YELP DATA ANALYSIS

ABSTRACT

Vaishali Siddeshwar

SCS_3252_019 Big Data Management Systems & Tools

Table of Contents

1. Objective

Yelp is a public crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores.

Yelp was founded in 2004 by former PayPal employees. By 2010 it had $30 million in revenues and the website had published more than 4.5 million crowd-sourced reviews.

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. The dataset has information about businesses across 11 metropolitan areas in four countries.

The goals of this project are:

1. To assess the public perception of restaurants on Yelp via exploratory data analysis in Spark. Understanding which attributes play a key role in making a flourishing business is essential in driving revenue growth for existing business owners, future business owners to make important decisions regarding new business or business expansion
2. To build a machine learning model which accurately predicts the rating by calculating the sentiment of the reviews.

2. Import and Clean Data

The data consists of 4 json files:

1. business.json contains business data including location data, attributes, and categories.
2. review.json contains full review text data including the user_id that wrote the review and the business_id the review is written for.
3. user.json contains user data including the user's friend mapping and all the metadata associated with the user.
4. checkin.json consisting of checkins on a business.

All the files are in s3 that was mounted locally for easy access. The below picture shows the number of samples in each dataframe.

```
number of rows in business table = 192609
number of rows in users table = 1637138
number of rows in reviews table = 6685900
number of rows in checkin table = 161950
```

The below picture shows the number of nulls in yelp_business table.

address	attributes	business_id	categories	city	hours	is_open	latitude	longitude	name	postal_code	review_count	stars	state
0	28836	0	482	0	44830	0	0	0	0	0	0	0	0

From the above code, "attributes" and "hours" have some null values. As I am not using them in EDA, I will drop them. The below picture shows the number of nulls in yelp_review table.

business_id	cool	date	funny	review_id	stars	text	useful	user_id
0	0	0	0	0	0	0	0	0

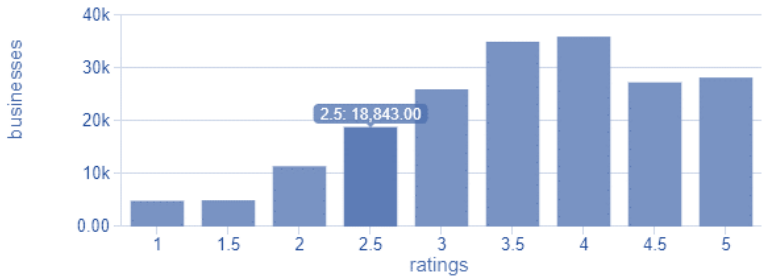
As review table has no nulls. No processing is done.

3. Data Analysis

The code for this section is at https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5818954073065180/1640118647309999/3954267115134911/latest.html

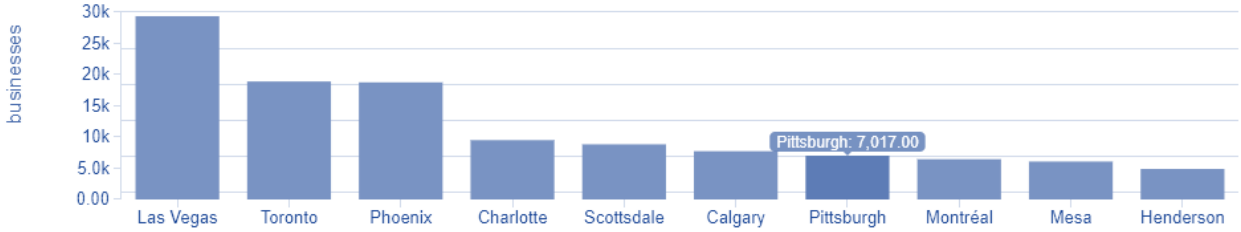
3.1. Rating Distribution

Below is the distribution of stars to understand its range and frequency. From the above graph we see that we can find the most of user give 3.5 or 4 stars to the different businesses.



3.2. Which city has the most reviews?

Below is a plot of top 10 cities that have maximum number of businesses. Notice that Toronto hosts the second highest number of businesses.



3.3. What is the average rating of a business in top 10 cities?

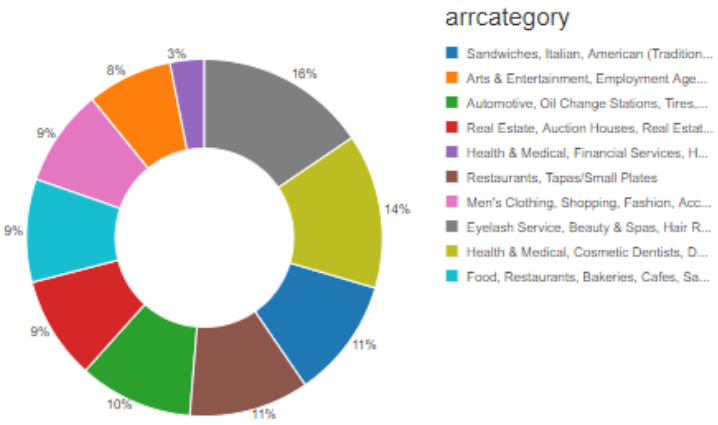
Below is the list of average rating for business in top 10 cities. We can see that an average rating for any business is 3.3 to 3.6 for most of the businesses in top cities.

city	reviewcounts	avgstars
Las Vegas	124686	3.7825216944965754
Phoenix	78424	3.733895236152198
Toronto	73215	3.5024653418015435
Charlotte	39964	3.61006155539986
Scottsdale	39601	3.9997727330117927
Calgary	30361	3.4300582984750174
Pittsburgh	28364	3.6431744464814555
Mesa	25117	3.6850937611975954
Montréal	24076	3.7410284100348896

3.4. Top 10 businesses that have good ratings

Below are the steps used to find the categories of businesses with good reviews.

1. flatten the categories from array of strings to multiple strings in yelp_business dataframe.
2. Use %sql command to generate the visualizations.



From the above graph, we see that the dataset is dominated by many positive reviews for restaurants and fast food joints and beauty salons. So, it would be good idea to invest in these businesses. from the above list we can infer that

1. Restaurants and Beauty Salons and restaurants have mostly positive reviews.
2. that most of the "Casino" are negatively reviewed.

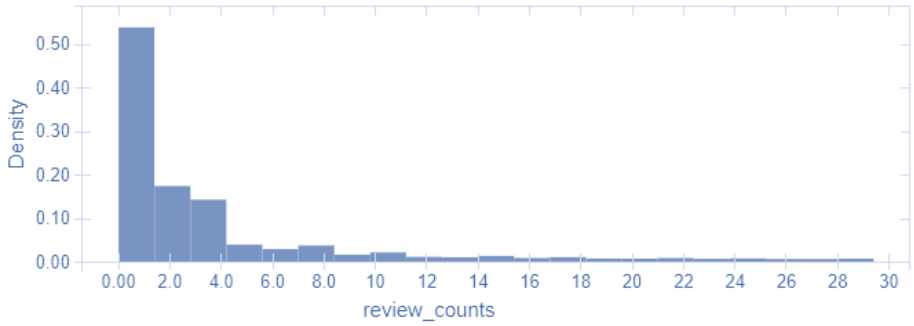
3.5.Which are the top 5 flourishing businesses in Toronto?

The same trend is seen in Toronto as well. Businesses in Food industry have lots of reviews and the average rating 2.2 to 3.4

name	reviewcounts	avgstars
Starbucks	155	3.41290322
Tim Hortons	98	2.73979591
Shoppers Drug Mart	60	2.88333333
Subway	48	2.3125
McDonald's	47	2.25531914

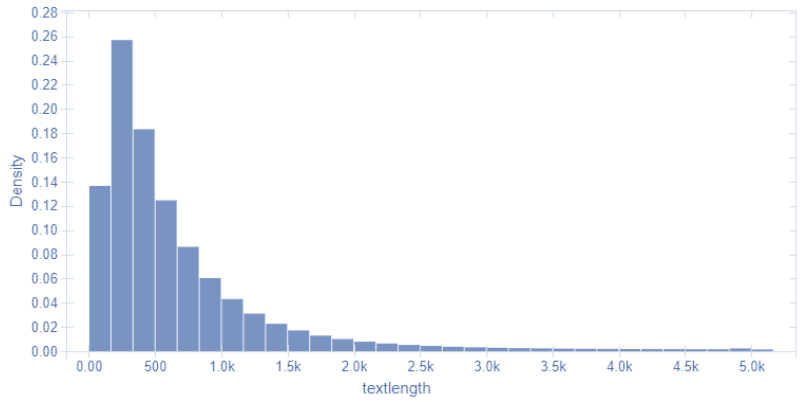
3.6.Average number of reviews by a user

From the below histogram it is clear that around 80% of the users write only about 2-4 reviews.



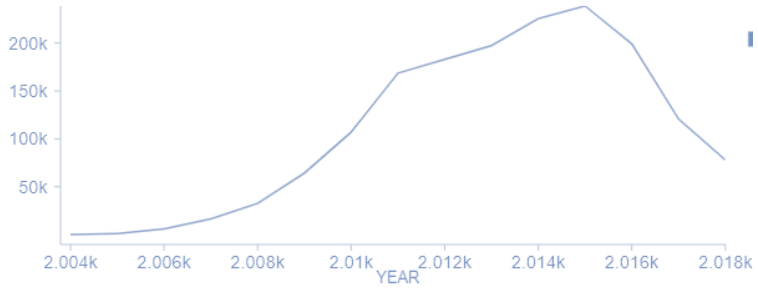
3.7.what is the average length of a review?

The above histogram counts the number of words in the review text and plots it's distribution. We can see that an average review is around 160 to 300 words.



3.8. Distribution of users that have been using yelp.

Below is the Distribution of users that have been using yelp. The users table has a "yelping_since" column which will be helpful for us here. Notice that user growth reaches its peak around 2014, plummeted around 2016, and been constantly dropping until 2017. As we can see, the growth is slowing down to about 10% YoY in 2017. Is 10% Year-on-year growth good or not? We cannot expect to get conclusion only from yelp data, we need to compare with other similar social media platform growth and also yelp's target. For many social media, user growth declining is expected due to market saturation.



3.9. Who are the top users on yelp and how many reviews have they given?

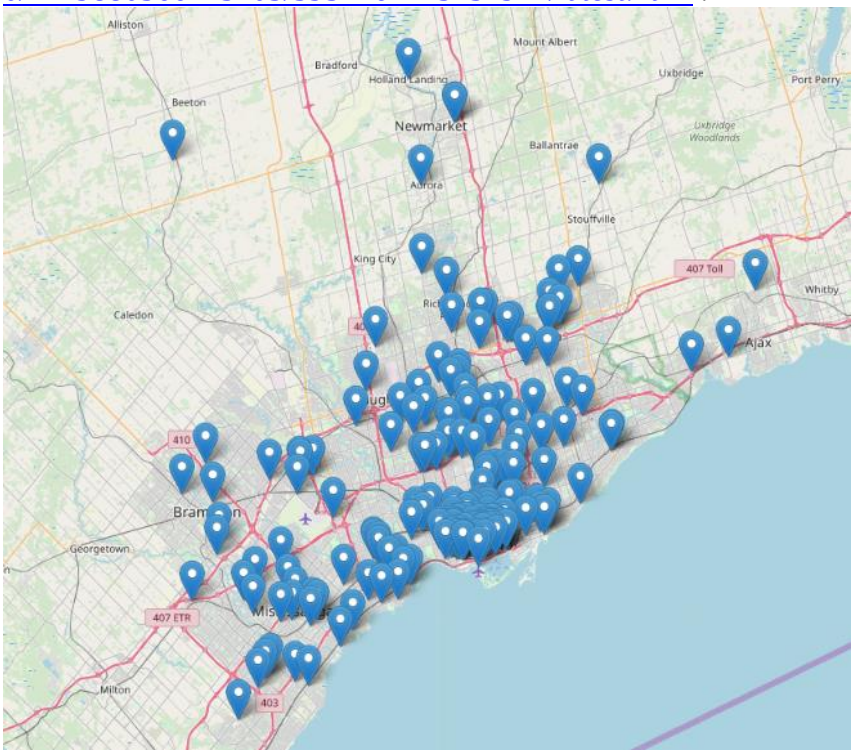
Now we can to find out if users give more reviews. For this analysis, we need to use the review dataset to find the date of review by each user. Below is the list obtained.

user_id	review_counts
CxDOIDnH8gp9KXzpBHJYXw	4129
bLbSNkLggFnqwNNzzq-Ijw	2354
PKEzKWv_FktMm2mGPjwd0Q	1822
ELcQDlf69kb-ihJfxZyL0A	1764
DK57YibC5ShBmqQl97CKog	1727
U4INQZOPSUaj8hMjLlZ3KA	1559
QJI9OSEn6ujRCtrX06vs1w	1496
d_TBs6J3twMy9GChqUEXkg	1360

It is interesting to see that a user has given 4129 users. Let us dig more into this.

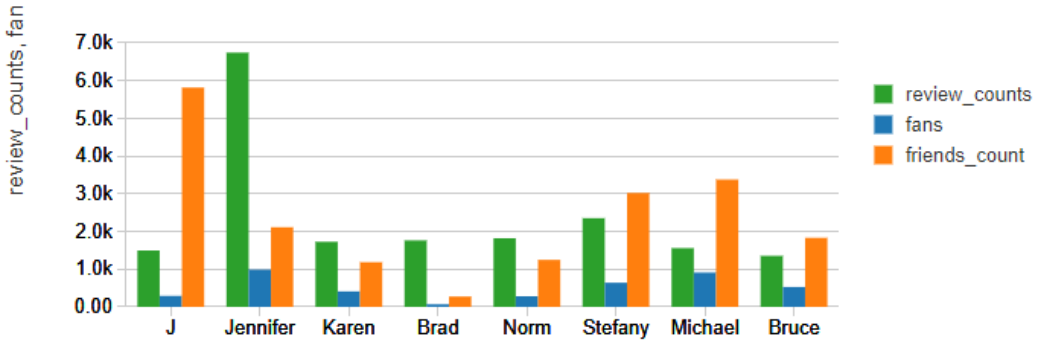
3.10. Knowing more about the top user with 4129 reviews.

It is interesting to see that a user has given 4129 users. Let us dig more into this. As the longitude and latitude of businesses are available, below is the map of locations in reviewed by this user. This suggests that most of the businesses reviewed by this person are right here in GTA. The code for this logic is in **folium_maps** notebook which is published at https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5818954073065180/4479880366423405/3954267115134911/latest.html .



3.11. Relationship between users' friends and review patterns

The users table also has friends and fans list. So, I think it would be good to see if we can notice any pattern between the user's reviews counts, number of friends and fans that he has.



From above graph we can see there is no defined relationship between user's fans, friends and review count. User reviewing for a business totally depends on user's character and/or satisfaction with business services.

3.12. Do user's friends influence business review?

I tried to find if there exists a user who gave a high rating to a business and has a friend who also gave a high rating to the same business. Here I have taken only 50 users out of all users into account because this task is very resource heavy. But, could not find such a user. So, I am not sure if I can say that user's friends influence business review.

3.13. Which are the top businesses within 10 miles radius of UofT?

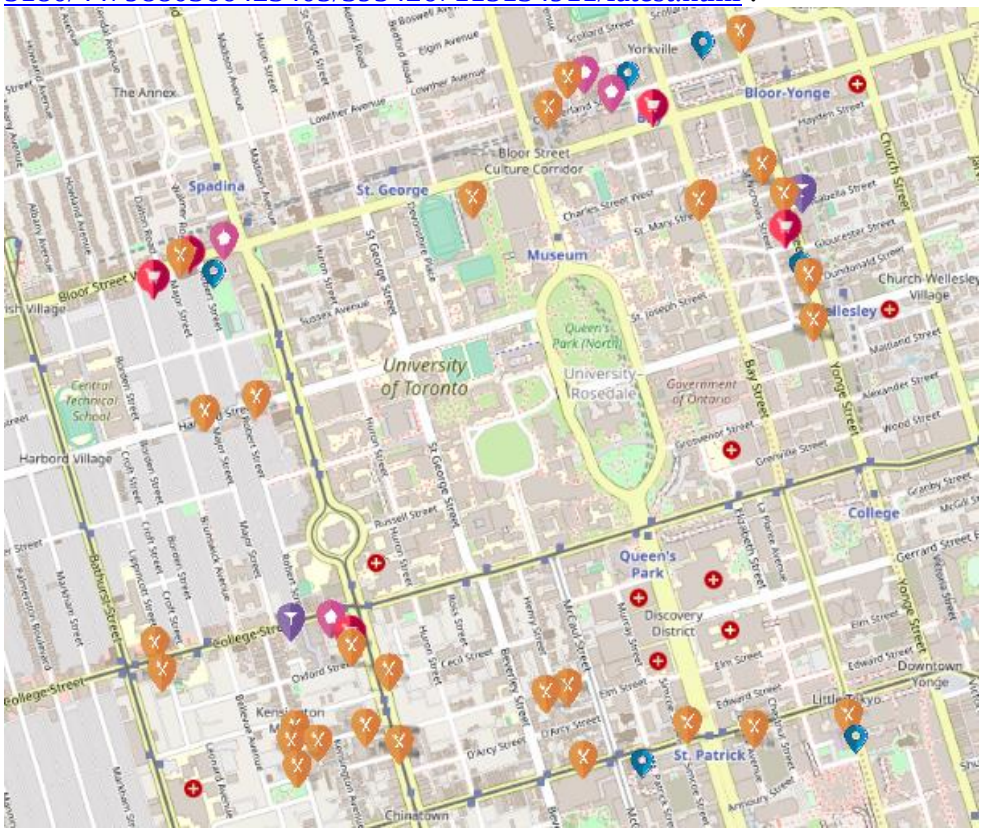
What are the top businesses within 10 miles of the University of Toronto St George? Center: University of Toronto - St.George Latitude: 43 39' 46.8" N, Longitude: 79 23' 44.8" W Decimal Degrees: Latitude: 43.663000, Longitude: --79.395764 The bounding box for this problem is ~10 miles, which we will loosely define as 10 minutes. So the bounding box is a square box, 20 minutes long each side (of longitude and latitude), with UofT at the center.

To do this, first load “yelp_academic_dataset_business.json” and collect “stars”, “categories”, “latitude”, “longitude”. For latitude, longitude explicit casting of float and double value was done to bypass the error of Type cast problem.

1. Flatten the categories, cast it as string, ignore NULL entries.
2. Calculated the 10 miles distance around the university using google distance measure feature and filtered the data for the specified bounds inside the %sql command.
3. Grouping was used to collect data for each category and then saved the details to a table for displaying geographical data.

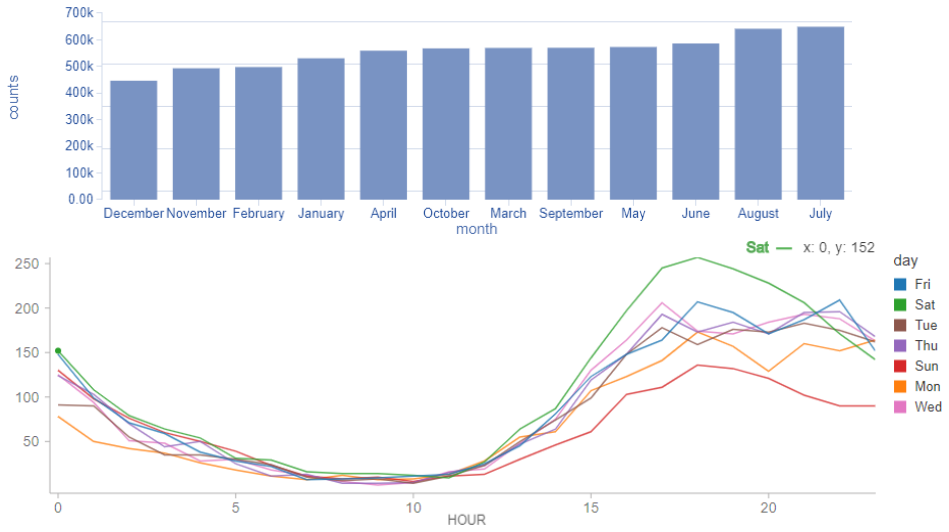
There exist nearly 50000 businesses within 10 miles radius of UofT. I have saved the locations of businesses reviewed by the top user in a table which I will be using to load a map. The code for populating the map is in **folium_map notebook** which is published at https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/581895407306

5180/4479880366423405/3954267115134911/latest.html .



3.14. Distributions of reviews by month and day.

Below 2 graphs are obtained by grouping the reviews table by month and day.



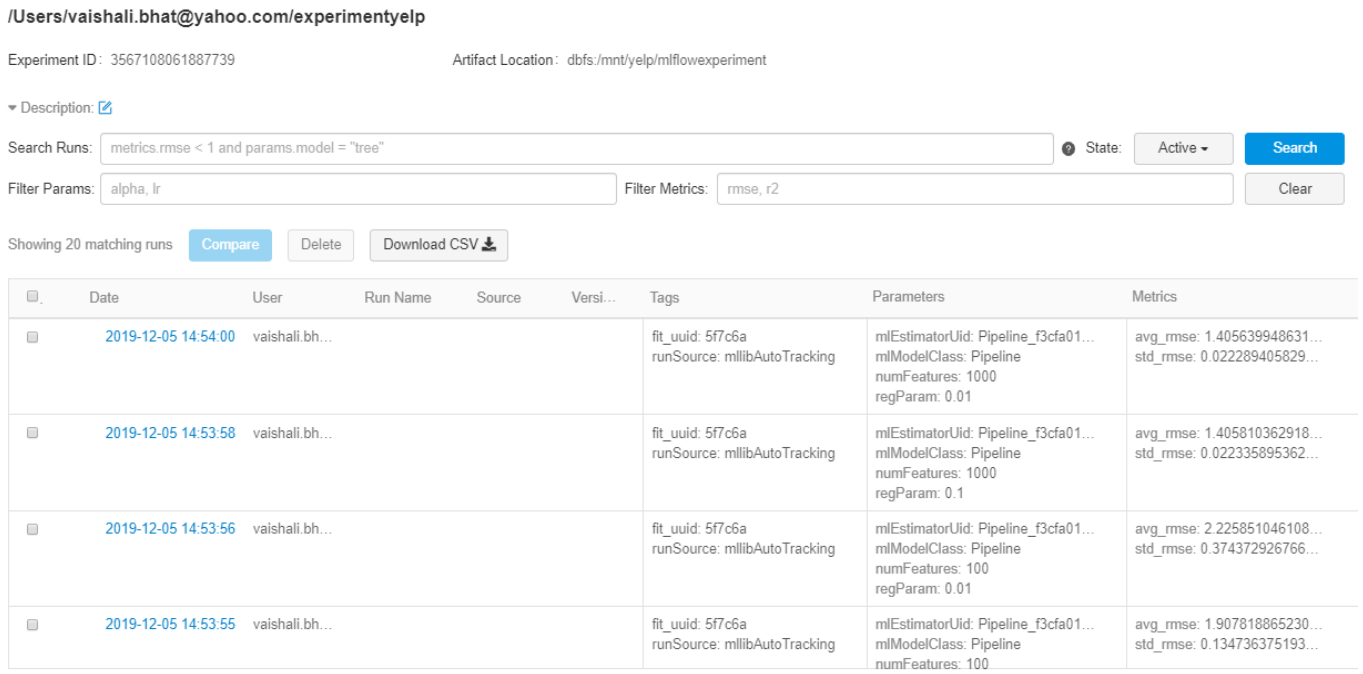
As expected, Fridays and Saturday between 3pm to 8 pm has the highest number of check-ins.

4. Predicting user rating using his/her review

The code for predicting the rating for a given review is helpful for businesses to improve. The code for prediction is in **yelp_sentiment notebook** which is published at https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5818954073065180/3567108061887726/3954267115134911/latest.html . I have used a Pipeline technique to chain 3 stages of data transformation and later perform Linear Regression. The steps followed are

1. The Pipeline is configured to have 3 steps - tokenizer, hashingTF, and Linear Regression.
2. A grid of Parameters to be optimized is also set. Currently, the parameters to be optimized are Learning rate and number of features. Totally there are 6 combinations of parameters are to be run on the train.

The MLFlow automatically logs the results of these combination and logs the desired metric. The metric that I had chosen here is RMSE. Below is the screenshot of the outcome of the experiments.



The best model has an RMSE of 0.004. As the data is very huge, I have not fine-tuned the model.

5. Conclusion

1. Las Vegas, Toronto and Phoenix have highest number of businesses.
2. Good idea to invest in Restaurants and beauty salons.
3. Users constantly dropping since 2016
4. Reviews are solely based on a user's satisfaction with business services.

6. References

https://www.kaggle.com/yelp-dataset/yelp-dataset/kernels
https://www.yelp.ca/toronto