# CSE-6242 Group 183 Final Report: Understanding Restaurant Success

Chris Skowronski, Rujuta Kortikar, Arjun Gheewala, Eric Yeon

## 1 Introduction

The purpose of our project was to create an application that could allow prospective restaurant owners to estimate their success given parameters such as location, price, and food categories. We planned to take ratings and reviews from existing businesses in the NYC area to determine success and then find the attributes that contribute most to that success. With this information, users can enter parameters to get a predicted level of success. We aim to give prospective restaurant owners more confidence and mitigate risk.

The reason this is important is that it creates new opportunities for restaurant owners to expand or open businesses with information that is not easily accessible to them otherwise. From the IBISWorld report, *"Industry Market Research, Reports, and Statistics"* [1], We see from industry market research that the number of restaurants in New York continue to grow in a post-Covid world. This means that there is a desire for new restaurant owners to start their own businesses, but it also means that New York is full of growing competition in the restaurant space. In order for new businesses to succeed, they must have informed and data-backed strategies from the get-go.

## 2 Problem Definition

The task at hand is to identify and classify restaurants in the NY vicinity as either successful or not. The threshold of success is determined by restaurant ratings with at least 3.5 stars and 20 reviews based on initial data analysis. Afterwards, the performance is evaluated by the classification accuracy: the number of restaurants predicted correctly out of all restaurants considered as a percentage.

## 3 Literature Survey

Before diving into the coding aspect of our project, we spent time doing a comprehensive literature review that would help us understand the existing landscape of predicting restaurant success. This is integral for our project because it helps us identify major gaps that can be addressed by our work.

First, we built an understanding around which similar studies had already been conducted. From this research, we found that in studies such as *"Customer Restaurant Choice: An Empirical Analysis of Restaurant Types and Eating-Out Occasions"* [3], where a broad survey was done to hone in on key factors for a restaurant's success, and *"Satisfaction and revisit intentions at fast food restaurants"* [11], a study done to find what satisfies customers most at fast food restaurants, the data is gathered by administering surveys. These studies helped inspire the initial subset of factors we considered. However our solution is more robust and scalable since it is an application rather than a survey.

We also leveraged our literature review as a way to identify potential risks to be aware of as we took on this project. For example, we see in *"Reviews, Reputation, and Revenue: The Case of Yelp.com"* [9] that Yelp reviews tend to be self-fulling and sales are inflated by the number of overall stars a restaurant has in their rating. A similar study, *"Characterizing non-chain restaurants Yelp star-ratings"* [7] concluded that aspects of a business matter differently depending on your Yelp review. This study only used text from reviews to score, where we aim to combine text and rating. Using Yelp reviews is also risky as it can contain fraud reviews, as found by *"Detecting fraudulent online yelp reviews using K-L divergence and linguistic features"* [4]. Here researchers used a Zipf distribution to improve the model for detecting fraudulent reviews. It may be useful for our initial data cleaning to explore using such techniques. In another study about fake reviews, *"Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud"* [10], researchers aimed to prove that businesses will leave fake reviews about themselves in order to boost ratings. Our product will try to focus solely on real reviews, and it is good to know that these are data quality issues to consider.

Finally, we conducted research to determine what the possible applications of this work could be. According to "Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes" *"Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes"* [2], the identification and rise of social media influencers are largely attributed through opinion mining (or reviews); however, this concept has yet to be applied in the food industry. Additionally, studies such

as *"Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction"* [5] attempts to attribute increased success to location. We want to improve on this by taking into consideration a larger number of factors, including hidden topics. *"Should I Invest it?: Predicting Future Success of Yelp Restaurants"* [8] attempts to do something similar to our goal, and uses both text and non-text data to predict success. We want to build on this idea and provide a way to interact with it. We found that an existing study, *"Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews"* [6], attempts to improve the sentiment lexicon analysis on reviews due to its findings that positive sentiment is usually overwhelmingly positive, while negativity is not as strong in negative reviews. Another study *"Generating Recommendation Dialogs by Extracting Information from User Reviews"* [12], attempts to parse reviews to create a useful questionnaire to find preference. These studies do only text analysis and do not provide any way to interact with the data.

Our full literature review helped us build a robust understanding of what work has already done in this space, which potential risks we have to acknowledge as we move forward, and gaps in studies done in the past that we can address with our project.

## 4 Proposed Method

### 4.1 Intuition

In order to have an application that can compete in the industry, it is essential for us to have aspects of the application that have not been replicated elsewhere. In particular, our project has the ability to be cutting edge by focusing on real reviews, leveraging both text and ratings, and by creating a more scalable and accessible application than existing surveys.

### 4.2 Description

The proposed method is to build a model using gradient descent and backpropagation that utilizes a neural network to best determine a restaurant's success. By fine-tuning the parameters and adjusting the weights of each node, we can minimize the error of the output as much as possible. In addition, the desired prediction results can be achieved while also taking into account the features that most heavily influence the restaurant. We plan to use multiple neural network structures and evaluate the performance of each before deciding the most optimal model.

The front end interface will include a map of New York (using Google Maps API) where users can view restaurants that are listed on Yelp. The main function will allow users to click a location on the map and input restaurant information in which a dialogue will inform the guest of the predicted level of success as a percentage. In addition, current restaurant owners will be able to change features from their own restaurant and see its prediction as well.

## 5 Experiments/Evaluation

### 5.1 Test bed

Our application is designed to answer several key questions in the restaurant industry. It is important to keep in mind that the primary focus of this project is to determine restaurant success. Using the Yelp API, we gathered restaurant data ranging including latitude, longitude, borough, restaurant type, cuisine, and price range. There are several questions that were highest priority for us, including:

- Will a restaurant be successful given a certain location, cuisine, and price range?
- Are there certain cuisines that are highly correlated with a restaurant's success?
- Are cheaper or more expensive restaurants more likely to succeed in New York City?

Answering these questions can help new restaurant owners build a data-driven strategy that is optimized for success and can help chain restaurants determine the most cost efficient ways to expand.

### 5.2 Description of Experiments and Observations

There are two main methods we employed and they both had important insights that we think are integral for new restaurant owners to internalize.

First, we did an exploratory data analysis of our Yelp API data. This allowed us to broadly understand the distribution of the variables we were looking at and any inherent skew. Largely, the main purpose of this part of our project was to determine any strong correlations between variables and restaurant success. We saw in our analysis that the variables that stood out most here did, in fact, align with the hypotheses we had when starting this project. Specifically, we see that number of positive reviews is heavily correlated with

| Model | Benefits | Mean Accuracy |
|---|---|---|
| **Logistic Regression** | Cheap to train and predicts quickly | ~65% |
| **Histogram-Based Gradient Boosting Classification Tree** | Works better for large datasets, more robust | **~70%** |

**Figure 1: Figure 1: Comparing accuracy of different ML Models.**

restaurant success. Interestingly, we also see that zip code plays an important role in success. Restaurants in more densely populated areas naturally have more competition, which leads to a lower expected value of success per restaurant. However, there are conflicting factors here as higher populations also have higher total potential traffic so while the average success rate is lower, the maximum success rate is much greater. The important insights from this work are that a new restaurant owner should be very aware of gaining numerous positive reviews when they first open and should also take notice of the neighborhood they are looking to open in.

The other method we leveraged was a modeling approach to land more causal insights rather than only depending on exploratory analysis. The initial method we chose to employ was a logistic regression in which we could determine which features have higher predictive power of determining restaurant success. Linear models are cheap to train and query fast to predict outcomes. We were particularly interested in seeing how these results would compare to those that we have highlighted earlier through our observational exploratory data analysis. However, after evaluating the model with cross-validation, the mean accuracy hovered around 65 percent. Therefore, we decided to use a more powerful model: a histogram-based gradient boosting classification tree which works well with large datasets. Since the data consisted of categorical and numerical features, we used ordinal encoding for the categorical variables to seamlessly feed into the classifier. This brought the mean accuracy score to above 70 percent.

Our final product displays a folium map backed by a Flask app. In addition to this, we leveraged the Yelp data we sourced from the Yelp API to train a classifier to take attributes about the business and predict a composite success score. Our final app displays a simple user interface where users can easily input data into our model and receive a success score for any given combination of inputs. We are very proud of the work we have done for this project and are excited about the various use cases of this work.

## 6 Conclusion and Discussion

All team members have contributed a similar amount of effort. With the help of machine learning, we can take some of the uncertainty out of opening a restaurant in New York. Using existing Yelp data, we confidently show that success can be predicted.

If given more time, we would love to continue making our model even more robust in order to instill full confidence in the restaurant owners that leverage our platform. We are also excited about the possibility of expanding this application to help new restaurant owners that are based outside of New York City.

The business applications of this work are vast. We aim to allow potential owners to use the tool to make important business decisions. To productionalize this application and catch trends over time, we would create a scheduled data pipeline to refresh the model with new data from Yelp. Deciding how to open a restaurant can be a difficult decision, and our goal is to prove that it does not have to be.

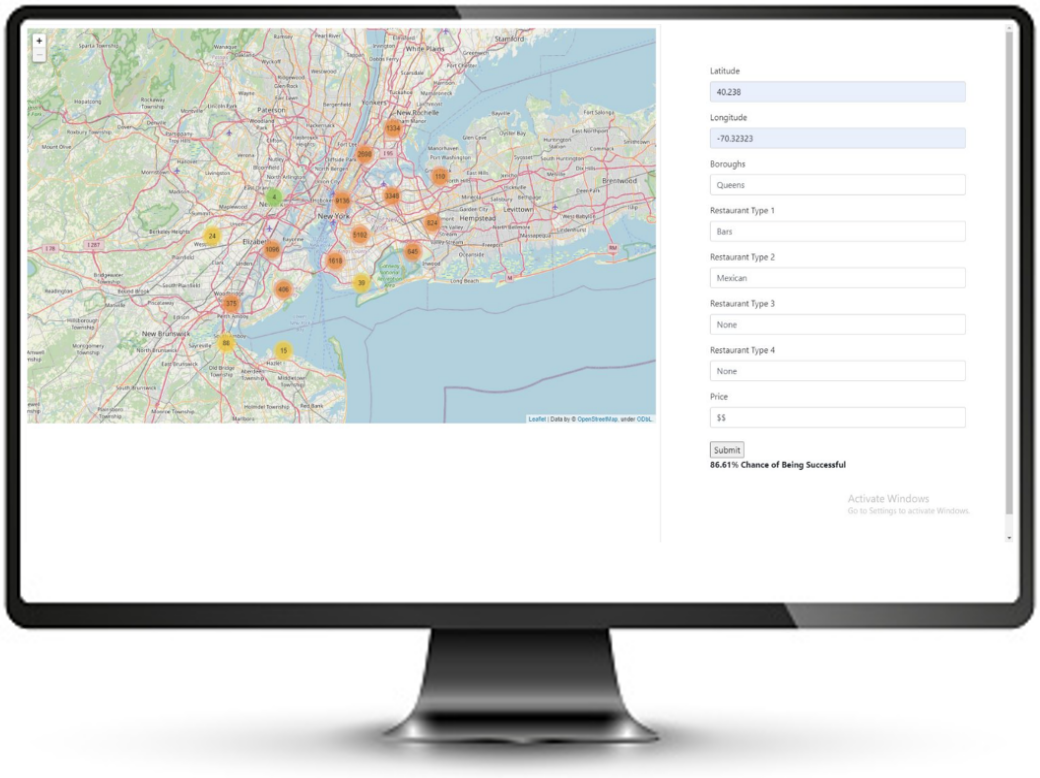


**Figure 2: Screenshot of our Restaurant Success predicting Flask app**

## References


[1] 2020. Industry market research, reports, and Statistics. https://www.ibisworld.com/industry-statistics/number-of-businesses/restaurants-in-new-york-united-states/

[2] Jaleel Alhadethy, Hamad and Refed. 2021. Sentiment Analysis of Restaurant Reviews in Social Media Using Naïve Bayes. In *Systems Analysis Modelling Simulation*. ARQII Publication, Iraq, 166–172. https://www.researchgate.net/publication/354990885_Sentiment_Analysis_of_Restaurant_Reviews_in_Social_Media_using_Naive_Bayes

[3] Bee-Lia Chua, Shahrim Karim, Sanghyeop Lee, and Heesup Han. 2020. Customer restaurant choice: An empirical analysis of restaurant types and eating-out occasions. *International Journal of Environmental Research and Public Health* 17, 17 (2020), 6276. https://doi.org/10.3390/ijerph17176276

[4] Christopher G. Harris. 2022. Detecting fraudulent online yelp reviews using K-L divergence and linguistic features. *Procedia Computer Science* 204 (2022), 618–626. https://doi.org/10.1016/j.procs.2022.08.075

[5] Longke Hu, Aixin Sun, and Yong Liu. 2014. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. *Proceedings of the 37th international ACM SIGIR conference on Research amp; development in information retrieval* (2014). https://doi.org/10.1145/2600428.2609593

[6] Hanhoon Kang, Seong Joon Yoo, and Dongil Han. 2012. Senti-lexicon and improved naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications* 39, 5 (2012), 6000–6010. https://doi.org/10.1016/j.eswa.2011.11.107

[7] Daniel Keller and Maria Kostromitina. 2020. Characterizing non-chain Restaurants' Yelp Star-Ratings: Generalizable findings from a representative sample of yelp reviews. *International Journal of Hospitality Management* 86 (2020), 102440. https://doi.org/10.1016/j.ijhm.2019.102440

[8] Xiaopeng Lu, Jiaming Qu, Yongxing Jiang, and Yanbing Zhao. 2018. Should I invest it?: Predicting Future Success of Yelp Restaurants. *Proceedings of the Practice and Experience on Advanced Research Computing* (2018). https://doi.org/10.1145/3219104.3229287

[9] Michael Luca. 2011. Reviews, reputation, and revenue: The case of yelp.com. *SSRN Electronic Journal* (Mar 2011). https://doi.org/10.2139/ssrn.1928601

[10] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp Review Fraud. *Management Science* 62, 12 (2016), 3412–3427. https://doi.org/10.1287/mnsc.2015.2304

[11] Amer Rajput and Raja Zohaib Gahfoor. 2020. Satisfaction and revisit intentions at fast food restaurants. *Future Business*

*Journal* 6, 1 (2020). https://doi.org/10.1186/s43093-020-00021-0

[12] Kevin Reschke, Adam Vogel, and Dan Jurafsky. 2013. Generating Recommendation Dialogs by Extracting Information from User Reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 499–504. https://aclanthology.org/P13-2089