ZOMATO BANGALORE RESTAURANT INSIGHTS AND PREDICTIVE MODELING FOR RATING.

25th JULY 2020

Detailed Report

OVERVIEW:

- Zomato is a well known food delivery startup which helps us get food of our choice rolled up at our doorstep just at fingertips.

- Zomato has a tie up with most of the restaurants around the world and has rich data of these restaurants and could give us great insights.

- Insights and features from this data can help us build a Predictive model to predict ‘Rating’ which plays a very important role in success of a restaurant.

BUSINESS PROBLEM TO SOLVE/GOALS:

- Gaining important insights and trends from Restaurant data.
- Understand what people like the most in a highly rated restaurant, how approx_cost_for_2, neighbourhood, locality, etc are related to ratings for a prospect restaurant.
- Understand factors that predict ratings using suitable model.
- Setting up Marketing strategies like personalized notifications, discounts etc. can be set up to attract an audience based on insights to get optimal results.

DATA COLLECTION AND WRANGLING

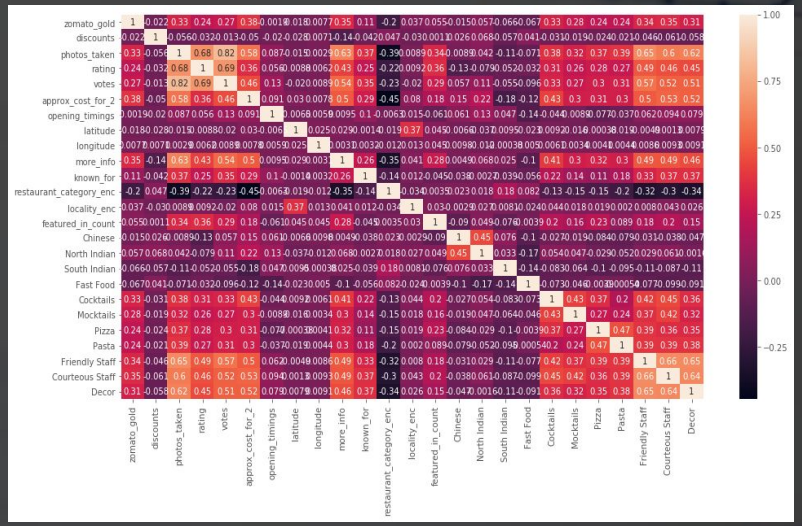
- The data is scraped from Zomato(https://www.zomato.com/bangalore) using the Python package ‘Beautiful soup’ as of Jan 2020.Ref: Web scraping
- Extracted clean data Ref: data csv

EXPLORATORY DATA ANALYSIS:

Insights: (Detailed EDA Ref EDA link)

- Quick bites and Casual dining are the most common of all restaurant categories.
- North Indian and Chinese are the most popular whereas Belgium and Portuguese are some of the rare cuisines.
- Sankey road, Lavelle road and Church street have highest average ratings.There is definitely locality playing a part in Restaurant rating.

Exploring Correlations and Trends:



Insights from Correlation and Pairplots:

- *Photos_taken* are positively correlated with *ratings* and *votes*, and negatively correlated with *approx_cost_for_2*.
- Features like *photos_taken*, *votes*, *approx_cost_for_2*, *featured_in_count*, *zomato_gold* tend to have positive correlation with restaurant *rating*.
- *Discounts* have a negative correlation with *ratings*.
- From the above restaurant_category distribution,we see that a few categories are very popular.
- Few localities and restaurant categories have relatively high *approx_cost_for_2*.

FEATURE ENGINEERING:

Some of the following feature engineering techniques were performed,

- Imputation
- Log Transform
- Encoding : Label Encoding, Multilabelbinarizer Encoding, Numerical Encoding, Target Encoding.

STATISTICAL INFERENCES: Ref link

We have performed some statistical tests using Bootstrap techniques to verify or have confidence of certain observations seen in Exploratory data analysis.

- Null hypothesis test resulted in positive correlation between approx_cost_for_2 and rating.
- Null hypothesis test resulted in Zomato Gold have high ratings compared to Non Zomato Gold restaurants.

PREDICTIVE MODELING FOR RATING: Ref link

Referring to the pair plots and correlation matrix, we have selected features which show certain relation with the target variable and we split the data into train and test.

To check linear relationships we tried with Linear Regression and further with decision trees which gave us the following results.

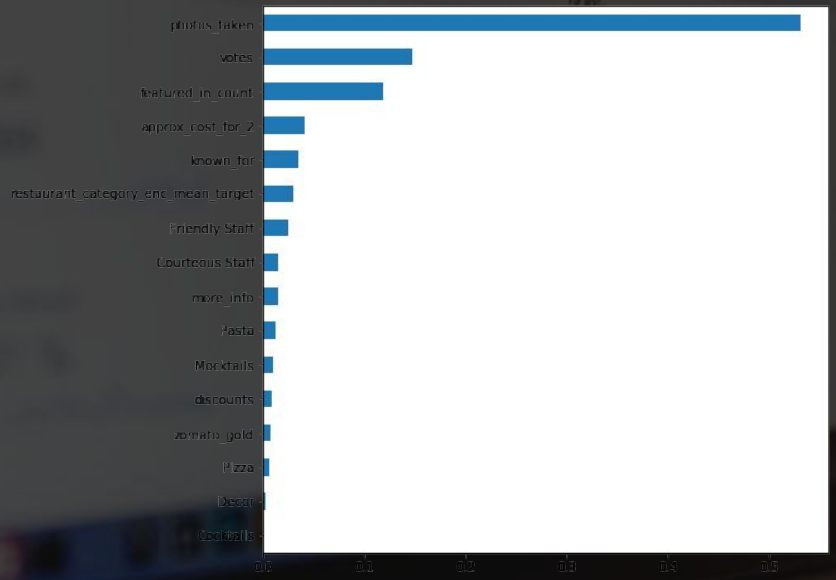
Further hyperparameter tuning using RandomizedSearchCV gave following results,

Linear Regression	**0.565**
Random Forest Regressor	**0.602**
XGBoost	**0.602**

We summarise that there is a good amount of noise in the data, also we need to include more features/factors and incorporate more data for better learning of the model.

Feature importance:

Important features of the XGBoost model,



From the plot we see that features like ***photos_taken, votes, featured_in_count*** and ***approx_cost_for_2*** are significant features for predicting "***Rating***" of a restaurant.