



Yelp Me Yelp You

Michael Barlow, Madeleine Crouch,
Khoa Le, Dalton Morrow



Our Questions

What attributes contribute to a successful business?

What external factors influence customer experience?

Can we find recurring problems in a business by analyzing review text?

Data Preparation

- Size of data necessitated random sampling
- Many attributes were very sparse
- Dimensionality Reduction

Tools Used

RapidMiner:

- Correlation Matrices
- Clustering
- K-Means
- Decision Trees
- Cleaning (Reduce dimensionality and split data)

Python:

- NLTK
- VADER package

Users

| Attributes | photos | list | funny | plain | note | writer | cute | stars | more | hot | profile | cool |
|------------|--------|------|-------|-------|------|--------|------|-------|------|------|---------|------|
| photos | 1.00 | 0.54 | 0.76 | 0.92 | 0.97 | 0.84 | 0.63 | 0.02 | 0.92 | 0.70 | 0.69 | 0.76 |
| list | 0.54 | 1.00 | 0.85 | 0.75 | 0.53 | 0.71 | 0.64 | 0.01 | 0.38 | 0.82 | 0.90 | 0.85 |
| funny | 0.76 | 0.85 | 1.00 | 0.91 | 0.72 | 0.90 | 0.83 | 0.01 | 0.55 | 0.97 | 0.94 | 1.00 |
| plain | 0.92 | 0.75 | 0.91 | 1.00 | 0.92 | 0.87 | 0.74 | 0.01 | 0.77 | 0.86 | 0.84 | 0.91 |
| note | 0.97 | 0.53 | 0.72 | 0.92 | 1.00 | 0.81 | 0.65 | 0.02 | 0.94 | 0.68 | 0.69 | 0.72 |
| writer | 0.84 | 0.71 | 0.90 | 0.87 | 0.81 | 1.00 | 0.84 | 0.02 | 0.76 | 0.88 | 0.89 | 0.90 |
| cute | 0.63 | 0.64 | 0.83 | 0.74 | 0.65 | 0.84 | 1.00 | 0.00 | 0.56 | 0.91 | 0.84 | 0.83 |
| stars | 0.02 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.00 | 1.00 | 0.02 | 0.01 | 0.01 | 0.01 |
| more | 0.92 | 0.38 | 0.55 | 0.77 | 0.94 | 0.76 | 0.56 | 0.02 | 1.00 | 0.51 | 0.56 | 0.55 |
| hot | 0.70 | 0.82 | 0.97 | 0.86 | 0.68 | 0.88 | 0.91 | 0.01 | 0.51 | 1.00 | 0.93 | 0.97 |
| profile | 0.69 | 0.90 | 0.94 | 0.84 | 0.69 | 0.89 | 0.84 | 0.01 | 0.56 | 0.93 | 1.00 | 0.94 |
| cool | 0.76 | 0.85 | 1.00 | 0.91 | 0.72 | 0.90 | 0.83 | 0.01 | 0.55 | 0.97 | 0.94 | 1.00 |

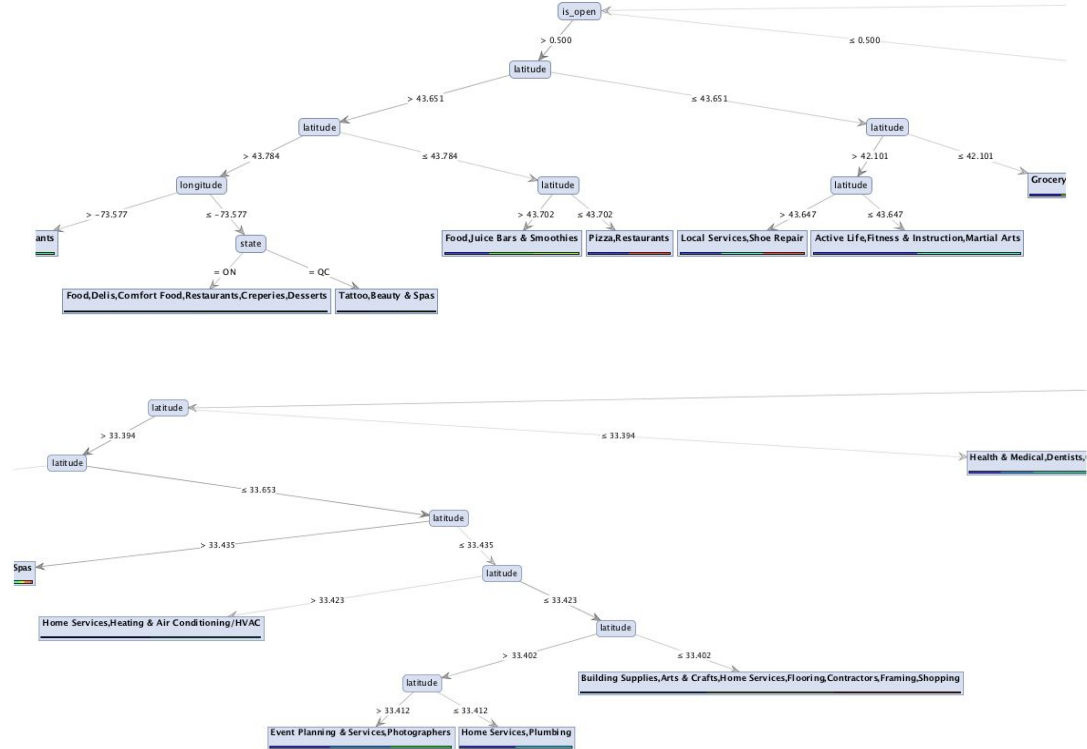
| Attributes | average_stars | total_compliments |
|-------------------|---------------|-------------------|
| average_stars | 1 | 0.0149127035 |
| total_compliments | 0.0149127035 | 1 |

The compliments that a user receives do not correlate with the star ratings that a user gives.

Compliments do correlate with other compliments.

Decision Tree

- Supervised learning
- Used categories as labels
- Is_open has the most information gain
- Large split on latitude and longitude



Clustering

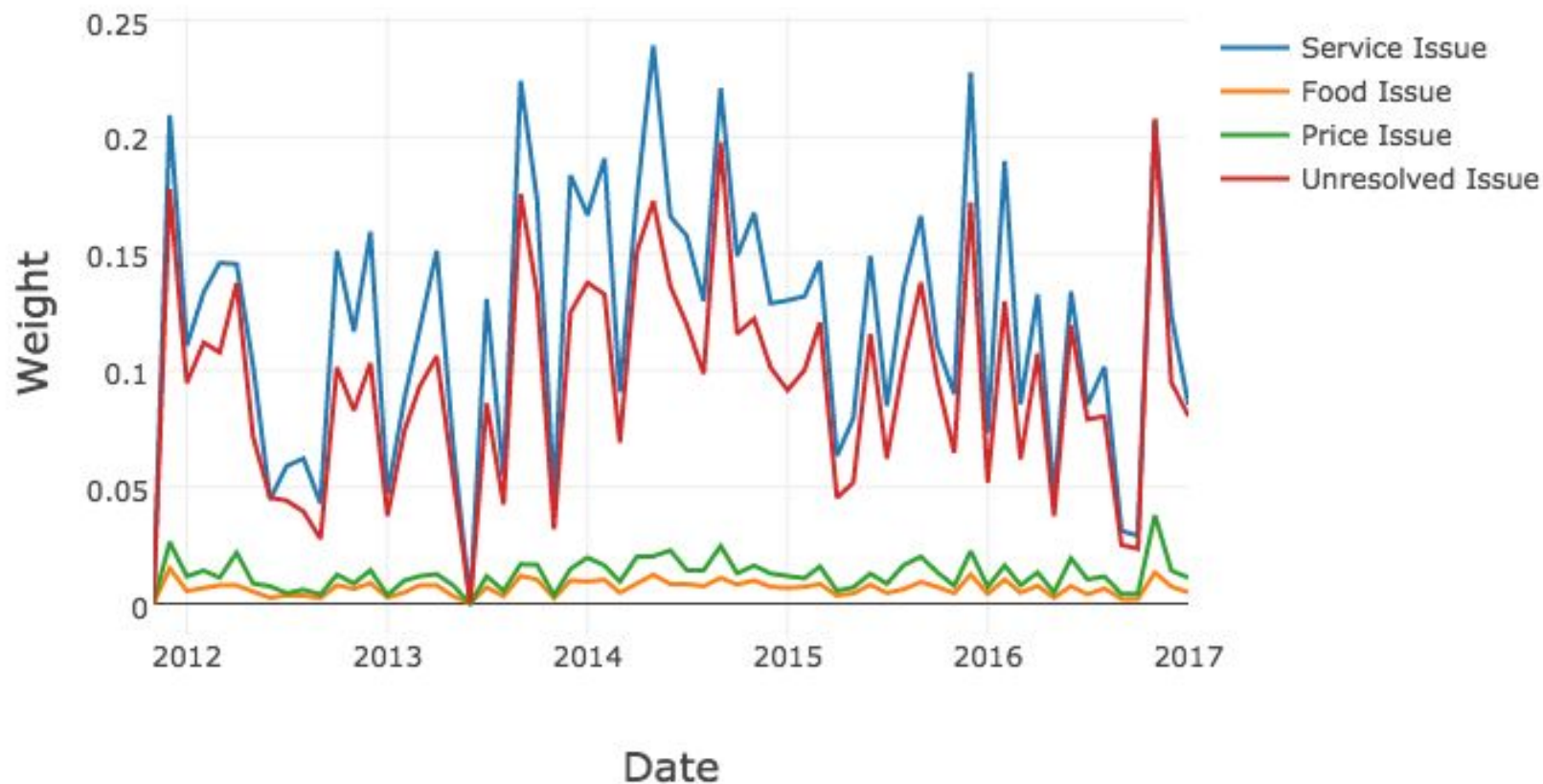
- K-means
- Clusters have similar average star rating
- Review_count is the main distinction
- Latitude and longitude once again are critical

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 |
|--------------|-----------|-----------|-----------|-----------|-----------|
| latitude | 36.746 | 34.532 | 42.765 | 35.911 | 35.989 |
| longitude | -103.384 | -113.111 | -72.059 | -111.101 | -115.026 |
| review_count | 206.414 | 18.012 | 14.472 | 777.515 | 2654.738 |
| stars | 3.784 | 3.704 | 3.561 | 3.884 | 3.692 |

Knowledge Gained

- Location
- Location
- Location
- The average star rating of an area all even out to ~3.7 stars
- Regular Yelp users are not evenly distributed
- Closed businesses cluster

Issue Distribution Over Time for Le Thai, Las Vegas



Applications

- New businesses can detect competition or trends
 - Decision tree indicates the category of a business is most discernible if they're open or closed
- Local businesses should adapt to the local community
 - Clusters average to nearly the same star rating. The 5 star system is relative to the best and worst in that area