High Level Design (HLD)

zomato

Restaurant Rating Prediction

CONTENTS

Abstract

Zomato started in 2008 underneath the name, ‘Foodiebay’ to begin with. Later in 2010, it had been renamed to ‘Zomato’. Constantly 2011, Zomato extended to increasingly urban regions the country over in Mumbai, Delhi NCR, Chennai, Bangalore, Kolkata and Pune. After that in the year 2012, the corporate extended working all around in various countries like the UAE, Qatar, Sri Lanka, UK, South Africa and Philippines. In the year 2013, Zomato had moved their organizations in Brazil, New Zealand, Turkey and Indonesia, with its applications and site open in various lingos isolated from English. After that in April 2014, Zomato impelled its organizations in Portugal Republic, trailed by Canada, Lebanon and Ireland around a similar time. The acquiring of Settled - based sustenance zone ‘Urban spoon’ signified the organization’s passageway into the United States, Canada and Australia, and conveyed it into direct test with ‘Wail’, ‘Zagat' and ‘Open Table’. With the introduction of .xxx zones in 2011, Zomato also impelled ‘zomato.xxx’, a site dedicated to finding spot to eat near to your territory. It later moved a print adjustment of the site substance named, ‘Citibank Zomato Restaurant Guide’, got together with Citibank in May 2012, at any rate later it was halted.

1.Introduction

1.1 Why Restaurant Rating Prediction Project?

Restaurant Rating has become the most commonly used parameter for judging a restaurant for any individual. A lot of research has been done on different restaurants and the quality of food it serves. Rating of a restaurant depends on factors like reviews, area situated, average cost for two people, votes, cuisines and the type of restaurant. The main goal of this is to get insights on restaurants which people like visit and to identify the rating of the restaurant. With this article we study different predictive models like Support Vector Machine (SVM),Random forest and Linear Regression, XG Boost, Extra tree Regressor, Decision Tree and have achieved a score of 94% with Extra tree Regressor.

1.2 Definition

1. **Database**: It is the collection of the information monitored by the system.

2. **IDE**: It is Integrated Development Environment. We have used VSCode JupyterNotebook as IDE it is very great platform for machine learning .

3. **Machine learning**: Machine learning is the concept that computer program can learn and adapt to new data without human interface.

2. General perspective

2.1 project perspective

Restaurant Rating prediction is based on the machine learning which will help us to detect the rating of any restaurant at any point of time .

2.2 Problem statement

The main goal of this project is to perform extensive Exploratory Data Analysis (EDA) on the Zomato Dataset and build an appropriate Machine Learning Model that will help various Zomato Restaurants to predict their respective Ratings based on certain features.

2.3 Proposed solution

The solution proposed here is a rating prediction model for all the restaurant the data is cleaned , after all the pre-processing our data is modelled then after the model is ready we used some of the function to predict the training scores, prediction, r2 score, Mean absolute error , mean squared error and root mean squared error. Similarly we can test different method like LinearRegression, DesicionTreeRegressor, RandomForestRegressor,GradientBoostRegressor,ExtratreeRegressor, BaggingRegressor,KneighboursRegressor and AdaBoostRegressor . for cross validation our model we used RandomizedSearchcv . At the end I have used Hyper tunning the model for much more accurate prediction.

2.4 Technical requirements

The project is basically machine learning & statistic intensive. We used Python for the implementation of the models & automation

1. Automated Script to Collect Historical Data

For any prediction/classification problem, we need historical data to work with. In this project, past restaurant dishes costs and types for each restaurant based on location collected on a daily basis is needed. Manually collecting data daily is not efficient and thus a python script was run on a remote server which collected prices daily at specific time.

2. Cleaning & Preparing Data

   After we have the data, we need to clean & prepare the data according to the model's requirements. In any machine learning problem, this is the step that is the most important and the most time consuming. We used various statistical techniques & logics and implemented them using built-in python packages.

3. Analysing & Building Models

   Data preparation is followed by analysing the data, uncovering hidden trends and then applying various predictive & classification models on the training set.

4. Merging Models & Accuracy Calculation

Having built various models, we now have to test the models on our testing set and come up with the most suitable metric to calculate the accuracy. Moreover, many a times, merging models and predicting a cumulative target variable proves to be more accurate.

2.5 Data requirements

Data requirement completely based on our problem statement

The basic structure of the script successfully extracts information from the Kaggle website and outputs a csv data file. Now an important aspect is to decide the parameters that might be needed for the flight prediction algorithm.

Kaggle returns numerous variables for each flight returned. However not all are required and thus we selected the following –

1. Online order
2. Book table
3. Votes
4. Rest_type
5. Approx. cost of two people
6. Listed_in(type)
7. Listed_in(city)

2.6 Tools used

Python programming language and framework such as numpy, pandas, seaborn, matplotlib, sklearn.

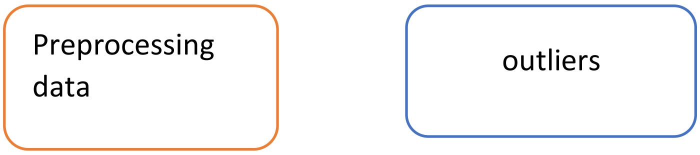
1. VS Code Jupyter is used as IDE.

   1. For visualization of the plot used as matplotlib, seaborn.

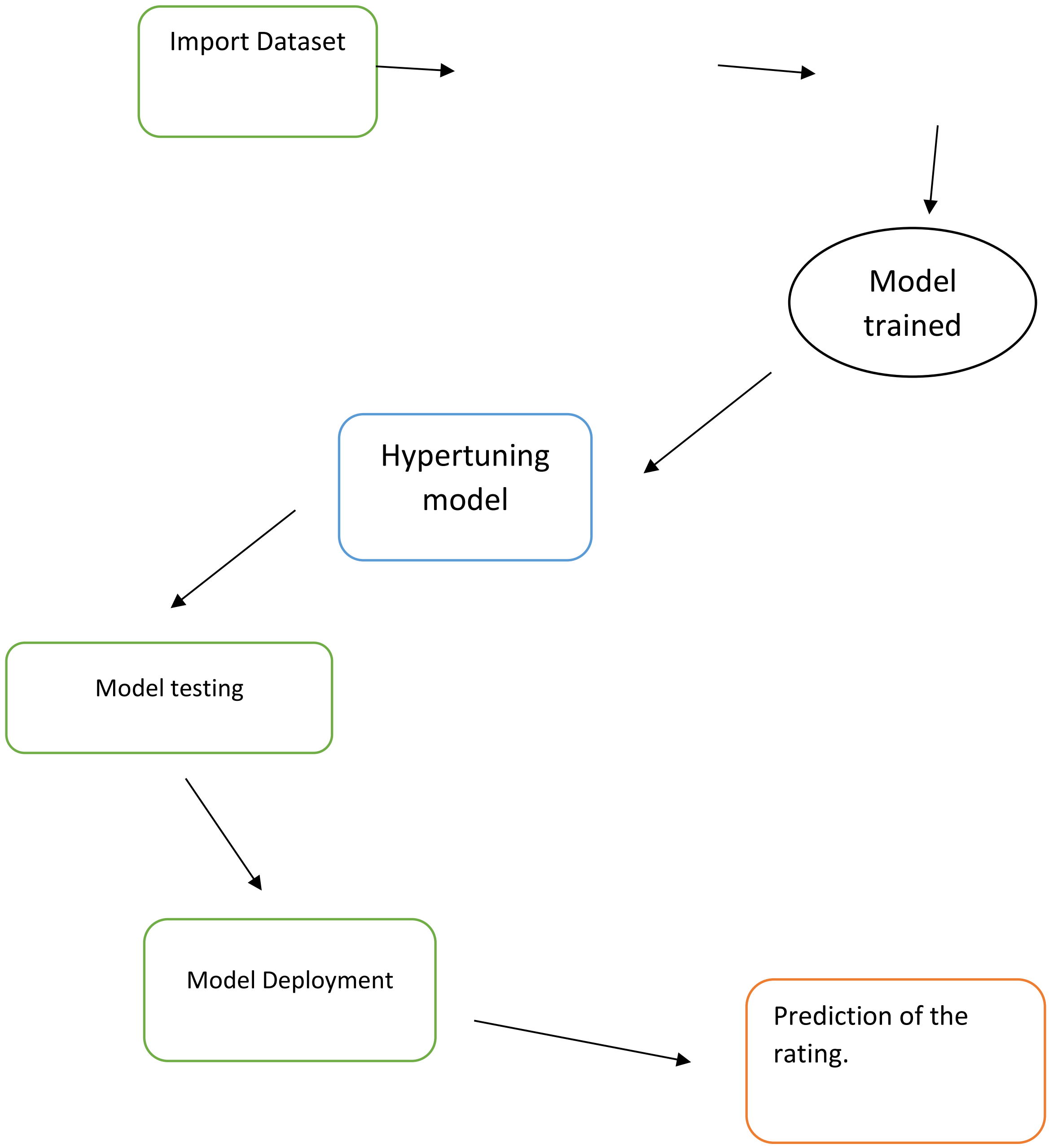
   2. Github is used as version control system.

3.Design details

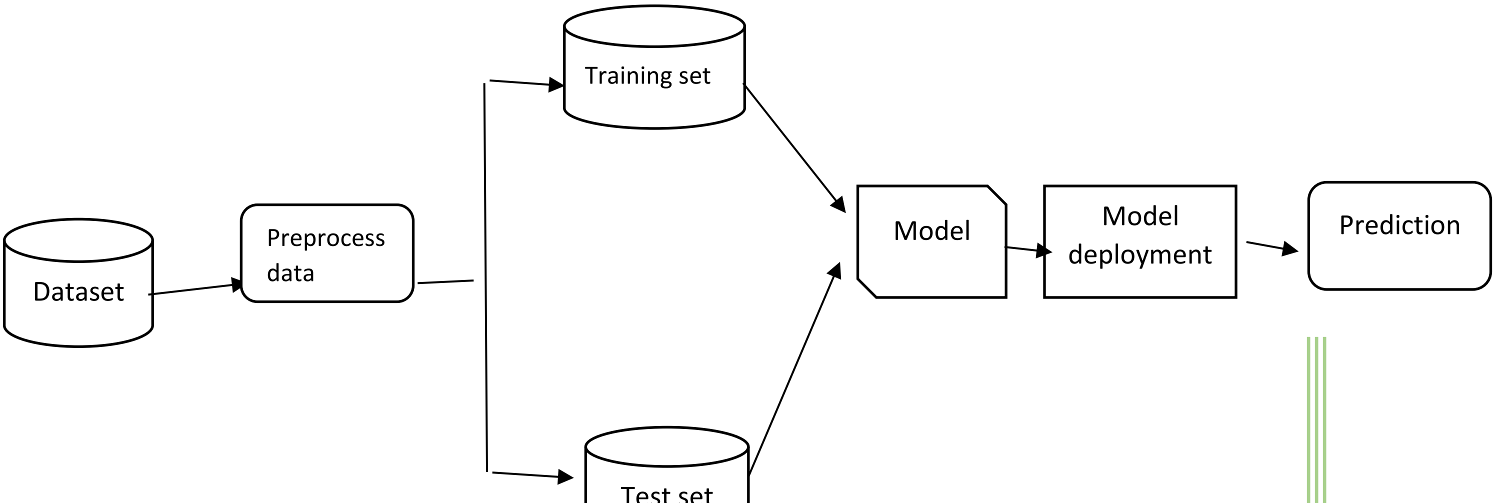
3.1 Process flow


Preprocessing data

outliers



3.2 model training and evaluation



3.3 Event Log

In this Project we are logging every process so that the user will know what process is running internally.

Step-By-Step Description:

- In this Project we defined logging for every function, class.
- By logging we can monitor every insertion , every flow of data in database.
- By logging we are monitor every step which may create problem or every step which is important in file system.
- We have designed logging in such a way that system should not hang even after so many logging's, so that we can easily debug issues which may arises during process flow.

3.3 Error Handling

Should error be countered , an explanation will be displayed as anything that falls outside the normal and integrated usage

4.Performance

1. First Records with null values were dropped from ratings columns and were replaced in the other columns with a numerical value

2. Values in the ‘Rating’ column were changed. The ‘/5’ string was deleted. For eg. If the rating of a restaurant was 3.5/5, it was changed to 3.5.

3. Using One-Hot encoding from sklearn library, encoding was done on columns like book_table,online_order,rest_type,listed_in(city).

4. We did not use any feature selection algorithms but eliminated some columns due to available domain knowledge and thorough study of the system.

6. Then I split the hole data set train-test split. After that I performed scaling on X_train and X_test.

7. After performing above step I was ready for model training. In this step, I trained my dataset on different Regression Learning algorithm LinearRegression, DesicionTreeRegressor,RandomForestRegressor,GradientBoostRegressor,ExtratreeRegressor,BaggingRegressor,KneighboursRegressor and AdaBoostRegressor). After training the dataset on different algorithms I got highest accuracy of 94% on ExtratreeRegressor.

8. After that I applied hyper-parameter tuning on all model which I have described above. Here also I got accuracy of 83% on test dataset by the same ExtratreeRegressor.So, we look go for default params of ExtratreeRegressor.

9. After that I saved my model in pickle file format for model deployment.

10. After that my model was ready to deploy. I deployed this model on cloud storage(heroku) and also dockerize this model.

2. Reusability

The code written and the component used should have the ability to reused with no problems.

3.Resource utilization

When the task is performed ,it will likely use all the processing power available until that function is finished

4.Future work

The greatest shortcoming of this work is the shortage of data. Anyone wishing to expand upon it should seek alternative sources of historical data, or be more methodical in collecting data manually over a period of time. Additionally, a more varied set of flights should be explored, since it is entirely plausible that airlines vary their pricing strategy according to the characteristics of the flight (for example, fares for regional flights out of small airports may behave differently than the major, well flown routes we considered here). Finally, it would be interesting to compare our system's accuracy against that of the commercial systems available today (preferably over a period of time).

6.Result

Sr.no	Algorithm	R2_score	Adj R2_score	RMSE	MAE	MSE
1.	ETR	0.94	0.94	0.10	0.02	0.01
2.	BAGR	0.91	0.91	0.12	0.05	0.01
3.	RF	0.91	0.91	0.12	0.05	0.01
4.	KNN	0.73	0.73	0.22	0.11	0.04
5.	GB	0.38	0.38	0.33	0.23	0.11
6.	DT	0.32	0.32	0.35	0.24	0.12
7.	LR	0.22	0.22	0.37	0.26	0.14
8.	ABR	0.05	0.05	0.41	0.34	0.17

*ETR - Extra Tree Regressor
*BAGR - Bagging Regressor
*RF - Random Forest
*KNN - k Nearest Neighbour
*GB - Gradient boosting Regressor

*DT - Decision Tree Regressor
*LR - Linear Regression
*ABR - Ada Boost Regressor

In this model, we have considered various restaurants records with features like the name, average cost, locality, whether it accepts online order, can we book a table, type of restaurant. This model will help business owners predict their rating on the parameters considered in our model and improve the customer experience. Different algorithms were used but in the end the final model is selected on Extra Tree Regressor which gives the highest accuracy compared to others.

7.Conclusion

This documents studies a number of features about existing restaurants of different areas in a city and analyses them to predict rating of the restaurant. This makes it an important aspect to be considered, before making a dining decision. Such analysis is essential part of planning before establishing a venture like that of a restaurant. Lot of researches have been made on factors which affect sales and market in restaurant industry. Various dine-scape factors have been analysed to improve customer satisfaction levels. If the data for other cities is also collected, such predictions could be made for accurate.