Restaurant Data Analysis

Varchaleswari Ganugapati
Dept. of Computer Science (of College of Arts and Sciences)
Georgia State University
Atlanta, USA
vganugapati1@student.gsu.edu

***Abstract*—Machine learning and an upsurge of advanced tools for data analysis has aided setting up different businesses a lot simpler and informed than ever before. Variety of insights about the demographics, culture, lifestyle, consumer demands, and ratings have enabled us to make informed decisions before implementing a business idea. In this project I analyse one such immediate and important feature of urban life: food and the entities that bring it to us. In this context I use a variety of data mining techniques to mine useful information from the data of a prominent restaurant aggregator and online delivery chain to predict the likelihood of success for a new restaurant based in Bengaluru region, Karnataka, India.**

***Index Terms*—Data mining, restaurant, demographics, cuisines.**

I. Introduction

Restaurants are the most familiar symbol of a vibrant and flourishing society in any town or city. In many ways they define the cultural ethos of the place. It is a place that unites people who come and share valuable time with loved ones or the ones who hold important business meetings that drive organizations forward. In order to set up a restaurant, three very important things to consider before are: 1) what the restaurant offers unique 2) Identifying the target consumers 3) How do the products offered impact the locality where the restaurant is being set up. If what is offered does not reach the right target audience, it is likely that the venture will fail or not be sustainable for the long run. In this project I try to answer various such questions by analyzing Zomato's data for the city of Bengaluru. Zomato Media Pvt. Ltd. is a leading Indian restaurant aggregator and food delivery start-up that provides information on menus and user-reviews for different restaurants across different cities in India and operates a delivery service for its client restaurants. I explore the factors contributing to potential success of setting up a restaurant with respect to cuisine, pricing service etc. based on the demographic needs and expectations using different data mining techniques.

II. Related Work

While there have been some really appreciable number of analysis and research groups working in this field of restaurant data analysis to offer best services or business to customer relationships, there are few which have been very interesting. One such work is available in [6] in which the purpose is to evaluate the customers' perceived consumer values in restaurant meal experiences and to compare the results with other studies on consumer values and service quality and with studies of meal experiences. [7] shows how Restaurants are characterised by predictable, seasonal factors and unpredictable, individual customer demand, which make it difficult for restaurateurs to attain efficiency. [8] This study sought to understand the views of executives at major U.S. restaurant chains regarding the process, motivation for, and challenges of offering healthier options on their menus.

III. Insight into the dataset

A. Source of data

The dataset used for this project is taken from kaggle. All copyrights for the data are owned by Zomato Media Pvt. Ltd.

The dataset in a nutshell is a collection of records that depicts hundreds of restaurants across the city of Bengaluru. Each row contains seventeen features each pertaining to different details of the restaurant such as its location, details of the items offered, pricing, the ratings given by consumers etc. A detailed description of the features is tabulated below:

B. Variables In the Dataset

- 'url': the url of the restaurant's website
- 'address': the postal address of the restaurant
- 'name': name of the restaurant
- 'online_order': If the restaurant accepts online order or not. (yes or no)
- 'book_table': If the restaurant provides table booking service or not.
- 'rate': Aggregated or overall ratings for the given restaurant.
- 'votes': Total number of ratings given for the restaurant.
- 'phone': Telephone/mobile number of the restaurant.
- 'location': The location in Bengaluru that the restaurant is situated in.
- 'rest_type': Type of the restaurant.
- 'dish_liked': Most liked dishes of the restaurant.
- 'cuisines': Different types of cuisines offered by the restaurant.
- 'approx_cost(for two people)': Cost of food , if ordered, for two people.
- 'reviews_list': Contains the list of tuples of different reviews provided by customers.
- 'menu_item': Range of menu/food items provided by the restaurant.
- 'listed_in(type)': Type of meal offered

- ’listed_in(city)’: contains the neighborhood in which the restaurant is listed

There are 51717 records in the data accurate till 15 March 2019 on Zomato. In this work, the target variable will be ‘rate’ using which we predict the aggregated ratings of the restaurant given the above mentioned scenarios and thereby determine whether a restaurant with the given features is likely to be successful or not.

IV. Data Cleaning/ Preprocessing

Data-intensive industries like banking, insurance, retail, telecoms, ecommerce have been relying on data for the improvement and growth of business. It has become very important to not only obtain the data but also clean it off any noises before using it. Data inconsistencies, redundancy or irrelevance causes poor results. If industries are looking for optimizing work and increasing their profits by using data, quality of the data is of prime importance. Poor quality data could entirely mislead and effect the interpretations made by the model, thereby leading to incorrect inferences.

For effective data analysis and machine learning models, it is essential to check the quality of the dataset at hand and ensure that unnecessary noise is eliminated. To analyse the relationships that exist between different variables. It is very important to consider the most relevant variables from the raw data before modelling our predictor. For instance, variables like ’url’, ’phone’, ’reviews_list’ which offer insignificant insights are least likely to be utilized. Therefore do not need to be leveraged for modelling the predictor. Please note that this project does not leverage the possibilities offered by language processing techniques for analysis. For instance, ’reviews_list’ is a set of ratings and text reviews given by customers for a restaurant. Since this needs some proficiency with specialized NLP frameworks, I will drop this feature and primarily consider the remaining variables for modelling.

A. *Checking for null values*

Null values are the values that occur due to the value at the given cell being missing or deliberately chosen to not to be mentioned. Among the fundamental checks, it is important to look whether the data has any null values. And also to understand if the null value convey some information or it was absent due to some man-made error like missing entries or it actually represents unavailability of data.

Certain features like **dishes_liked** have over twenty eight thousand null values. This means over half the records in our dataset do not have a legitimate entry for this feature. It is not sensible to fill or guess an entry for over half the dataset. After label encoding the variable appears to have some correlation to other predictors. Figure 1 encapsulates the correlations of different predictors. Therefore I chose to use a flag variable or label for the null values rather than replacing the values with mean or entirely removing the columns. The **menu_item** variable also has over 30000 records with empty list. Once again, I chose to use this variable by encoding null values with some flag as replacing menu_item with any central tendency does not make sense and in turn can mislead the model or increase the bias. **name** variable though sounds nominal, has a significance in the sense that that a particular restaurant could have multiple outlets. It is therefore important to keep the variable to maintain that relationship.

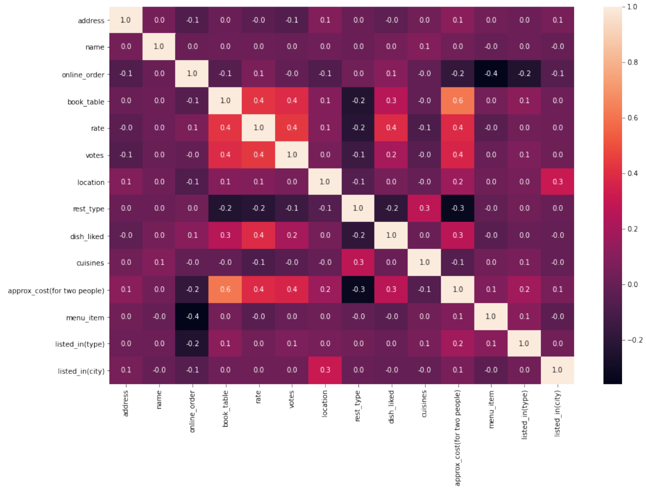


Fig. 1: Correlation matrix

B. *Checking for duplicates*

The data was checked for any duplicates and it was found to have sixty six duplicate rows overall which were dropped. Removal of redundant values could help reduce the complexity in data regarding processing time. It will also help us to avoid irrelevant insights due to redundancy

C. *Structural Errors*

While preprocessing, data could exist in structurally different form but mean the same, i.e. if we enter a name as Varchala, VARCHALA, varchala or vARCHALA all mean the same but are structurally different. This will cause misinterpretation and will lead to irrelevant insights. Hence, such errors should be handled. **Rate** has unwanted /5 character set in the structure and **vote** has a comma separated pricing value like 2,000. It was important to rectify these errors.

Target variable **rate** and independent variable **votes** are expected to be of float or integer datatype but were observed to be existing as object type with some unimportant characters such as ”/5” , ”,” , ”-” , ”NEW” present in them. These columns were cleaned from the noise and then were changed into the relevant data type.

D. *Label Encoding and Scaling*

Machine learning algorithms will perform better in terms of accuracy when the data is represented in a numeric format (or rather require so) instead of categorical. This is attributed to the fact that a computer understands only numbers in the first place. Label encoding is a primary step in any model training that uses categorical predictors. All the categorical data in the dataset is encoded by assigning some numeric labels. Since linear regression and other models use only numerical data, this is an important data preprocessing step.

The predictors in the dataset were label encoded in order to follow the convention and help train the models better.

Scaling is necessary in order to perform principal component analysis or linear regression however it is not required for models such as random forest. We will see more about these three models in the coming sections as they play a role in our analysis.

V. Data Exploration

A. Target variable: rate

The target variable **rate** shows a slightly left skewed normal or Gaussian distribution, shown in fig. 2. The rate values are symmetric near the mean = 3.7.

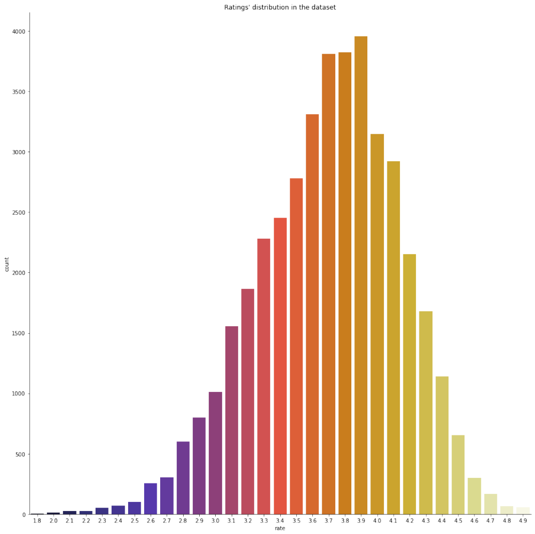
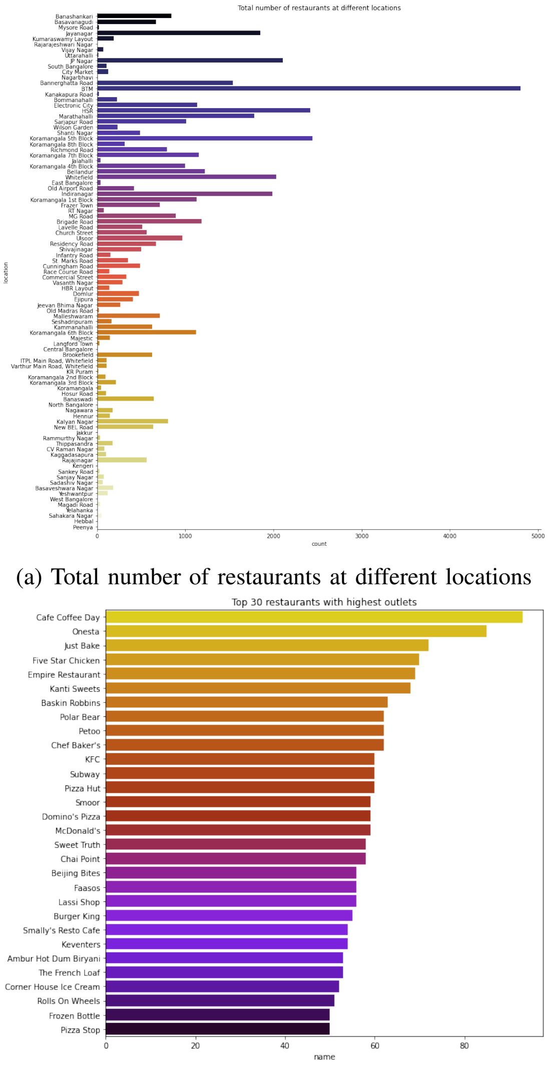


Fig. 2: Ratings' distribution

B. Predictors

Figure 3(a) shows the total number of restaurants at different locations and it is observed that BTM area has the highest number of restaurants. Figure 3(b) shows the top thirty restaurants with the highest number of outlets, a list topped by the coffee chain Cafe Coffee Day.

It can be clearly inferred that places like BTM, even with the highest number of restaurants did not necessarily house the most successful restaurants and likewise Cafe Coffee Day with highest outlets is the not the highest rated restaurant among the candidates. Rather the highest rating = 4.9 has shown a different trend for the choice of restaurants and places as shown in figure 4. This trend helps us to infer that the increase in outlets does not imply better ratings but is also based on the location and the competition offered for various services offered matters the most. Figure 5 shows the variation of top 30 restaurants based on total number of people who have rated the restaurants vs cuisines, location and type. This trend reveals that the most visited (rated) restaurants (which have also been the most rated) as shown in figure 4. This backs the assertion that the ratings and vote predictors are positively correlated as also observed in figure1. Amongst all, Big Brewski Brewing Company has been the most rated and voted restaurant and that means that not only did people rate it high but also it was the most reviewed restaurant.



(a) Total number of restaurants at different locations

(b) Top 30 restaurants with highest outlets

Fig. 3: Count of Restaurants at different locations and its outlets

Figure 6 shows what percentage of restaurants offer online ordering of food and table booking options. Based this it is clear that more than 50% of restaurants offer online food ordering and only around 40% restaurants offer table booking.

The ratings for different restaurants have shown a significant linear relationship with the approximate cost for two people. This trend conveys that restaurants which have been mid-range to slightly costly have never disappointed customers. This inference could probably be attributed to the fact that costly places tend to offer better services and ambience. However, the highest rated restaurants have not been pricier than Rs 2000/- for two people in their pricing as shown in figure 7. So

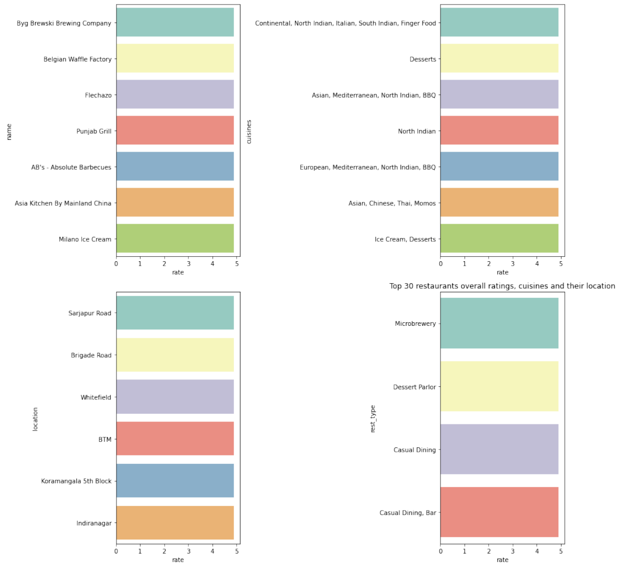


Fig. 4: Top 30 restaurants' rating vs location,restaurant type and cuisines

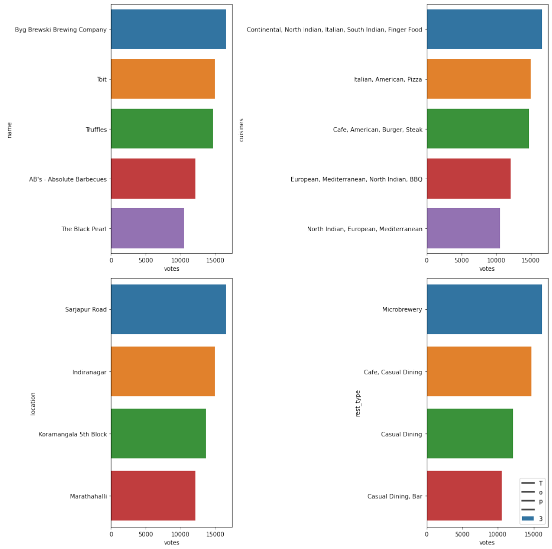


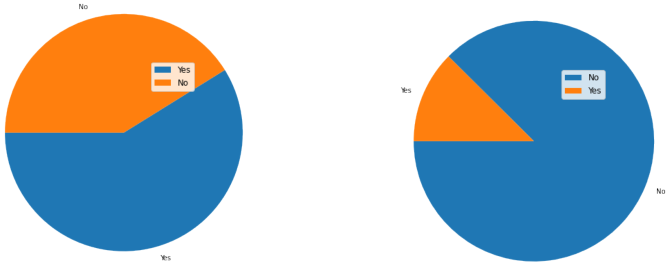
Fig. 5: Top 30 restaurants' vote vs location,restaurant type and cuisines

while the service may get better with increase in the prices, the general satisfaction quotient is still retained by relatively mid-tier restaurants, possibly through uniqueness in other aspects like content, taste etc.

VI. Model

A. Principal Component Analysis (PCA)

A brief introduction of PCA can be termed as a data exploration technique that involves finding 'principal components', which are a collection of direction vectors. A particular direction vector denotes a line that best fits the data and is orthogonal to the remaining principal components. PCA is generally used to find these orthonormal bases and then change the basis of the data to these bases (often times using only a few than all). This technique is used in exploratory data analysis and predictive data modelling. It is most commonly used for dimensionality reduction by projecting the data onto a few principal components while preserving much of the variance of the actual data.



(a) Online Ordering (b) Table Booking

Fig. 6: Proportion of Restaurants that offer online ordering and table booking

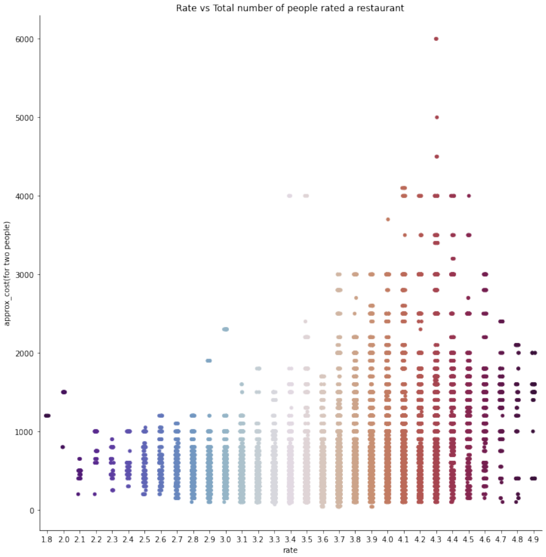
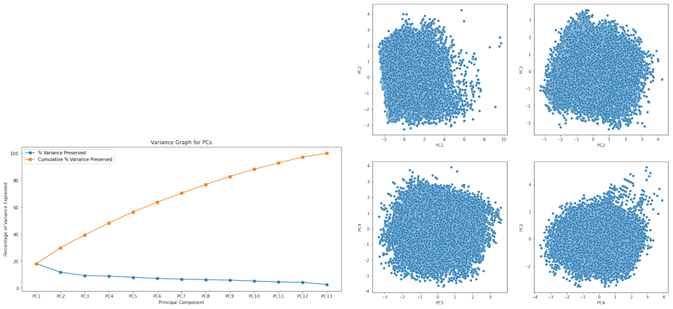


Fig. 7: Pricing vs Rating

In this work, PCA was used to summarize the dataset with a smaller number of representative variables that collectively explain most of the variability in the original dataset. The intention here was to explore if a feature is more dominant in describing the data as compared to other features in the dataset. In such a case we can use the feature which describes the data 'almost' perfectly (almost because usually one feature does not describe 100% of the outcome in the real world scenarios) and avoid the features which have the least significance in the outcome.

The analysis as shown in figure 8(a), however, conveys that most of the individual features are not very highly correlated as the first principal component could explain not more than 20% of the variance in the data. Still, PCA was useful for visualizing the data and as will be seen later, Random Forest



(a) Variance Graph

(b) Spread of the observations in 2-d plane using four different principal components

Fig. 8: PCA

Regressor is applied on all the thirteen PCs that were obtained to analyse if the PCs did any better job in predicting the outcome.

B. Linear Regression

We now look at another machine learning model called Linear Regression to model our data. Linear regression is a supervised machine learning technique that models the relationship between a response variable and one or more explanatory variables. All the linear relationships are modelled using linear predictor functions. In other words, a linear regression model assumes the relationship between the response variable and the independent variables is linear. It is predominantly used in predicting and forecasting tasks. There are multiple variations in the implementation of linear regression details for which can be found in the references.

The data did not clearly show any linear trend among the features except for the fact that there did exist some appreciable correlations with predictors as was shown in figure 1. This motivated me to consider multiple linear regression for building a prediction model to ascertain if the data was linear in some way. The results however, were not encouraging. With this it could be concluded that linear models like linear regression failed to fit well with our dataset and that any relation has been mostly non linear.

C. Random Forest

Until this point all our analysis only conveyed the non linearity of the data and so it became important to apply a non-linear model to be used for prediction. Decision trees have been widely used for modeling non linear relationship among the training observations and making the required predictions of responses. They work by dividing the predictor space i.e, the set of possible values for X1, X2,...,Xp into J distinct and non-overlapping regions, R1, R2,..., RJ. For every observation that falls into the region Rj, the same prediction is made, which is simply the mean of the response values for the training observations in Rj. Decision trees can end up overfitting the data with a very complex structure. They tend to perform quite well on training data but usually fail in accuracy and efficiency during validation. Ideally, our goal should be to minimize error due to bias and variance. This is where Random forests come to the rescue. A random forest is a collection of decision trees whose results are aggregated into one final outcome. Random Forests reduce variance by training on different samples of the data by choosing subsets of decision trees and also leverage the possibility of choosing subsets of features to train a set of sub trees in comparison to others in the entire model. This random choice of features helps in reducing correlation between different base trees, hence, limiting the error due to bias or variance. This very reason in accordance with the non linearity of data was the biggest motivation to choose Random forest regressor model for the original dataset and then perform the predictions to evaluate the results.

Random Forest regressor on the original dataset has shown impressive results on test data compared to all other models as shown in Table 1.

TABLE I: Accuracy Table

Test	Models		
Score	***RF on PCs***	***Linear Regression***	***Random forest***
R2 Score	0.58383619	0.35219669	0.9113043
RMSE	0.27809065	0.34695712	0.12806398

VII. Cross Validation

A consistent practice in machine learning and data mining is dividing datasets into two halves, Training set and Test set. This process helps us to train the model on a subset of the available samples and then test it on an entirely different subset from the same sample space so that the model is tested for its performance on unseen data. The important question is, does different combinations of training and test subsets produce different results and thus lead to varying accuracy? Yes indeed, that is exactly what is caused due to some bias or variance in the subsets of data while training and testing. The most noble approach to validate the dataset is by using cross validation. Cross validation is an evaluation/validation approach that is applied on the dataset by dividing it into several folds and then iteratively applying train and test evaluations on all of these folds. K-fold cross validation is one such technique which divides the dataset into k folds, holds one subset for testing and performs training on the remaining k-1 subsets. This process continues until every subset has had the chance to become test set at least once. K-fold CV was applied on the random forest regressor to evaluate its overall performance. It has shown that the model's average accuracy using 10 fold CV was **87.53%**.

VIII. Conclusion

Random forest is the most suitable model among all the techniques applied for the prediction of the target variable based on the limited set of features that were chosen for the model to predict the success of a restaurant startup. The architecture of the model is designed such that it can be generalized for application across different datasets from different cities across India provided the set of features remain

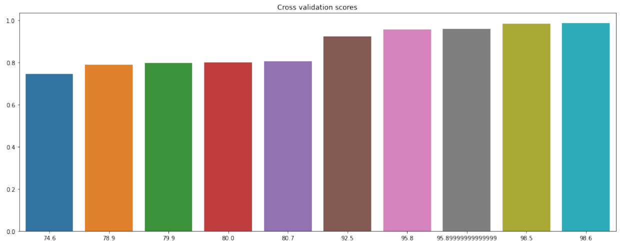


Fig. 9: 10 fold cross-validation on Random forest regressor

the same. However, to scale the model to cater to higher dimensions requires a rework on evaluating the model using random forest as the algorithm tends to perform poor with an increase in the number of variables beyond certain threshold.

The results from this model will aid the prospective restaurant owners and stakeholders to make informed decisions before establishing a restaurant chain in or across Bengaluru city. The scope for further development will be to create a foundation for a scalable model that can generalize parameters to successfully work across different cities and achieve good results with higher dimensions of data.

IX. REFERENCES

1) Dataset: Zomato media PVT. LTD. and kaggle.
2) Gareth James,Daniela Witten, Trevor Hastie, Robert Tibshirani: An Introduction to Statistical Learning
3) Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
4) https://en.wikipedia.org/wiki/Linear_regression
5) https://en.wikipedia.org/wiki/Principal_component_analysis
6) Consumer values among restaurant customers,International Journal of Hospitality Management,Volume 26, Issue 3,2007,Pages 603-622,ISSN 0278-4319, https://doi.org/10.1016/j.ijhm.2006.05 004.
7) Mhlanga, O. (2018), ”Factors impacting restaurant efficiency: a data envelopment analysis”, Tourism Review, Vol. 73 No. 1, pp. 82-93.
8) Karen Glanz, Ken Resnicow, Jennifer Seymour, Kathy Hoy, Hayden Stewart, Mark Lyons, Jeanne Goldberg, How Major Restaurant Chains Plan Their Menus: The Role of Profit, Demand, and Health,American Journal of Preventive Medicine, Volume 32, Issue 5, 2007, Pages 383-388, ISSN 0749-3797,