

*Note: Citations are in brackets [], answered Heilmeier questions are in bold parentheses ()*

## **Introduction**

Selecting the right location is a critical yet difficult decision for new restaurant owners (4). The location plays a key role in the restaurant's performance and longevity, and selecting the right place should be a decision backed up with data and certainty (5). Our target with the project is to create a tool that restaurant owners would be able to use to test different locations given user-generated reviews, demographics, and population data using a predictive model. (1)

## **Literature Survey**

**Factors Influencing Restaurant Success:** Our feature engineering derives its groundwork from past research that provides multiple baselines to predict restaurant success. Specifically, intense competition and poor locations have been identified as failure reasons [13], while factors such as ethnic specific cuisine have been attributed to success [4]. These factors interact with a location's demographic and socioeconomic environment to directly influence the food environment [8]. These research efforts validate our core hypothesis: modeling variables such as competitor saturation and demographic alignment is significant in predicting restaurant success. However, some limitations of the above research include the reliance on qualitative or small-scale survey data [4, 13], along with focuses on public health rather than retail survival [8]. Our project's approach addresses these shortcomings through large-scale, quantitative data to predict restaurant success.

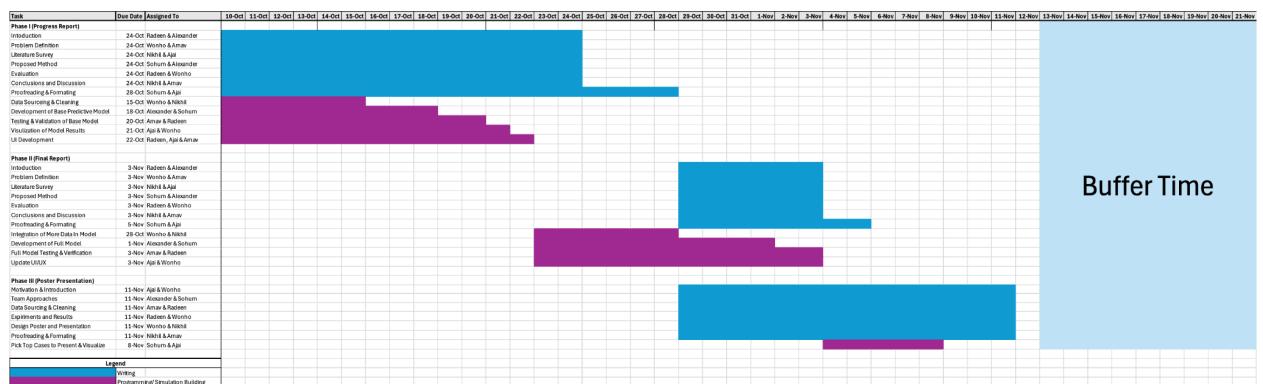
**The Predictive Power of User-Generated Content (UGC):** Current literature establishes that user-generated content (UGC) is a strong predictor of restaurant success, with studies showing that signals like customer rating trends [5], review sentiment [1, 2, 6, 18], photo volume [20], and user mobility [10] can forecast a restaurant's survival. This research is useful as it validates the use of the Yelp dataset alongside feature engineering to predict restaurant success; however, these paper's primary shortcoming is that these models are reactive, predicting the success of an existing business using its own historical data [5, 10, 17, 20], making them unsuitable for task of evaluating a new, undeveloped location. (2) Furthermore, these studies are constrained to a certain geographical area [17] and lack additional context, a gap we fill by integrating demographic data. Starakiewicz et al. [15] incorporates demographic data into an XGBoost model and uses SHAP values for explainability, which is highly useful as it validates our choice of model and provides a methodology for explaining why a location is favorable. A potential shortcoming is that this method still presents a reactive model, which our approach will adapt this framework to a proactive model that evaluates and predicts new locations rather than existing restaurants.

**Optimal Business Site Selection:** Past research in optimal business site selection provides a strong foundation for our project, particularly focusing on the important role of ML techniques. [3] particularly shows how data from Location-Based Social Networks such as Yelp, are valuable in identifying profitable retail locations. A key takeaway here was the predictive power of leveraging data made by users that is tied to locations. [12] and [14] also share an ML-based approach, using clustering and classification to recommend retail locations. A key differentiator, however, was that they looked at detailed factors like customer density, mobility patterns, and competition. While these studies excel as a proof-of-concept for UGC and ML-based techniques, [16] expands on this foundation by creating a scalable ML pipeline, using retail-specific geospatial data and mobility data, to screen locations. A key issue with these prior works is that they focus on the general retail industry, rather than restaurants specifically. They also do not distinguish between different types of restaurants. Looking particularly in the restaurant industry, by using Yelp's public dataset, [19] found that making predictions and decisions about optimal restaurant location depended heavily on the category of restaurant. Our approach will address this by developing subtype-specific predictive models that operate at a finer granularity, such as cuisines and demographics. After further research, what predicting restaurant success really looks includes using models like logistic regression, artificial neural networks [11], and LSTMs (for survival rate predictions) [9]. This led us to understanding a restaurant's ability to survive, which was supported by studies in both temporal modeling

and interpretability. Looking at Long-Short Term Models (LSTMs) [9], we can use the models to predict the survivability of a restaurant based on the characteristics of a business's commercial district. It was concluded that modeling customer traffic, over time, played a significant role in determining the future of a business's survivability. One limitation was the geographic constraints to Seoul, and it collected aggregate survival rates. Taking this concept a step further, Restaurant Survival Prediction and Explanation (RSPE) [18] focuses on predicting which restaurants will survive, seeking explanations through graph neural networks to identify the connection between users and restaurants. This method works adjacent to a tool that uses attention as a way to highlight and summarize the reviews. For our approach, we will build on these insights to expand to a wide variety of areas and regions. Our goal would be to achieve high accuracy and generalizability. (3) These studies are valuable since they show that restaurant viability is quantitatively modelable. Other approaches, like MCDM and AHP/TOPSIS, give structured methods to compare criteria and weights for potential sites [7, 21]. One major downside of predictive models is their reactive nature and the need for restaurants' prior historical data to predict the future [9, 11]. (2) Similarly, MCDM frameworks can be hard to keep subjective and scalable. For our model, we will address these gaps by providing a proactive, data-driven 'Opportunity score' at the zipcode level, allowing us to evaluate new locations before restaurant establishment.

**Our Approach:** Our novel approach will synthesize user-generated and demographic data to create subtype and location-based models to show the best locations to open a new restaurant. Unlike prior work, we'll calculate an opportunity score for opening a new restaurant of a specific subtype at the zipcode level and overlay these scores on a choropleth map (3). For our primary dataset, we will use the [Yelp Open Dataset](#) which contains business attributes, user reviews, and more across 11 metropolitan areas. We will combine this dataset with [US Census Bureau](#) data to calculate metrics such as competitor saturation, rating, and diversity. We can add weights to ethnic groups for their specific ethnic cuisine. The core model will be the XGBoost Classifier model, trained on historical data to predict the probability of a given restaurant being open for 5 years. (5) We will utilize the probabilistic prediction and normalize it across the entire metropolitan area as an opportunity score from 1 to 100 per zip code. Due to the size of the datasets and number of variables, the alignment of data may be a risk. A benefit of this complexity is the potential accuracy of predicting restaurant success in various conditions. (6) Another risk is the inconsistency and subjectivity of user-based Yelp data, as certain areas may have bias or lower sample size. However, using user-based data provides a practical and dynamic evaluation. Regarding cost and duration, the dataset will be downloaded and stored on Google Drive, significantly offloading the storage cost. The training of the XGBoost model and processing of the data will be done on free computation tools such as Google Colab. (7) In terms of the time, it is outlined in the section 5, but the data collection, cleaning, and feature engineering will take an estimated time of 2-3 weeks, with the model development and frontend deployment taking around a month. (8, 9)

## Plan of Activities



All team members have contributed a similar amount of effort

## References

1. McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In Proceedings of the 7th ACM conference on Recommender systems (RecSys '13), pages 165–172. ACM. <https://doi.org/10.1145/2507157.2507163>
2. Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2017). A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews. *Journal of Quality Assurance in Hospitality & Tourism*, 18(4), 465–492. <https://doi.org/10.1080/1528008X.2016.1250243>
3. Damavandi, A. J., Mahmoodi, D., & Nasrabadi, A. M. (2021). Utilizing LBSN data for optimal retail store placement. *Expert Systems with Applications*, 183, 115389. [https://eoge.ut.ac.ir/article\\_75927\\_77b93d9b7e4d889d8e7dbeb42190a294.pdf](https://eoge.ut.ac.ir/article_75927_77b93d9b7e4d889d8e7dbeb42190a294.pdf)
4. Agarwal, R., & Dahm, M. J. (2015). Success factors in independent ethnic restaurants. *Journal of Foodservice Business Research*, 18(1), 20–33. <https://www.tandfonline.com/doi/abs/10.1080/15378020.2015.995749>
5. Christof Naumzik, Stefan Feuerriegel, and Markus Weinmann. 2021. I Will Survive: Predicting Business Failures from Customer Ratings. *Marketing Science* 41, 1 (Jan. 2021), 188–207. DOI: <https://doi.org/10.1287/mksc.2021.1317>.
6. S. Jia. 2018. Behind the ratings: Text mining of restaurant customers' online reviews. *International Journal of Market Research* 60, 6 (Dec. 2018), 561–572. DOI: <https://doi.org/10.1177/1470785317752048>
7. Javaid, A., Iqbal, M., & Akhtar, P. (2021). A hybrid AHP/TOPSIS-based method for business store site selection. *International Journal of Management and Decision Making*, 20(4), 356–372. [https://www.researchgate.net/publication/340812777\\_An\\_AHPTOPSIS-Based\\_Approach\\_for\\_an\\_Optimal\\_Site\\_Selection\\_of\\_a\\_Commercial\\_Opening\\_Utilizing\\_GeoSpatial\\_Data](https://www.researchgate.net/publication/340812777_An_AHPTOPSIS-Based_Approach_for_an_Optimal_Site_Selection_of_a_Commercial_Opening_Utilizing_GeoSpatial_Data)
8. Larson, N. I., Story, M., & Nelson, M. C. (2009). Neighborhood environments: Disparities in access to healthy foods in the U.S. *American Journal of Preventive Medicine*, 36(1), 74–81.e10. <https://doi.org/10.1016/j.amepre.2008.09.025>
9. Lee, S., & Lee, J. (2024). LSTM-based survival rate prediction model for restaurants using public commercial district data. *Computers, Environment and Urban Systems*, 110, 102018. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12244476/>
10. Lian, J., Zhang, F., Xie, X., & Sun, G. (2017). Restaurant survival analysis with heterogeneous information. *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 993–1002). <https://doi.org/x/10.1145/3041021.3055130>
11. Llewellyn, G., & Mun, S. G. (2023). Prediction of restaurant site success using logistic regression and artificial neural networks. *Journal of Hospitality and Tourism Technology*, 14(2), 301–318. [https://arodes.hes-so.ch/record/13136/files/Llewellyn\\_2023\\_prediction\\_restaurant\\_site\\_success.pdf](https://arodes.hes-so.ch/record/13136/files/Llewellyn_2023_prediction_restaurant_site_success.pdf)
12. Lu, J., Zhou, C., & Chen, Y. (2024). Retail store location screening: A machine-learning-based approach. *Computers in Industry*, 159, 104857. <https://www.sciencedirect.com/science/article/abs/pii/S0969698923003715>
13. Parsa, H. G., Self, J. T., Njite, D., & King, T. (2005). Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3), 304–322. <https://doi.org/10.1177/0010880405275598>

14. Pathak, A., Joshi, R., & Kumar, P. (2020). A novel approach for business store site selection using machine learning on Foursquare data. *Procedia Computer Science*, 171, 1401–1410. ([PDF](#)) [a-novel-approach-for-business-store-site-selection-IJERTCONV8IS05014](#)
15. Starakiewicz, T., Nowak, M., & Kowalski, R. (2025). Predicting restaurant survival using nationwide Google Maps data: An interpretable XGBoost approach. *Applied Spatial Analysis and Policy*. <https://www.sciencedirect.com/science/article/pii/S095070512500245X>
16. Ting, C. Y., & Jie, M. Y. (2021). Location profiling for retail-site recommendation using machine learning. *Journal of Retail Analytics*, 7(2), 112–125. <https://www.atlantis-press.com/article/125980663.pdf>
17. Vallapuram, A. K., Nanda, N., Kwon, Y. D., & Hui, P. (2021). Interpretable business survival prediction. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 99–106). <https://doi.org/10.1145/3487351.3488353>
18. Wang, Y., Li, Z., & Zhao, D. (2022). Explainable restaurant survival prediction (RSPE) via graph neural networks. *Expert Systems with Applications*, 200, 117094. <https://aclanthology.org/2022.emnlp-main.216.pdf>
19. Priyadi, A., Lande, N. M., Faradilla, A., Hasan, M., & Widianti, E. (2025). Identifying Key Features in Yelp Data for Success in Different Types of Restaurants. INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 9(1), 33-45. <https://doi.org/10.29407/intensif.v9i1.23476>
20. Zhang, M., & Luo, L. (2022). Consumer-posted photos as a leading indicator of restaurant success. *Journal of Marketing Research*, 59(3), 512–530. <https://pubsonline.informs.org/doi/10.1287/mnsc.2022.4359>
21. Shaikh, S. A., Memon, M., & Kim, K.-S. (2021). A Multi-Criteria Decision-Making Approach for Ideal Business Location Identification. *Applied Sciences*, 11(11), 4983. <https://doi.org/10.3390/app11114983>