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

Amir Abbas Sadeghian
amirabs@stanford.edu

Hakan Inan
inanh@stanford.edu

Andres Nötzli
noetzli@stanford.edu

ABSTRACT

Here goes nothing.

1. 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. [**?**] Our hypothesis is that this balance is a function of the size of the cluster. The bigger the cluster grows, the more customers start to search for a restaurant within this cluster, increasing the number of total customers of the cluster and emphasizing the symbiosis of the restaurants living in the cluster. At the same time, however, the number of potential customers is limited and as it nears saturation, the competitive nature of the relationship between the restaurants grows stronger. The goal of our project is to test this hypothesis by finding clusters of restaurants in a dataset and to model how the individual restaurant is affected by being part of the cluster. Machine learning will play a crucial role in this process, from finding clusters of restaurants to model the impact on the individual restaurant. We will start by comparing clusters of different sizes but ideally we would like to find clusters where we can observe growth over time and the effect on the restaurants over this time period. If the project proves to be successful, it has the potential to offer a unique insight in the relationship between restaurants in such clusters. To design and evaluate our model, we are planning to use the Yelp dataset.

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

2. THE DATASET

Yelp is a website where users review businesses like restaurants. The dataset contains data for more than 40000 businesses and more than 1 million of reviews. There is a rich set of attributes for each business and there is additional data like the number of checkins that can be used to model the popularity of a place.

Two data points that are missing from the dataset are the opening and the closing date of a business. To compensate for the lack of information, we use a simple heuristic: We assume that the business opened on the date of the first comment and we assume that it has been closed on the date of the last comment if the last comment more than 2 months older than the newest comment in the dataset. We argue that this is a reasonable choice because our project requires us to look at businesses with a reasonable number of ratings and in these cases the opening/closing date should be reasonably close to the date of the first and the last review. 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

252,898 users

320,002 business attr.

403,210 tips

31,617 check-in sets

1,125,458 reviews

3. PREPROCESSING

3.1 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 **??** depicts an example of a moving average for a business over time. In addition, we filter out businesses with a low number of ratings.

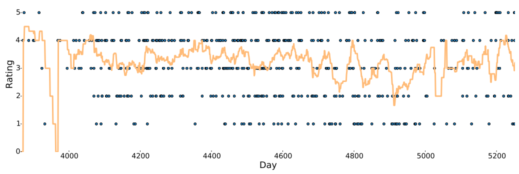

Figure 1: Moving average of ratings for a specific business

3.2 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 **??**.

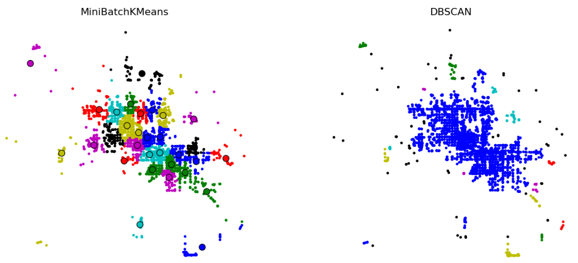

Figure 2: K-Means and DBSCAN clustering

As we previously described in the introduction, we are interested to look at clusters of businesses. The first step in our project is thus to find a good way of clustering businesses. We found that using zip codes to group businesses is ineffective as groups of businesses often span zip code boundaries. We experimented with multiple clustering methods and ended up using k-means clustering with the geographical location as features because we are interested to observe interaction between businesses that are physically close to each other. Using k-means we are taking the advantage of clustering close businesses together and also putting far businesses that are not influencing each other into different clusters. In this case we can assume that the businesses in two different clusters are independent, and only businesses in one cluster influence each others behaviors. We will have a brief overview on the different clustering techniques used and a benchmark that reasons which one works the best in the case of our study.

4. FEATURE SELECTION

In various domains, like text learning, image classification, and specially cases where there are many features compared to data samples feature selection techniques are used. The feature sets selected for our model plays an important role to define a better feature similarity measure which can lead to improvement of our prediction algorithms and also finding correlations between different businesses. Once feature which is used to find the correlations of business ratings in section 5, is the moving average of ratings stars. This feature give us an understanding of how a business was performing in a period of time and how the users rated that specific business in that time.

5. CORRELATION OF RATINGS

After getting the clusters, we analyze the correlation of ratings between different businesses in the same cluster. To do so, we first apply a Gaussian filter to the time series of ratings to smooth out local fluctuations in ratings. The correlation of the ratings is then given by:

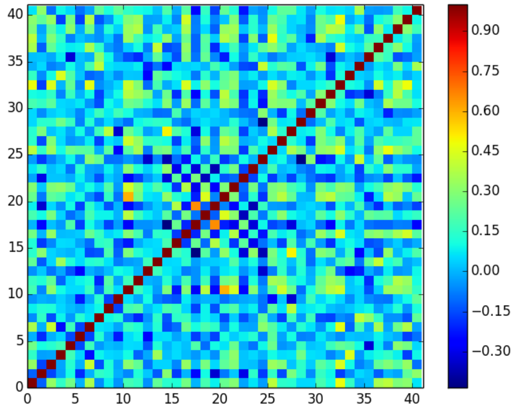

Figure 3: Correlation between ratings of businesses in the same cluster

$$\rho_{XY} = \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}}$$

$$E(X) = \frac{1}{n}\sum_{t=0}^{n} r_t$$

Where r_t is the filtered rating. In this step we discard all businesses with a low number of ratings because they provide not enough signal to get a good estimate for the correlation.

As Figure 3 shows, there are a couple of cases with strong correlation.

6. 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.

7. MODELS

All the analysis in the project was based on using the pairwise features outlined above for predicting pairwise metrics to be defined in what follows. Specifically, in this section we introduce 4 different metrics which we will henceforth call ”pairwise impact metrics”.

NOTE: Anything below is within a cluster

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.

The mean ratings were calculated as follows:

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

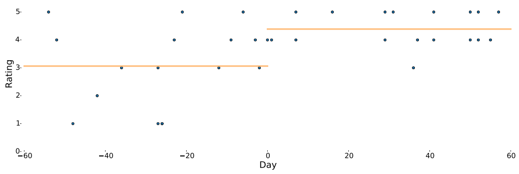

Figure 4: Example mean analysis for a pair of businesses

$E_{after}(b) = \frac{1}{R_a} \sum_{x: d_0 \leq d_x \leq d_0 + M} r_x(b)$

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

For this analysis, we needed a date of opening for the businesses. However, we didn't have the true opening dates in the dataset, and we estimated them to be the dates of the first review for the businesses.
The pairwise impact metric in the conditional mean analysis was determined as $E_{before} - E_{after}$.

Trend Analysis

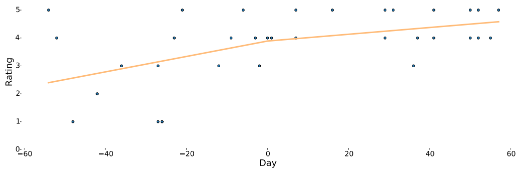

Figure 5: Example trend analysis for a pair of businesses

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.

First, we estimated the opening date of the businesses as explained in the previous subsection. Then, for each pair of businesses we fit two lines for the ratings of the older business around the origin (the estimated opening date of the newer business) within a specified time window, imposing that the lines touch at the origin . Specifically, we are solving the following least squares problem:

$$\begin{bmatrix} x_{before} & 0 & 1 \\ 0 & x_{after} & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ c \end{bmatrix} = \begin{bmatrix} y_{before} \\ y_{after} \end{bmatrix},$$

where x_{before} (x_{after}) is a vector of the days of filtered ratings of the older business before (after) the newer business opens, y_{before} (y_{after}) is a vector of filtered ratings of the older business before (after) the newer business opens, s_1 (s_2) is the slope of the line fitted to the ratings of the older business before (after) the newer business opens, and c is the common intercept for the two lines. One thing to note here is that the elements of x_{before} and x_{after} are shifted such that the last element of x_{before} is 0 and the first element of x_{after} is 1.

The pairwise impact metric was determined to be the difference in the angles of the two slopes.

General Trend Analysis

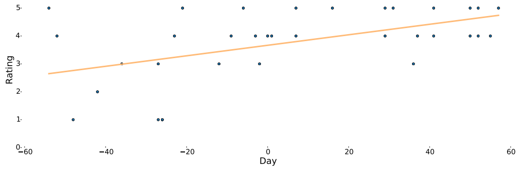

Figure 6: Example general trend analysis for a pair of businesses

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.

The method to apply was very similar to that for the trend analysis, with the distinction being that for general trend analysis we fitted a single line for the whole time window and calculated a single slope. Mathematically, we calculated the least square solution to the following equation:

$$\begin{bmatrix} x_{before} \\ x_{after} \end{bmatrix} \begin{bmatrix} s \\ c \end{bmatrix} = \begin{bmatrix} y_{before} \\ y_{after} \end{bmatrix},$$

with everything except for s is as defined in the previous subsection. s is the slope to the fitted line for the whole time window .

Correlation Analysis

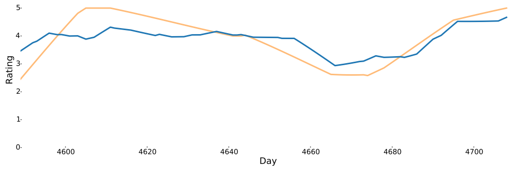

Figure 7: Example of correlation analysis for a pair of businesses

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.

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.

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.

8. CONCLUSION

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.

9. FUTURE WORK

We have already performed different types of clustering on the data and chose the best one based on the geographical location of the businesses. It would be interesting to perform other types of clustering techniques based on the features distance metrics described in the features section, and study how the correlated businesses are geographicly located. This would allow us to analyze the structure of a cluster, e.g. whether a given cluster consists of many similar businesses (for example Chinatown in San Francisco) or whether a cluster is heterogenous (for example the Great Mall).
Since we have used very common business attributes like geographic distance, open hours, business types, and similar common features for our studies, we have the advantage to do the same experiments on similar datasets. In the future, it would be interesting to see if similar or even better observations can be made on different datasets, where more information on the businesses is available. In addition it would be interesting to look interactions between businesses other than the opening of a new business.

		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