YELP RESTAURANT RECOMMENDER ENGINE

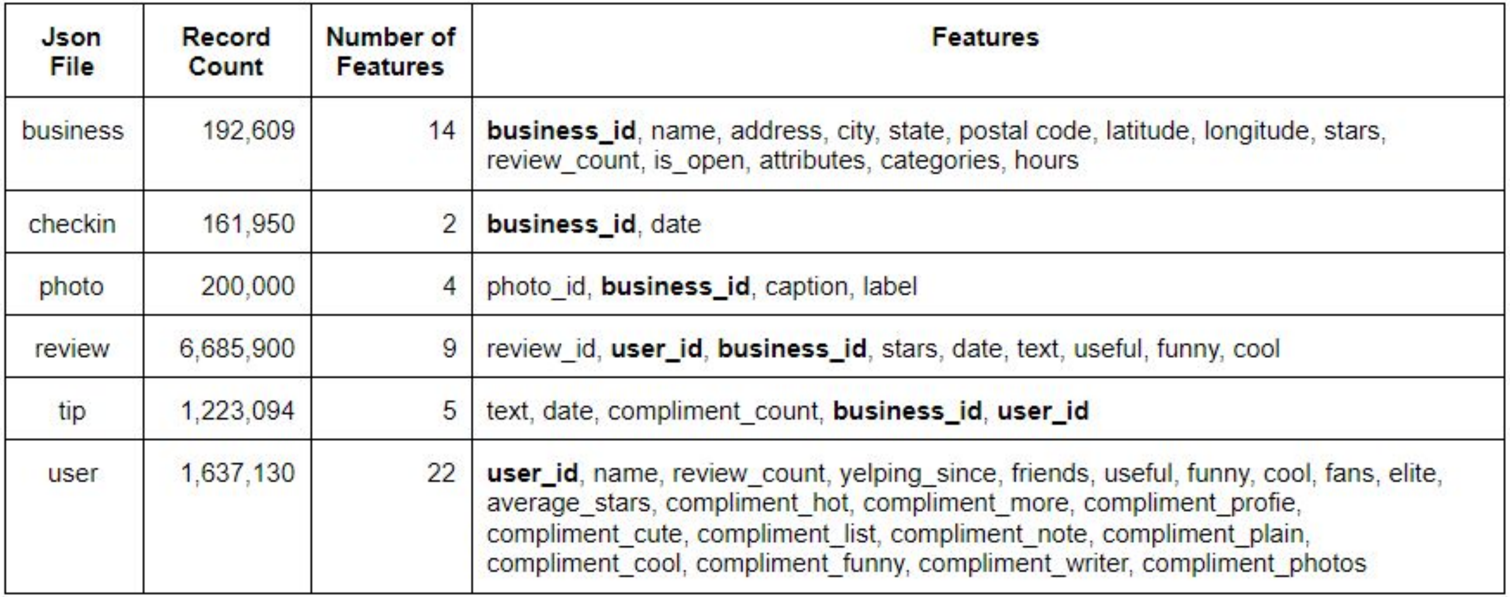
Capstone Project II | Nikki Seegars

PROBLEM STATEMENT

In any given city there are many restaurants that locals and tourists can patronize. In order to streamline users choices, a good recommender engine for Yelp, an online crowd sourced review website, will help to keep and drive more users to its platform.

DATA SOURCES

Yelp has a public dataset available on its website for data analysis competitions and other academic purposes.



Json File	Record Count	Number of Features	Features
business	192,609	14	**business_id**, name, address, city, state, postal code, latitude, longitude, stars, review_count, is_open, attributes, categories, hours
checkin	161,950	2	**business_id**, date
photo	200,000	4	photo_id, **business_id**, caption, label
review	6,685,900	9	review_id, **user_id**, **business_id**, stars, date, text, useful, funny, cool
tip	1,223,094	5	text, date, compliment_count, **business_id**, **user_id**
user	1,637,130	22	**user_id**, name, review_count, yelping_since, friends, useful, funny, cool, fans, elite, average_stars, compliment_hot, compliment_more, compliment_profie, compliment_cute, compliment_list, compliment_note, compliment_plain, compliment_cool, compliment_funny, compliment_writer, compliment_photos

DATA WRANGLING

The following steps were taken to clean up the data:

- Loaded json files to pandas DataFrame
- Convert json to csv and save to file server
- Import csv files into PostgreSQL database
- Write query to summarize data
- Load query into DataFrame
- Update state designation for records incorrectly categorized

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was conducted by visually exploring the data to gather insights and trends from summary statistics.

Below are the features that were initially reviewed:

- Open and closed restaurants
- Restaurant count by state
- Average star ratings by metropolitan area
- Restaurant types (categories)
- Star ratings by increasing count

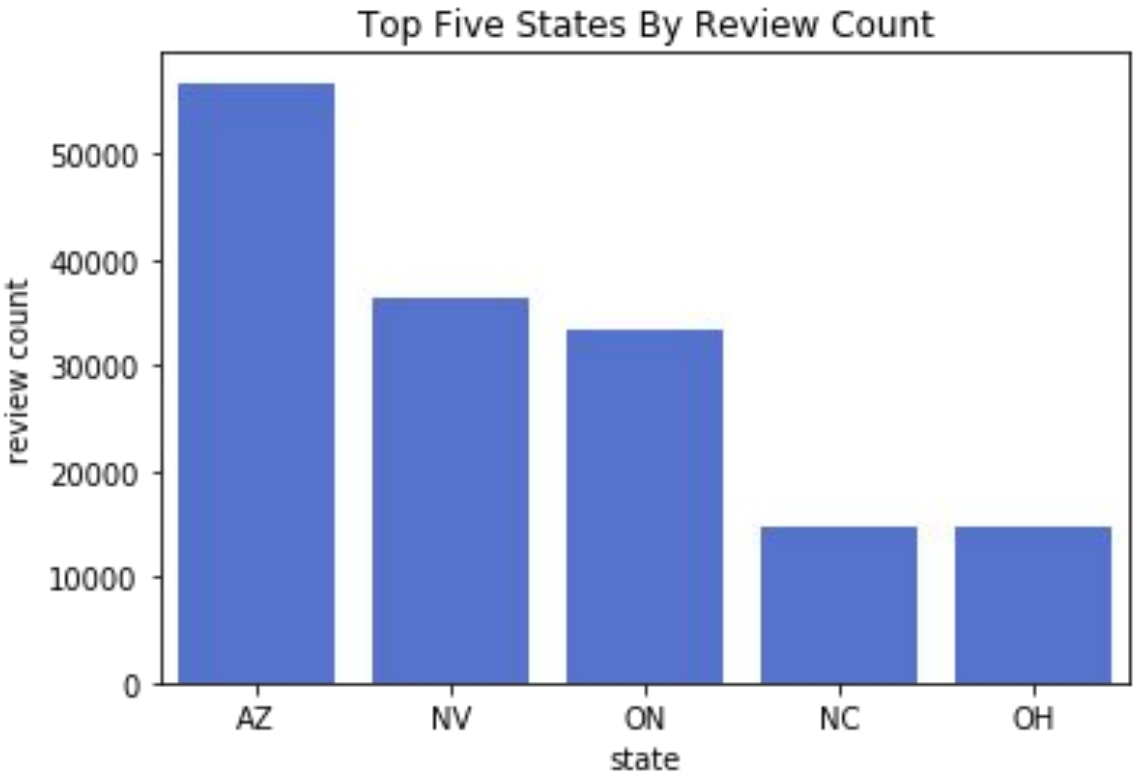
EXPLORATORY DATA ANALYSIS

Some restaurants are closed and were not included in the analysis.

193k total restaurants

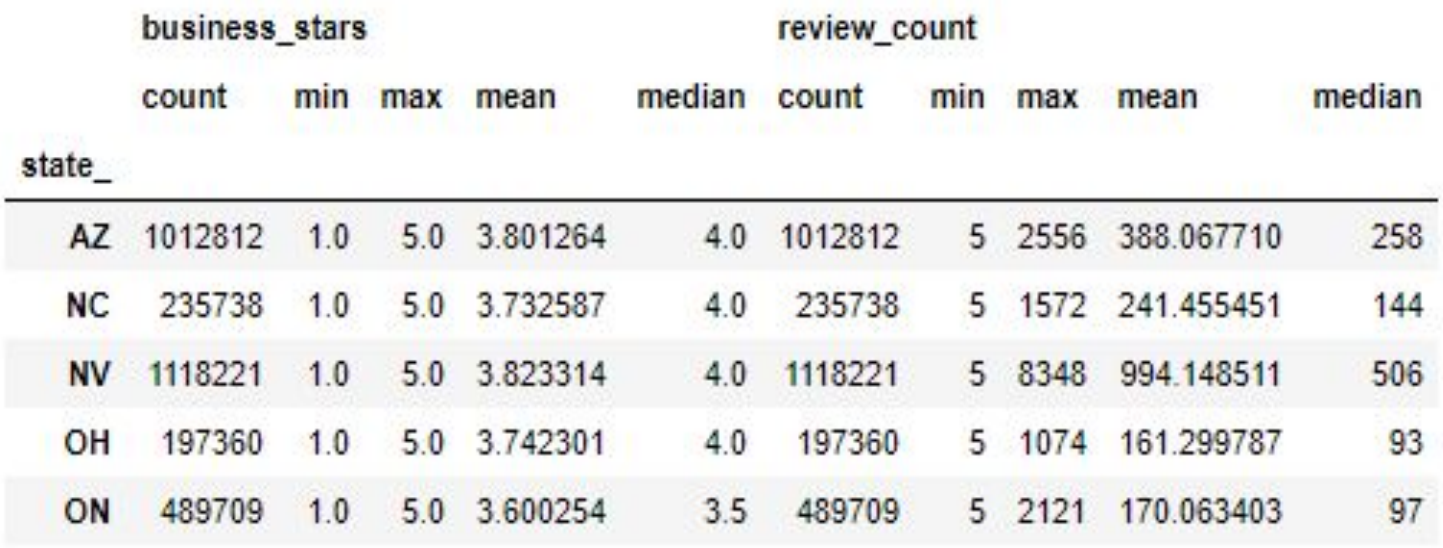
82% of restaurants are open

EXPLORATORY DATA ANALYSIS



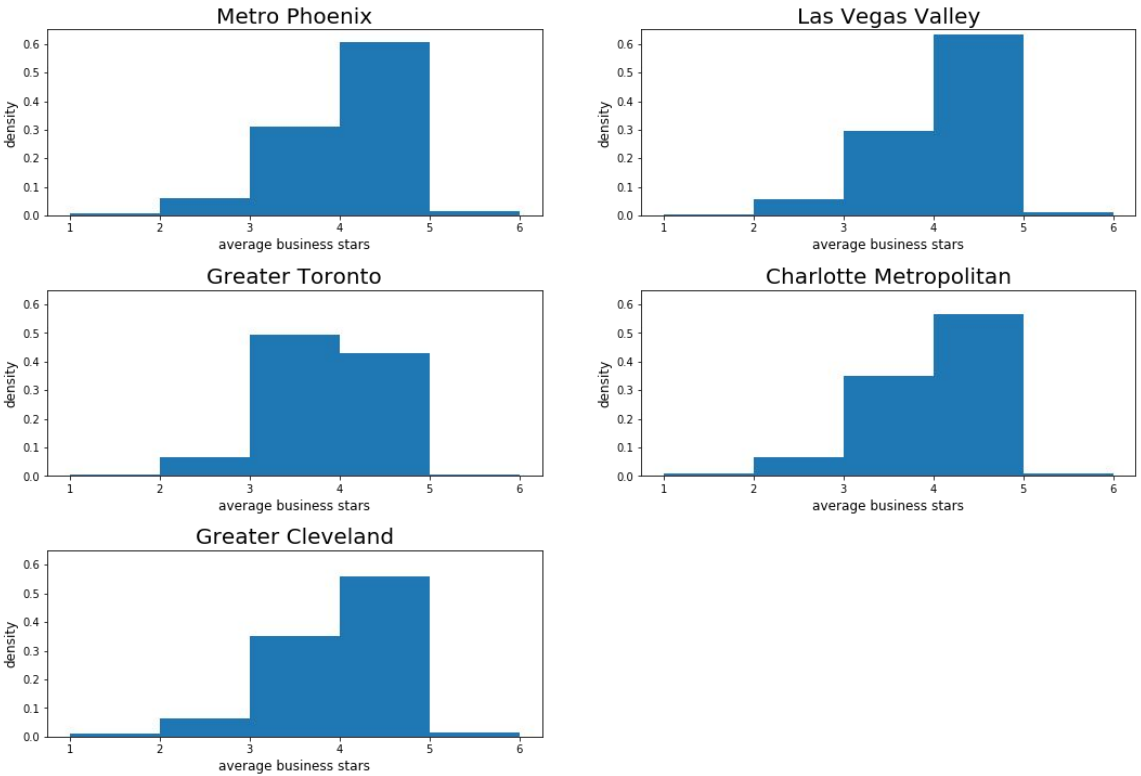
EXPLORATORY DATA ANALYSIS

The mean star rating across metropolitan areas is between 3.6 and 3.8 stars with Toronto (ON) having a slightly lower mean than the other metropolitan areas.



	business_stars					review_count				
	count	min	max	mean	median	count	min	max	mean	median
state_										
AZ	1012812	1.0	5.0	3.801264	4.0	1012812	5	2556	388.067710	258
NC	235738	1.0	5.0	3.732587	4.0	235738	5	1572	241.455451	144
NV	1118221	1.0	5.0	3.823314	4.0	1118221	5	8348	994.148511	506
OH	197360	1.0	5.0	3.742301	4.0	197360	5	1074	161.299787	93
ON	489709	1.0	5.0	3.600254	3.5	489709	5	2121	170.063403	97

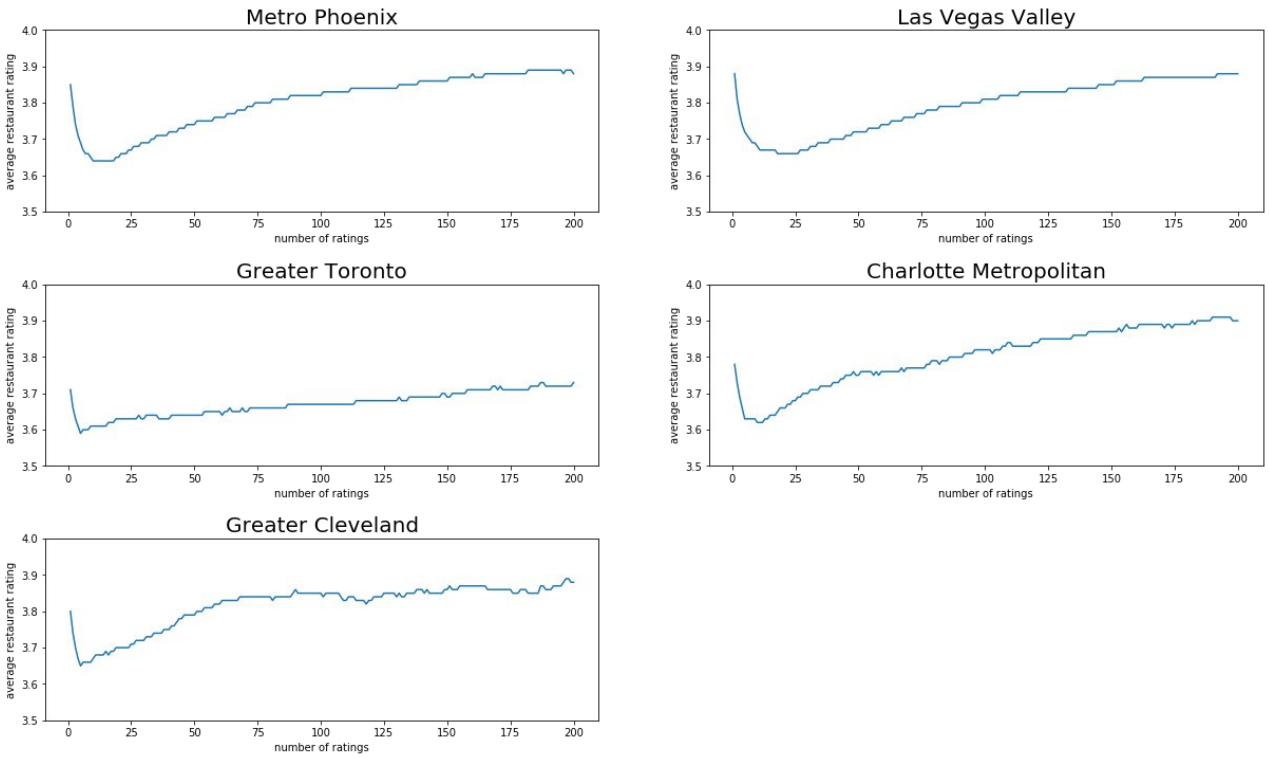
EXPLORATORY DATA ANALYSIS



Yelp Average Restaurant Star Ratings by Metro Area

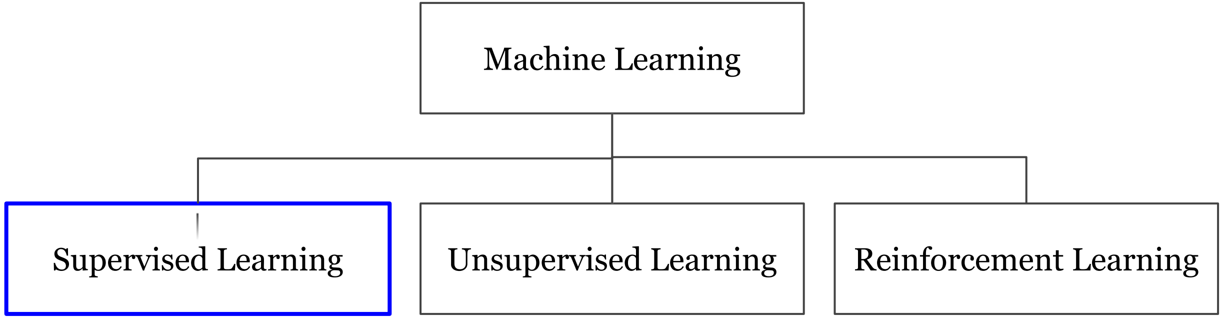
EXPLORATORY DATA ANALYSIS

Yelp Star Ratings Running Average by Metro Area



MACHINE LEARNING

Machine learning is a branch of artificial intelligence that uses algorithms to learn without explicitly being programmed.



Machine Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

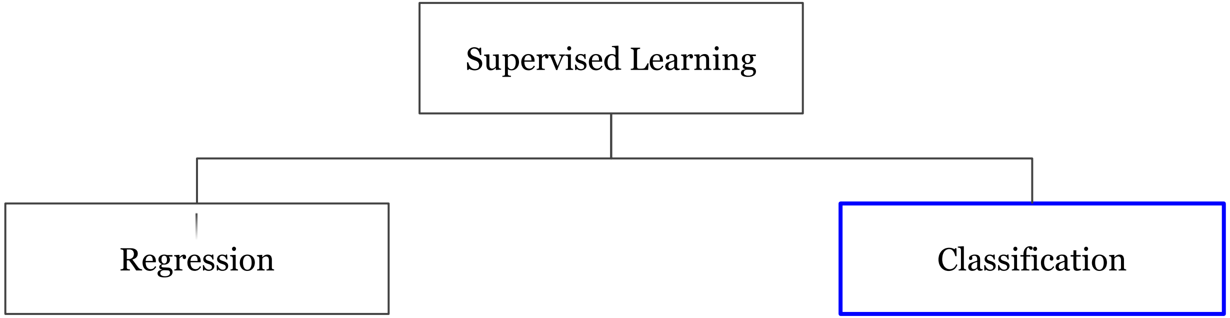
output category labels are known

output category labels are not known

Interacts with the environment and learning is derived from past experiences

MACHINE LEARNING

Of the two types of supervised learning, classification is the appropriate choice for assigning each restaurant to a star rating of 1-5 with 1 being least favorable, and 5 being more favorable.



numerical and used to
predict for a
continuous value

categorical, used to
predict for discrete
classes

MACHINE LEARNING

Due to the very large number of records in the dataset, a random sample of four categories (Italian, Japanese, Mexican and Burgers) was used create the model.

MACHINE LEARNING

Four classification models were considered and measured using precision and recall.

Algorithm	**Precision (weighted average)**	**Recall (weighted average)**
Custom Weighted Average	0.34	0.30
Surprise BaselineOnly	0.44	0.24
Surprise Matrix Factorization SVD	0.55	0.25
Surprise KNNBasic	0.38	0.27

Precision measures the proportion of positive identifications that are correct: True Positive /(True Positive + False Positive)

Recall measures the proportion of actual positives correctly identified: True Positive / (True Positive + False Negative)

MACHINE LEARNING

The Custom Weighted Average algorithm was the first to be executed and measured.

$$\text{weighted average (prediction)} = \frac{\sum_{u=1}^{n} (\textit{ratings for movie, } r_u)\ x\ (\textit{cosine similarity for user, } u)}{\sum_{u=1}^{n} \textit{cosine similarities, } u}$$

$$\text{cosine similarity} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\ \sqrt{\sum_{i=1}^{n} y_i^2}}$$

MACHINE LEARNING

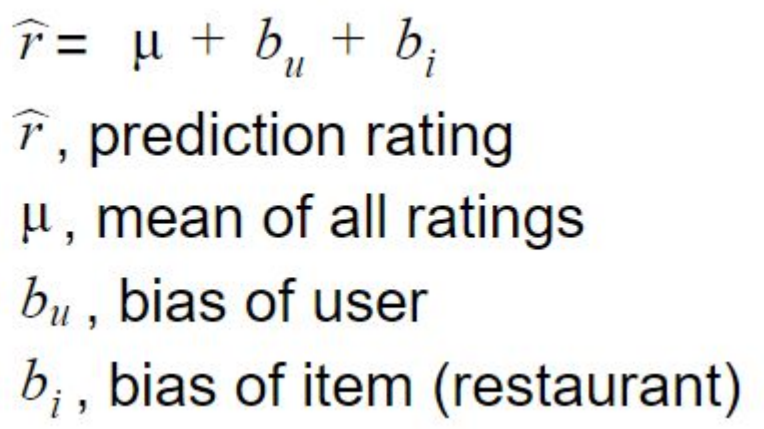
The Surprise library in scikit learn offers several built-in algorithms for recommender engines.

The predictions algorithms used for this project were:

- BaselineOnly
- Matrix Factorization SVD
- KNNBasic

MACHINE LEARNING

The BaselineOnly algorithm makes prediction by using the mean of all ratings and adding biases for both the user and item (restaurant).



$$\hat{r} = \mu + b_u + b_i$$

$\hat{r}$, prediction rating

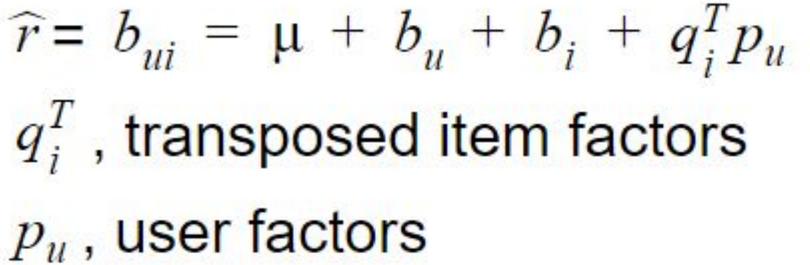
μ, mean of all ratings

b_u, bias of user

b_i, bias of item (restaurant)

MACHINE LEARNING

The Matrix Factorization SVD algorithm adds on to the BaselineOnly algorithm using matrix multiplication and SVD that allows for highly dimensional and complex data to be reduced to a lower dimensional space to help find better features for data classification.



$$\hat{r} = b_{ui} = \mu + b_u + b_i + q_i^T p_u$$

q_i^T , transposed item factors

p_u , user factors

MACHINE LEARNING

The KNNBasic is a non-parametric algorithm that makes a classification decision based on its proximity to known data.

$$\widehat{r_{ui}} = \frac{\sum\limits_{v \in N_i^k(u)} sim(u,v) \cdot r_{vi}}{\sum\limits_{v \in N_i^k(u)} sim(u,v)}$$

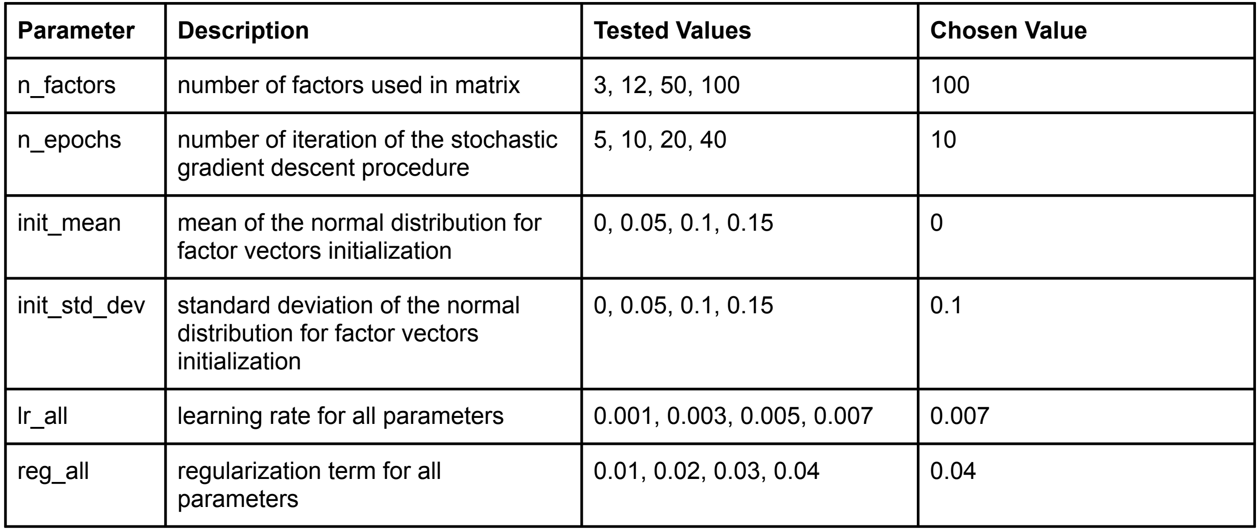
MACHINE LEARNING

Surprise Matrix Factorization SVD was chosen as the algorithm to use for this recommender engine since its precision score for was higher than all the rest of the algorithms. The recall scores were all closely clustered together for all the algorithms.

Algorithm	**Precision (weighted average)**	**Recall (weighted average)**
Custom Weighted Average	0.34	0.30
Surprise BaselineOnly	0.44	0.24
Surprise Matrix Factorization SVD	0.55	0.25
Surprise KNNBasic	0.38	0.27

MACHINE LEARNING

The following parameters were tuned for the surprise matrix factorization SVD with the chosen values returning a precision of 0.56 and a recall of 0.24.



Parameter	Description	Tested Values	Chosen Value
n_factors	number of factors used in matrix	3, 12, 50, 100	100
n_epochs	number of iteration of the stochastic gradient descent procedure	5, 10, 20, 40	10
init_mean	mean of the normal distribution for factor vectors initialization	0, 0.05, 0.1, 0.15	0
init_std_dev	standard deviation of the normal distribution for factor vectors initialization	0, 0.05, 0.1, 0.15	0.1
lr_all	learning rate for all parameters	0.001, 0.003, 0.005, 0.007	0.007
reg_all	regularization term for all parameters	0.01, 0.02, 0.03, 0.04	0.04

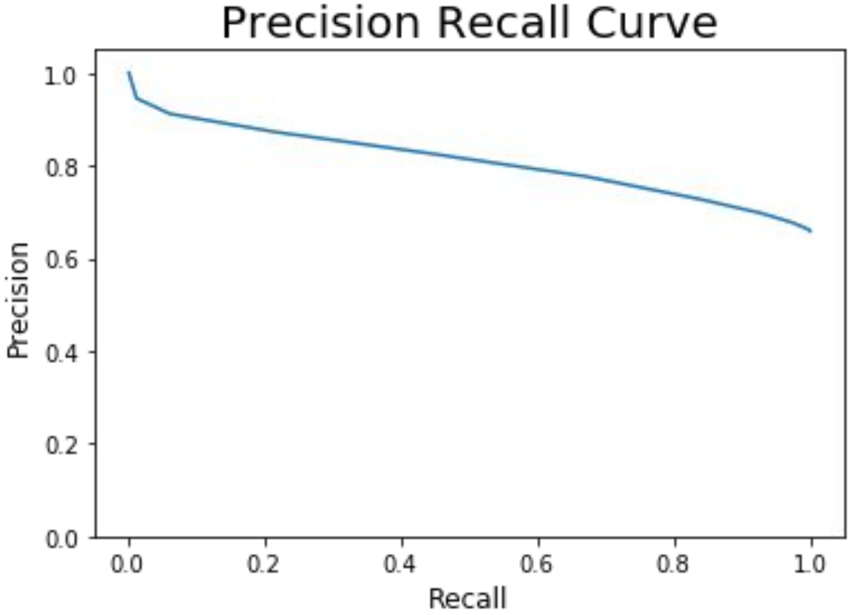
MACHINE LEARNING

Since the purpose is to recommend items, each of the ratings were converted to binary labels, 0 and 1 with 1 indicating a recommendation.

- Each actual label in the test set was converted to a 1 if the user's rating was >= 4
- 3.5 was used as the threshold for the predicted label to determine if the restaurant was counted as a recommendation
- The updated precision was 0.67 and recall was 0.69

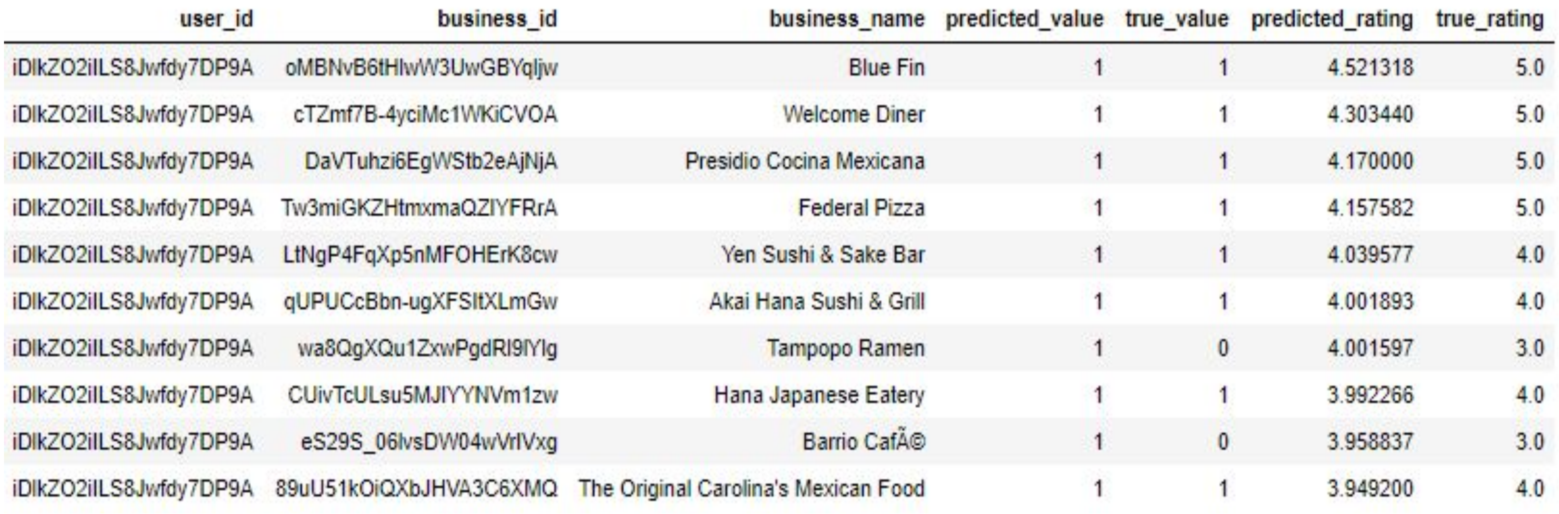
MACHINE LEARNING

The precision recall curve using many thresholds between 1-5 resulted in an area under the curve of 0.80798.



MACHINE LEARNING

Finally, the restaurant recommendation were given sorted by highest to lowest prediction rating.



user_id	business_id	business_name	predicted_value	true_value	predicted_rating	true_rating
iDIkZO2iILS8Jwfdy7DP9A	oMBNvB6tHlwW3UwGBYqljw	Blue Fin	1	1	4.521318	5.0
iDIkZO2iILS8Jwfdy7DP9A	cTZmf7B-4yciMc1WKiCVOA	Welcome Diner	1	1	4.303440	5.0
iDIkZO2iILS8Jwfdy7DP9A	DaVTuhzi6EgWStb2eAjNjA	Presidio Cocina Mexicana	1	1	4.170000	5.0
iDIkZO2iILS8Jwfdy7DP9A	Tw3miGKZHtmxmaQZIYFRrA	Federal Pizza	1	1	4.157582	5.0
iDIkZO2iILS8Jwfdy7DP9A	LtNgP4FqXp5nMFOHErK8cw	Yen Sushi & Sake Bar	1	1	4.039577	4.0
iDIkZO2iILS8Jwfdy7DP9A	qUPUCcBbn-ugXFSItXLmGw	Akai Hana Sushi & Grill	1	1	4.001893	4.0
iDIkZO2iILS8Jwfdy7DP9A	wa8QgXQu1ZxwPgdRI9lYlg	Tampopo Ramen	1	0	4.001597	3.0
iDIkZO2iILS8Jwfdy7DP9A	CUivTcULsu5MJIYYNVm1zw	Hana Japanese Eatery	1	1	3.992266	4.0
iDIkZO2iILS8Jwfdy7DP9A	eS29S_06lvsDW04wVrIVxg	Barrio CafÃ©	1	0	3.958837	3.0
iDIkZO2iILS8Jwfdy7DP9A	89uU51kOiQXbJHVA3C6XMQ	The Original Carolina's Mexican Food	1	1	3.949200	4.0

CONCLUSION

- A subset of restaurant categories (Italian, Japanese, Mexican and Burgers) was used for this recommender engine but additional categories can be added, however additional computing resources will be required
- The user-restaurant matrix was very sparse (~99.9%) and makes the recommender engine less accurate with precision and recall scores for star ratings
- Focusing only on if a restaurant should be recommended (prediction >=3.5) improved the precision and recall scores in this model