Learning From Yelp

David Nichols
Department of Computer Science, Stanford University

Project Summary

Using machine learning techniques, this project aims to learn, primarily from reviews, what drives a restaurant to obtain a high Yelp rating, whether it be key terms such as 'service' or 'freshness', location, or some other variables.

Background / Data

Yelp is the leading platform for business reviews and ratings. In particular, it is considered the go-to source when trying to decide on a restaurant to eat at. Due to that, from the perspective of restaurant owners/managers, it is critical to understand the feedback they are getting from Yelp in order to position themselves to earn high ratings, and thus drive customers to their restaurants.

The core dataset for this project consists of:
- 23,000 Restaurant business listings with basic attributes
- 1.5 million text reviews for these restaurants with 1-5 point ratings

Machine Learning Problem Formulation

For the machine learning problem, the primary **features** are:

- User submitted text reviews
- Retaurant category (French/Chinese/Italian/etc)
- Restaurant Location
- Opening Hours

The **response variable** for this problem:

- 1-5 star rating awarded by the reviewer. This was collapsed into positive (4 or 5 stars) and negative (1 to 3 stars) as the average rating is 3.5.

Feature Engineering

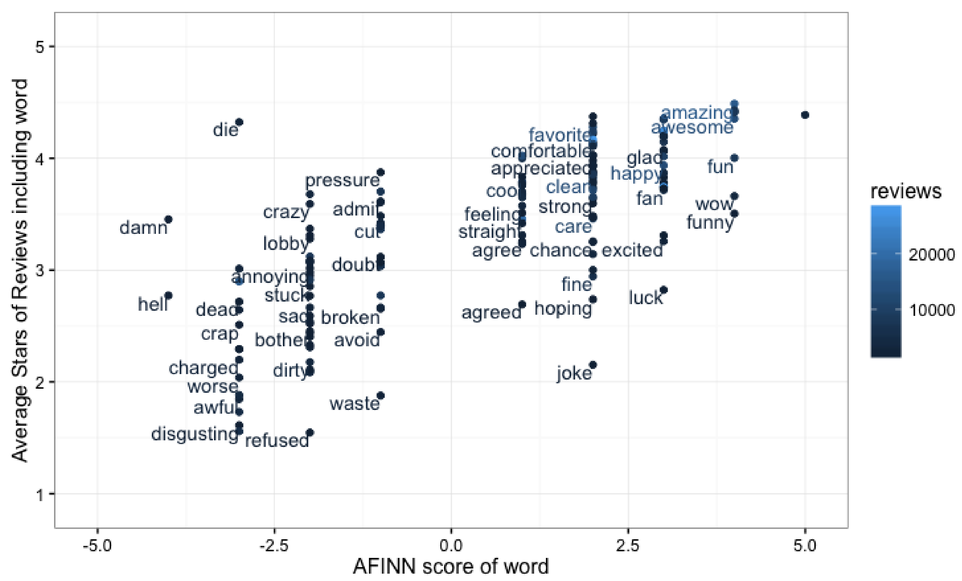
While some of the features came straight from the raw data, a few had to be pre-processed, most notably the reviews and location data:

Natural Language Processing - Given the user-generated text reviews, one of the goals was being able to infer sentiment information from the reviews in order to predict how the user will rate the restaurant. The bag of words approach to NLP was used which boils a review into the words that it contains. The sentiment was then inferred using two metrics - learning the average review rating that the word appears in in a Yelp training set, and using an independent measure of sentiment known as AFINN. By blending these two sentiment metrics, a balance of contexts was used for determining sentiment of a review.

K-means clustering - The dataset included lat/long coordinates of each restaurant, but on its own, that's not particularly useful information. K-means clustering was performed on the locations in each metro area to get a sense of how relatively dense of an area the restaurant is in. The within-cluster dissimilarity was normalized and used as the measure of density in the restaurant's area.

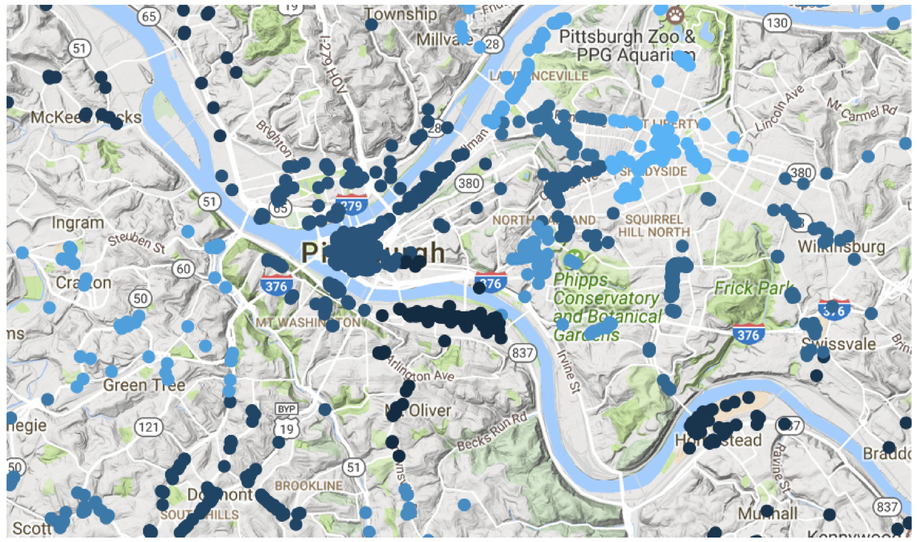
Natural Language Data

Figure: Average rating of reviews containing word vs. AFINN values*



Location Clustering

Figure: K-means clusters applied to Pittsburgh-area restaurants



Preliminary Findings

Table: Prediction accuracy by model

Model	Training Accuracy	Test Accuracy
Linear Regression	0.665	0.648
Ridge Regression	0.659	0.651
Logistic Regression	0.658	0.649
Gaussian Discriminant Analysis	0.685	0.678
SVM	0.671	0.661

- GDA currently has best performance
- Regression models doing well on binary response

Discussion & Further Research

Remarks:

- From initial experimentation with ML methods, it certainly appears possible to perform significantly better than random.
- Sentiment scoring from the reviews is consistently an important predictor of rating.
- Restaurant categories certainly play a role in determining ratings given to restaurants.

'Positive' Categories	'Negative' Categories
French	Tex-Mex
Delis	Buffets
Polish	Fast Food
Vegan	Chicken Wings

Further research:

- By applying more sophisticated Natural Language Processing techniques (possibly deep learning) to the review data, one would imagine that model performance could be improved.
- Modeling networks of users and possibly applying collaborative filtering could help predict user ratings.
- Using computer vision techniques, there is a set of restaurant photos that accompanies the dataset that could be used to build additional features for the model.

**Inspired by work by David Robinson*

david.nichols@stanford.edu