Predicting Success of Businesses Based on Internal/External Factors

Yuan Ma, Li Yi, Takahiko Tsuchiya, Vinish Chamrani, Devang Mistry
October 4th, 2015

1. Problem Definition

How do local businesses succeed? Common reasons may be the quality of service, prices, etc. However, besides these "internal" factors, there may be external factors such as the location of business, traffic, demographics, climate and crime rate. We try to find relationships between the success of businesses and these external factors, using Yelp's business dataset as the primary data source.

2. Heilmeier Questions

1. Our objective is to find the endogenous and exogenous factors contributing to the success of a business. We propose a model to find the direct factors leading to the success of a business based on Yelp's business dataset [15], geographic and demographic data. The focus is to help new business owners succeed in this competitive industry.

2. Current works are more focusing on the Yelp dataset itself rather than some exogenous concerns, which may not be revealed by endogenous reviews/comments. Also, some studies are geographically limited to the US or certain areas, but our dataset includes five countries which may give us a new perspective of the universal and individual feature of a certain factor. Lastly, we will predict the success of business based on both endogenous and exogenous factors.

3. We will focus on finding correlations using Yelp's business data and integrating it with other data sources, such as Google Places API, CityGrid API and US census data. We will also extract external factors from the review text, as well as dominant internal factors. This will help us develop a model that will give a clear picture or which factors.

4. Our analysis results may be useful for business owners who are trying to find out what they need to do in order to be more successful, maybe Yelp can come with a report which it can provide to business owners that will help them to improve their business.

5. With the models we will develop, Yelp could create a new analytics service for business owners: provide business advice before it's too late! And we can also provide those advice to existing business owners.

6. The only risk is that we may not find a promising correlation between those factors and the success of a certain business.

7. The system will be build with free-of-cost technologies such as Weka, OpenNLP, ArcGIS, and D3JS. So except for AWS credits, there are no costs apart from the time and effort committed by the five team members over a ten week period.

8. The project can be done approximately in 10 weeks given we not face a major roadblock.

9. At the end of the fifth week, a trained model generated from Yelp dataset will be complete. The model will be used to predict whether a business is successful against the ratings and reviews of business. At the end of the tenth week, we should be done with the analysis and evaluation of our assumption with some visualization.

3. Survey / Literature Review

We have reviewed the literature related to our approach of using external factors, and other useful analytic methodologies. Park's research has shown that the demographics of a location is pivotal in driving the success of local businesses. Some of the demographic data that he has cited includes age, ethnicity and income [8]. Bakhshi, Kanuparthy and Gilbert discuss the effect of demographics on restaurant reviews. Their study has shown that education level shows a strong positive effect on the number of reviews. More specifically, neighborhoods with high percentage of college degrees (greater than 50%) are highly likely to have restaurants with high number of reviews [3]. Bhuyan, Blisard, et al. discuss the importance of income, age and household structure in their study of consumer spending at fast-food and full service restaurants [16]. We can use their data as a reference in out correlation model to understand how each of these factors contributes to the success of a restaurant.

There are also studies that relate the location of the business with the success of it. Athiyaman, A proposed a model for deciding the optimal location for a new coffee shop at a University campus using spatial interaction theory and customer density estimates [2]. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V. and Mascolo

introduced a mechanism to predict the popularity of retail store using a dataset collected from Foursquare in New York [6]. Qu, Y. and Zhang, J proposed a new framework to analyze the relationship between customers' check-in information to a business site and the revenue of it [10]. And Yu, Z., Zhang, D. and Yang, D discuss an approach to get geographic data from exploiting user-generated contents from location based social networks, like Foursquare and Yelp [12].

Yelp dataset includes "business closed" values which may suggest a business failure. However, as pointed by Brian Headd [5] in their research that one third of the closure were successful were shut down due to factors such as retirement, another venture etc. We will look into closures and try to find why those businesses failed. In this age of globalization, although due to advancement of technology, traditional barriers of location have been broken but we still see clusters i.e. similar businesses being present in vicinity of each other [9] hinting that location is most important factor in survival/success of business. Apart from location, there are six other factors [11] viz. (listed in order of importance) customer orientation, quality products, efficient management, capital accessibility and marketing strategy that contributes to success of business. So, we will try to look into this finding and see what other factors apart from location are important that contribute to success/failure of business.

Review-text mining is also a common approach for business analysis, in which various classification and quantification approaches are taken. For example, Liu et al. discusses methods of extracting and classifying significant phrases in terms of popularity, concordance, informativeness, and completeness, using various NLP techniques [7]. Fan and Khademi propose regression approaches using word-frequency-based features mapped to Yelp review scores, showing linear correlations between the significant words and reviews [4]. The work by Abidin et al. introduces a method of sentiment analysis which separates factual, emotional and meta information in a HTML text-mining context [1], which may be applied to our task of extracting different categories of Yelp review contents.

Lastly, in terms of analysis method, Cohen, Jacob, et al. introduced multiple regression/correlation analysis as an approach that whenever a quantitative variable, the dependent variable is to be studied as a function of, or in relationship to the independent variables [13]. But this approach is heavily depended on the normality of data. Therefore, Bolker, Benjamin M., et al. proposed the use of generalized linear model to analyze nonnormal data [14]. In our study, we use multiple regression and pearson correlation test to figure out if there is a correlation between our chosen factors and success of businesses, measured by star ratings and number of reviews. Furthermore, we build generalized linear model which allows for response variables that have error distribution models other than a normal distribution, and see if such model would better fit our data.

4. Timeline / Plan

We have 9 weeks to complete the project. Below is the timeline and the responsibility of each member.

Week 1-2		Week 3-6		Week 7-9	
Data Collection	Data Pre -processing	Modeling	Prediction	Visualization	Evaluation
Li	Yuan	Vinish	Li	Taka	Yuan
Vinish	Taka	Yuan	Devang	Devang	Vinish

5. Appendix: Heilmeier Questions

1. What are you trying to do? □Articulate your objectives using absolutely no jargon.
2. How is it done today; what are the limits of current practice?
3. What's new in your approach; why it will be successful?
4. Who cares?
5. If you're successful, what difference will it make?
6. What are the risks and payoffs?
7. How much will it cost?
8. How long will it take?
9. What are the midterm and final "exams" to check for success?

References

[1] Abidin, S.Z., Omar, N., Radzi, M.H. and Haron, M.B. 2011. Quantifying Text-based Public's Emotion And Discussion Issues In Online Forum. International Journal of New Computer Architectures and their Applications (IJNCAA). 1, 2 (2011), 428–436.

[2] Athiyaman, A. 2010. Location decision making: the case of retail service development in a closed population. Academy of Marketing Studies (2010), 13.

[3] Bakhshi, S., Kanuparthy, P. and Gilbert, E. 2014. Demographics, weather and online reviews: A study of restaurant recommendations. Proceedings of the 23rd international conference on World wide web (2014), 443–454.

[4] Fan, M. and Khademi, M. 2014. Predicting a business star in Yelp from its reviews text alone. arXiv preprint arXiv:1401.0864. (2014).

[5] Headd, B. 2003. Redefining business success: Distinguishing between closure and failure. Small Business Economics. 21, 1 (2003), 51–61.

[6] Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V. and Mascolo, C. 2013. Geo-spotting: Mining online location-based services for optimal retail store placement. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (2013), 793–801.

[7] Liu, J., Shang, J., Wang, C., Ren, X. and Han, J. 2015. Mining Quality Phrases from Massive Text Corpora. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (2015), 1729–1744.

[8] Identification of Site Selection Factors in the U.S. Franchise Restaurant Industry: An Exploratory Study: 2002. *http://scholar.lib.vt.edu/theses/available/etd-01112002-135621/*. Accessed: 2015-10-04.

[9] Porter, M.E. 2000. Location, competition, and economic development: Local clusters in a global economy. Economic development quarterly. 14, 1 (2000), 15–34.

[10] Qu, Y. and Zhang, J. 2013. Trade area analysis using user generated mobile location data. Proceedings of the 22nd international conference on World Wide Web (2013), 1053–1064.

[11] Wijewardena, H. and Zoysa, A.D. 2005. A factor analytic study of the determinants of success in manufacturing SMEs. *Faculty of Commerce - Papers (Archive)*. (Jan. 2005), 1–11.

[12] Yu, Z., Zhang, D. and Yang, D. 2013. Where is the largest market: Ranking areas by popularity from location based social networks. Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC) (2013), 157–162.

[13] Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.

[14] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3), 127-135.

[15] Yelp Dataset Challenge: http://www.yelp.com/dataset_challenge. Accessed: 2015-10-04

[16] The Demand for Food Away from Home: Full-Service or Fast Food? USDA Economic Research Service - AER829: *http://www.ers.usda.gov/publications/aer-agricultural-economic-report/aer829.aspx*. Accessed: 2015-10-04.