BIG DATA TOOLS 2 GROUP PROJECT

databricks

Submitted By (GROUP 9) –

BHURJI Prineet Kaur

BRUNO Théau

M G Bindhu

Table of Contents

1. Project Definition

1.1. Case Background

The case picked up for this project falls under the category of negative impacts that were faced by many industries post covid-19 pandemic lockdowns. While it hit badly almost every industry somehow or the other, there were few industries that became hard victims of this situation. One of such industry and our topic of project was the leisure industry.

Many entities like the restaurants, bars, cafes, and other social gathering spots had no option but to take the hit. While some had to permanently close, few tried to cope with the situation by using innovative ways to keep their customers with them. Some used ways to enhance their online delivery channel while some used online tools like Yelp (Yelp.com) to understand their customers better and service them accordingly.

1.2. About the Company

Yelp is an American public company headquartered in San Francisco, California. Yelp's website, Yelp.com, is a crowd-sourced local business review and social networking site. The site has pages devoted to individual locations, such as restaurants or schools, where Yelp users can submit a review of their products or services using a one-to-five-star rating system. Businesses can also update contact information, hours, and other basic listing information or add special deals. In addition to writing reviews, users can react to reviews, plan events, or discuss their personal lives.

1.3. Problem Statement

Through this project our aim was to study the data provided by Yelp and help them identify patterns of behaviour that are more likely relatable to a business's capability to adopt innovative and digital transformations. To build a prediction model for their Data Science team to identify and predict what factors lead some businesses to start doing delivery or takeout for the first time after the first lockdown. Further, for their Business Development team think and come up with creative ways of how Yelp could target their communications to these selected set of businesses. All this can help them use our findings as a foundation to build their advertising solutions strategy.

2. Project Approach

We were provided with 6 datasets – **Business, Checkin, Covid, Reviews, Tip and Users**

2.1. Table Descriptions

COVID dataset

Shape – `(19053, 9)`

The dataset has variables – *`Call To Action enabled, Covid Banner, Grubhub enabled, Request a Quote Enabled, Temporary Closed Until, Virtual Services Offered, business_id, delivery or takeout, highlights`*

USERS dataset

Shape – `(305084, 22)`

The dataset has variables – *`name, review_count, useful, user_id, average_stars, compliment_cool, compliment_cute, cool, elite, fans, friends, funny etc`*

REVIEWS dataset

Shape – (500000, 9)

The dataset has variables – *business_id, cool, date, funny, review_id, stars, text, useful, user_id*

BUSINESS dataset

Shape – (19018, 58)

The dataset has variables – *address, attributes.AcceptsInsurance, attributes.AgesAllowed, attributes.Alcohol, attributes.Ambience, is_open, latitude, longitude, name, postal_code, review_count, stars, state etc*

CHECKIN dataset

Shape – (1990914, 2)

The dataset has variables – *business_id, date*

TIP dataset

Shape – (124161, 5)

The dataset has variables – *business_id, compliment_count, date, text, user_id*

2.2. Steps Included

2.2.1. DATA IMPORTING

Imported all datasets one by one and dropped all rows with null values.

2.2.2. DATA PROCESSING

COVID dataset –

```
+----------------+--------------------------+
|count(business_id)|count(DISTINCT business_id)|
+----------------+--------------------------+
|           19053|                     19018|
+----------------+--------------------------+
```

- Checked for the duplicate values and dropped them.
- Did One-hot encoding with TRUE as 1 and FALSE as 0
- Renamed the target columns to label for the machine learning purpose.
- Created the target table to merge.
- Created the covid table features subset.

BUSINESS dataset –

```
+--------+-----+
|Delivery|count|
+--------+-----+
|    None|   78|
|   False| 3312|
|    null|13639|
|    True| 1989|
+--------+-----+
```

```
+-------+-----+
|TakeOut|count|
+-------+-----+
|   None|   12|
|  False|  599|
|   null|12694|
|   True| 5713|
+-------+-----+
```

- Renamed all the columns for better business understanding.
- Analysed the Delivery column to check for the count of each of the delivery types.
- Analysed the TakeOut column to check for the count of each of the TakeOut types.
- Replaced NaN with False on the columns TakeOut and Delivery.
- Filtered the business_id which never did Delivery and TakeOut.
- Checked for the duplicate values and realized there was nothing to drop here.

```
+----------------+--------------------------+
|count(business_id)|count(DISTINCT business_id)|
+----------------+--------------------------+
|           13153|                     13153|
+----------------+--------------------------+
```

- Filled all missing values with NaN and did One-hot encoding with TRUE as 1 and FALSE as 0

TIP dataset –

```
+----------------+--------------------------+
|count(business_id)|count(DISTINCT business_id)|
+----------------+--------------------------+
|          124161|                     12578|
+----------------+--------------------------+
```

- Checked for the duplicate values and dropped them.
- Converted the datetime type to a numerical value.
- Did feature engineering, for each business created min & max count and tip average time of tip

REVIEW dataset –

- Checked for NaN values and dropped columns that were not needed.
- Converted Date to a proper format.
- Created recency column using most_recent_date col with spark.
- Rounded the avg_starts values.

CHECKIN dataset –

- Dropped all rows with NaN values.
- Cleaned the Date to extract the Year values and further created year & month columns.
- Did feature engineering (using SQL) to create min max count for Year values.

2.2.3. MERGING THE DATASETS

Performed merging for "checkin_features", "reviews_cleansed", "tip_final", "covid_features", "business_target" by joining them all on "business_id"

2.2.4. MODEL BUILDING

Performed **feature selection** using ChiqSelector, thereby reducing the features from 83 to 50. Further, used following 3 algorithms to build the prediction model –

- Decision Tree

```
Test Error = 0.0434993
DecisionTreeClassificationModel: uid=DecisionTreeClassifier_64647fdb789d,
depth=5, numNodes=35, numClasses=2, numFeatures=50
```

- Random Forest

```
Test Error = 0.0393179
RandomForestClassificationModel: uid=RandomForestClassifier_c51cb7fc7083,
numTrees=10, numClasses=2, numFeatures=50
```

- Gradient Boosting

```
Test Error = 0.0448113
GBTClassificationModel: uid = GBTClassifier_fdfc933edb29, numTrees=10,
numClasses=2, numFeatures=50
```

We went ahead with performing a **k-fold Cross validation (with k=10)** and the final AUC scores for the Models used were as follows –

Model	AUC Score
Decision Tree	0.56409
Random Forest	0.51155
Gradient Boosting	0.54906

3. Main Conclusions

The Yelp and covid datasets provide a quantitative look at how restaurants are managing during the pandemic.

Even though because of pandemic many businesses have faced huge amount of loss. There were still new openings especially for the businesses like restaurants, food trucks, food delivery companies, etc. Likewise, many businesses reopened and adapted to the new situation.

We can see in the below insights and its quite possible that even during the pandemic situation people still prefer to have the restaurants open and if the businesses are ready to adapt then they can still earn profit and can stay connected to the customers.

The data from Yelp shows that these restaurants have been almost 50% more likely to close compared to other restaurants during the pandemic, but there is still a chance if they are ready to adapt to new situation.

There has still been an economic impact but is manageable for many small-scale restaurants who are ready for new delivery/Take out.

We can see in the insights, when it comes to business that were open during covid, New York shows the lowest count that might be because of the frequency of covid hits in the state. This could be improved by taking appropriate measures and following some of the business actions recommended below 5.

It is also one of the reasons why Arizona has been on the top list in business open during covid situations 6.

4. Relevant Insights

Analysing the count of businesses that were open during covid situation state wise. In the below image, we can see Arizona has highest number of businesses open and New York to be the lease.

state	all
AZ	4270
SC	63
QC	473
NV	2884
WI	382
CA	3
NC	996
IL	111
OH	940
PA	700
NY	1
ON	1837
AB	493

Features importance ranking: From our best model, this is the average number of stars that have the most important impact with 9.61. The 5 first features represents only 40% of importance in the model, we do not have major features with a huge impact which explain the low AUC.

	Feature_number	Feature_importance	Column_name
7	7	9.61	avg_stars
0	0	8.26	count_checkin
6	6	7.93	sum_useful
8	8	7.76	review_recency
21	21	6.89	num_of_categories
3	3	5.17	checkin_diff
14	14	5.13	Call To Action enabled
2	2	3.88	max_year
43	43	3.17	AppOnly_Missing
10	10	2.61	max_date_diff
13	13	2.42	diff_date_diff
4	4	2.41	sum_cool
9	9	2.14	min_date_diff
81	81	1.99	state_ON
35	35	1.98	BikeParking_Missing

Ranking delivery per state: As mentioned above about the businesses, Arizona has highest ranking.

state	new	all
AZ	104	4270
SC	4	63
QC	27	473
NV	51	2884
WI	11	382
NC	34	996
IL	2	111
OH	37	940
PA	21	700
ON	76	1837
AB	22	493

Percentage of new delivery/TakeOut comparing to the total number of business per state: South Carolina has highest percentage of new delivery/take out.

state	new	all	percentage
AZ	104	4270	2.4355971896955504
SC	4	63	6.349206349206349
QC	27	473	5.708245243128964
NV	51	2884	1.768377253814147
WI	11	382	2.8795811518324608
NC	34	996	3.413654618473896
IL	2	111	1.8018018018018018
OH	37	940	3.9361702127659575
PA	21	700	3.0
ON	76	1837	4.137180185084377
AB	22	493	4.462474645030426

Ranking with the best improvement: The businesses in South Carolina showed highest percentage of improvement compared to other stated mentioned below.

state	new	all	percentage
SC	4	63	6.349206349206349
QC	27	473	5.708245243128964
AB	22	493	4.462474645030426
ON	76	1837	4.137180185084377
OH	37	940	3.9361702127659575
NC	34	996	3.413654618473896
PA	21	700	3.0
WI	11	382	2.8795811518324608
AZ	104	4270	2.4355971896955504
IL	2	111	1.8018018018018018
NV	51	2884	1.768377253814147

Number of reviews per state: The businesses in state Nevada got the highest number of review count and New York being the least.

state	sum(review_count)
NV	119892
AZ	90327
ON	23716
NC	14510
OH	10686
PA	8388
QC	7073
WI	5130
AB	4127
IL	1186
SC	467
CA	24
NY	3

5. Recommended Business Actions

- ***Communicate effectively (Take out/New delivery):***
  Updating the business profile, with location changes (if any), modified hours, locations on google which can appear on google search, information about their needs, if the restaurant offers Take out/ delivery.

- ***Social Advertisements/Online presence:***
  Advertising about their business on social media like on Facebook, Instagram, Television, Banners, etc.

- ***New Social Media channels:***
  Opening account and creating a new channel for updates on platforms like YouTube to gain customers attention.

- ***Offering Rewards:***
  Like using an application to order the food instead of website, offer on multiple meals, etc.

- ***Assuring Strict Regulations that are followed:***
  Assuring the customers about the strict regulations and precautions that is performed for their safety in the businesses which helps in grabbing customer's attention.

- ***Play around with colours:***
  When used effectively, they can affect consumer behaviour too. A quick look at different restaurant ads around you will reveal a trend. You will notice that ads for fast food chains tend to use brighter colours that reflect energy.

- ***Be selective with the photos you upload:***
  People eat with their eyes. When it comes to food photography, there are a lot of variables that go into taking a great shot.

6. Sources

- Yelp — Wikipédia (wikipedia.org)
- https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_City
- https://www.usnews.com/news/best-states/arizona/articles/2021-01-03/arizona-hits-new-daily-high-in-covid-19-cases-but-no-deaths
- https://kjzz.org/content/1665257/arizona-covid-19-cases-hit-4-month-low
- https://www.wordstream.com/blog/ws/2017/03/03/food-advertising
- https://www.lightspeedhq.com/blog/restaurant-marketing-how-to-attract-customers/