RESTAURANT RATING PREDICTION

Problem Statement

- Nicole has a restaurant at BTM in Bangalore, India.
- she is thinking of joining Zomato as new restaurant to increase her sales.
- but before joining she would like to know if her restaurant will be able to get at least rate 4 over 5.
- she doesn’t think of moving her restaurant since she just renovated.
- Other than that, any suggestion can achieve the goal?

Objective

- Using Machine Learning technique to predict her restaurant rating at Zomato platform.
- And identify the important features that for high rating.

Agenda

- Methodology
  - Datasets, Models, Metrics, Tools
- Process Workflow
  - EDA
  - Data preparation
  - Data analysis
  - ML model training/evaluation
- Metrics
  - R square, MSE, MAE
- Conclusion
  - How it helps with business case
  - Recommendations
  - Interesting insight
- Future Opportunities
- Appendix

Methodology

- Source of dataset
  - From Kaggle
  - Zomato is a platform that provides information, menu and user-reviews of restaurant as well as food delivery options from partner restaurant in select cities.
  - This data is scraped and accurate to that available on the Zomato website until 15 March 2019.

- Model, Metrics and Tools
  - Supervised Machine Learning Regression Problem
  - Model: Linear Regression, Lasso, Decision Tree and Random Forest Regression.
  - Metrics: R square, MSE, MAE.
  - Tools: mssql, powerbi, jupyter notebook, python, pandas, numpy, matplotlib, seaborn, scikit learn, etc

Process Workflow

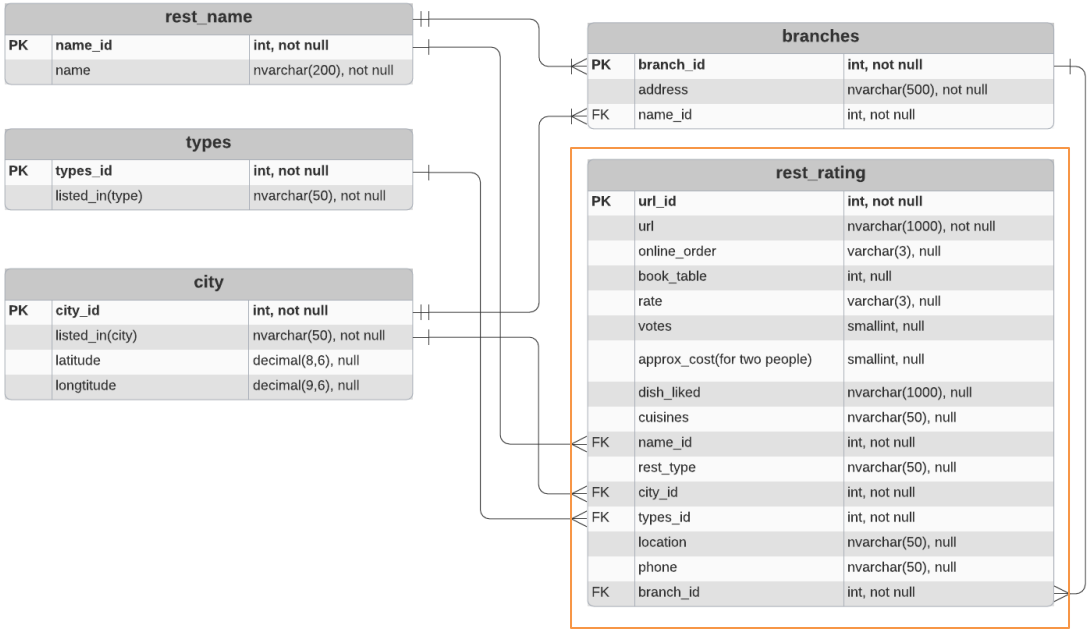
Extract Transform Load (ETL)

	feature	null	dtype	uni_value	max_len
0	url	0	object	51717	538.0
1	address	0	object	11495	346.0
2	name	0	object	8792	159.0
3	online_order	0	object	2	3.0
4	book_table	0	object	2	3.0
5	rate	7775	object	64	6.0
6	votes	0	int64	2328	0.0
7	phone	1208	object	14926	34.0
8	location	21	object	93	29.0
9	rest_type	227	object	93	29.0
10	dish_liked	28078	object	5271	134.0
11	cuisines	45	object	2723	86.0
12	approx_cost(for two people)	346	object	70	5.0
13	reviews_list	0	object	22513	1284117.0
14	menu_item	0	object	9098	24897.0
15	listed_in(type)	0	object	7	18.0
16	listed_in(city)	0	object	30	21.0

- 17 columns 51717 rows.
- url is the unique value.
- In the ‘rate’ column (initially look like 4.1/5) and contained ‘-’ and ‘New’. I will remove the value, change the type to int and treat it as new restaurant that will be predicted later.
- Drop ‘review’ column, since there is already ‘rate’ column.

Process Workflow

Entity Relationship Diagram (ERD)



- Create an ER diagram.
- Use **pyodbc** to store the tables in ms sql database.
- For rating prediction, I will use rest_rating table.

Process Workflow

Exploratory Data Analysis (EDA)

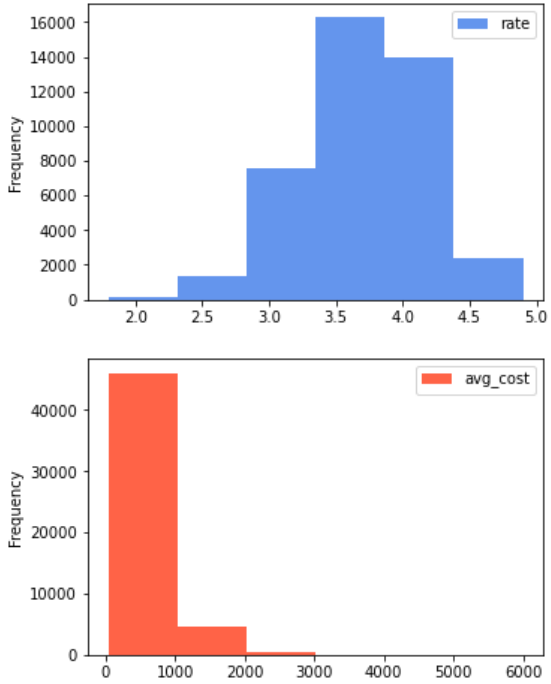
	feature	null	dtype	uni_value	max_len
0	url_id	0	int64	51717	0.0
1	url	0	object	54747	538.0
2	online_order	0	object	2	3.0
3	book_table	0	object	2	3.0
4	rate	10052	float64	31	0.0
5	votes	10027	float64	2327	0.0
6	phone	0	object	14927	34.0
7	location	21	object	93	29.0
8	rest_type	227	object	93	29.0
9	dish_liked	28078	object	5271	134.0
10	cuisines	45	object	2723	86.0
11	avg_cost	346	float64	70	0.0
12	menu_item	0	object	3095	8000.0
13	name_id	0	int64	8743	0.0
14	types_id	0	int64	7	0.0
15	city_id	0	int64	30	0.0
16	branch_id	0	int64	11466	0.0

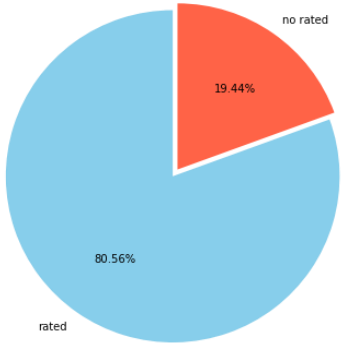
- So here is some information from the rest_rating table.
- Remove url, phone, name_id, branch_id.
- I will remove those location and city also, since Nicole has no intention to change location.
- I don't need menu and dish_liked as well.
- After removing unnecessary data, there are still some null values which I will treat it later.

Process Workflow

Exploratory Data Analysis (EDA)

- Rate normal distribution.
- No rate data is about 20%. For this portion will keep it for prediction.
- In the average cost, most of the price are within 1k rupee, so for null value I will replace it with median.
- I will remove votes because of the multicollinearity.





Process Workflow

Exploratory Data Analysis (EDA)

types_id	meal_type
1	Buffet
2	Cafes
3	Delivery
4	Desserts
5	Dine-out
6	Drinks & nightlife
7	Pubs and bars

	online_order	book_table	types_id	avg_cost	rest_type	cuisines	rate
0	No	Yes	1	3000.0	Fine Dining	Continental, North Indian, Italian, Chinese	4.0
1	No	Yes	5	3000.0	Fine Dining	Continental, North Indian, Italian, Chinese	4.0
2	Yes	No	3	250.0	Quick Bites	North Indian, Fast Food, Street Food	3.9
3	Yes	No	5	250.0	Quick Bites	North Indian, Fast Food, Street Food	3.9
4	Yes	No	3	250.0	Quick Bites	North Indian, Fast Food, Street Food	3.9
...	...	...	...	...	...	...	...
51712	No	No	3	NaN	Takeaway, Delivery	Continental, Italian, Steak, American	4.1
51713	No	No	3	200.0	Food Truck	Fast Food	3.4
51714	No	No	3	200.0	Food Truck	Fast Food	3.4
51715	No	No	5	200.0	Food Truck	Fast Food	3.4
51716	No	No	5	200.0	Food Truck	Fast Food	3.4

51717 rows × 7 columns

- This is the table after extracting the attributes for machine learning.
- There are 7 types means Buffet, Café, Delivery, Dessert, Dine-out, Drink & nightlife and Pubs and bars.
- Restaurant type is more detail, for example, Quick Bites, Bakery, Casual Dining, Food Court, Kiosk, Sweet Shop, etc.

Process Workflow

Exploratory Data Analysis (EDA)

	types_id	avg_cost	rate
count	51717.000000	51371.000000	41665.000000
mean	3.807375	555.431566	3.700449
std	1.140839	438.850728	0.440513
min	1.000000	40.000000	1.800000
25%	3.000000	300.000000	3.400000
50%	3.000000	400.000000	3.700000
75%	5.000000	650.000000	4.000000
max	7.000000	6000.000000	4.900000

- From the statistical view, the average price are within 40 – 6000 rupee and the average is around 555 rupee.
- For rate are within 1.8 – 4.9 and the mean of the rate is around 3.7.

Process Workflow

Data preparation

- Replace null in the ‘avg_cost’ column by median value.
- Replace null in ‘rest_type’ and ‘cuisines’ column by random value.
- Label encoding for categorical data.
- After excluded new restaurant from the training set, the remaining data is 41,665.
- Then allocated 80% for training and 20% for testing.

Process Workflow

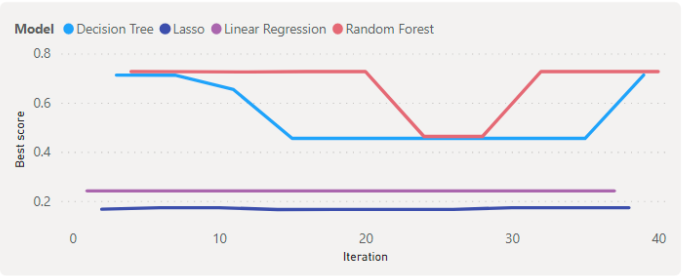
Machine Learning model training/evaluation

- Use Grid Search to find the best parameters and the best R square score.

41665
Total training data

- Linear Regression
- Lasso
- Decision Tree Regression
- Random Forest Regression

BEST SCORE MODEL by TIME SERIES



BEST SCORE VS TRAINING TIME

BEST SCORE MODEL

Random Forest
0.73

Decision Tree
0.71

Linear Regression
0.24

Lasso
0.17

CROSS VALIDATION GRID SEARCH

best_params	best_score	model	train_time
{'max_features': 'auto', 'n_estimators': 500}	0.73	Random Forest	547.14
{'n_estimators': 500}	0.73	Random Forest	327.67
{'n_estimators': 600}	0.73	Random Forest	357.58
{'max_features': 'auto', 'n_estimators': 600}	0.73	Random Forest	551.56
{'max_features': 'auto', 'n_estimators': 200}	0.73	Random Forest	156.54
{'max_features': 'auto', 'n_estimators': 60}	0.72	Random Forest	63.05
{'criterion': 'friedman_mse', 'max_depth': None, 'splitter': 'best'}	0.71	Decision Tree	4.41
{'criterion': 'mse', 'max_depth': None, 'splitter': 'best'}	0.71	Decision Tree	4.29
{'criterion': 'mse', 'max_depth': None, 'min_samples_leaf': 1, 'splitter': 'best'}	0.71	Decision Tree	4.16

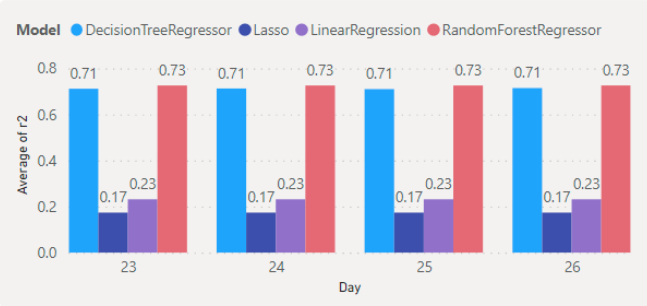
Process Workflow

Machine Learning model training/evaluation

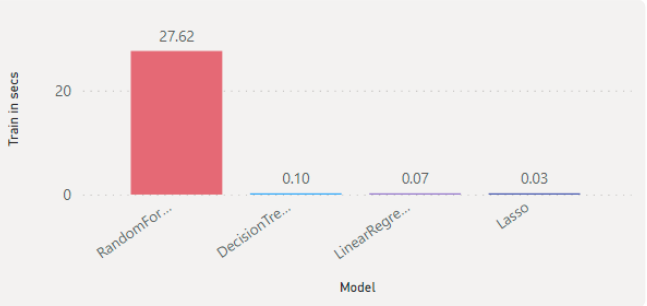
33332
Train data

- Splitting 80%, 20% to train data and test data. There are 33,332 data is in the training set.
- Random forest and decision tree are the top 2 accuracy model.
- Using Random Forest Regression as the best model for prediction.

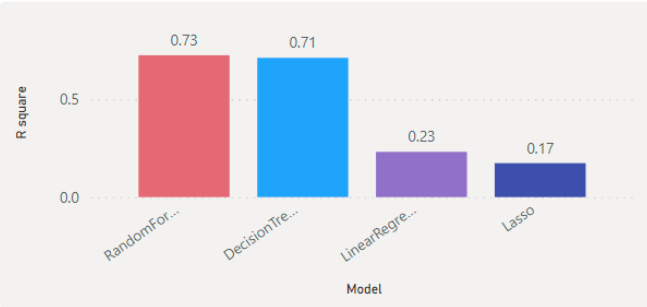
R SQUARE by TIME SERIES



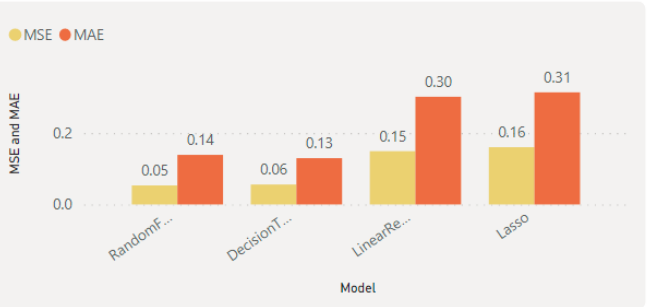
AVERAGE TRAINING TIME by MODEL



R SQUARE by MODEL



MSE & MAE by MODEL



Process Workflow

Machine Learning model training/evaluation

8,333
Test data

10,052
New restaurants

3.51
Average of pred_rate

ACTUAL VS PREDICTED

number	actual	predicted
0	4.00	4.00
1	3.10	3.23
2	3.90	3.86
3	3.10	3.35
4	3.40	3.74
5	3.90	3.90
6	3.70	3.74
7	3.50	3.50
8	3.40	3.40
9	4.20	4.20
10	3.70	3.74
11	3.90	3.72
12	3.80	3.64
13	4.40	4.40
14	4.10	4.10
15	3.90	3.44
16	4.30	4.25
17	3.20	3.68
18	3.70	3.72
19	3.60	3.07
20	3.60	3.53

NICOLE's RESTAURANT INFO.

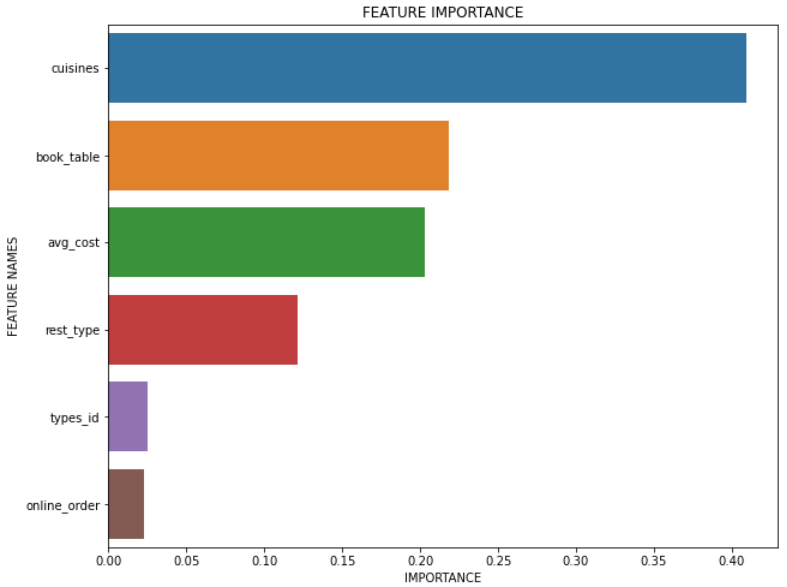
cuisines	online_order	book_table	type	avg_cost	rest_type
American, Cafe, Continental, French, Burger, Mexican, Desserts, Pizza	No	0	Buffet	450.00	Bakery, Sweet Shop

- Using Random Forest Regression as the best model.
- tested the 8333 data.
- The average rating of new restaurants is 3.5,
- Nicole’s restaurant rate prediction is 3.7.

NICOLE's RESTAURANT RATE PREDICTION

Conclusion

- The predicted rate 3.7 which is lower than Nicole's expectation.
- In order to get higher rate, I suggest her to reference the feature importance and make some adjustment to her restaurant's data.
- Like for example, she can fine tweak on her restaurant cuisines; consider to provide the 'book table' service or even increase her food price since the mean of the avg_cost is around 555 rupee.



Future Opportunities

- Inverse encoding.
- Coming up will look into more detail on the existing features, there is 2723 unique value in the cuisines columns, I wonder if label encoding is the best approach to encoding my categorical data.
- Scrape the data from different country, so I can analyst by region.
- I would explore more on the review column in order to better understand the customer feedback.

Appendix

- https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
- Wiki: https://en.wikipedia.org/wiki/Zomato
- tools: sql server data tool, ms sql, ssms, jupyter notebook, python, ssis(etl),

external ssis installation through vscode => https://www.mssqltips.com/sqlservertip/6481/install-sql-server-integration-services-in-visual-studio-2019/

- internal ssis installation => https://www.mssqltips.com/sqlservertip/6635/install-ssis/

connect python to sql server using pyodbc: https://datatofish.com/how-to-connect-python-to-sql-server-using-pyodbc/

- ssis tutorial: https://www.youtube.com/watch?v=0ikNnenDyNw

ssms import data from CSV File through ssis: https://www.youtube.com/watch?v=wozqnFbjyYc

- Feature importance: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e
- https://www.kaggle.com/thiagopanini/predicting-the-success-of-a-restaurant/?scriptVersionId=42278583
- https://medium.com/analytics-vidhya/zomato-bangalore-restaurant-analysis-and-rating-prediction-101fd635ab15
- https://towardsdatascience.com/zomato-bangalore-data-analysis-6ee83652890f
- https://medium.com/@shubh1795/starting-a-new-restaurant-in-bangalore-heres-what-you-should-know-e53bbce55a8
- https://datatofish.com/pie-chart-matplotlib/

Special Thanks

Aditya
Gouri
Tong Wei
Fredy
David Liu
Adrian
An Ting
Aza
Billy
Chris
Kenny
Linda
Noel
Michelle
Zaleha

Thank you