Comprehensive Restaurant Rating

Abhishek Rath
Chanakya Valluri
Smitha Eshwarahalli Ramesh
Sudarshan Aithal

Part I

Predicting Yelp Restaurant Ratings based on Business Attributes

- Past research shows that reviews and ratings play a direct role on the demand a business receives
- Encourage higher ratings through targeted alterations to their offerings
- Can be extended to any business
- We expect physical amenities, such as whether the business has a parking lot will have more of an impact than atmospheric variables, such as whether or not the business is romantic or casual

Feature Extraction

- The Yelp data used in this study originally contained 192609 records with 14 variables.
- To draw the sample for analysis, first all the non useful variables were dropped.
- The sample was narrowed down to 163773 records with 6 variables after dropping records with any of the 6 variables missing.
- Additionally, the dataset was filtered to only include records belonging to “Restaurants” categories. This narrowed the dataset considerably, bringing it down from the initial 163773 records to **57176.**

Sample data

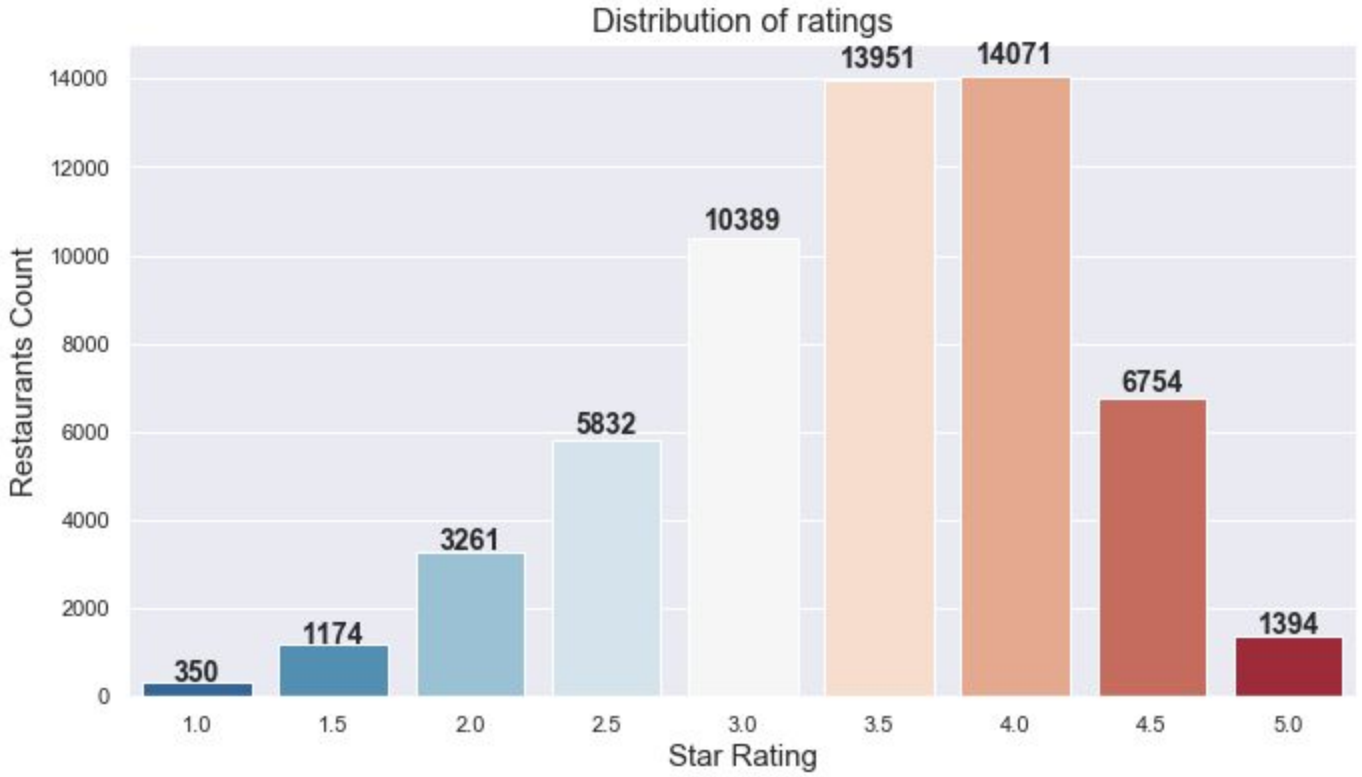
	attributes	business_id	categories	name	review_count	stars
0	{'RestaurantsReservations': 'True', 'GoodForMe...	QXAEGFB4oINsVuTFxEYKFQ	Specialty Food, Restaurants, Dim Sum, Imported...	Emerald Chinese Restaurant	128	2.5
1	{'GoodForKids': 'True', 'NoiseLevel': 'u'avera...	gnKjwL_1w79qoiV3IC_xQQ	Sushi Bars, Restaurants, Japanese	Musashi Japanese Restaurant	170	4.0
2	{'RestaurantsTakeOut': 'True', 'BusinessParkin...	1Dfx3zM-rW4n-31KeC8sJg	Restaurants, Breakfast & Brunch, Mexican, Taco...	Taco Bell	18	3.0
3	{'RestaurantsPriceRange2': '2', 'BusinessAccep...	fweCYi8FmbJXHCqLnwuk8w	Italian, Restaurants, Pizza, Chicken Wings	Marco's Pizza	16	4.0
4	{'OutdoorSeating': 'False', 'BusinessAcceptsCr...	PZ-LZzSlhSe9utkQYU8pFg	Restaurants, Italian	Carluccio's Tivoli Gardens	40	4.0

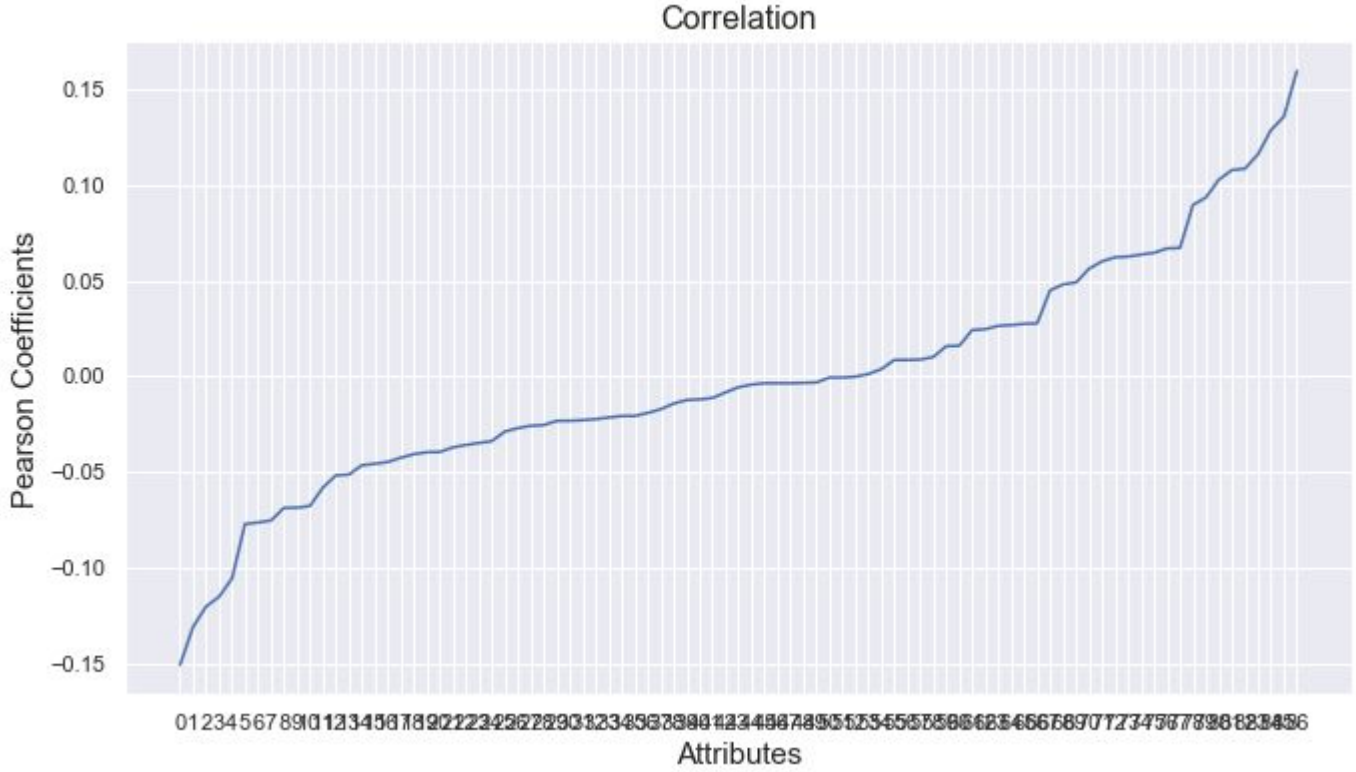
Data Analysis

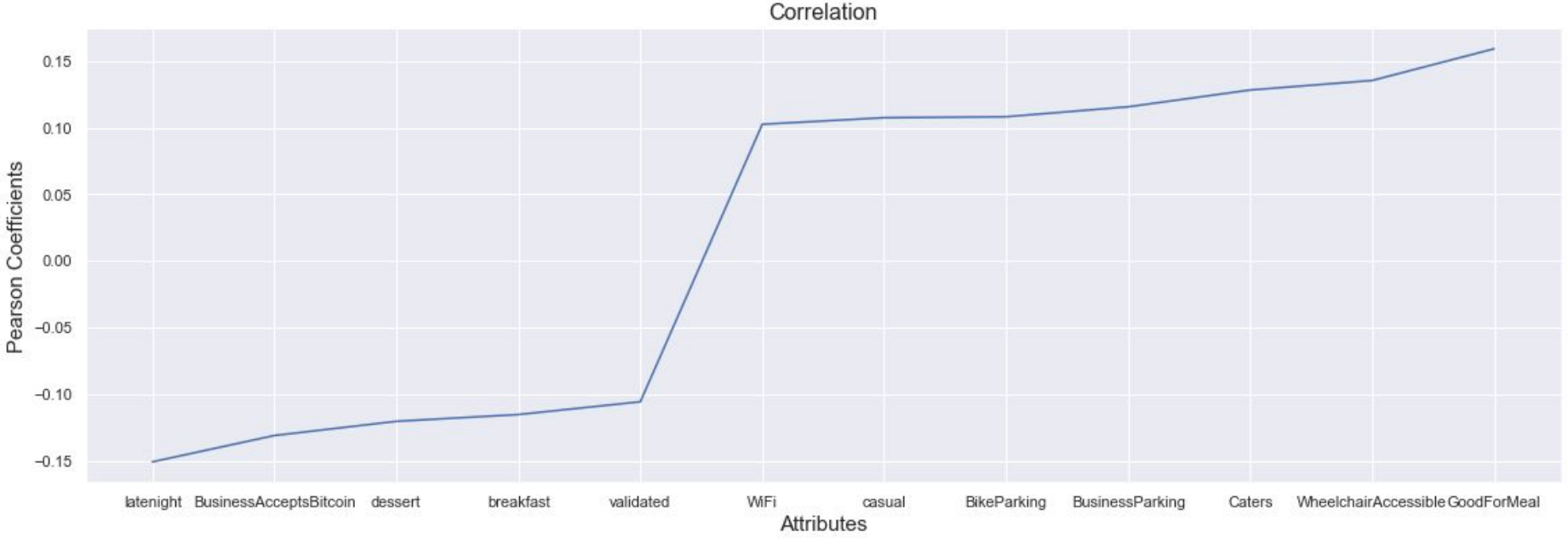
- All the attributes in the dataset were extracted. Certain attributes such as alcohol, were split up into types.
- For instance, the variable, “Alcohol”, would have the type, “none”, “beer_and_wine”, or “full_bar”. Some variables, such as “Ambience”, had subtypes, like “hipster”, “divey”, and “trendy”. Subtypes listed within these variables had binary options, true or false.
- An individual column for each variable was created. Each column would portray the variables as either having or lacking the certain attribute in a binary format (0 or 1)
- This resulted in 87 attributes

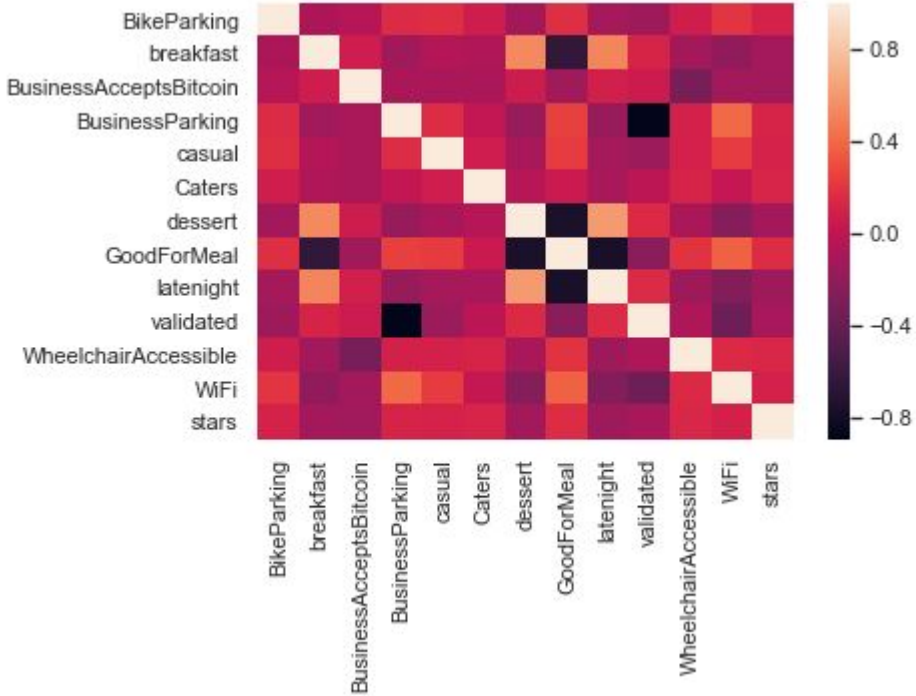
Attribute list

```
['AcceptsInsurance', 'africanamerican', 'AgesAllowed', 'Alcohol', 'Ambience', 'asian', 'background_music', 'BestNights', 'BikeParking', 'breakfast', 'brunch', 'BusinessAcceptsBitcoin', 'BusinessAcceptsCreditCards', 'BusinessParking', 'ByAppointmentOnly', 'BYOB', 'BYOBCorkage', 'casual', 'Caters', 'classy', 'CoatCheck', 'coloring', 'Corkage', 'curly', 'dairy-free', 'dessert', 'DietaryRestrictions', 'dinner', 'divey', 'dj', 'DogsAllowed', 'DriveThru', 'extensions', 'friday', 'garage', 'gluten-free', 'GoodForDancing', 'GoodForKids', 'GoodForMeal', 'halal', 'HappyHour', 'HasTV', 'hipster', 'intimate', 'jukebox', 'karaoke', 'kids', 'kosher', 'latenight', 'live', 'lot', 'lunch', 'monday', 'Music', 'no_music', 'NoiseLevel', 'Open24Hours', 'OutdoorSeating', 'perms', 'RestaurantsAttire', 'RestaurantsCounterService', 'RestaurantsDelivery', 'RestaurantsGoodForGroups', 'RestaurantsPriceRange2', 'RestaurantsReservations', 'RestaurantsTableService', 'RestaurantsTakeOut', 'romantic', 'saturday', 'Smoking', 'soy-free', 'straightperms', 'street', 'sunday', 'thursday', 'touristy', 'trendy', 'tuesday', 'upscale', 'valet', 'validated', 'vegan', 'vegetarian', 'video', 'wednesday', 'WheelchairAccessible', 'WiFi']
```


Distribution of ratings
Restaurants Count
Star Rating
0
2000
4000
6000
8000
10000
12000
14000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
350
1174
3261
5832
10389
13951
14071
6754
1394

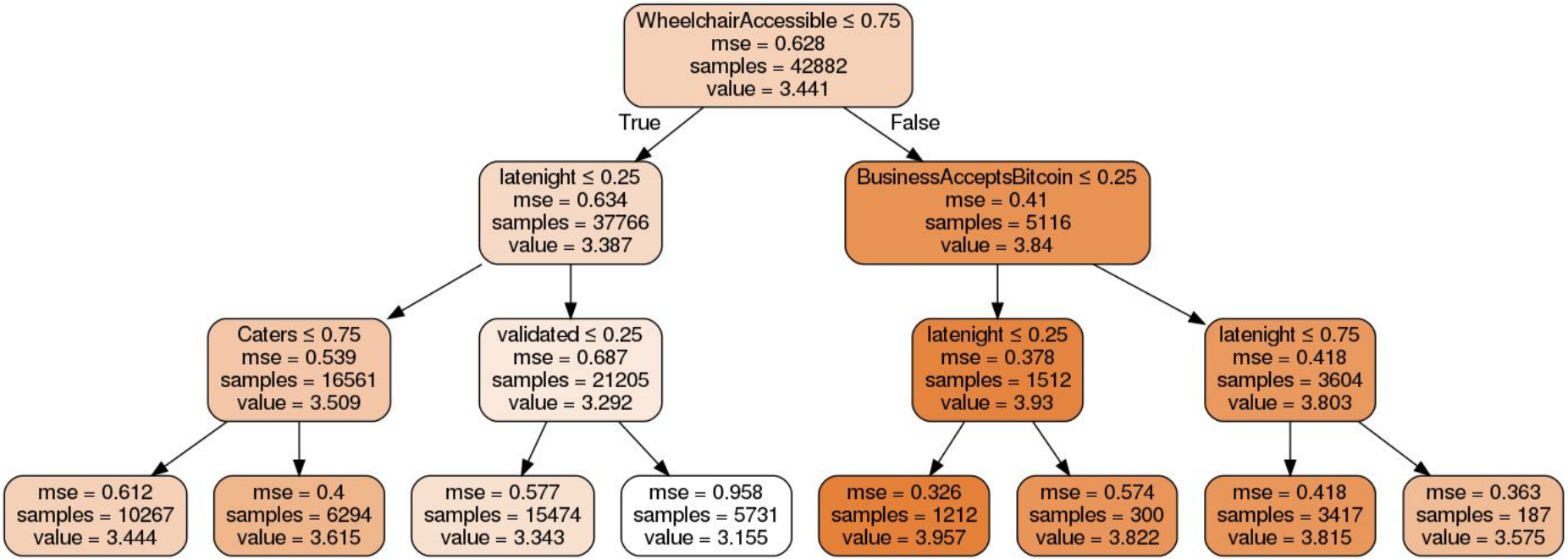

Correlation
Pearson Coefficients
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
Attributes


Correlation
Pearson Coefficients
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
latenight
BusinessAcceptsBitcoin
dessert
breakfast
validated
WiFi
casual
BikeParking
BusinessParking
Caters
WheelchairAccessible
GoodForMeal
Attributes


BikeParking
breakfast
BusinessAcceptsBitcoin
BusinessParking
casual
Caters
dessert
GoodForMeal
latenight
validated
WheelchairAccessible
WiFi
stars
0.8
0.4
0.0
−0.4
−0.8

Approach

- Linear regression and Decision tree regression models were built
- The variables used in modeling were limited to those identified with a statistically significant correlation to business rating after correlation analysis


WheelchairAccessible ≤ 0.75
mse = 0.628
samples = 42882
value = 3.441
True
False
latenight ≤ 0.25
mse = 0.634
samples = 37766
value = 3.387
BusinessAcceptsBitcoin ≤ 0.25
mse = 0.41
samples = 5116
value = 3.84
Caters ≤ 0.75
mse = 0.539
samples = 16561
value = 3.509
validated ≤ 0.25
mse = 0.687
samples = 21205
value = 3.292
latenight ≤ 0.25
mse = 0.378
samples = 1512
value = 3.93
latenight ≤ 0.75
mse = 0.418
samples = 3604
value = 3.803
mse = 0.612
samples = 10267
value = 3.444
mse = 0.4
samples = 6294
value = 3.615
mse = 0.577
samples = 15474
value = 3.343
mse = 0.958
samples = 5731
value = 3.155
mse = 0.326
samples = 1212
value = 3.957
mse = 0.574
samples = 300
value = 3.822
mse = 0.418
samples = 3417
value = 3.815
mse = 0.363
samples = 187
value = 3.575

Part II

Predicting Yelp Restaurant Ratings based on User Reviews

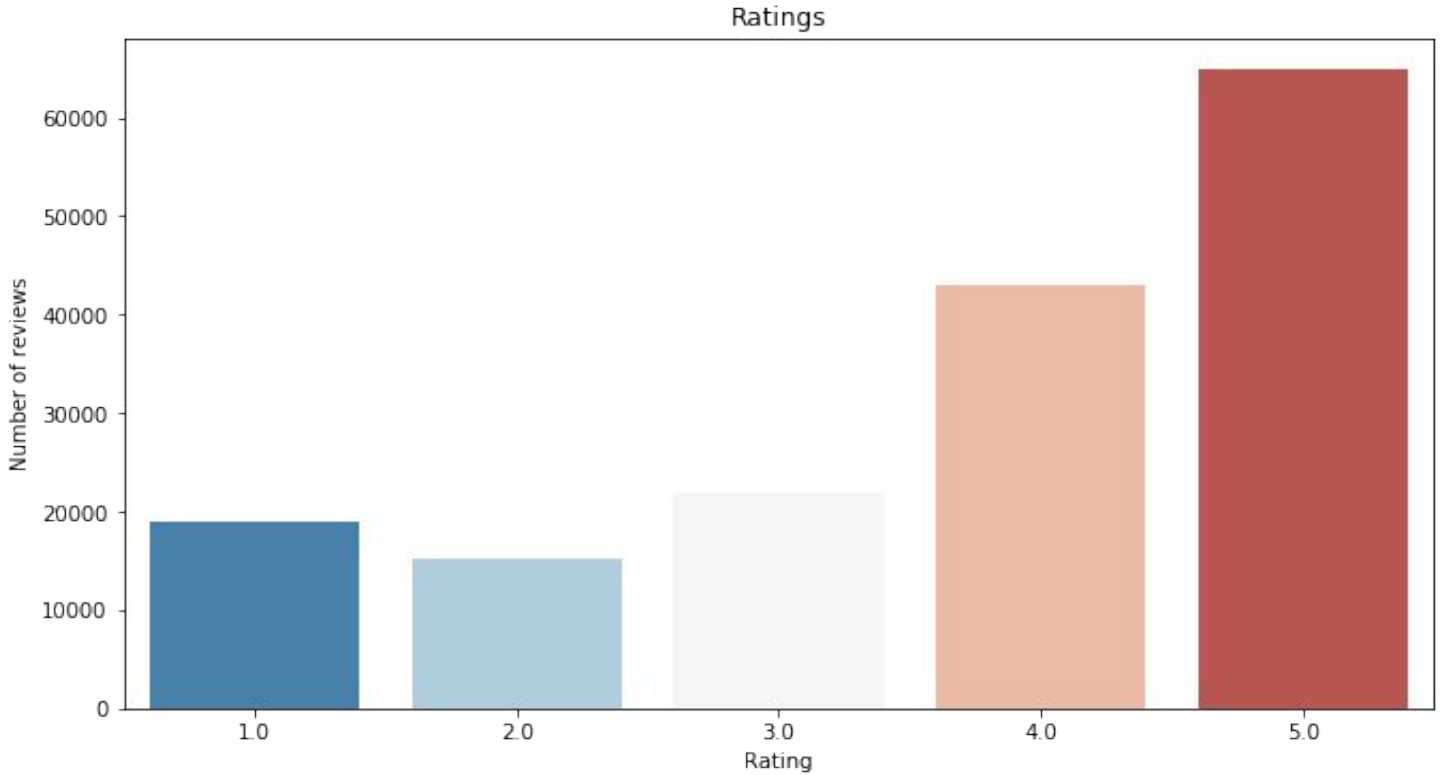
- The Review.Json file originally contained approximately 5,200,000 user reviews.
- We processed Review.json file to extract only restaurant reviews.
- The file was narrowed down to 164045 reviews with 5 variables of interest, the star ratings(1 to 5).
- We are balancing the dataset to make sure no biased selection occurs during testing and training. The dataset is scaled to 15267 ie total number of 2 star ratings which is the least.
- To reduce the computation time, TF/IDF is used to vectorize the text data .
- Linear SVM is used to classify and predict the outcomes.

Review.Json Overview

Out[25]:

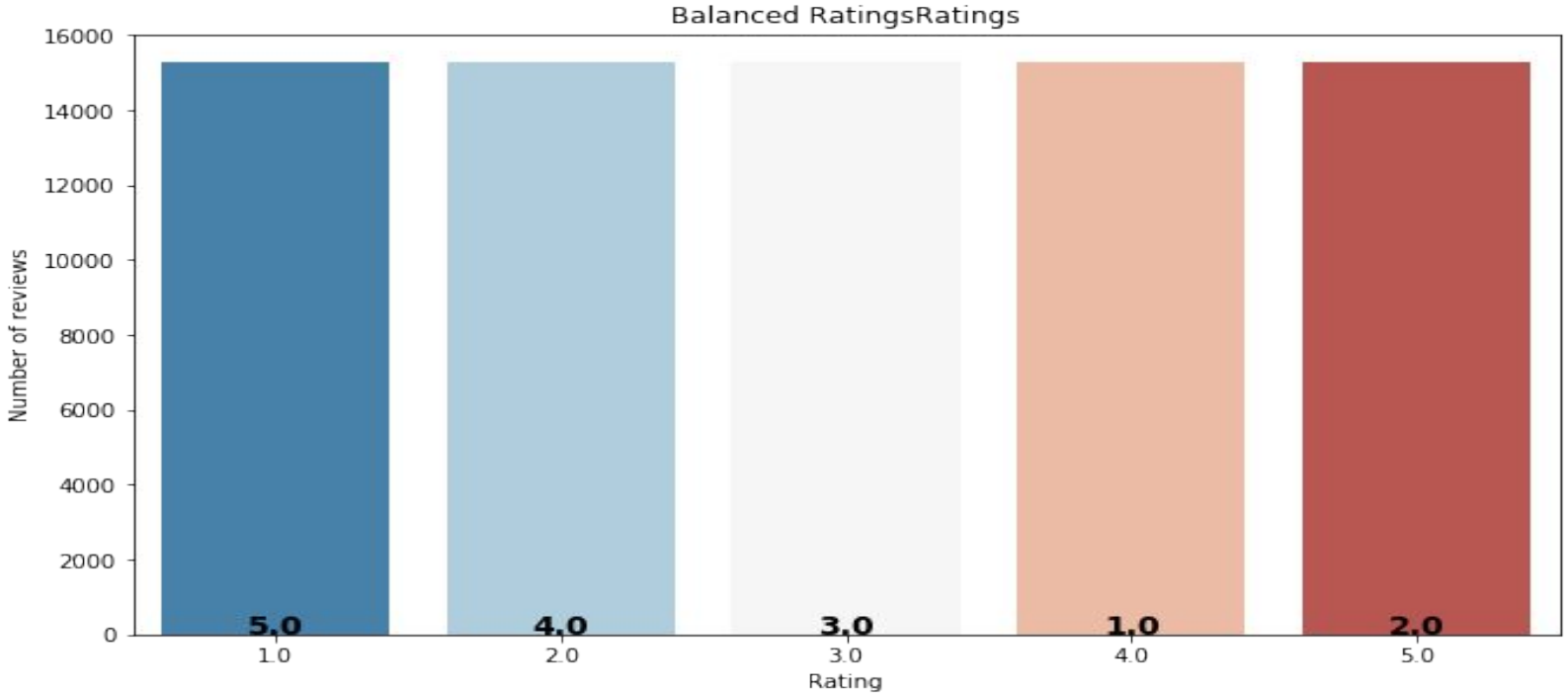
	unnamed	business_id	cool	date	funny	review_id	stars	text	useful	user_id
0	NaN	business_id	cool	date	funny	review_id	stars	text	useful	user_id
1	0.0	ikCg8xy5JIg_NGPx-MSIDA	0.0	2018-01-09 20:56:38	0.0	yi0R0Ugj_xUx_Nek0-_Qig	5.0	Went in for a lunch. Steak sandwich was delici...	0.0	dacAIZ6fTM6mqwW5uxkskg
2	1.0	eU_713ec6fTGNO4BegRaww	0.0	2013-01-20 13:25:59	0.0	fdiNeiN_hoCxCMy2wTRW9g	4.0	I'll be the first to admit that I was not exci...	0.0	w31MKYsNFMrjhWxxAb5wlw
3	2.0	3fw2X5bZYeW9xCz_zGhOHg	5.0	2016-05-07 01:21:02	4.0	G7XHMxG0bx9oBJNECG4IFg	3.0	Tracy dessert had a big name in Hong Kong and ...	5.0	jlu4CztcSxrKx56ba1a5AQ
4	3.0	zvO-PJCpNk4fgAVUnExYAA	1.0	2010-10-05 19:12:35	1.0	8e9HxxLjjqc9ez5ezzN7iQ	1.0	This place has gone down hill. Clearly they h...	3.0	d6xvYpyzcfbF_AZ8vMB7QA

Stars vs Number of reviews



Balanced Stars vs Number of Reviews

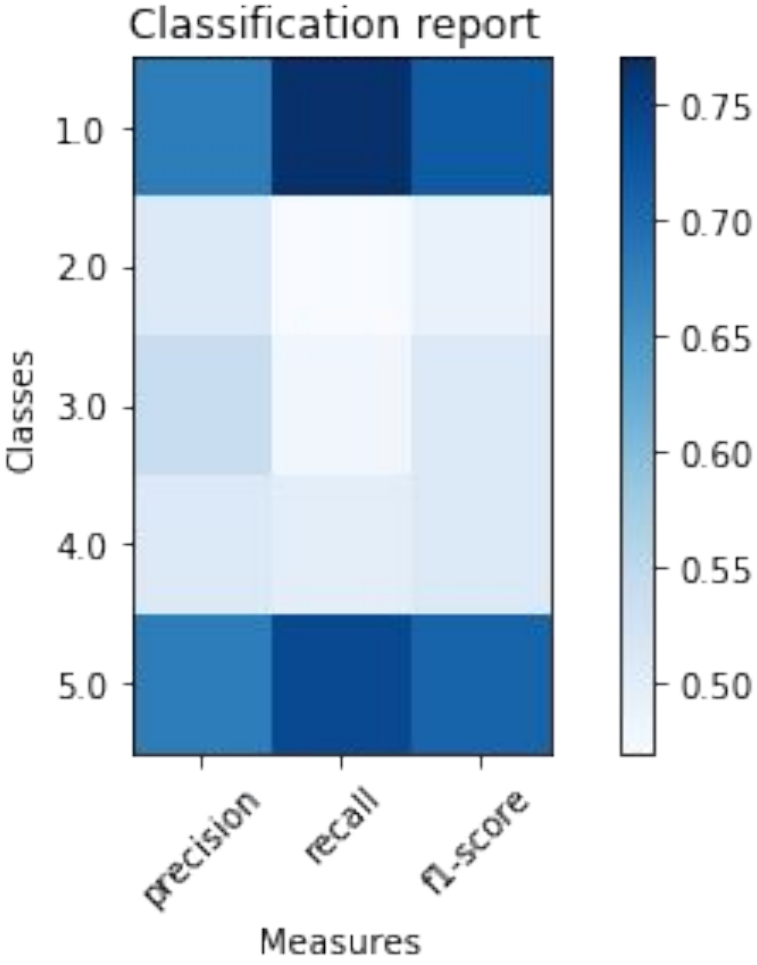
Balancing the data scales the Number of ratings down to 15267 for all ratings



Classification Report

Accuracy Score of our algorithm was 0.5921559287046961 i.e. Approximately 60% Accurate

	precision	recall	f1-score	support
1.0	0.68	0.77	0.72	5100
2.0	0.51	0.47	0.49	4957
3.0	0.54	0.48	0.51	5100
4.0	0.51	0.50	0.51	4964
5.0	0.68	0.74	0.71	5070
micro avg	0.59	0.59	0.59	25191
macro avg	0.58	0.59	0.59	25191
weighted avg	0.59	0.59	0.59	25191


Classification report
1.0
2.0
3.0
4.0
5.0
Classes
precision
recall
f1-score
Measures
0.75
0.70
0.65
0.60
0.55
0.50

Results

- Model that fairly predicts business rating for restaurants based on business attribute
- Aims to find attributes that have a significant correlation to business rating
- RMSE ~ 0.74
- Extracting restaurant ratings from review.json was hard and time consuming.
- Determining the influencing factors for the model was tricky.
- Model that predicts Restaurant ratings based on Review.json using SVM classifier with 59.21% accuracy

Conclusion

- Analysis of a pooled set of restaurant records, reviews and their attributes produces important insights about the general rating behavior of consumer
- This information is valuable to businesses, as they may be able to identify the most impactful features on rating, effectively implement or remove them, and potentially raise future ratings

Future Research

- Attempt to find greater accuracy by splitting the attributes into two model types, with the first type consisting mainly of amenities while the second type focusing on the ambience
- Analysis including the variables that were provided by the Yelp dataset that were excluded from our sample and other variables in the attribute section that were not considered in this piece.
- Try out other modeling techniques to improve accuracy