Client Project:

New Light Technologies

Dmitriy Pavlov, Brian Pyron, Thomas Rubio
ATL - DSI 6

Overview

- Client Problem
- Data Science Problem
- Data Collection
- Feature Engineering
- Exploratory Data Analysis
- Modeling
- Conclusion

Client Problem:

Our client, New Light Technologies, tasked us with determining the affluence of a ZIP code using price data from the popular restaurant and business review platform Yelp.

Data Science Problem:

Are we able to utilize the available data on Yelp in order to predict the affluence of a given zip code?

Data Collection: Zillow's price per square foot was used as proxy for affluency of a zip code

Feature	Type	Description
regionname	object	region zip codes
state	object	states to which zip codes belong
price_per_sqft	int	median home values per sq ft of each region

https://www.zillow.com/research/data/

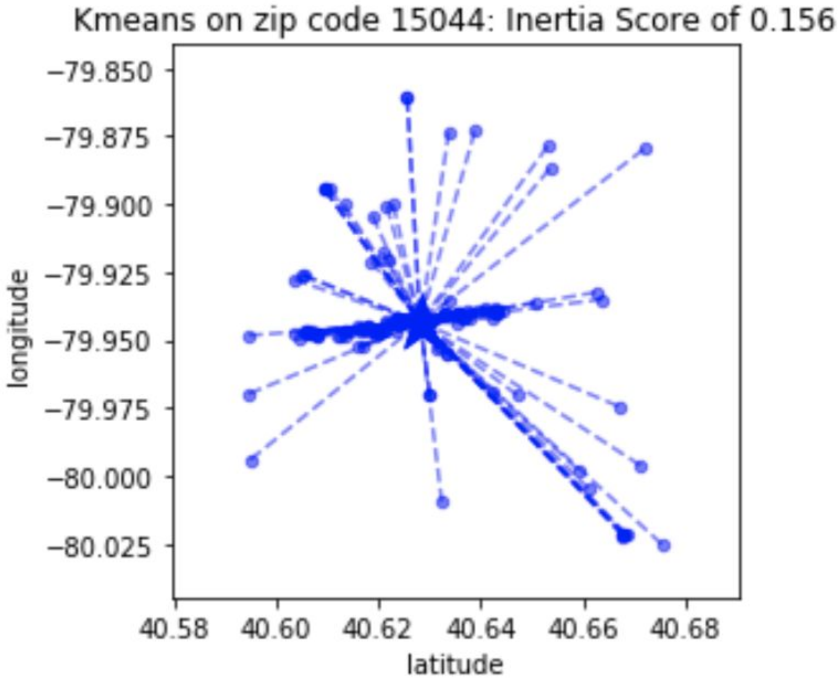
Data Collection: Yelp's public dataset was used as predicting features of our model

Feature	Type	Description
postal_code	int	business zip codes
categories	object	categories under which businesses fall
is_open	float	whether or not businesses are still open
latitude	float	latitudes of all businesses
longitude	float	longitudes of all businesses
review_count	int	number of yelp reviews each business received
stars	float	average number of star ratings each business received

The features of of each business were sum aggregated by zip code in order to obtain X for our model.

https://www.kaggle.com/yelp-dataset/yelp-dataset

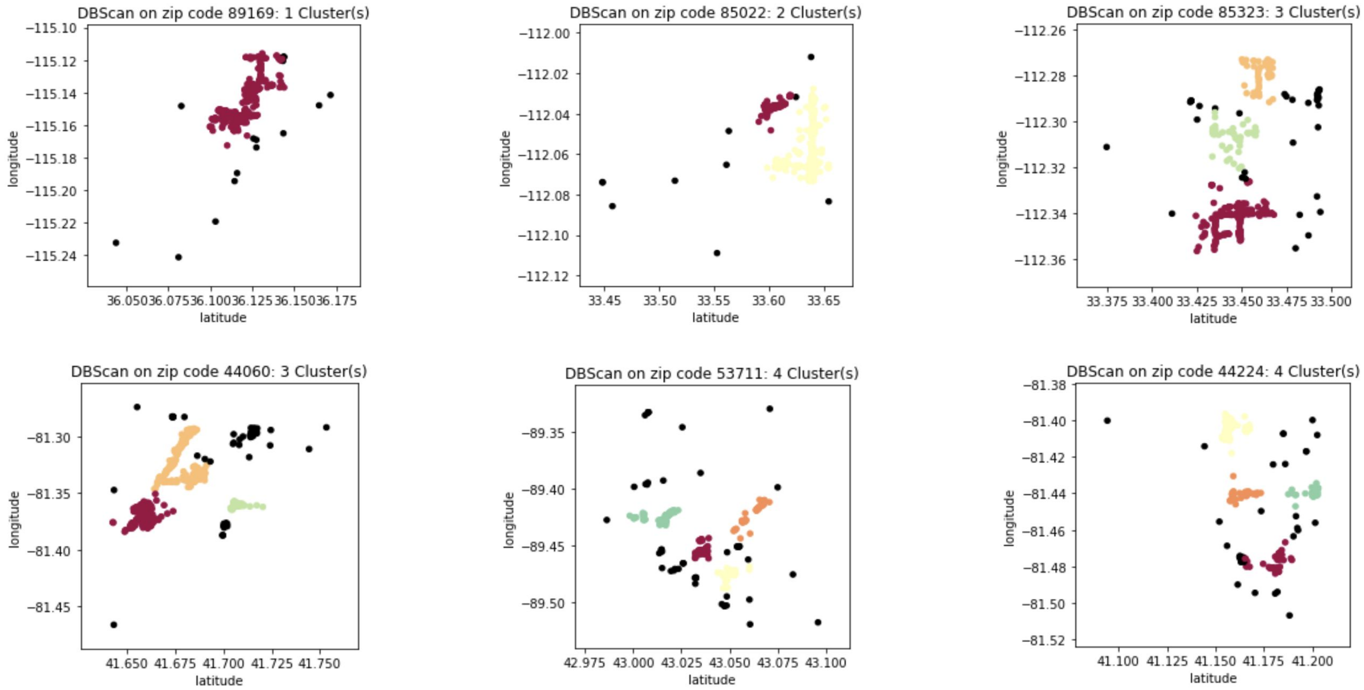
Feature Engineering: To capture the density of businesses per region, we use the inertia score of K-means on business latitudes and longitudes



Feature Engineering: To determine where there are higher concentrations of businesses, a manual grid search is performed on DBSCAN parameters

	eps	min_samples	correlation	train_score	test_score
0	0.002	10	0.238531	0.61558	0.46251
1	0.002	20	0.325427	0.6009	0.46562
2	0.002	30	0.352071	0.63136	0.47918
3	0.002	40	0.373147	0.63715	0.46429
4	0.002	50	0.32825	0.64074	0.45972

Feature Engineering: In some regions, there tend to be higher concentrations of businesses, the number of which is captured for analysis

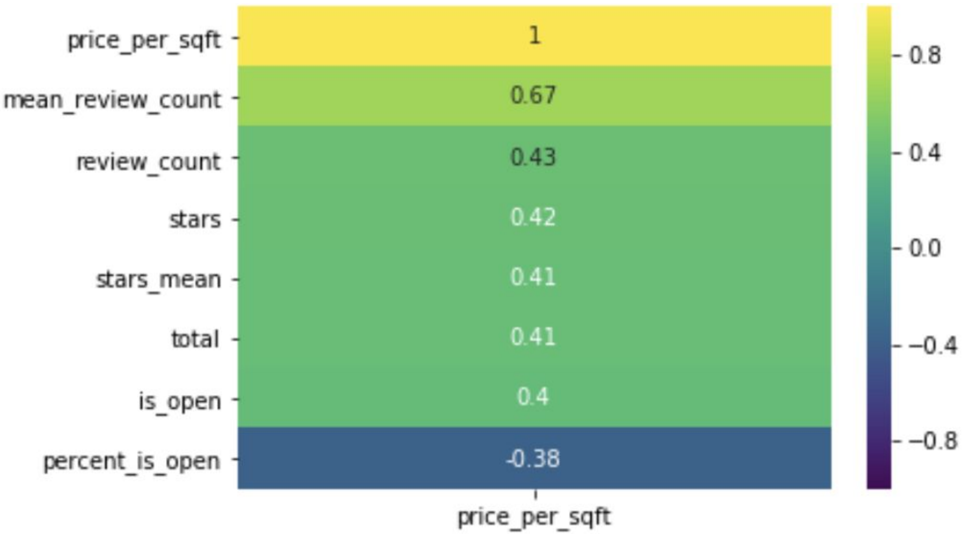


Feature Engineering: A word vectorizer was used to create business categories for each zip code

	categories
1	Chicken Wings, Burgers, Caterers, Street Vendo...
3	Insurance, Financial Services
5	Coffee & Tea, Food
8	Mexican, Restaurants
9	Flowers & Gifts, Gift Shops, Shopping
12	Bars, Sports Bars, Dive Bars, Burgers, Nightli...
17	Shopping, Fashion, Department Stores
18	Financial Services, Check Cashing/Pay-day Loan...
19	American (Traditional), Food, Bakeries, Restau...
20	Home Services, Masonry/Concrete, Professional ...

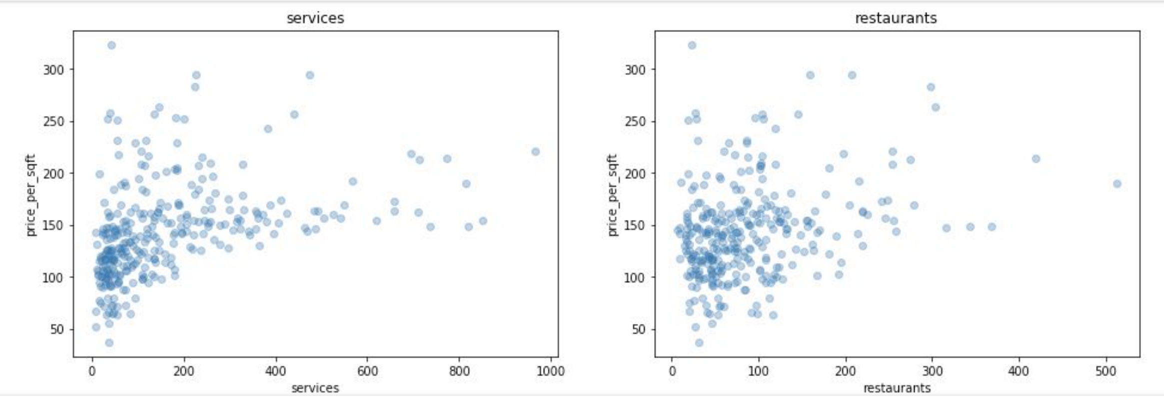
	3d	abatement	acai	accessories	accountants	acne	active
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

EDA: We are getting some signal from the features provided by Yelp



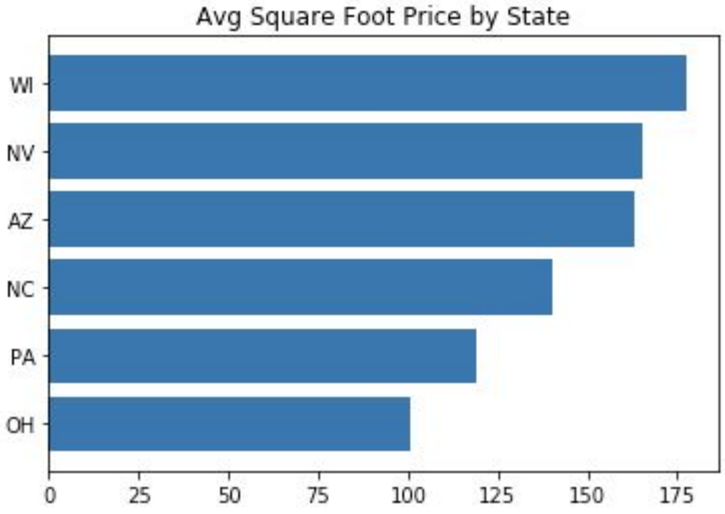
EDA: With over 1,000 business types and lack of clear correlation to price to we will let the model pick the features

	price_correlation
price_per_sqft	1.000000
active	0.392080
life	0.389524
fitness	0.375817
instruction	0.373343
estate	0.361504
real	0.360322
arts	0.358712
centers	0.349595
coffee	0.348906



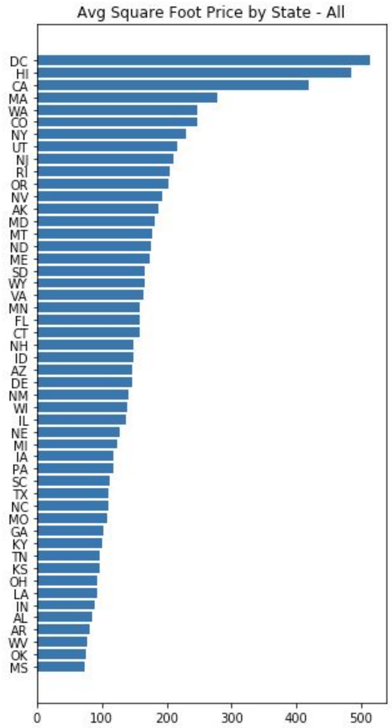
EDA: There is a significant variation in housing prices across the states, this was also the case with our data

Yelp Data

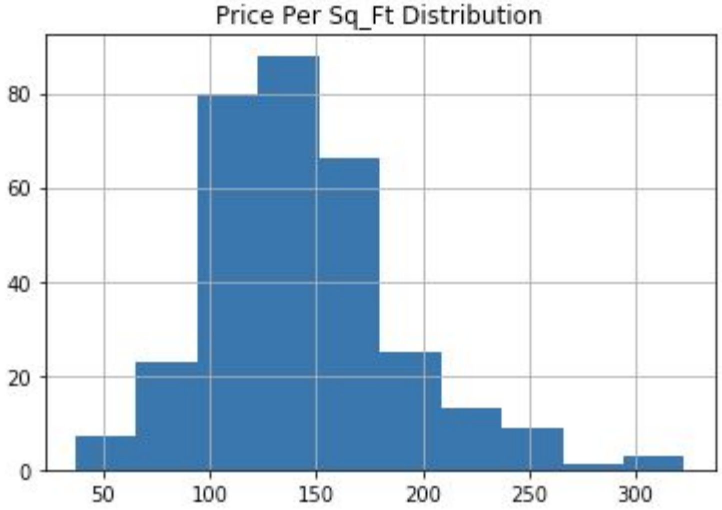


All USA

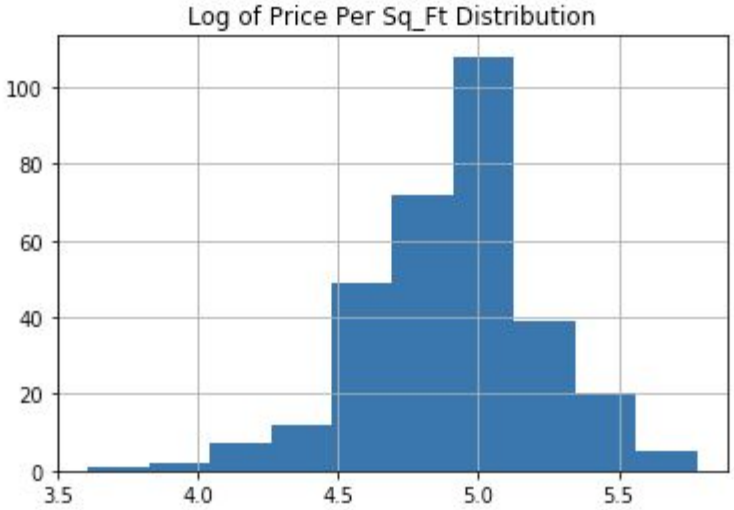
count	50.000000
mean	166.205246
std	92.799310
min	72.481752
25%	108.959880
50%	147.674705
75%	185.541217
max	512.809524



EDA: Log transformation of square foot prices helps normalize our target y for modeling



Log



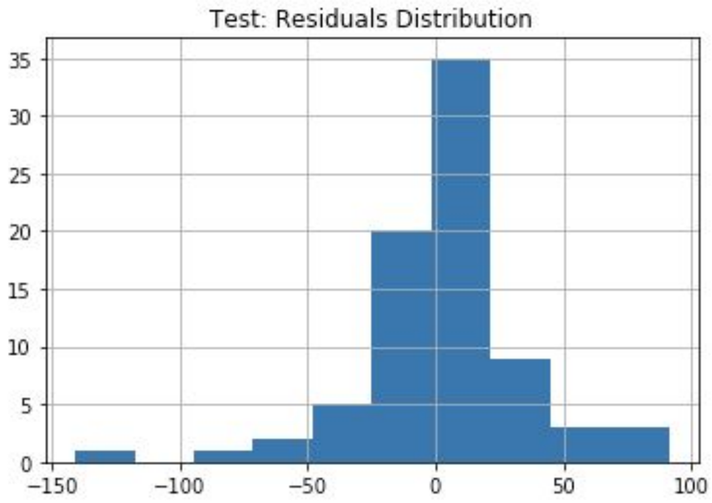
Modeling: Despite overfitting the training data Lasso model performed the best on the data

<u>Lasso:</u>

Data:
Input features: 1375
Useful features: 43

Performance:

Train r2:	0.69
Test r2:	0.53
Cross_Val r2:	0.52



This is a high variance model and overfits our training data, but provides promising directional results. If we had more information about sampling of the data and more data we could improve model's performance.

Model: The affluence of a zip are driven by business types, state, and business quality

Business Quality

Business Types

Increasing Value

	variable	coefs
1360	stars_mean	0.079056
1374	state_WI	0.032503
1366	review_count_log	0.029110
576	hepatologists	0.023682
809	new	0.023654
1246	train	0.021867
580	hiking	0.017902
1353	yoga	0.016268
100	barre	0.015430
702	laundry	0.012915

Decreasing Value

	variable	coefs
1372	state_OH	-0.077278
1373	state_PA	-0.044538
1348	wings	-0.031615
1362	percent_is_open	-0.029757
1113	soul	-0.011465
581	himalayan	-0.009008
797	musicians	-0.007976
396	dumpster	-0.005258
691	land	-0.004793
228	chicken	-0.001356

States

Conclusion:

- Yelp business data does contain signal to predict neighborhood affluence
- Gathering a more robust dataset from yelp could significantly improve model performance
  - Sparse data coverage
  - Possibility of selection bias
  - Incomplete business data for zip codes
  - Lack of business dollar signs
- Next Steps
  - Data transformation of business categories
  - Obtain a more robust dataset
  - Build a platform for model utilization