Predicting Restaurant Success Based on Review Technology and ZIP Code Level Demographic Data

Alexander Nelson-Groocock, Sitara Uppalapati, and Jerry Yang

February 2020
(Revised: May 2020)

1 Summary

Americans are spending a larger amount of their income on eating at restaurants than ever before. The National Restaurant Association estimates 2019 sales at roughly $863 billion, close to double the $590 billion in sales during 2010. As a result, many technological restaurant review services have arisen to help eager consumers find their optimal dining experience; whether that be determined by cost, menu options, or other factors. As such, this paper seeks to answer if restaurant review technology data can be used as an indicator of restaurant success in the future. Our paper discusses a novel solution to predicting restaurant success through the creation of a comprehensive dataset and the use of various machine learning models. Existing literature that seeks to measure restaurant success focuses solely on restaurant-specific data to create their explanatory variables. This paper demonstrates that ZIP Code specific demographic and density data contribute important information to predicting restaurant success. This paper also shows that year one and year two review counts are strong indicators of the number of year three reviews for a restaurant, hinting at the importance of a "ramp-up period" in the restaurant industry.

2 Data Collection and Cleaning

To encompass the numerous aspects of a restaurant, we gathered data from a variety of public data sources. We looked at both restaurant-specific information and surrounding area demographic data to provide important contextual information about a restaurant. The first obstacle was figuring out how to gather restaurant-specific data in a well-structured way. At first, we attempted to manually gather this data through extensive web crawling. However, this process was computationally expensive and would take months. As such, we decided to reference structured restaurant information through the Yelp API. We examined restaurant features such as the ambiance, noise level; outdoor seating availability; also more obvious attributes such as cuisine served and price range. Existing literature that measured restaurant success focused solely on restaurant-specific factors to form their explanatory features. We believe that this ignores an integral aspect of a restaurant: the community that surrounds it. The surrounding community has the most access to a restaurant's doors. By including data from the Internal Revenue Service (IRS) and US Census Bureau (USCB), we began to account for the different aspects of a community that form, presumably, the primary customer base of a restaurant.

An important question that we initially struggled to answer was how to quantify success. Obvious answers, such as revenue were impractical as we were largely dealing with private companies that do not publish financial data. We conceptualized success in the restaurant industry as gradient in nature. Such an understanding is consistent with the sentiment that while some restaurants remain open, they are not necessarily successful. Some restaurants turn large profits while others barely scrape by. The solution was to use the number of reviews for a restaurant as a metric of success. The number of reviews on Yelp captures a gradient and, in our view, is a proxy for quantifying customer demand. While such a metric is understandably imperfect, we felt confident that errors would be random or not significant enough to obfuscate potential conclusions. We opted to use ZIP Codes as our restaurant location feature as opposed to city or county

indicators. This is because of the diversity in taste that exists within urban centers both in demographics and food preferences. The variability in consumer taste might not be identifiable at the city level. The use of ZIP Codes allows us, in some capacity, to increase the level of specificity of the model. We then created a restaurant density factor for each ZIP Code which represented the number of Yelp-registered businesses in a ZIP Code and subsequently filtered out ZIP Codes with less than 100 businesses. This is illustrated in the following code:

```
lowDensityIndex = []
for row in dataframe.index:
if int(dataframe.loc[row,"density"]) < 100:
lowDensityIndex.append(row)
highDensData = dataframe.drop(lowDensityIndex, axis = 0)
```

The information we gathered from Yelp provided data included over 190,000 businesses (not all of which were restaurants). We then used this code to filter out non-restaurants from the Yelp data:

```
for x in biz.index:
    if str(biz.loc[x,"categories"]).find("Food") != -1 or str(biz.loc[x, "categories"]).find
    ("Restaurants") != -1:
        restaurants = restaurants.append(biz.loc[x])
```

This process cut down the number of businesses in the dataset by 61.28% to 74,587. We made a "density score" of each attribute so that a restaurant's attributes were viewed comparatively to their surroundings rather than in isolation. This was engineered on the hypothesis that a restaurant's success is partially based on the quality of surrounding restaurants. For example, if there are five restaurants in a ZIP Code that serve alcohol compared to if there was just one, then perhaps the latter will receive (most likely) more business because they are not competing for customers. By tracking the number of restaurants in a ZIP Code that share the same attributes with each other, the model is able to create a density score for how unique each restaurant is to its respective ZIP Code. The next step in our project was pulling the data associated with selected ZIP Codes from the USCB and IRS. This included data such as average gross income; distribution of ethnicity, ages, and gender; while also looking at the total population for the area. Data collection was a continuous process of adding and removing attributes, cleaning and filtering the data, etc. Our final dataset included 37,775 restaurants, with 258 attributes.

3 Preliminary Analysis

Taking a cursory look at the dataset made (which included both the explanatory and dependent variables), we noticed a few trends. We visualized the relationship between all three years of a review counts. This allowed us to see if there was a strong relationships between the variables.

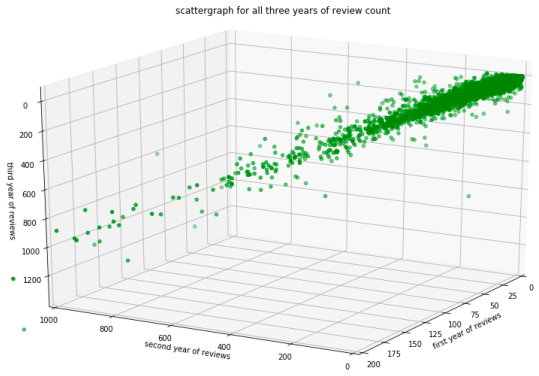

Figure 1: Three dimensional plot of of three year review count

The strong linear relationship in this graph (especially between years two and three) indicates that a restaurant needs a "ramp-up" period before it can establish itself inside of a community as either good or

bad. It is our belief that it takes time for restaurants to create a reputation for themselves and thus it makes logical sense that the number of year two reviews is a better predictor of the number of year three reviews than year one.

4 Predictive Model and Results

With the finalized data, we used regression and neural network models. Our first approach was the simplest one. Namely, a linear regression model.

```
X = df[df.columns[3:-1]]
y = df[  Year   3   ]

from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X, y)
dfN = df
dfN["predicted"] = model.predict(X)

import matplotlib.pyplot as plt
dfN[["Year 3", "predicted"]].plot(alpha=0.7)

from sklearn.metrics import r2_score
r2 = r2_score(dfN["Year 3"], dfN["predicted"])
print(r2)
```

This in-sample regression resulted in an r-squared value of 83.68% and the following graph:

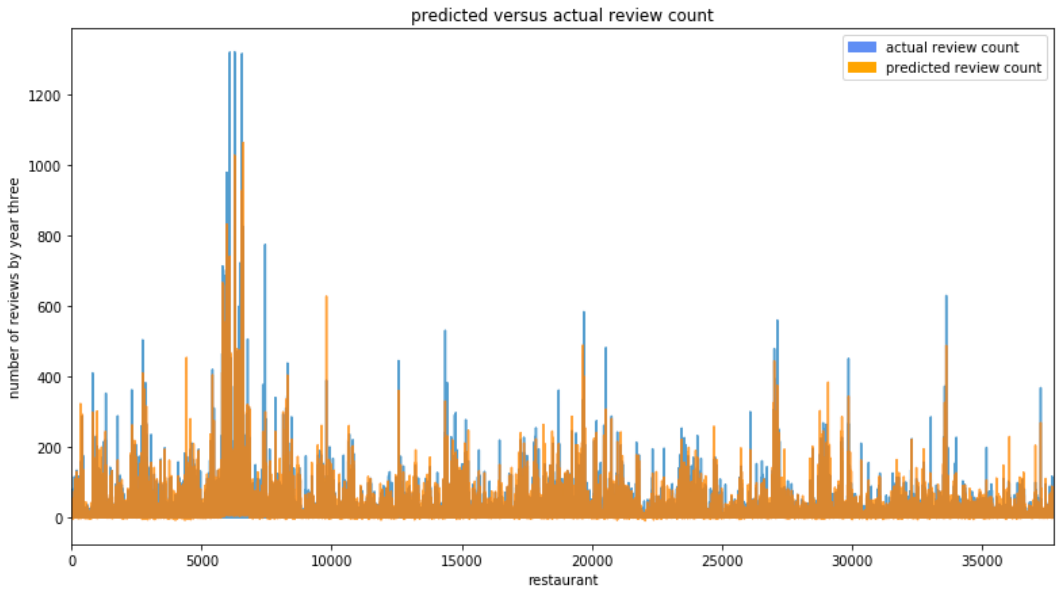

Figure 2: The performance of the linear regression predict the number of reviews

However, we believe it's also important to analyze the data out-of-sample as well, to see the accuracy when evaluating a new restaurant that was not previously analyzed. To that end, we randomized an 80–20% training-test split and evaluated the mean squared error. Over 100 iterations, the average mean-squared-error was 178.5. Furthermore, we wanted to see if there were any non-linear relationships among the inputs and outputs by using a neural network. We built 30 different Sequential neural network models that each had three dense layers. The different models were used to test various combinations of activation functions and numbers of nodes (32, 64, 128). We used out-of-sample testing and compared the mean-squared errors for the different networks as a way to try and find the best model. Ultimately, we found that a model with a

linear activation function and 128 nodes performed the best. This model achieved the lowest mean squared error value of 1186.18. After finding the best activation function and the number of nodes, we optimized the number of epochs; 60 epochs performed best with a mean squared error of 920.2376. A linear activation function creates an output proportional to the input by multiplying inputs by weights. We also performed linear regressions and achieved similar results indicating that though the relationships between the variables initially seem complex, they appear to be more linear than previously assumed.

5 Discussion

We believe that the success of the model is practically significant. Existing literature is able to explain up to 69% of the variance in star-count, and our analysis expands on this by looking at review count. The model indicates that the value of a restaurant goes beyond just the restaurant itself but also includes the community that surrounds it. To demonstrate the importance of years one and two review count on the model we performed a linear regression on the same dataset but excluded years one and two review counts. The r-squared value dropped to 0.28 (a 68% decrease). Here is the graph:

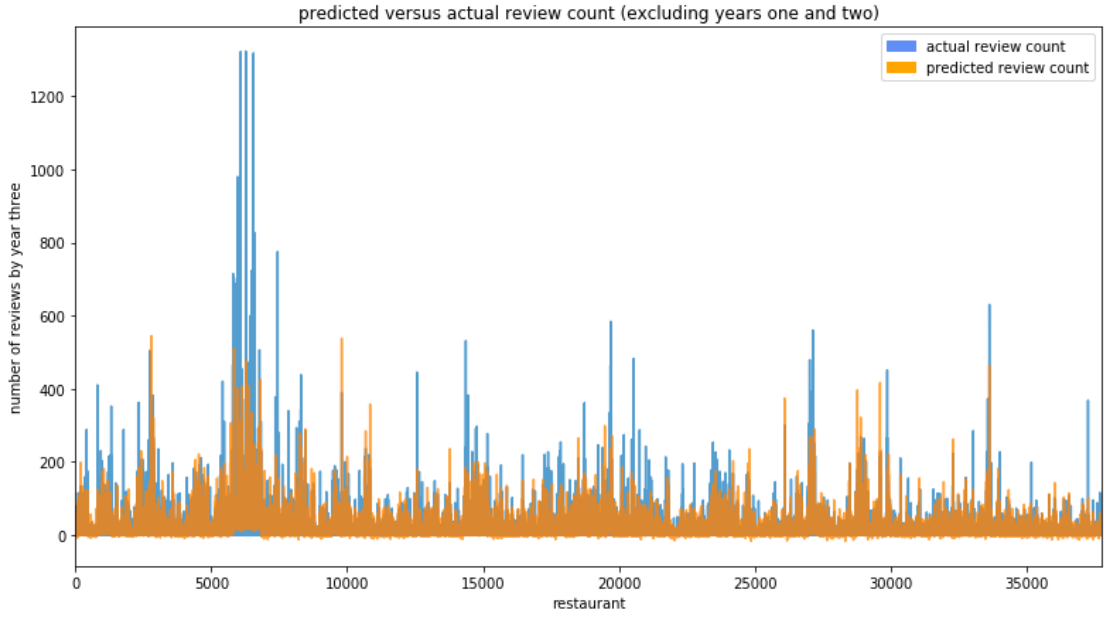

Figure 3: Linear regression model with year one and two review counts removed

For the purpose of our study, this highlights an important part of the restaurant industry, a fundamental lack of predictability. However, visual analysis of the graph suggests that there is still information to be gleaned. The local maxima of actual review count (the blue lines) seem to match the local maxima of predicted review count (the orange lines) when disregarding the order of magnitude. Specifically, when looking at restaurants with an abnormally high number of predicted reviews those restaurants generally correlate to the abnormally high number of actual reviews. For example, this can be seen in the chart between the restaurant business index 5000 and 10000. We hypothesize that the best performers in both predicted and actual review count categories will match each other.

To statistically verify this hypothesis we then decided to find the percentile match between the predicted versus the actual number of reviews in a test set after training. We ordered the restaurants by the number of reviews (high to low) for both predicted and actual. If our hypothesis was correct, there would be a notable degree of similarity between the top quartiles. The percent similarity between the top 10% of predicted and actual review count was 51%. Because of this we believe the model is still informative even without including the number of year one and two reviews. Including the review count for years one and two increased the

similarity score to 78%.

This analysis makes our results especially relevant to investors because even if we are not able to accurately predict large spikes in review count (absent years one and two) at least initially, we are able to predict successful restaurants. This means, at least theoretically, that restaurant investors would invest initially into restaurants that we deem most probable to be in the top percentile of their ZIP Code. Then as time progressed into year one and two (and their review counts became subsequently available) invest more into restaurants that the model predicts will have large spikes in their review count.

This leads us to confirm our previous conclusion that it takes time for a restaurant to establish itself inside of a community (a ramp-up period). This is because if review count spikes, one can see those spikes in years one and two because of the highly linear relationship between years one and two with respect to year three (observed earlier).

6 Next Steps

We believe that this research makes a step forward in understanding what makes a restaurant successful in today's world. Our research demonstrates that including demographic and community factors contribute to understanding the success of a restaurant. Work can still be done to understand how tastes and preferences change inside of a community. One area of particular interest is the effect of food delivery services on restaurant success. Food delivery services now make up a majority of restaurants' sales and thus their value cannot be understated. Other methods could also be employed to get more informative results. Our current model did not use any normalization or scaling, feature selection, or non-linear machine learning methods. Our project does offer insight into the way the restaurant industry works and some predictive power for selecting successful restaurants.