the authors wanted to predict the rating of a neighbouring restaurant of an already rated restaurant. The reason behind this is, there is a high possibility of their likeliness to go nearby that restaurant more and more. They came to a decision that there is a weak correlation between the rating of a business and its neighbour. They used two kinds of factors (intrinsic and extrinsic) of latent factor model for deciding this. Using geographical location model, they have achieved much less error compared to the state of art models like social MF, biased MF and SVD++. They have used matrix factorization for this task. They talked about three observations. Firstly, most businesses have neighbours within a short geographical distance from their five locations. Secondly, observations are weakly positively correlated, and lastly, the observation is for all type of business. They also used factors like business category, popularity, and review content. As future work, they have shown interest in investigating the influence of geographical neighbourhood in POI recommendation and sentiment analysis of business reviews.

III. Data Preprocessing

In our research, we have worked with four different kinds of machine learning algorithms on trial and error basis to see what fits best with our model. Same algorithms with different parameters have been operated to boost up the accuracy further. These four algorithms gave us different results. Before going to the trial and error part, we need to do preprocessing and cleaning which are described below.

A. Dataset

In the field of machine learning, data acquisition/selection is very crucial. We used a renowned dataset [14] from the website called YELP. This site is a USA-based site, storing different kinds of business information from 2005.

B. Cleaning

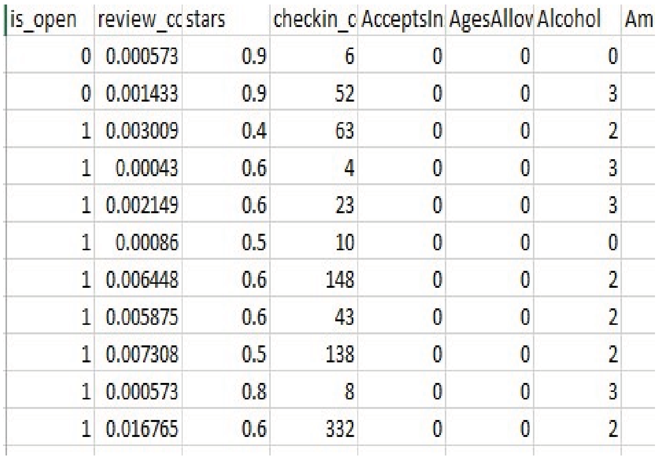
Yelp dataset contains six sub-datasets: business, checkin, tips, review, photos and user. For our work business and check-in, subsets were taken into consideration. In the business subset, we have 15 columns with 156635 instances of businesses. The columns that we have in our business subset are address, attributes, business id, categories, city, hours, is open, latitude, longitude, name, neighborhood, postal code, review count, stars and state. In the check-in subset, we have 135148 rows and 2 columns: business id and time. In total, we have 16 columns as business id, which is common in both of these subsets. There are some businesses without its data in the check-in subsets. We merged these two subsets by matching their business id. Among these columns, there are some irrelevant columns which are not necessary for our inferential work. Business id was kept for the identification of the business. Address, postal code, latitude, longitude, neighborhood, hours, is open, and the name is not necessary since these are for the identification of the business. Thus, we are left with 8 columns. Category column denotes the type of the business (i.e. restaurants, shop etc.). As our concern is regarding the restaurant business, we picked the businesses where there is the word 'restaurant' in the category. Thus, we ended up with 49536 restaurants and dropped the category column as it was not necessary anymore. The interesting and tricky part is, there are 2 nested columns (attributes and time columns) among these 7. In attributes columns, there are 82 nested columns which define the specific attributes (i.e. car parking, alcohol etc.) of a restaurant. We flattened these columns to get rid of these 82 nested columns. The time column contains the time of check-in at different times of a day on an hourly basis. For example, 7 pm to 7.59 pm is counted as an hour and check-in count for this particular range is stored with respect to it. Our concern was to find the total check-in count of certain business and to do so, we aggregated this hourly check-in data into a single value of total check-in count. Lastly, we dropped the time column as the aggregated check-in count was calculated. Thus, we ended up these 88 columns.

C. Handling missing values

There were some columns with missing values which could cause the anomaly in our dataset. For example, missing data can introduce a substantial amount of bias, making the handling and analysis of the data more arduous and can create reductions in efficiency and most importantly, can cause errors which need to be handled. For this purpose, we followed different approaches for different types of values. To deal with the missing values of the star column, we have used imputation. Imputation is a process to substitute the missing values with available information. We took the average of the whole column and substituted the missing values. For the ease of classification, we rounded those average values to their nearest value to maintain similarity with dataset columns value (1, 1.5, 2 and so on). Basically, for all the values which are numerical like review count, we have used imputation. The next type of values is the categorical values. In the case of this type of columns, we substituted the missing values with the most frequent ones. Thus, we cleaned our data for our work. As we mentioned earlier, the motive of our work is to assist a person in choosing a suitable place for a restaurant opening. To obtain this, we have used the existing restaurant's data that we have cleaned. Through inference, we can let the user know the suitable places. In the previous paragraph, we discussed the procedure for cleaning the data. After cleaning the data, we were left with 88 features or in other words columns.

D. Feature selection

Lastly, the features we took into consideration were is open, review count, check-in count, Accepts Insurance, Ages Allowed, Alcohol, 9 features related to ambience (casual, classy and so on), BYOB, BYOBCorkage, 7 features related to Best Nights (Friday, Monday and so on), Bike Parking, 5 features related to business Parking (garage, lot and so on), Business Accepts Bitcoin, Business Accepts Credit Cards, ByAppointmentOnly, Caters, Coat Check, Corkage, 6 features related to Dietary Restrictions (gluten-free, vegetarian and so on), Dogs Allowed, Drive Thru, Good For Dancing, Good For Kids, 6 features related to Good For Meal (breakfast, brunchand and so on), 8 features related to Hair Specializes In (African American, Asian and so on), Happy Hour, Has TV, 9 featuresMusic (DJ, background music and so on), Noise Level, Restaurants Delivery, Restaurants Table Service, outdoor seating, Open 24 Hours, Restaurants Counter Service, Restaurants Attire, state, city, Restaurants Take Out, Wheelchair Accessible, Wi-Fi, Restaurants Good For Groups, Restaurants Price Range, Smoking, Restaurants Reservations, Restaurants Table Service. After that, we did a feature selection which also played a vital role. There are different methods for feature selection and among those, we have used the chi-square test [14]. From the 88 features, the star is our output label and we excluded this from our feature selection procedure. Chi-square test is a statistical tool for measuring the importance of features. For this purpose, we set the value of alpha of chi-square to 0.001 to strongly discard the possibility of a null hypothesis. Many experts suggest the value of alpha, which we have used. After performing the chi-square test, based on the value of alpha as mentioned before, some features got eliminated such as check-in count, Accepts Insurance, Ages Allowed etc. as those features chi square test value were below the value of alpha which we have set. Eventually, after doing all of the processes we left with 75 features (there is an extra column which is for the business identification); and data is similar to the snapshot of figure 1 (a snapshot is given because of the space limitation; actual data consists of 75 columns/features and 49536 instances). For location identification, we used the business id to find the location of that respective place from the actual business sub-dataset.



is_open	review_cc	stars	checkin_c	AcceptsIn	AgesAllov	Alcohol	Am
0	0.000573	0.9	6	0	0	0	
0	0.001433	0.9	52	0	0	3	
1	0.003009	0.4	63	0	0	2	
1	0.00043	0.6	4	0	0	3	
1	0.002149	0.6	23	0	0	3	
1	0.00086	0.5	10	0	0	0	
1	0.006448	0.6	148	0	0	2	
1	0.005875	0.6	43	0	0	2	
1	0.007308	0.5	138	0	0	2	
1	0.000573	0.8	8	0	0	3	
1	0.016765	0.6	332	0	0	2	

Fig. 1. Snapshot of the data after preprocessing and feature selection

IV. Working Procedure, Result and Analysis

Before going to suggest the things that mentioned earlier, we need to have a look for a suitable algorithm that can predict the rating for the maximum time.

We began our work with a simple linear regression model. We used the Linear Regression model from Scikit learn [15], where we got a 15.14% accuracy on the training set which is better than the random guess of 10%. We also got 15.8% accuracy using 10-fold cross validation which is similar to the training set, which proves that there was no overfitting. This algorithm will not work for sure since we are doing a classification problem whereas linear regression is for continuous values.

TABLE I. SCORE TABLE

Name	**Accuracy Score**	**Precision Score**	**Recall Score**	**Cross Validation**
Decision Tree	60.48%	49.10%	60.48%	61.00%
Decision Tree (pre-sort)	60.50%	49.08%	60.50%	60.76%
Logistic Regression	61.30%	37.58%	61.30%	61.87%
Non-Linear SvM	97.02%	95.29%	97.25%	97.01%

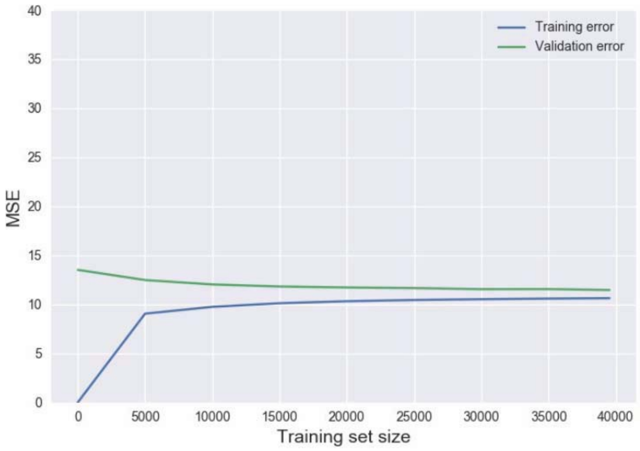


Fig. 2. The learning curve for decision tree

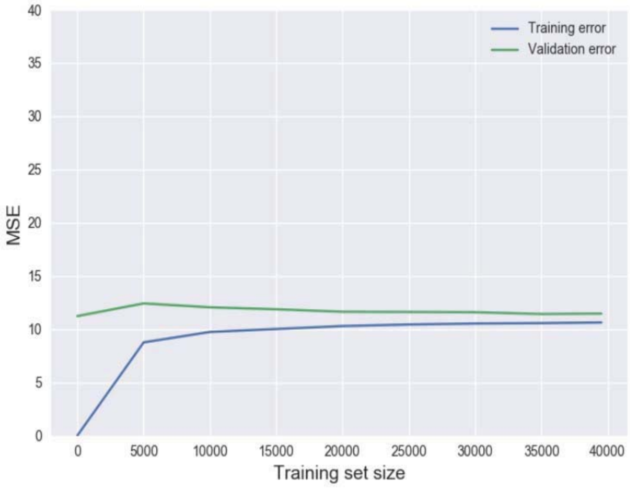


Fig. 3. The learning curve for decision tree with presort

Next, we used the Decision Tree. Here we got better results and the accuracy percentage was 60.4%. Overfitting was also controlled as the cross-validation is almost similar to the accuracy score which handles one of the problems of the decision tree. The problem with the decision tree is that it does not work with out-of-sample values on most of the occasion. Thus, the decision tree does not work very well. Increasing the data will not increase the accuracy as the training and validation error goes parallel [Figure 2].

After that, we tried to use a variation of the decision tree (Presort) with the hope that we might find the best splitting condition. But, we were still unable to find the best splitting although we found a slightly better splitting condition and thus increased the accuracy by around 0.7%. In this algorithm, the increase in data size will not improve the accuracy for the same reason as figure 2 [Figure 3].

Later, as it is a classification problem, we have used logistic regression. In this case, accuracy increased by around 0.6%. It is valid on this dataset as the cross-validation score is almost similar to the accuracy score [follow table 1].

SVM is one of the widely used algorithms for classification. Using this algorithm our accuracy increased drastically to 97.02%. Our Cross-validation score is 97.01%, which gives us the proof that it is not due to overfitting.

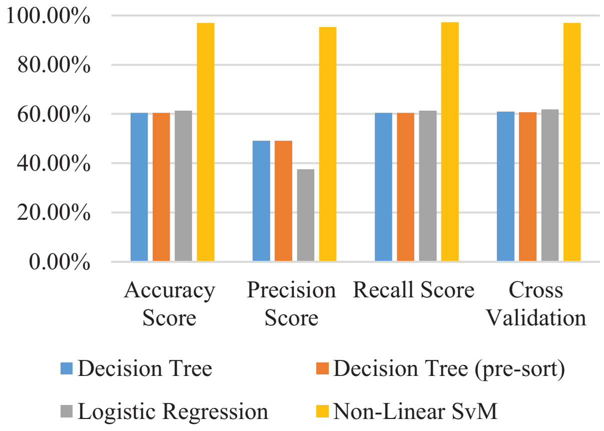


Fig. 4. Comparison of different approaches

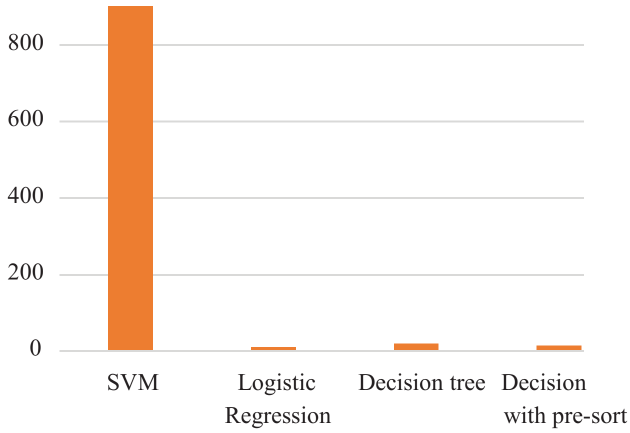


Fig. 5. Time comparison graph

SVM is the algorithm which gives us accuracy to an expected level. It outperforms others algorithms in the context of precision, recall and accuracy score. [Figure 4] In case of time complexity, support vector machine is on the slower side [figure 5]. Since time is not a huge concern for our purpose, we preferred the support vector machine.

In our task, we have used 75% of data as for training purpose and 25% for test purpose. Support vector machine was able to correctly classify the rating of 12014 from 12384 restaurants (test set).

Our model has performed far better than the other works which have been mentioned in the literature review section. Now according to our aim for this work, entrepreneurs can put the features that they are thinking to keep for their restaurants and can predict an expected average rating of their restaurants; for example, we got around 97% accuracy here in this research predicting the rating for unknown samples. Thus, they (investors or entrepreneurs) can justify their idea from the previous data and eliminate the risk of losing money and can invest accordingly.

Later, as suggesting the suitable place was also the aim of our work, we took this 12014 number of restaurants and tried to find the top 5 cities for establishing the restaurant. Restaurants rating between 3.5 and 5 are considered to be the good one. Thus, we tried to find out the cities with the most restaurants in it with stars between 3.5 and 5. Correctly classified restaurants numbers of respecting ratings (3.5, 4, 4.5,5) are given in table 2.

TABLE II. CORRECTLY CLASSIFIED STARS

Stars	Correctly classified
5	302
4.5	1432
4	3123
3.5	1131

From these correctly classified restaurants, we tried to find the city with most restaurants in this range, which is known as the quantity approach. The top 5 cities for establishing a restaurant is shown in table 3.

TABLE III. TOP 5 CITIES BASED ON MOST RESTAURANTS

City name	Restaurants number
Charlotte	30
Las Vegas	29
Toronto	20
Phoenix	16
Richmond Heights	15

Rating of 3.5 and a rating of 5 are not the same. For example, city A and B could have 50 and 40 restaurants respectively. From a straight view, it may be derived as we should prefer city A over B. However, city B could have all of its restaurants of rating 5 where city A has a few 3.5 rated restaurants. In this particular scenario, the entrepreneurs should select city B over A. In short, the number of restaurants does not reflect the suitable place perfectly. So, we came up with the idea of weighting or biasing where we tried to calculate a score (restaurants with rating 3.5 got multiplied by the number restaurants having the rating of 3.5 and so on). In this score, higher rating restaurants will get a higher score. Here too, we tried to find the top 5 cities to establish a restaurant business. From this scoring, the top 5 cities have been found and these are listed below in table 4.

TABLE IV. TOP 5 CITIES BASED ON SCORE

Name Of city	Rating 3.5	Rating 4	Rating 4.5	Rating 5	score
Las Vegas	5	3	13	9	133
Toronto	3	5	7	5	87
Phoenix	3	2	8	3	69.5
Montreal	2	3	3	7	64.5
Mississauga	6	3	3	3	61.5

We found that Las Vegas, Toronto, Phoenix, Montreal and Mississauga are the top 5 cities to set up a restaurant business. Interestingly, the most preferred city of table 3, Charlotte, has been excluded along with another one, Richmond Heights, while we take the scores for suggesting the location. Lag Vegas jumped on top with a score of 133, clearly ahead of the second preferred city, Toronto (scored 87) with an almost same number of restaurants (follow table 4).

V. FUTURE WORKPLAN AND CONCLUSION

Thus far, we only considered the business and check-in sub-datasets. Later, we plan on adding the review sub-dataset to get more insights into the restaurant business. We are also thinking of adding the user object to analyze the behaviour of users and have a look at the user's preference level to get more insights in the context of the restaurant business. Lastly, we are also considering to predict where a user will go if he/she goes to another state based on their previous preference. To conclude it all, in our work, we have shown that by using 75 variables, we can predict the rating of a business using the SVM algorithm on the data collected from YELP. We did not use a group of complex variable combinations which keeps our model simple and with good accuracy. To accomplish this, we took all of those features to predict the average rating of a business, in which we achieved 97.02% accuracy. We have also shown comparative analysis among different classification algorithms. Finally, we predicted the expected rating from the given features of a restaurant. From the comparative analysis, we found that for our dataset, SVM is the best in terms of accuracy. We hope that our work will facilitate the business people in taking decisions regarding setting up a restaurant business. We have only used YELP as the data source of our research as, to the best of our knowledge, no other similar source was found during the initial phase. Now, we are planning to test it with other similar sources.

REFERENCES

[1] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in The 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, 2012, pp. 1097–1105.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision", in The IEEE conference on computer vision and pattern recognition, Caesars Palace, Las Vegas Valley, Nevada, United States, 2016, pp. 2818--2826.

[3] J. Zeng, F. Li, H. Liu, J. Wen and S. Hirokawa, "A Restaurant Recommender System Based on User Preference and Location in Mobile Environment", in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Kumamoto, Japan, 2016, pp. 55-60.

[4] T. Osman, M. Mahjabeen, S. Psyche, A. Urmi, J. Ferdous and R. Rahman, "Adaptive food suggestion engine by fuzzy logic", in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 2016, pp. 1-6.

[5] C. Carnaghan, "Business process modeling approaches in the context of process level audit risk assessment: An analysis and comparison", International Journal of Accounting Information Systems, vol. 7, no. 2, pp. 170-204, 2006.

[6] S. Khatwani and M. Chandak, "Building Personalized and Non Personalized recommendation systems", in 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, India, 2018, pp. 623-628.

[7] L. Anitha, M. Devi and P. Devi, "A Review on Recommender System", International Journal of Computer Applications, vol. 82, no. 3, pp. 27-31, 2013.

[8] A. Rashid, I. Albert, D. Cosley, S. Lam, S. McNee, J. Konstan and J. Riedl, "Getting to know you: learning new user preferences in recommender systems", Proceedings of the 7th international conference on Intelligent user interfaces - IUI '02, pp. 127-134, 2002.

[9] K. Lunkad, "Prediction of Yelp Rating using Yelp Reviews", 2015.

[10] A. Wang, W. Zeng, J. Zhang, "Predicting New Restaurant Success and Rating with Yelp", 2016.

[11] M. Yu, M. Xue and W. Ouyang, "Restaurants Review Star Prediction for Yelp Dataset", 2015.

[12] L. Hu, Y. Liu, A. Sun, "Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction, 2014 Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, Queensland, Australia, 2014, pp. 345-354.

[13] "Yelp Dataset". Yelp.com, 2017. [Online]. Available: https://www.yelp.com/dataset.

[14] M. L. McHugh, "The Chi-square test of independence", 2013, Biochemia Medica, 23(2), page:143–149.

[15] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.