Strength in numbers? Modelling the impact of businesses on each other

An exploration of the Yelp dataset

Amir Sadeghian & Hakan Inan & Andres Nötzli

PhD Students, Stanford University

Introduction

In many cities, there is a small number of streets with a lot of restaurants. Being in a street like this is a double-edged sword for the individual restaurant. On one side, it is valuable because it gets them the attention of potential customers for free. On the other hand, the restaurants are competing for customers with similar needs and the offerings are not free from overlap. When a new business opens in a cluster, this delicate balance is disturbed. The goal of this project is to model the impact of a new business on the existing businesses. Our hypothesis is that the new business has an impact on the perception of customers of existing businesses. With increased competition, customers have to reevaluate existing businesses taking into account the new options. We use customer ratings as a proxy for the value of a business and to observe this reevaluation.

Main Objectives

1. Business Clustering
2. Impact of a new business on a cluster
   - Propose and test impact models
   - Use machine learning techniques to predict the impact

Dataset

Yelp is a website where users view and review businesses like restaurants. The dataset contains data from several cities and there is a rich set of attributes for each business.

- 42,153 businesses
- 320,002 business attr.
- 31,617 check-in sets
- 252,898 users
- 403,210 tips
- 1,125,458 reviews

Preprocessing

Running average of ratings

The running average of ratings plays an important role when predicting the correlation of two businesses. The raw user ratings are highly noisy and relatively sparse. Figure 1 depicts an example of a moving average for a business over time. In addition, we filter out businesses with a low number of ratings.

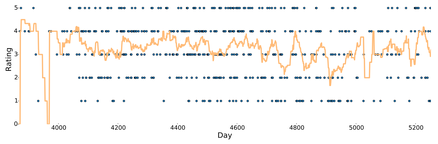

Figure 1: Moving average of ratings for a specific business

Business Clustering

We checked different clustering algorithms and for the same number of clusters, K-Means based on the geographic locations of businesses had the best result. The result of two clustering algorithms results are shown in Figure 2.

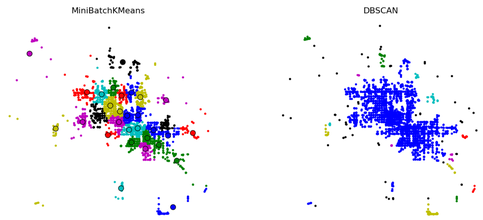

Figure 2: K-Means and DBSCAN clustering

Models

Features

We found that a combination of geographical distance, price range difference, age, and business categories work best as features to describe the relationship between two businesses. The categories encode attributes of a restaurant like the type of food that it serves.

Conditional Mean Analysis

Hypothesis : Opening of a new business has an impact on the mean of ratings of the businesses nearby.
Proxy : Calculate the mean ratings of the nearby businesses before and after a new business opens, and get a comparative metric.
Expected results : The change in the conditional means of the existing businesses may be predicted using the attributes of the existing businesses and the new business.

$$E_{before}(b) = \frac{1}{R_b} \sum_{x:-M+d_0 \le d_x \le d_0} r_x(b) \qquad E_{after}(b) = \frac{1}{R_a} \sum_{x:d_0 \le d_x \le d_0+M} r_x(b),$$

d_x = day of the review x d_0 = opening day of the new business
r_x = rating of review x M = number of days to average over

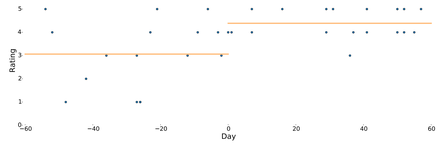

Figure 3: Example mean analysis for a pair of businesses

Trend Analysis

Hypothesis : Opening of a new business has an impact on the trends of ratings of the businesses nearby.
Proxy : Fit separate lines for the ratings of a business both before and after a new business opens in the neighborhood. Calculate a metric based on the difference in the slopes of the two lines.
Expected results : The change in the trends of the existing businesses with respect to the launching of a new business in the neighborhood may be predicted using the attributes of the existing businesses and the new business.

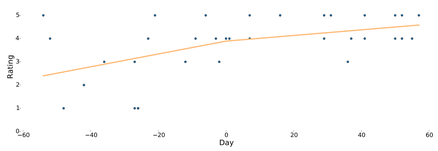

Figure 4: Example trend analysis for a pair of businesses

General Trend Analysis

Hypothesis : The exact opening date is not known and since prediction is noisy, the trend analysis might fail. The general trends of the existing businesses around a rough estimate of the opening time of a new business may reflect (with less noise compared to the trend analysis) the impact of the new business on them.
Proxy : Fit a single line for the ratings of a business around the estimated opening date of a newly opened business in the neighborhood. Determine if the business has an increasing or a decreasing trend based on the slope of the line.
Expected results : The general trends of the existing businesses around the launching date of a new business in the neighborhood may be predicted using the attributes of the existing businesses and the new business.

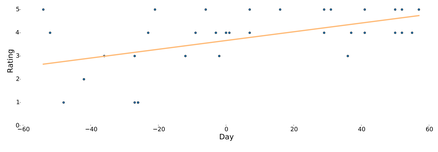

Figure 5: Example general trend analysis for a pair of businesses

Correlation Analysis

Hypothesis : None of the previous approaches provided an adequate metric. For our last model we opted to use a correlation metric which reflects the relationship of two businesses over a long period of time.
Proxy : Compute the correlation of ratings over time.
Expected results : The correlation can be predicted using the attributes of the existing businesses and the new business.

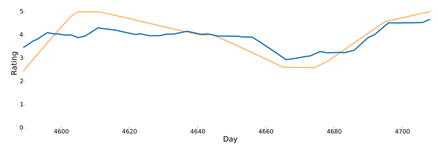

Figure 6: Example of correlation analysis for a pair of businesses

Results

The following table contains the mean training score and the 10-fold cross-validation score for predictions in a single cluster. The cluster consists of 147 businesses which corresponds to 10730 pairings. We did two types of experiments: (a) classification of positive vs. non-positive outcomes for all models and (b) classification of significant vs. insignificant outcomes (absolute value bigger than a certain threshold) for all models.

		sig/insig classif.		pos/neg classif.	
		60 days	90 days	60 days	90 days
Conditional Mean	SVM rbf	0.84/0.61	0.82/0.60	0.84/0.60	0.85/0.65
	Logistic Regression	0.61/0.52	0.60/0.52	0.56/0.49	0.56/0.49
Trend Analysis	SVM rbf	0.83/0.62	0.82/0.59	0.84/0.60	0.81/0.58
	Logistic Regression	0.60/0.55	0.58/0.48	0.57/0.49	0.56/0.51
General Trend Analysis	SVM rbf	0.84/0.62	0.80/0.62	0.81/0.60	0.83/0.63
	Logistic Regression	0.59/0.52	0.58/0.54	0.57/0.49	0.58/0.49
Correlation Analysis	SVM rbf	0.85/0.81		0.86/0.81	
	Logistic Regression	0.84/0.82		0.86/0.83	

Table 1: Training and 10-fold cross-validation score for predictions

Note: We do not present the results for linear, poly and sigmoid kernels because the results were significantly worse than SVM with the rbf kernel.

Conclusions

We tried to predict the interaction of two businesses with multiple models. The quality of our predictions turned out to be relatively low for most models. Possible reasons are:

- The rating data is noisy and sparse at the same time. Most businesses don't have ratings every day and the variance of the ratings for a given time period is high.
- The training score of SVM is high in general but the cross-validation score is low in a lot of cases which may be a hint that SVM is overfitting.
- For the conditional mean analysis and the trend analysis, we assume that the opening date of the business is close to the first review submitted for the business. This might not always be the case.

We achieved good results when predicting correlation and this shows that our approach has merit. We also observed that SVM generally outperformed logistic regression in cases other than correlation analysis.