Yelp Restaurant Review

By: Muhammad Faiz Bin Mohd Puad

AGENDA

01 INTRODUCTION & PROBLEM STATEMENT

02 EXPLORATORY DATA ANALYSIS & MODEL PREPARATION

03 CONCLUSION & RECOMMENDATION

04 REFERENCES

INTRODUCTION

Emergence of platform like Yelp help business owner i.e. restaurant may continually improve through study of diners' eating experiences and review of the cuisine and service. Reevaluate its business plan, which has a significant economic value.

From customer P.O.V, they may steer clear of uncomfortable situations by eating experience and learn more about regional specialties by reading reviews of the cuisine.

PROBLEM STATEMENT

Newly open Restaurant need to face with potential hype of food's trend of preference by customers due to cultural preference [4]. It can even be influenced by geographical aspect [4]. Thus, it is most likely to lose significant revenue due to inadequate research for start up [1].

EXPLORATORY DATA ANALYSIS

- Business df
  - 150,346 rows & 14 columns
  - Missing Values:
    - Attributes : 13,744
    - Categories : 103
    - Hours : 23,223

Table 1.0 : Business df features

Features	Datatype	Null Values
business_id	string	0
name	string	0
address	string	0
city	string	0
state	string	0
postal_code	string	0
latitude	float	0
longitude	float	0
stars	float	0
review_count	integer	0
is_open	integer	0
attributes	string	13744
categories	string	103
hours	string	23223

- Review df
  - 2 million rows & 9 columns

Table 2.0 : Review df features

Features	Datatype	Null Values
review_id	string	0
user_id	string	0
business_id	string	0
stars	float	0
useful	integer	0
funny	integer	0
cool	integer	0
text	string	0
date	string	0

EXPLORATORY DATA ANALYSIS

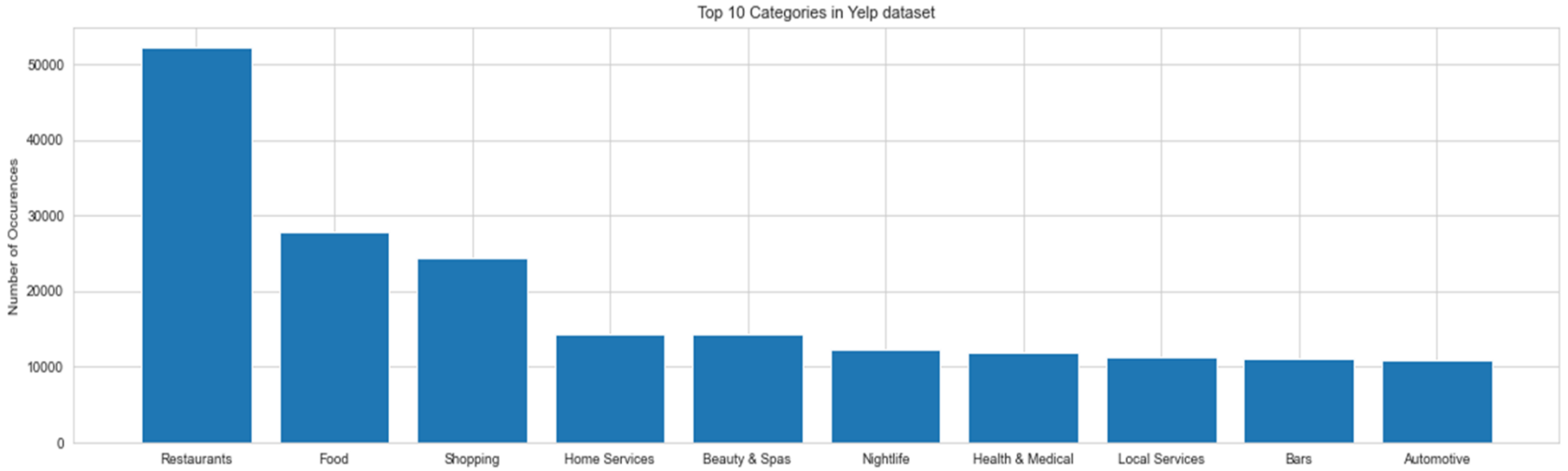


Figure 1.0 : Top 10 Categories of Industry In Business df

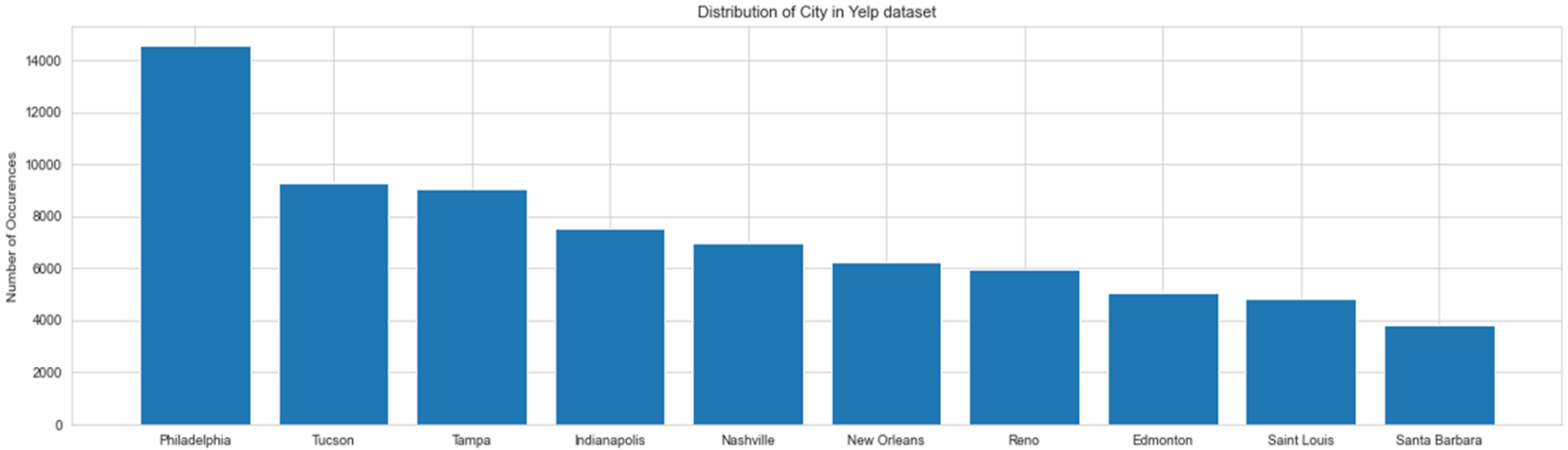


Figure 2.0 : Top 10 Cities With Highest Contain of Industry In Business df

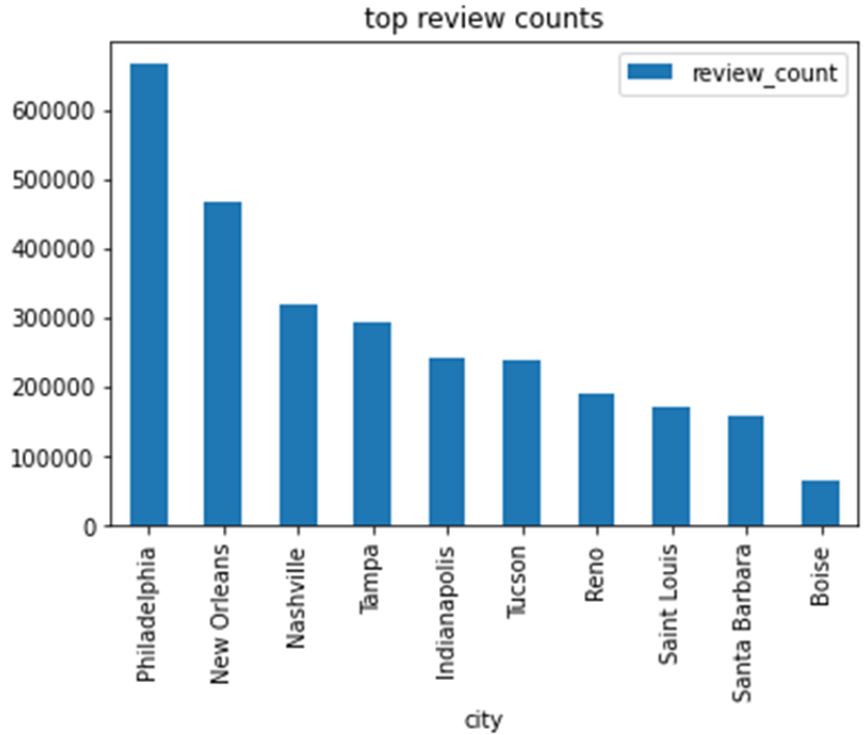


Figure 3.0 : Top 10 Cities With Highest Review Count In Business df

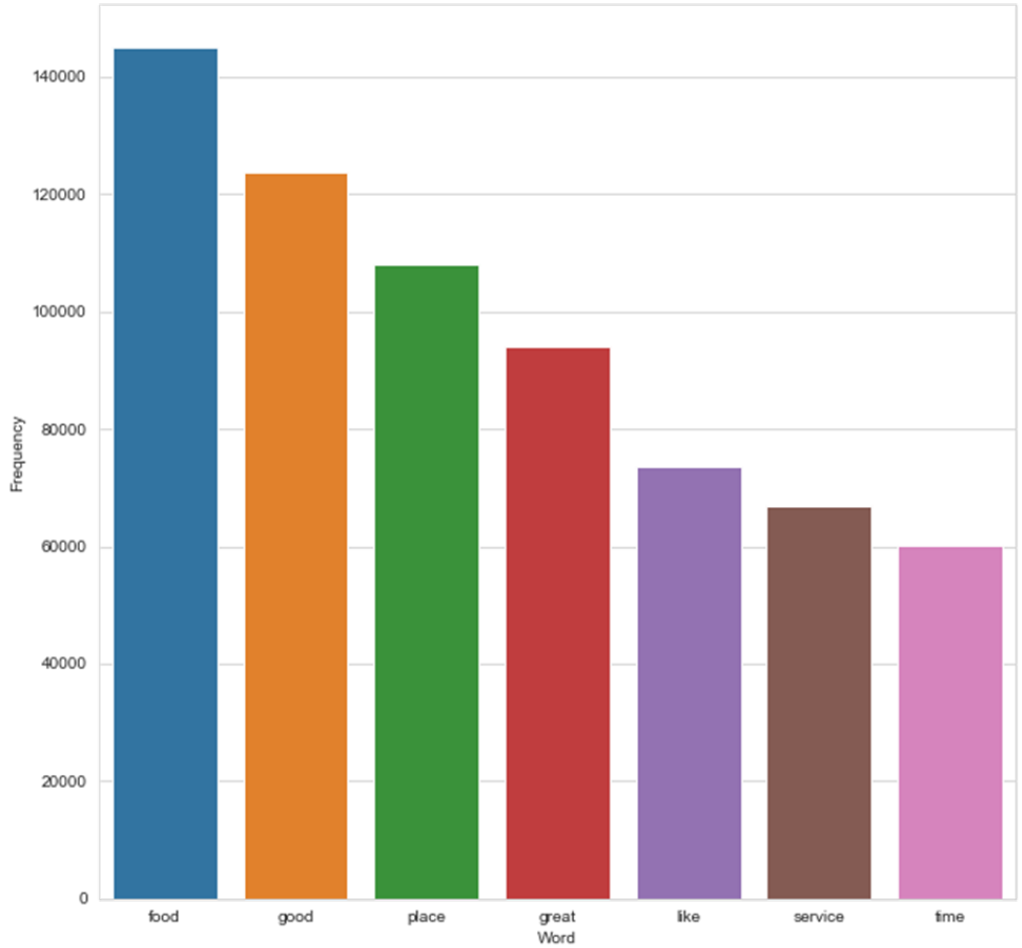


Figure 4.0 : Top Frequent Word Used In Review df

Model Preparation

Missing values replace with ‘unknown’ :

- Attributes - 13,744
- Categories – 103
- Hours – 23,223

Final dataframe:

- Business df : 150346 rows, 10 columns
- Initial Review df shape : 2,000,000 rows & 9 columns
- Final Review df shape : 895930 rows, 6 columns

Preprocess NLP including:

- Lemmatizing,
- Remove stopwords, symbols, indentation

```
This is nice little Chinese bakery in the hear...
This is the bakery I usually go to in Chinatow...
A delightful find in Chinatown! Very clean, an...
I ordered a graduation cake for my niece and i...
HK-STYLE MILK TEA:  FOUR STARS\n\nNot quite su...
```

```
this is nice little chinese bakery in the hear...
this is the bakery i usually go to in chinatow...
a delightful find in chinatown! very clean, an...
i ordered a graduation cake for my niece and i...
hk style milk tea: four stars not quite sure w...
```

Model Preparation

Table 3.0 : Business df features

Features	Datatype	Null Values
business_id	string	0
name	string	0
address	string	0
city	string	0
state	string	0
postal_code	string	0
latitude	float	0
longitude	float	0
stars	float	0
review_count	integer	0
is_open	integer	0
attributes	string	13744
categories	string	103
hours	string	23223

Extract subset of attributes
Convert to columns
And dummified

```
'BusinessAcceptsCreditCards', 'BusinessParking', 'BikeParking',
'RestaurantsPriceRange2', 'RestaurantsTakeOut', 'ByAppointmentOnly',
'WiFi', 'Alcohol', 'Caters', 'RestaurantsReservations',
'RestaurantsGoodForGroups', 'RestaurantsAttire', 'HasTV', 'Ambience',
'GoodForKids', 'GoodForMeal', 'NoiseLevel', 'DogsAllowed', 'HappyHour',
'WheelchairAccessible', 'RestaurantsTableService', 'Smoking', 'Music',
'GoodForDancing', 'BusinessAcceptsBitcoin', 'Corkage', 'BYOBCorkage',
'BestNights', 'CoatCheck', 'BYOB', 'DriveThru', 'Open24Hours'],
```

Model Preparation

A. Predict Rating Using Reviewer' comment

1. Check unbalance ranking
2. Perform SMOTE to balance
3. Naïve Bayes & Random Forest Model
4. X = 'text' features
5. Y = 'reviewer rating' label
6. Train Test Split → train: 75% & test: 25%
7. Transformers: Count Vectorizer & Tfidf Vectorizer

Table 4.0 : Value Count of Each Rating

Rating	Total
1	30476
2	15609
3	14853
4	33046
5	92750

Model Preparation

A. Predict Rating Using Reviewer' comment

Table 5.0 : Estimators with transformers and their parameter setup

Model Name	Transformer	Parameters	Values	Train Score	Test Score
Multinomial NB	Tfidf Vectorizer	max_df	0.9	0.70	0.69
		max_features	10000		
		min_df	2		
		ngram_range	(1, 2)		
		stop_words	None		
Multinomial NB	Count Vectorizer	max_df	0.9	0.69	0.66
		max_features	10000		
		min_df	3		
		ngram_range	(1, 2)		
		stop_words	english		

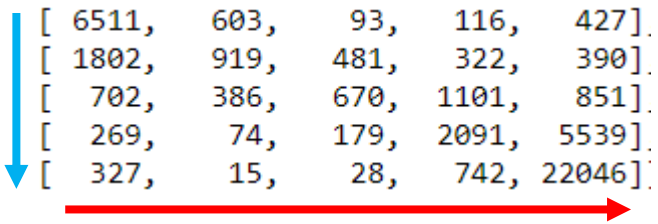
Model Evaluation

A. Predict Rating Using Reviewer' comment

Table 6.0 :Naïve Bayes Model with Tfidf Transformer score

	precision	recall	f1-score	support
1.0	0.68	0.84	0.75	7750
2.0	0.46	0.23	0.31	3914
3.0	0.46	0.18	0.26	3710
4.0	0.48	0.26	0.33	8152
5.0	0.75	0.95	0.84	23158
accuracy			0.69	46684
macro avg	0.57	0.49	0.50	46684
weighted avg	0.65	0.69	0.65	46684

Table 7.0 :Confusion Matrix of Naïve Bayes Model



Model Testing

A. Predict Rating Using Reviewer' comment

```
# Prediction example
print('Actual Rating:', df.iloc[652]["stars"])
new_text = [df.iloc[652]["text_fix"]]
print('Review: ', new_text[0])
text_features = tfidf_vect.transform(new_text)
nbPredict = nbModel.predict(text_features)
rfPredict = rfModel.predict(text_features)
print('-------------------------------------------------------')
print('Naive Bayes Prediction: ', nbPredict[0])
print('Random Forest Prediction: ', rfPredict[0])
```

```
Actual Rating: 5.0
Review:  seriously the best breakfast i have had in a long time it is pricey but everything here in sb is breakfast bun, choc c
roissant and the yogurt dear jesus excellent cappuccino as well go there now
-------------------------------------------------------
Naive Bayes Prediction:  5.0
Random Forest Prediction:  5.0
```

Figure 5.0 : Snapshot of Model Testing on Test Dataset

Model Preparation

B. Predict Rating Using Business Attributes

1. Create new features that contain text polarity and Label each reviewer's rating into:
   - Less than 3 is negative
   - More than 3 is positive
   - Equals 3 is neutral

Table 8.0 :Sentiment Polarity Score Applied On Review df

neg	neu	pos	compound	polarity	business_id	score
0.043	0.723	0.234	0.939	positive	MTSW4McQd7CbVtyjqoe9mw	4.0

2. Transform categorical columns to numerical using label encoder

Table 9.0 :Transformed Review df's attributes columns

review_id	review_stars	...	Corkage_None	Corkage_True	Open24Hours_False	Open24Hours_True
BXQcBN0iAi1IAUxibGLFzA	4.0	...	0	0	0	0

Model Evaluation

B. Predict Rating Using Business Attributes

Table 10.0 : Score of Models on Not Normalized Rank Data

Classifier	F1	Precision	Recall	Accuracy
GradientBoostingClassifier	0.361	0.374	0.392	0.392
LogisticRegression	0.358	0.380	0.394	0.394
RandomForestClassifier	0.320	0.375	0.378	0.378

Table 11.0 : Score of Models on Normalized Rank Data

Classifier	F1	Precision	Recall	Score Train	Score Test
GradientBoostingClassifier	1.000	1.000	1.000	1.000	1.000
LogisticRegression	1.000	1.000	1.000	0.998	0.997
RandomForestClassifier	0.721	0.658	0.805	0.811	0.809

Table 12.0 :Count of Rating Based on Polarity

Polarity	Rating	Total
Negative	1	15776
	2	15964
Neutral	3	15488
Positive	4	16959
	5	16071

Model Preparation

C. Recommender System

1. Combine attribute from Business dataset with review id
2. Attributes about 116 columns used
3. Drop duplicate business name
4. Change name as business index

Table 13.0 : Dataframe With Index As Label And Attributes As Columns

name	overall_stars	review_count	is_open	review_stars	useful	garage	street	validated	lot	valet
St Honore Pastries	4.0	80	1	4.0	0.0	0	1	0	0	0
Core de Roma	5.0	12	1	5.0	0.0	0	0	0	1	0
Wawa	3.0	56	1	4.0	0.0	0	1	0	0	0
Golden Chopstick Chinese Restaurant	3.0	137	1	3.0	0.0	0	1	0	0	0
Rita's Italian Ice	4.0	7	1	5.0	0.0	0	0	0	1	0

Model Testing

C. Recommender System

```
df_sim['Turning Point of North Wales'].sort_values(ascending=False)
```

```
name
Turning Point of North Wales    1.0
Apollo's Family Pizzeria        1.0
Port of Subs                    1.0
Ricuras de Venezuela            1.0
Testa's Bakery                  1.0
```

Figure 5.0 : Recommend of Similar Restaurant Without Attributes Influence

```
df_sim['Turning Point of North Wales'].sort_values(ascending=False)
```

```
name
Turning Point of North Wales                    1.000000
Elevation Burger                                0.999713
Oregon Diner                                    0.999706
Persian Grill                                   0.999688
Bacco Italian Restaurant                        0.999666
```

Figure 6.0 : Recommend of Similar Restaurant With Attributes Influence

CONCLUSION & RECOMMENDATION

1. In multiclass classification prediction, gradient bossting classifier perform the best with F1 score of 1 .
2. For prediction using text input, Multinomial Naive Bayes perform at moderately average at 0.69.
3. Increasing bag of word could help increasing Multinomial performance but computational heavy. Bias in review itself is also factor that reduce the score [3].
4. For future work, is a possibility to combine multiclass approach which use attributes and combine with bag of words to further increase model performance to predict rating.
5. This is shown in recommender system too as more variance introduce(attributes introduce), more relevant restaurant are recommended.

REFERENCES

1. Failory.com. What Was Sprig? Available online: https://www.failory.com/cemetery/sprig (accessed on 10 May 2022).
2. Adak, Anirban & Pradhan, Biswajeet & Shukla, Nagesh. (2022). Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review. Foods. 11. 1500. doi:10.3390/foods11101500.
3. Y. Jiang, "Restaurant Reviews Analysis Model Based on Machine Learning Algorithms," 2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID), 2020, pp. 169-178. doi: 10.1109/MSIEID52046.2020.00038.
4. H.L. Meiselman, R. Bell, in Encyclopedia of Food Sciences and Nutrition (Second Edition), 2003

THANK YOU