Inferring Future Success of a Business Based on Internal and External Factors

Harika Sabbella, Mohammad Adil, Sugeerth Murugesan

May 30, 2014

Abstract

In order to predict if business will do well in a city/town, we must uncover the properties of the location and the tastes of the user. For example, a not-so-good Indian restaurant in very popular area will be more likely to prosper than a very good restaurant in a not so very happening place. We can visualize the businesses given in the Yelp dataset, which contains 335,000 reviews texts written by at least 70,000 users, in heat map which intuitively lets us know the most happening place within a community. One way to measure the success of a business is to gure out the number of positive reviews a business will get based on the number of positive reviews it currently has. In our work, we aim to develop techniques and visualizations that dene the future success factor of a business. Later, we aim to detect shy users who were paid by Yelp to write reviews.

Overview

The success of a business, without doubt, depends on many outside factors not pertinent to the just the business itself (those factors that are outside the control of the business)–some of these factors include time of the year, the current economy, the weather, and the customer base. One such outside factor which influences the success of a business is the success of businesses within the same community. Our hypothesis is that the success of a business depends not only on the quality of the products the business itself offers, the satisfaction of its employees and customers, and the profit it generates, but also on the success of the neighboring businesses. Our idea which involves integrating the success of neighboring businesses is novel and different from approaches before which were based simply on just the internal factors of the business at hand–these approaches did not consider any external factors. In order to make a sound determination of the future success of a business, such external factors most obviously need to be considered! In our case, we have chosen to consider the success of the neighboring businesses.

Literature Review

The paper Inferring Future Business Attention mentions that there are two categories of features: metadata (longitude/latitude) about business and description of reviews (number of stars/number of reviews). By performing what the authors call sentiment analysis on the reviews, we can determine the features of a business and the positivity or negativity associated with these features. Sentiment analysis involves [mining] the most frequently occurring keywords among all restaurant reviews and [using] the counts of the sentiments for each of these keywords as features [1]. Each feature is associated with an adjective and the adjective is determined to be positive or negative. Then, features which are to be used to determine the positivity or negativity of a review are picked using PCA. Principal component analysis (PCA) is a tool that reduces dimensionality of a feature space–in other words, they reduced the number of features in the working set and predicted future business using these new features. Three ways to collapse the data to extract out certain features:

1. Univariate Feature Analysis– look at the prediction quality (meaning how much would including this feature help us in predicting the total number of future reviews) of each feature individually; use the best n of these in the prediction model

2. Greedy Feature Removal – Involves [removing] the worst performing feature until a certain performance is achieved or until there are a specified number of remaining features left [1].

3. Naive approach – A naive approach to feature selection uses an exhaustive search over all possible subsets of features to pick the subset with the smallest test error [1].

The work involved running a number of experiments to predict the future number of reviews using simple reviews (using features about the business metadata) first and then using review text features second. By the collapsing methods mentioned above, they were able to choose a few of these different features (to eliminate what they call noise), and built a regression model (which relates the current user reviews to future number of reviews) and then predicted the number of future reviews using regression model. Specifically, the Support Vector Regression (SVR) was used as regression model.

We have built upon the algorithms and methods suggested in the Inferring Future Business Attention paper and come up with a way in which we can account for the external factors that affect the success of a business.

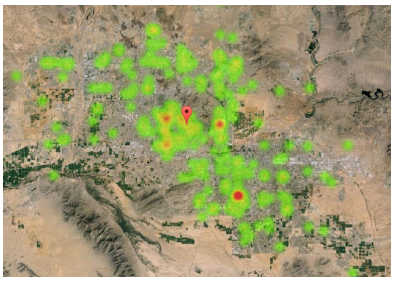

Figure 1: Heatmap showing the number of businesses in Phoenix, Arizona. Red spots depict areas with a large number of businesses.

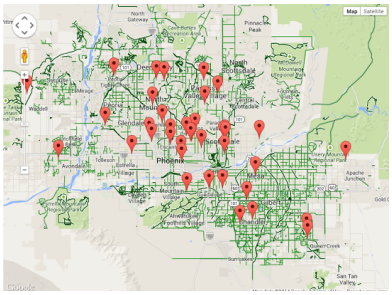

Figure 2: Visualization depicting the corresponding businesses in heatmap visualization given above

Initial Results

In this section,we will be describing a few methods that we have used to generate initial results. We begin with a simple context visualization of the yelp dataset in Arizona, Phoenix. We generated a heatmap visualization, which defines clusters of businesses that are located close to one another. Presumably, the areas colored in red are the areas where there are a large number of businesses. With the overall idea of the placement of the businesses in Arizona, Phoenix, we used two techniques for data analysis and generation of a regression model–semantic analysis and location analysis.

Semantics Analysis

The general approach that we used for the semantic analysis is:

- Given a business, P, we would like to derive a value, S, that indicates Ps current success; in order to derive the success factor:
  - (Step 1) Derive a value q which indicates how positive Ps reviews are
  - (Step 2) Derive values [b1 ,b2 ,..., bn] where b1...bn indicate how positive the reviews of the n neighboring businesses of P are

  - (Step 3) Integrate q and [b1 ,b2 ,..., bn] to derive a value S which is the success factor for P

- Given a business, P, we would like to predict how likely it is that P would succeed in the future
  - (Step 4) Take the success factor derived earlier, build a regression model, and derive a future success factor value based on its location

Given a business, P, derive a value, S, that indicates Ps current success (based on qualities of the business itself as well as the influence of the surrounding businesses). Then, using S, and a regression model figure out the future success factor of P

To figure out the number of positive reviews that a business, P, currently has, we first extract all the associated reviews for a particular business from the Yelp review dataset. Then, we tokenize these reviews into words. Using the Python Natural Language Toolkit, we can assign each word into its part of speech. Every time, we encounter a word that is a noun, we push it into the stack. In order to build a noun phrase, if the immediately occurring words after we encounter a single noun are also nouns, we push them onto the stack. We keep the latest occurring noun/noun phrase in the stack. When we encounter a word that is an adjective, we associate the noun/noun phrase in the noun stack with this adjective. At the end of this process, we have a counter which indicates the number of times a noun/noun phrase is seen throughout the reviews and the all associated adjectives of the noun/noun phrase. More concretely, here is a sample of the noun adjective dictionary:

Noun: [Number of occurrences, associated adjectives, Number of (+) adjectives, Number of negative (-) adjectives]

Chicken: [1, 'best', 0, 0],

Thai Pan: [7, 'divine, only, new, dish, wonderful, authentic, best', 0, 0],

Thai curries: [1, 'Chinese', 0, 0],

place: [3, 'dark, good, great', 0, 0],

restaurant: [2, 'next, clean', 0, 0]

Then, we built a list of adjectives that have a positive meaning and a list of adjectives that have negative meaning using Wordnet. To build these lists, we start out with a small list of positive adjectives such as great or wonderful and recursively derive all the synonyms of these words to generate get a somewhat complete list of positive and negative adjectives.

Using these positive and negative adjective lists, we can derive whether the given noun/noun phrase is associated with a positive adjective or a negative adjective. We increment the appropriate positive and negative counter for each noun/noun phrase. For example, using the same example as above, we would now have the following noun adjective dictionary:

Chicken: [1, 'best', 1, 0],

Thai Pan: [7, 'divine, only, new, dish, wonderful, authentic, best', 4, 0],

Thai curries: [1, 'Chinese', 0, 0],

place: [3, 'dark, good, great', 2, 1],

restaurant: [2, 'next, clean', 1, 0]

As a side node, for the noun Thai pan the total count of the positive and negative adjectives does not add up to 7 because the adjectives new, dish, and only are not particularly positive or negative words.

At this point, we have a list of all the good features (or nouns/noun phrases) associated with a particular business and all the negative features associated with the business. We sort the noun adjective dictionary by the number of occurrences of the noun and we pick the top fifteen most occurring noun/noun phrases (using the number of occurrences counter) and using the positive and negative counters, we make a determination of the success of the business–we say that a business is successful if we see that the number of positive noun/noun phrases associated with the top fifteen occurring noun phrases is greater than the number of negative noun/noun phrases. More specifically, the value q = (positive / (positive + negative)) * 100.

Location Analysis

Next, we derive values [b1 , b2 ,..., bn] where b1...bn indicate how positive the reviews of the n neighboring businesses of P are. We can calculate the distance between business P and bi using haversines formulae for the great circle distance between two points, used widely for navigation. Then, we find the positivity of reviews bi for each neighboring business Pi. In order to integrate q and [b1, b2,...,bn] to derive a value S which is the success factor for P, we use a kernel smoothing function to calculate the success factor. A smoothing kernel defines a set of weights for each x:

$$f(x) \ = \sum_{i=1}^{n} W(x) * S(x).$$

The weights, $W_i(x)$ are described by a density function with scaling parameter that adjust the size and form of weights near x. We use a gaussian function on the distance from the business to be studied for computing the weight density. This means that businesses close to the one under study have a higher impact on the success compared to businesses farther away. Also, by adjusting the parameters of the gaussian, we can change how quickly this effect declines with distance. In the final step we combine the sum of the contribution of neighbouring businesses with the success factor of the business in attention:

$$S(x) = \alpha * S_i + \beta * \sum_{i=1}^{n} W(x) * S(x)$$

Where

$$W(x) = \alpha * \exp(-(\pi - b))^2/2(c) \; and \; \alpha = 1 - \beta$$

Next Steps

We still need to build a regression model to determine the future success factor value for P. Also, we would like to detect fishy users who can be defined as those users who post many reviews in a short period of time. In order to detect these reviews, we could either look at the timestamps of the reviews or we could look at certain grammatical patterns within the reviews to determine if the review was posted by a fishy user. Eliminating fishy reviews from the analysis would yield accurate success factors for businesses.

References

[1] Bryan Hood, Victor Hwang and Jennifer King *Inferring Future Business Attention.*

[2] A J Smola and B Scholkopf. 2004, *A Tutorial on Support Vector Regression.* Statistics and computing (2004).

[3] G A Miller, R Beckwith, and C Fellbaum. 1990. Introduction to Wordnet: An on-line lexical database* *International Journal.* Addison Wesley, Massachusetts, 2nd edition, 1994.