Market Research and Business Opportunities Based on Yelp Data

Team #105: Michael A Fronda, Hal Zhang, Vladimir Orlov, Shivani Narahari, Abhishek S Mishra, Anthony J Mazza

Our project is based on Yelp dataset [link]. Yelp dataset provides information on businesses in several North American metropolitan areas and their customers, with inter-relationships and influence scores. It also includes relationships between businesses and their customers through reviews, check-in, and tip records.

With this data, our objectives are to research businesses in a selected geographical area, their categories, and services; to analyze their customer base; to identify key customer groups, traits and preferences; and to study the relationship between businesses and their customers. These insights will help us to suggest relevant business opportunities for the given geographical area **(Q1).**

As of today, suggesting a business opportunity for a given area or neighborhood is not a publicly available feature of Yelp. Yelp is limited to serving relationships between customers and businesses, helping businesses improve their service based on customers' feedback, and helping customers find relevant service providers **(Q2).**

The novelty of our approach is to extract and measure market demand based on customer feedback records and to evaluate market supply based on existing businesses and their popularity. Using market supply and demand, our project can help to identify potential business opportunities in a selected geographical area. And we want to provide proper visualizations as the last part of our effort is to clearly communicate the insights produced like a heatmap of supply and demand based on selected areas. This approach has a good chance of success, as there is enough relevant data to calculate the gap between market supply and demand. **(Q3).**

Those who care about our project are businesses looking to expand, local entrepreneurs, and investors (**Q4**). Derived market research insights can aid current and future businesses and local communities in the selected area to make data-driven business decisions. We can measure success by performing a natural experiment on new businesses who have and have not used our approach, to see if there is any difference in the success of the businesses **(Q5).**

The key project risk is the difficulty of testing calculated market supply and demand, as it involves economic intricacies and data beyond the scope of this project. The payoff of course is improved return on investment (ROI) for those who decide to apply the insights derived. There are no monetary costs for this project, and we plan to complete it within 2 months. The midterm check is to have data preparation completed and key modeling techniques agreed upon. The final check is to have market supply and demand estimation and interactive visualization. Please check our work breakdown summary table below for more details **(Q6,7,8,9).**

In our literature review, we aim to identify relevant models, analysis methodologies, and visualization techniques for our project. Initially, we draw on Geron's [6] practical insights into machine learning models, tools, and techniques, seeking to enhance these foundations by applying pertinent models to Yelp data. Tayeen et al.'s [11] exploration of the impact of location data on business review performance offers a springboard for us to investigate how such data influences demand.

Singh et al.'s [1] work with Yelp data using AzureML and SparkML also serves as a valuable reference, which we plan to expand upon by translating text analysis into insights on supply and demand. Cormen et al. [5] provide a mathematical foundation for graph analysis methods, which we will utilize to develop algorithms for our data analysis. While Brath et al. [3] discuss the business implications of graph analysis, we aim to delve deeper into the analytical aspect.

Our methodology will also include customer and business clustering, leveraging k-means and Gaussian Mixture Models as discussed by Bishop et al. [4] and Sinaga et al. [19]. Our goal is to compare these algorithms and identify the most effective ones for yielding meaningful results.

Glaeser et al. [10] findings on Yelp data's ability to explain business demand variations prompt us to focus on dense urban areas. We also plan to construct graphs of customers and businesses, refining McClanahan et al.'s [2] centrality analysis with the addition of natural language processing (NLP). Our review analysis will incorporate NLP to discern genuine from fake reviews, a challenge given the unlabeled nature of our data, which rules out classification methods suggested by Sihombing et al. [16] and Lin et al. [20].

Instead, we will explore clustering and other techniques to derive quantitative insights from reviews, similar to Deho et al.'s [15] word embedding approach. We aim to enhance Jiang et al.'s [14] word2vec application by shifting from clustering to sentiment classification, further refining Mutinda et al.'s [13] lexicon and CNN methods to focus on Yelp reviews exclusively.

Our analytical framework includes assessing various models and selecting one based on performance metrics, with the end goal of creating a composite score to reflect business demand. This approach is supported by Li et al. [7] and Hu et al.'s [8] research on the impact of online reviews on demand, which we will augment by considering the volume of reviews and the influence of popular reviewers.

Future demand predictions will be informed by time series analysis, drawing inspiration from Pereira et al.'s [9] work on hotel demand, but adapted to Yelp data. We also plan to predict business opportunities and star ratings, using classification methods informed by Jiang, A., & Zubiaga, A.'s [18] research, enhanced with our NLP techniques.

The culmination of our project will be a heatmap visualization of supply and demand scores, an approach that builds upon Wang et al.'s [12] E-comp map, extending it to include predictions of business opportunities rather than comparisons.

In conclusion, our analysis of Yelp dataset is novel, as we will build composite scores for supply and demand based on the proxy of number of reviews and review sentiment using NLP. Our visualization will also be unique, as it will be a heatmap of supply and demand based on selected areas and type of business. This is exciting because it could provide businesses with crucial data to identify their next business opportunity.

Work breakdown table summary: all teams members contribute equally to the project

Task	**Start Date**	**End Date**	**Team Member(s) Leading**
Project Proposal	2/13/2024	2/21/2024	All
Project Literature Survey	2/21/2024	2/26/2024	All
Proposal Slides and Video	2/26/2024	2/29/2024	Anthony, Michael
Data Processing	3/1/2024	3/9/2024	Vladimir, Shivani
Modeling	3/9/2024	3/27/2024	Abhishek, Hal
Visualizations	3/27/2024	4/10/2024	Shivani, Michael, Anthony
Final Deliverables	4/10/2024	4/19/2024	All

Citations:

1. Singh, R., Woo, J., Khan, N., Kim, J., Lee, H. J., Rahman, H. A., ... & Gudigantala, N. (2019). Applications of machine learning models on yelp data. *Asia Pacific Journal of Information Systems*, *29*(1), 117-143.
2. McClanahan, B., & Gokhale, S. S. (2016). Centrality and cluster analysis of Yelp mutual customer business graph. In 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) (pp. 592-601). https://doi.org/10.1109/COMPSAC.2016.79.
3. Brath, R., & Jonker, D. (2015). *Graph analysis and visualization: discovering business opportunity in linked data*. John Wiley & Sons.
4. Bishop, C. (2006). Pattern recognition and machine learning. Springer.
5. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to algorithms (3rd ed.). MIT Press.
6. Geron A. (2022) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 3rd Edition, O'Reilly Media.
7. Xiaolin Li, Chaojiang Wu, Feng Mai, The effect of online reviews on product sales: A joint sentiment-topic analysis, Information & Management, Volume 56, Issue 2, 2019,Pages 172-184, ISSN 0378-7206.
8. Hu, N., Liu, L. & Zhang, J.J. Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Inf Technol Manage 9, 201–214 (2008).
9. Luis Nobre Pereira & Vitor Cerqueira (2022) Forecasting hotel demand for revenue management using machine learning regression methods, Current Issues in Tourism, 25:17, 2733-2750, DOI: 10.1080/13683500.2021.1999397.
10. Glaeser, E. L., Kim, H., & Luca, M. (2017, November). Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity (Working Paper No. 24010). National Bureau of Economic Research. http://www.nber.org/papers/w24010.
11. Tayeen, A. S. M., Mtibaa, A., & Misra, S. (2020). Location, location, location! quantifying the true impact of location on business reviews using a Yelp dataset. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19) (pp. 1081–1088). Association for Computing Machinery. https://doi.org/10.1145/3341161.3345334.
12. Wang, Y., Haleem, H., Shi, C., Wu, Y., Zhao, X., Fu, S., & Qu, H. (2018). Towards easy comparison of local businesses using online reviews. Computer Graphics Forum, 37, 63-74.
13. Mutinda, James, Waweru Mwangi, and George Okeyo. 2023. "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network" Applied Sciences 13, no. 3: 1445. https://doi.org/10.3390/app13031445.
14. Jiang, R., Liu, Y., Mentor, K., & McCann, B. (2015). A General Framework For Text Semantic Analysis And Clustering On Yelp Reviews.
15. Deho, O., Agangiba, A. W., Aryeh, L. F., & Ansah, A. J. (2018). Sentiment analysis with word embedding. In 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST) (pp. 1-4). IEEE. https://doi.org/10.1109/ICASTECH.2018.8506717.
16. Sihombing, A., & Fong, A. C. M. (2019). Fake review detection on Yelp dataset using classification techniques in machine learning. In 2019 International Conference on Contemporary Computing and Informatics (IC3I) (pp. 64-68). https://doi.org/10.1109/IC3I46837.2019.9055644
17. Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. Sustainability (Basel, Switzerland), 11(19), 5254-. https://doi.org/10.3390/su11195254
18. Jiang, A., & Zubiaga, A. (2018). Leveraging aspect phrase embeddings for cross-domain review rating prediction. PeerJ. Computer Science, 5, e225-. https://doi.org/10.48550/arxiv.1811.05689

19. Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means clustering algorithm. IEEE Access, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796
20. Lin, T.-Y., Chakraborty, B., & Peng, C.-C. (2021). A study on identification of important features for efficient detection of fake reviews. 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 429–433. https://doi.org/10.1109/ICDABI53623.2021.9655845

Dataset Link: https://www.yelp.com/dataset/documentation/main