STUDYING THE AVERAGE COST FOR TWO AT RESTAURANTS ASSOCIATED WITH ZOMATO

Akshat Dubey
Birla Institute of Technology, Mesra
imh10035.18@bitmesra.ac.in

Anindya Sadhukhan
Birla Institute of Technology, Mesra
imh10036.18@bitmesra.ac.in

Ankit Tewari
Artificial Intelligence Engineer
Knowledge Engineering and Machine Learning Group
ankit.tewari@estudiant.upc.edu

July 2, 2019

ABSTRACT

We have used the data-set on Zomato Restaurants which was available to public, and shared by an user on: " https://www.kaggle.com " .This, data-set contains restaurant's ID,City,Country,Cuisines offered, Average Cost for two,Currency,Aggregate rating(in float data-type), Rating text(Excellent,Very Good, Good, Average, Poor, Not Rated) and some other relevant data. We have tried to visualize some bar graphs and other plots, and find some general trends to make a conclusion. Then we have used KNN-regression method to make predictions and finally we have made conclusions.

1 Introduction

Zomato is a restaurant search and delivery network which provides food services for millions of users every month. Zomato let's users search restaurants, get recommendations, add reviews, photos and such. It also provides an at a glance overview of restaurants. There is online ordering is almost all serviceable areas.

This is an Exploratory Data Analysis of the Zomato Dataset with respect to India using Python. Here, we have looked into the data, visualise it, and make some conclusions.. Ultimately, we have gone for prediction of , average cost for two persons, when they have their meal in a restaurant on the basis of their rating category(Excellent,Very Good, Good,Average,Poor, Not Rated).

Ultimately, we have gone for prediction, checked the R2_score and RMSE of our model and at the last we have dicussed the results and have made conclusions.

2 Methods and Experiments

2.1 Data Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Through plots, we can predict some points directly although approximately. In our project we have used many plots to predict some basic but useful predictions.

2.2 Linear Regression:

Linear regression is used for finding linear relationship between target and one or more predictors. Linear linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight. We have used this to predict "Average Cost for two". The equation of the linear line is given by:

$$Y(pred) = b0 + b1 * x$$

where Y is the variable we want to predict and x is the independent variable, b0 is the intercept and b1 is the slope.
The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

$$Error = \sum (actual_output - predicted_output)^2$$

2.3 KNN for regression:

KNN stands for K-Nearest-Neighbors. It is a machine learning algorithm used for prediction. It can be used for both classification and regression. The algorithm uses 'feature similarity' to predict values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. Here, K refers to the number of neighbors which closely resembles the points in the training set. The K is choosen in such a way that RMSE(Root Mean Squared Error) is minimum and R2_score is maximum. We will focus on KNN for regression.
Working of KNN algorithm:

1. First, the distance between the new point and each training point is calculated. The most used way of calculating distance is by using Euclidean Distance.

$$EucledianDistance = \sqrt{(y_i - x_i)^2}$$

2. The closest k data points are selected (based on the distance).
3. The average of these data points is the final prediction for the new point.

We thought of using KNN on our data-set because the Linear Regression doesn't always work and what if the correlation between the variables is very low.
We will use the inbuilt KNN model through scikit-learn package of python.

3 Results

We have tried to predict the "Average Cost for two" and the rating of the restaurants from the currency using KNN for regression.
We have following results:
Figure 9 shows the relation between R2_score and RMSE value.We can observe that the minimum RMSE value, the R2_score is maximum. Also, we can note it down that the minimum RMSE is corresponding R2_score of approx 0.68.
The table 1 shows the RMSE value for different values of K, we can observe that the minimum RMSE value is corresponding K=2. Also, from the Figure 8 it can be verified.
The table 2 shows the maximum R2_score corresponding different values of K. The maximum accuracy is corresponding K=2. Also, from the Figure 7 it can be verified. The accuracy score of our KNN model is 68.92% which is a nice score.
To calculate the accuracy score we have the following formula:

$$AccuracyScore = R2\_score * 100$$

We also tried using the Linear regression but the correlation values were very small. The R_score for our linear regression model was 0.09, and the accuracy score came out to be 9% which is very low. Predictions made from the data visualization are:

1. Figure 1:

- Indian city has maximum number of Zomato restaurants.
- Zomato has its presence in 23 countries but the most important country is India.

2. Figure 2:

- New Delhi has the highest number of restaraunts associated with Zomato with a count of more than 5000.
- Gurgaon and Noida are behind New Delhi with count of more than 1000 restaurants associated with Zomato.

3. Figure 3:

- Restaurants providing only North-Indian cuisines are the highest in number with count of approximate 850.
- Restaurants providing both Chinese and North Indian and restaurants providing only Chinese are behind the resturants providing both North Indian with a count of approx 450 and 380 respectively.

4. Figure 4: A bar graph showing the number of restraunts which provide online delivery.

5. Figure 5:

- Restaurants with ratings 3.3 are highest in number.
- Near about 30 restaurants are unrated. May be they are new.
- There are no restaurants having rating less than 3,which may show that Zomato doesn't collaborate with restaurants having ratings less than 3.

6. Figure 6:

- Cafe Coffee Day has maximum number of restaurants associated with Zomato in INDIA followed by Domino's Pizza and Green Chick Chop.

After the successful implementation of KNN-regression algorithm, it is clear that for various currency & based on rating, average cost for two varies when they go for a meal in a restaurant. For Indian rupees the average cost for two based on rating are approximated as follows-
Poor-Rs.550
Average-Rs.575
Good-Rs.300
Very good-Rs.1400
Not rated-Rs.475
So,with the above prediction any customer can be sure of how they have to spend for their required quality of foods on Indian Rupees

3.1 Figures

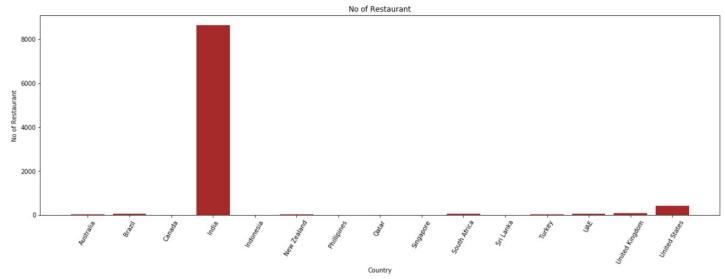

Figure 1: Number of restaurants in different countries

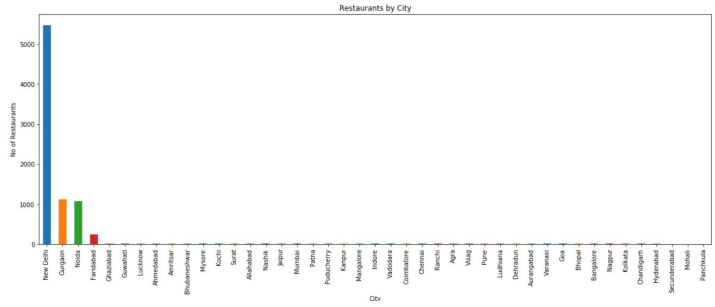

Figure 2: Number of restaraunts in different Indian cities

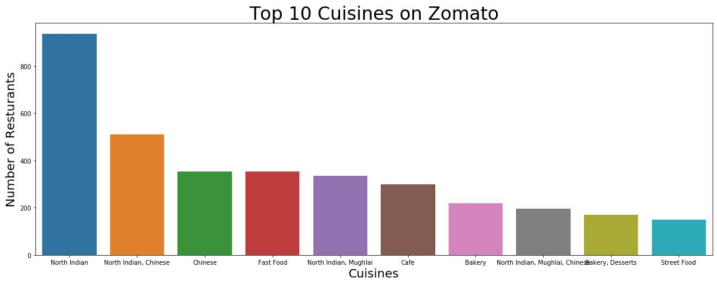

Figure 3: Number of restaurants offering different cuisines

Figure 4: Number of restaurants having online deliveries

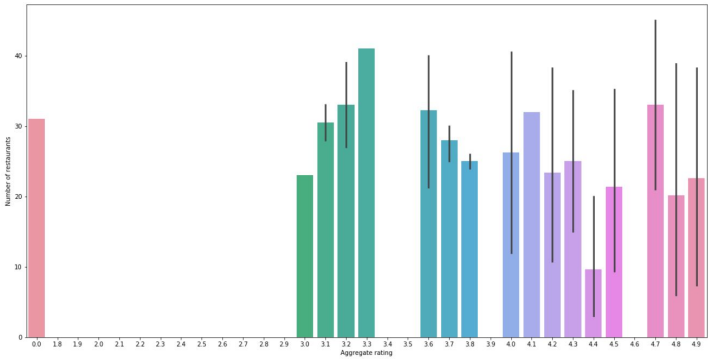

Figure 5: Number of restaurants vs aggregate rating

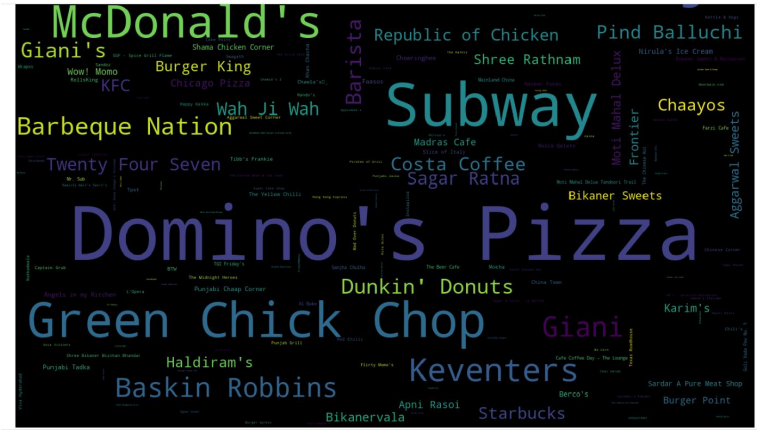

Figure 6: Word cloud

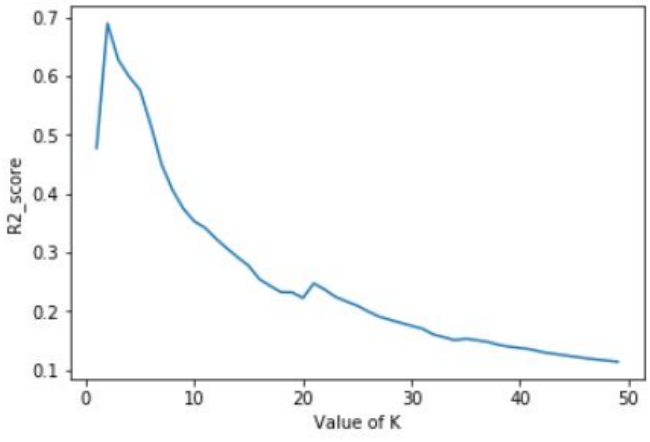

Figure 7: Line plot of R2_score vs value of K

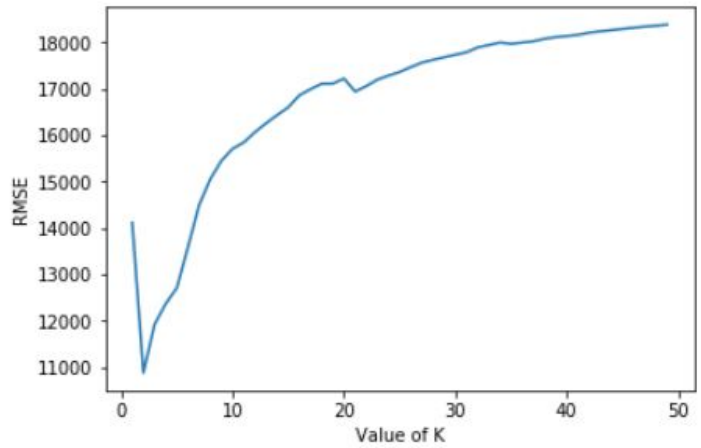

Figure 8: RMSE vs different values of K

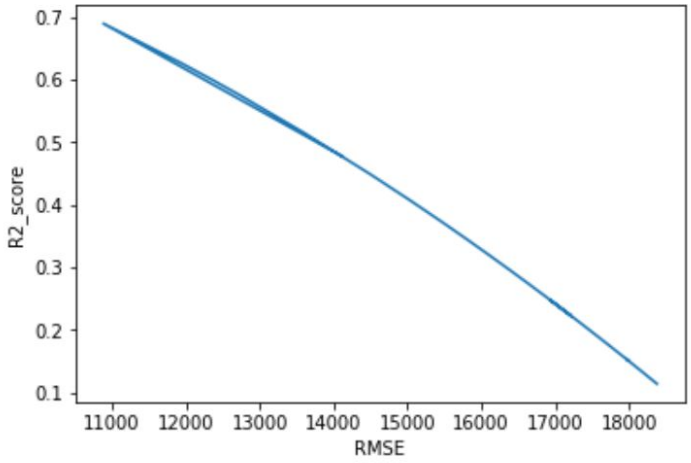

Figure 9: RMSE vs R2_score

3.2 Tables

Table 1: RMSE vs K

Value of K	RMSE
1	14113.66
2	10883.00
3	11922.81
4	12370.25
5	12710.56
6	13587.39
7	14494.73
8	15051.40
9	15447.80
10	15705.94
11	15841.87
12	16061.68
13	16253.57
14	16430.80

Table 2: R2_score vs K

Value of K	R2_score
1	0.4773
2	0.6892
3	0.6270
4	0.5985
5	0.5761
6	0.5156
7	0.4487
8	0.4056
9	0.3738
10	0.3527
11	0.3415
12	0.3231
13	0.3068
14	0.2916

4 Discussions

We, first tried to visualize our data by plots and graphs. Through graphs and plots we were able to make quick decisions and some basic predictions.
We then checked for the correlation values of the variables using [DataFrame.corr()] method of Python, from the correlation table we came to know that there was a poor correlation among the variables. To, have a check we tried to fit and train a linear regression model and this came out to be our worst choice. The R2_score of this linear regression model came out to be 0.09, which resulted into an accuracy score of 9% which was very poor. Our Linear regression model was just 9% accurate, which meant that out of 100 predictions approx 9 were right. So, we dropped our idea of using linear regression.
Then we proceeded towards KNN for regression. We divided our data into train and test data with the help of 'train_test_split' method of 'sklearn.model_selection'. We then trained the KNN model and predicted values for different values of K. Also, we checked for RMSE and R2_score for different values of K. The RMSE value for K=2 was lowest and R2_score corresponding K=2 was greatest and it was around 0.6892 which resulted into an accuracy

score of 68.92%. This score meant that our model was 68.92% close to the actual values, which is much better than the accuracy score of our linear regression model which was approx 9%.

5 Conclusions

We have predicted the average cost for two people when they have their meal in a Zomato associated restaurant from the currency they use for payment. This project can be used to learn that although the Zomato has its presence in 23 countries but the most important country is India.Zomato should try to focus on other countries too if it wants to capture a great proportion of market and also wants to compete with its competitor such as Uber Eats. Zomato will face ups and downs whenever the value of Indian currency will change on international scale. We came to know that Zomato associated restaurants are categorized into Excellent,Very Good,Good, Average, Poor and Not-rated according to the ratings received by country. A customer can know about the average price which a Zomato associated restaurant should charge on the basis of rating text. In the coming future we can compare the difference between average price for two on the basis of the rating and also monitor a change in the average price for two when the value of the Indian currency alters in the international market.

6 Acknowledgement

This was our first project and we would like to thank our mentor Ankit Tewari, who guided us on how to use the KNN for regression, even in the most of the confusing problems which contains majority of qualitative variables. He helped us in building different two models for this models namely, Linear Regression and KNN for regression algorithm. He also made us compare two models and told us the difference between the two models and why the KNN for regression model is better than the linear regression model.

References

1. Data-set: https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants.
2. Notebook: https://www.kaggle.com/akshat0007/project-0001
3. https://www.antoniomallia.it/on-implementing-k-nearest-neighbor-for-regression-in-python.html
4. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics).