College of Computing and Informatics
University of North Carolina at Charlotte

A project report on

Impact of Competition on Local Businesses.

Project Guidance By
Dr. Minwoo Lee

"Team5"

Aditya Rathi Amit Shetty Mohit Varma Gaurav Yadav

Introduction:

When a new business opens in an area, it makes an impact on the ongoing trends of other businesses in the locality. The main goal of this project would be to calculate the impact of new business on the existing local businesses. New restaurant options would certainly put customers in a dilemma and one of the things that customers usually do is re-evaluate their options between new and existing businesses. We believe that restaurants thrive in a competitive environment if they have a better knowledge of key factors that will contribute to their success. This is the main motivation behind creating a predictive model through this project. The way through which we plan to achieve this is by analyzing customer reviews and footfalls on restaurants before and after a new one opens in the vicinity. The analysis is expected to provide us a pattern that we will use for training a model.

Background/Related Work:

Our initial idea for studying how businesses can impact each other came from a Harvard Business school whitepaper on the impacts of local economies in clusters [2]. The paper discusses the importance of competition in local economies to stimulate growth in the modern era. The authors have concluded by studying the changes business has made when another one is introduced in the same cluster. The paper has given a very good explanation of how clusters can play a vital role and have compelled us to explore the topic further. To understand how customers can make an impact on the way a restaurant conducts its business we studied a survey [3] conducted by the University of Lancaster. The paper shows how in an area where restaurants serve the same type of cuisines, business owners have accentuated other facilities at their restaurants. This is true since to survive today restaurants sell not only food but also the experience of eating at their establishment. This paper was very interesting since as restaurant-goers we could relate as to how apart from price, we look at other factors to decide where to eat.

We researched work related to the business field that will not only be intuitive to learn but also will have a real-life application. The related research was done in (Amir Abbas Sadeghian, Hakan Inan & Andres Notzli). This paper forms the basis of our project since it arrives at the same conclusion as the previous papers discussed but goes into technical details of how machine learning can help make sense of this information. This paper explains the way impact can be calculated using hypothesis and working towards deriving conclusions based on it. The hypothesis made is that the opening of a new business has an impact on the mean of ratings of the businesses nearby. The paper then discusses the implementation of this hypothesis by calculating business ratings before and after a new business has been opened. Besides, the paper also implements clustering techniques such as DBSCAN and K-Means to cluster existing businesses. One of the cons of using the DBSCAN model as per the reference paper used is that it is not as good in predicting the impact compared to K-means. [1]

While working on this project we have come to revise the scope of our data by limiting our data to include only restaurants instead of all establishments recorded by Yelp. By doing so, we are focusing on what each restaurant has to offer apart from its cuisines that will bring in customers in a particular area. Each attribute column in our dataset will play a crucial role in determining a restaurant's success. Our goal is to see how a restaurant's fare changes when a new one opens in the same locality.

Method:

So far, the approach followed takes us through the following steps:

1. Preprocessing:

The Yelp dataset is extremely large and is a huge collection of unstructured data. The data consists of information about not only restaurants but also bakeries, bars and slaughterhouses. The primary challenge here has been cleaning the data since the data is in JSON and contains key-value pairs within

key-value pairs which is not consistent across the board. A majority of the development time has gone into solving this issue and cleaning the data. The results have been a dataset containing information about restaurants only.
The next step after getting rid of the obvious invalid entries has been to mold the data into something that can be easily interpreted by our model. Certain complex attributes have been
broken down by taking the most common attributes for all restaurants. We used one-hot encoding to show which perks a certain restaurant provides.

2. Running Average Model:

This step aims to calculate how the average rating of a restaurant changes over time. To achieve this, data is extracted from a separate data file containing individual reviews with their ratings. Ratings are linked to the businesses with a unique business ID generated by Yelp. Reviews from a certain business are grouped and sorted by the date when the review was posted and plotted on a graph. The average reviews need not be explicitly calculated as they are already included in the primary data file.

3. Calculate Footfall:

Yelp provides footfall information for a certain business by recording the date and time a customer checks in. We will be using this information to group the data by restaurants, then by date (by day or month) to get an accurate understanding of how the footfall of a restaurant changes over time.

4. Model Implementation:

The distance matrices of various features (extracted in preprocessing such as ratings, footfalls, and pricing) can be calculated. This will give us a pairwise difference between the features of restaurants. Using this information, we identify patterns in the dataset and use it for training.
This approach is unlike (Amir Sadeghian et al.) since we expect to find new insights such as the decision to open a new restaurant, the correct timing, attributes to be taken into consideration depending on the geolocation and customer preferences captured from the Yelp dataset [3]. Furthermore, we have gained new insights from the data such as the attributes that contribute more to a new business impact by clustering them as per the geolocation and customer preferences that weren't previously considered using this model.

5. Estimation:

We make our predictions by first calculating the average performance of a restaurant in that area (called Ebefore). Once we get our results we wait for a certain amount of time (which can range from a few weeks to a couple of months). Post which we collect data on the ratings provided by the customers (called Eafter) and train our model accordingly. This is in stark contrast to the approach used by (Amir Sadeghian et al.) where a cluster-based approach was used to estimate the business' success. We are focusing on using feature engineering to drive our predictions.

Benefits of using this approach:
While one benefit of using this model is impact estimation, this method can also be reverse engineered to show which attributes affect the success of a restaurant/business per geolocation. The other benefit is, this is a plug and play model where different combinations of restaurant attributes can be used to estimate our target variable and at the same time, the input attributes will tell us which areas of the restaurant business the owners have to focus on based on their geography.

6. Experiments:

Data:

Yelp Datasets which include JSON dumps of business feature attributes, user reviews and ratings, and geolocation data.

Experiment 1: Running Average Method:

To understand how a restaurant performs over time, it is important to understand how customers feel about it. Whether or not it is their first choice when deciding to eat out. If they decide to visit the business, it is in the interest of the latter to give the best service to get a good rating. We calculated the average ratings of a restaurant over time to see how they perform.

Steps to calculate the Running Average:

1. We calculated the number of Check-in for the restaurant and based on Checking the restaurant was chosen.

Out[16]:

	Business_Id	Count_Checkin
0	5LNZ67Yw9RD6nf4_UhXOjw	46384
0	IZivKqtHyz4-ts8KsnvMrA	38277
0	Wxxvi3LZbHNIDwJ-ZimtnA	32343
0	EI4FC8jcawUVgw_0EIcbaQ	30098
0	RESDUcs7fliihp38-d6_6g	28872
0	DkYS3arLOhA8si5uUEmHOw	25835
0	eAc9Vd6loOgRQolMXQt6FA	25612
0	hihud--QRriCYZw1zZvW4g	25416
0	uGupeWqih0yIcCg8anM1PA	22790
0	K7lWdNUhCbcnEvI0NhGewg	22225

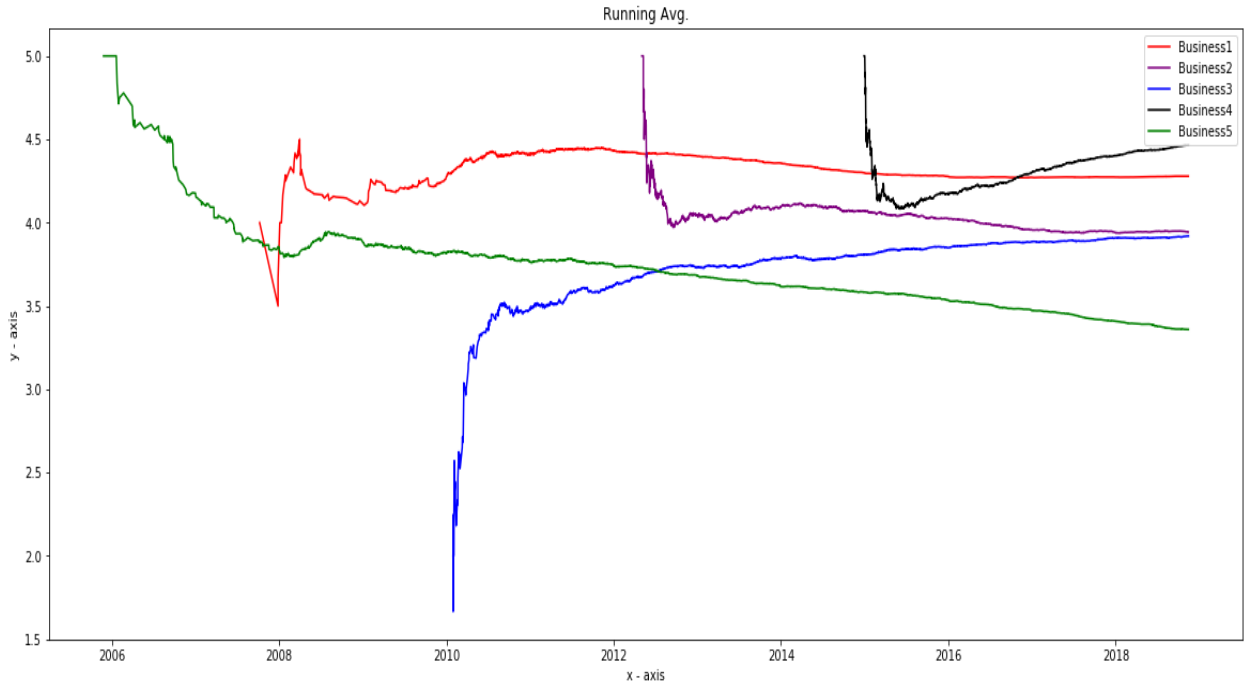
2. The next was to select the restaurant for our experiment with maximum check-in in the geographical location

3. We are using the below formula to calculate the running average of all the restaurants we selected.

$$\frac{1}{n}\sum_{i=0}^{n} R_i$$

where, n is total number of reviews
R_i is the current review

Graph:
Y-Axis: Is the Running Average of Rating
X-Axis: Is the Date

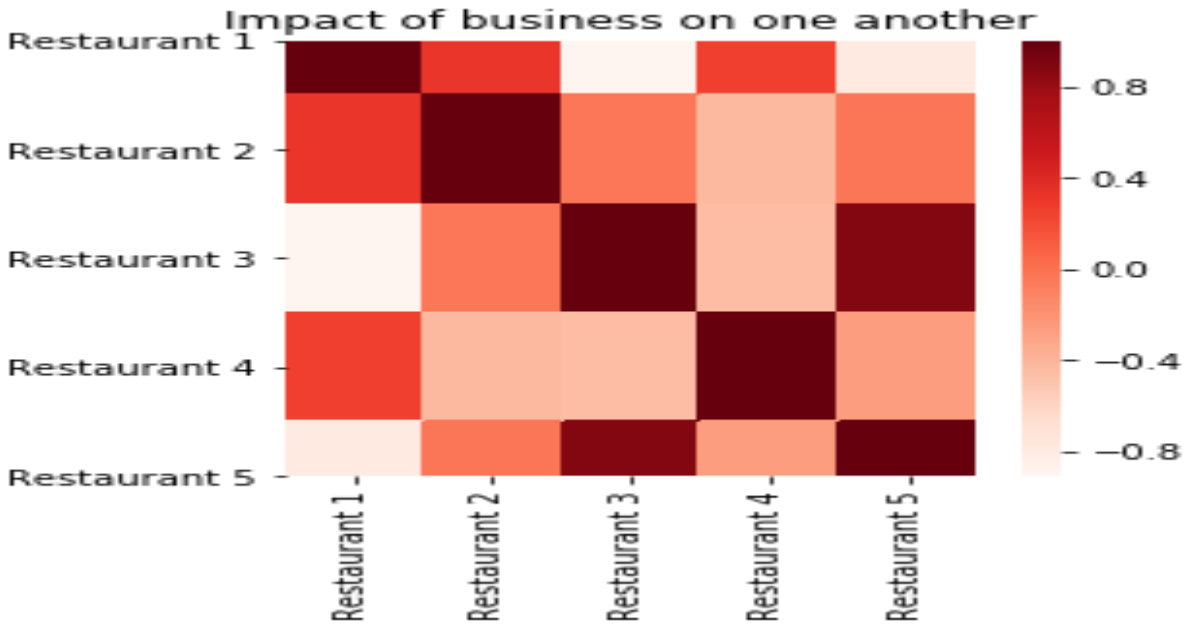


The graph shown above can give us a bird's eye view about a restaurant's performance and give key insights such as the effect of a new business on existing competition. How much of an effective new business can also be studied from this one graph. We can see in some cases how the introduction of a new business may cause the ratings

Experiments 2: Impact Matrix

Unlike what was followed in the research of this project, we followed a different approach to calculate the impact of restaurants on each other. The research calculated impact by using the correlation coefficient which was derived by taking differences between features such as prices, sanitation ratings, parking. We instead, used slope array to figure out the impact. The impact matrix gives us a clear depiction of how two restaurants affect each other. Using slopes between two data-points, an array of slopes was calculated for each rating graph. We then used these arrays of slopes to figure out how each restaurant affected other restaurants. This was done in an incremental manner. We followed the same concept and plotted a correlation matrix of impact. The result has been shown in the following graph.

We used a polyfit function to figure out slopes and intercepts of a line between two data-points. The polyfit function internally uses the same formula Y2-Y1/ X2-X1 for slope calculations.



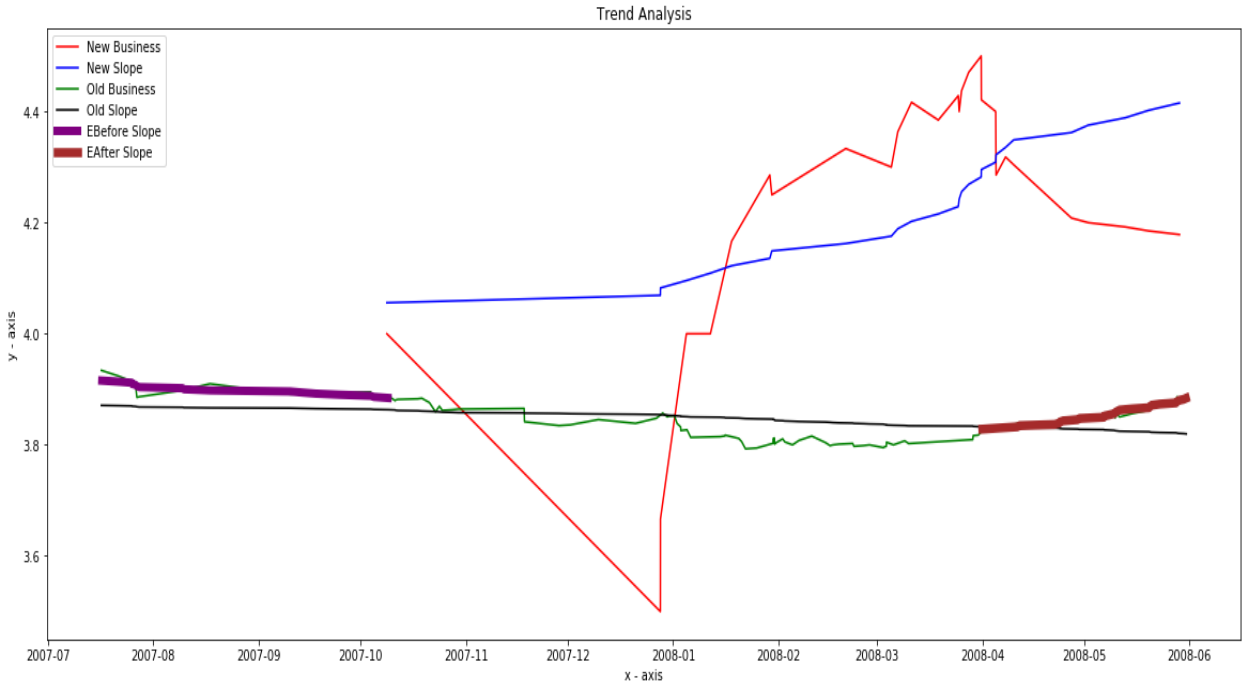
Experiments 3: Trend Analysis

In trend Analysis we are calculating the trend of the business before the new business is opened in the vicinity i.e. called EBefore and we are calculating the trend after the business is opened. The impact of new business on local markets may not be immediate and can take some time. Therefore, the period after which, EAfter is to be calculated was given manually and represented in the equation by M. This EAfter is the actual business that is affected after the new business is opened in the vicinity.

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

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

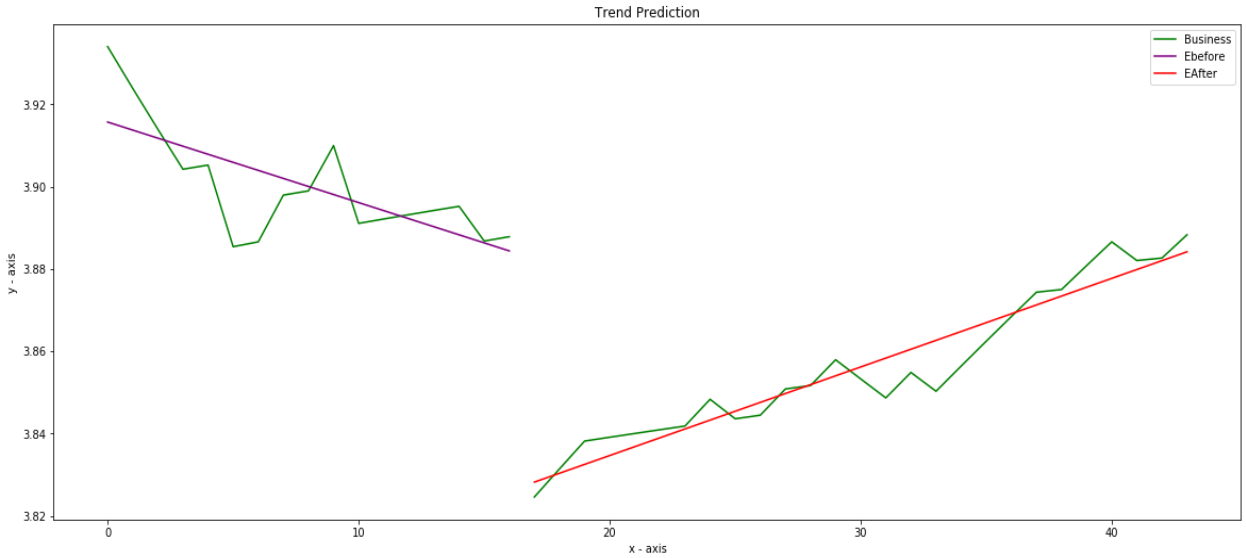
Graph: -
X-Axis: Date
Y-Axis: Running Average of Ratings



Experiment 4: Prediction

The trend analysis of slopes was derived from the dataset. This analysis provided us with an array of slopes that we eventually used to train our model using least square regression. The least-square regression helped us determine a line. The trend analysis of slopes was derived from dataset. This analysis provided us with array of slopes which we eventually used to train our model using least square regression. The least square regression helped us determine a line through our predicted review points. In the graph shown below, the after rating is the predicted result. The individual data point you see (on green graph was the predicted slope) and then least square regression was used to draw a line to represent the future trend.

X-Axis: Day at which the reviews are taken. (The dates are number from zero since that is when the prediction has started, and every consecutive day is in increments of one)
Y-Axis: Rating of Review



Differences from existing research:

Refer below sections to understand more about the differences from existing research we have performed in this project.

- Experiments 2: Impact Matrix
- Experiments 3: Trend Analysis

Reflections (Response to the instructor's feedback):

1. *this is not "group name"*

Team5 is the name decided by the TA's in our poster presentation session and was the name by which we were rated. We have decided to go with that name.

2. *Author's name (XXX et al.)*

We have taken that into consideration and have made the necessary changes to our report. We couldn't do the format (xxx et al., <year of publication>) since the year of publication is not mentioned in the Google Scholar search results. Refer author's Google scholar page (https://scholar.google.com/citations?user=ZqbHW0gAAAAJ&hl=en&oi=sra)

3. *Typical pipeline. explain your uniqueness*

We have made changes to our pipeline and divided our model creation step (in the midterm report) into different parts that takes into account how our approach differs from the source paper. We have also mentioned the benefits of building the model in a different way.

4. *this must be your difference. Please highlight your reasoning and possible pros*

We have made the necessary changes in the model building section. please refer the previous point

5. *Do you mean 5*

we missed that mistake in our last report. We have proofread our final report and such errors should not happen again.

6. *please make sure to further clarify your method especially your difference.*

Please refer experiment 2 and 3. We have made that the primary focus of writing this report and based on the progress we have made since our last project; we believe we are in a better position to explain the model implementation and prediction in a better way.

References:

[1] Sadeghian, Amir Abbas, Hakan Inan, and Andres Nötzli. "Strength in numbers? Modeling the impact of businesses on each other.
[2] Porter, Michael E. "Clusters and economic policy: Aligning public policy with the new economics of competition." ISC White Paper, November (2007)
[3] Auty, Susan. "Consumer choice and segmentation in the restaurant industry." Service Industries Journal 12.3 (1992): 324-339