CX4242/6242 | Team 22 - CHEF'S KISS

Ajai Singh, Alexander Steele, Arnav Patidar, Nikhil Dasari, Radeen Dixon, Sohum Gaitonde, Wonho Choi

Introduction

Selecting the right location is a critical yet difficult decision for new restaurant owners. We have created a model and accompanying visualization tool that predicts the long-term survivability of a restaurant based on demographic data and user reviews. Tools such as this support the growth and sustainability of the restaurant industry, which is vital to local economies. Restaurants are often the largest source of entry level jobs, are crucial to tourism, as well as local culture. Data informed tools that support the long term success of this industry help promote more sustainable urban development, limiting economic waste and financial loss. Our model and visualization tool demonstrate an excellent use case of technology as a means to benefit communities through cultural preservation and economic improvement. Extension of this analysis beyond the restaurant industry also introduces novel use cases of machine learning.

Problem Definition

The vast majority of entrepreneurs select locations based primarily on qualitative metrics, such as intuition and community sentiment. Qualitative metrics alone fail to provide the best insight into location selection, which has been shown to be one of the best predictors of long-term restaurant success. Current predictive models fail to properly provide insight into the location selection process as they lack the ability to provide predictions for restaurants that do not yet exist and do not present results in a visualizable way that is usable by a non-technical business owner. Implementation of data-driven tools in the restaurant industry would allow for quantitative measurements of a restaurant's success to be calculated prior to opening, creating an ecosystem that can take qualitative and quantitative factors into account to determine optimal location. We aim to address this issue by predicting the probability of long-term restaurant survival before an owner opens their shop, by ZIP-code. We are using publicly available information - Yelp's business and review data and U.S. Census demographics data - to find given areas' "opportunity score", which we define as the probability that the new restaurant will survive over time. Our tool provides an opportunity score for a user defined city and restaurant type.

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.

Proposed Method

Our proposed method aims to create a data-driven tool to generate "opportunity scores" for opening a new restaurant. This score, rooted in the probability of a 5 year survival, is calculated at the zip code level and is specific to a restaurant subtype. Current literature focusing on restaurant survival is largely reactive, forecasting the survival of existing businesses [5, 10, 18, 20]. We aim to be proactive and hypothesize that the success of a new restaurant is a function of location-based context. Specifically, success is grounded by the socioeconomic landscape of the area as well as the existing competitive landscape. By training a model on the historical success of thousands of restaurants against these location-related features, we can create an XGBoost model that predicts the success of a hypothetical new restaurant.

Key Innovations:

1. Proactive prediction: Unlike existing work, which reactively predicts the success of already established restaurants, our prediction framework only uses inference-time available features. These include features like subtype, price range, zip code, and derived location features and notably exclude performance metrics of established restaurants like stars, review count, and business age. We use historical data from other restaurants to create our features at the location level during training, but a restaurant's own performance metrics are not used during inference, in order to meet the proactive and predictive nature of our project. Our design uniquely focuses on providing utility for prospective restaurant owners, not established businesses.
2. Subtype and zipcode granularity: We uniquely create 100+ features by aggregating restaurants by both zip code and subtype, compute 5 metrics per group including average stars, total count,

average price, median reviews, and median age. As an important data engineering step, we then pivot this data in order to portray features in a wide format, such as Italian_average_stars. This captures existing competition at a very fine and specific granularity.

3. Normalizing competition features: We create engineered competition metrics such as competition density (which we calculate as, same subtype competitor divided by zip code population) as well as market share of competition (which we calculate as same subtype competitors divided by total restaurants). This normalization of competition by market size is especially important because it allows us to feasibly make comparisons across zip codes and different market sizes.

Outlining our pre-processing flow, our data comes from two primary sources, particularly the Yelp Open Dataset and the US Census Bureau's ACS 5-Year Data. The Yelp JSON is parsed into a dataframe and filtered to restaurants only. We process this review data to extract an engineered business age metric (which we calculate from first to last review date). This is used to create our target variable, namely whether a restaurant survives 5+ years. Zip code is used as the primary key for joining the location-based features. The user-input features include the primary cuisine (Italian, Mexican, Indian, etc) as well as price point (1-4). We extract key demographic features such as population, median age, and race per zip code from the Census data. We then aggregate all restaurants for each zip code and calculate the average star ratings, median review counts, average price range, and median business age. As discussed above, we then further granulate the competition features into per subtype and zip code metrics, engineer competition features like competition density and market share, and ensure the model uses features that are only suitable for inference time (notably not using performance metrics).

We use an XGBoost Classifier as our predictive model passing in our features. We do an 80/20 split and optimize for the ROC-AUC score. The model is trained to predict the binary outcome of whether or not the restaurant survives 5 years. For the visualization, we map the outcomes of the model to a choropleth map user interface. Users will be able to control various inputs such as the restaurant cuisine and price range, and based on the inputs, the model will produce a respective output using the zip-code aggregated data. The application will display a map where users can hover over various zip codes to display the "investment score." Below, we show our current map UI that contains model outputs for the Philadelphia area. This will be further expanded to accommodate for other states and zip-codes.

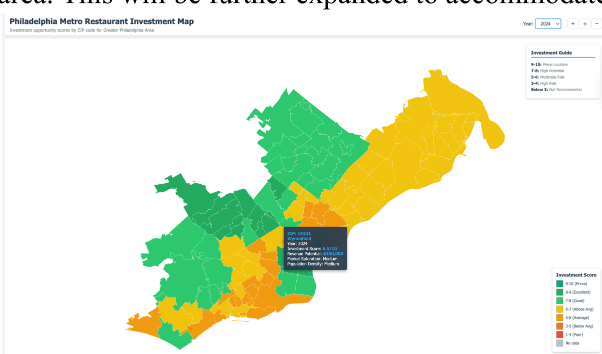


Figure 1. Map of Philadelphia, State Level

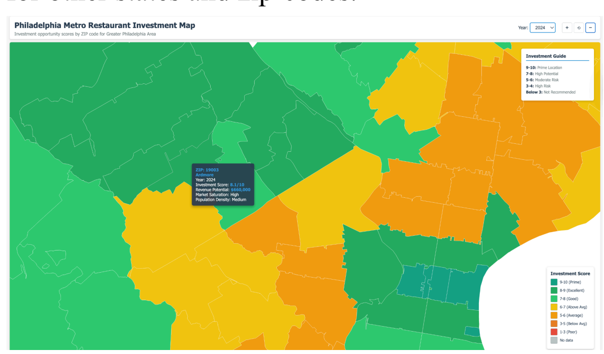


Figure 2. Map of Philadelphia, Zip Code Level

Evaluation

Currently, our testing is performed through visualizations and metrics on our jupyter notebook. The notebook contains the logic for processing our Yelp and Census datasets, connecting these with the existing XGBoost model to predict five year survivability. For our data cleaning pipeline, we address the relevancy and accuracy of our processing. We employ empirical methods such as displaying the heads of dataframes and metrics such as the dimensions along with the variable distributions to ensure consistency. For our current XGBoost model, our experiments address the overall accuracy of the model along with the significance of different features to the accuracy. This experiment is performed through an 80/20 train/test split of the dataset, computing predictions for the five-year survivability with default hyperparameters. The notebook records metrics such as accuracy, ROC-AUC, and precision. We also

compute the importance of the different input features through XGBoost's built in feature score. Afterwards, we finetune the model through 25 iterations of randomized search CV, which tests against different hyperparameters and returns the best combination. We observe that the model performs overall reasonably, achieving ~66% accuracy along with a ROC-AUC value of 0.70. The ROC-AUC value indicates that the model's binary classification of survival surpasses random probability. We also observe that the top features include the median business age in the zip codes and price range. With finetuning, the model improves in the respective values by around 0.1% and 1% for the accuracy and ROC-AUC values respectively, which is below our initial expectations. However, the ability for the model to classify consistently across hyperparameters, shown by the finetuning, may also indicate the lack of data scaling.
Future experiments and questions to answer include the zero-shot performance against unseen locations and times. In other words, we aim to test the learning and generalizability of the XGBoost model from one location or time to another. Our experiment will include various splits in the test dataset along with cross-region and cross-temporal generalization. For example, we can train the model on a specific location/time, and test the model for a range that it has not been trained on. We will evaluate all metrics to compare its performance, and then manipulate the hyperparameter and feature weights respectively. Moreover, we aim to make the model usable for potential restaurant owners and provide an end-to-end model that is comprehensive. This will include both user-testing where we input various factors into a UI and output a result in an existing map, along with verification that the model and visualizations are correctly connected through debugging statements.

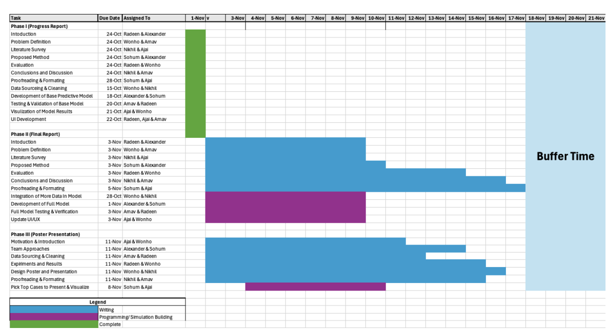
Conclusions and Discussion

The results of the model were summarized in an "opportunity score", which encapsulates the relative strength of an investment to prospective entrepreneurs. A visualization of this score across zip codes was developed to accompany the model, allowing non-technical entrepreneurs to easily incorporate data-driven quantitative metrics alongside more abstract factors. The model demonstrated that demographic and socioeconomic data can be correlated to restaurant outcomes using user-generated reviews on Yelp, achieving an accuracy of 66% and an ROC-AUC of 0.7 when predicting the 5-year survivability rate of restaurants. Although robust the model is susceptible to a lack of generalizability across different geographic regions. Factors like regional biases, uneven accessibility of data, and time dependences could all limit the implementation of the model across different regions. Future development will improve on the models ability to source data across different regions through the use of different data sets, such as google maps, and also improve the models ability to incorporate time trends by integrating real-time data sources. The front end UI/UX will also be improved to create more intuitive and powerful visualization tools. As a whole the model creates a tool for prospective entrepreneurs to make more informed decisions when picking a geographic region, demonstrating a textbook example of the symbiotic relationship between data visualization, analytics, and economics.

Plan of Activities

Old Plan of Activities

Revised 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
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
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.orgx/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
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