Predicting the Success of New, Upcoming Restaurants

Vamshi Chenna, Shiva Deviah, Koustav Mukherjee, Revanth Rayala
(chenna, deviah, koustavm, vrayala) @usc.edu
University of Southern California

Abstract

When opening a new restaurant, it is critical to know what areas to best invest in to make your venture successful. Oftentimes, this is not straightforward to determine, because it depends on a myriad of factors such as the location, the category (cuisine), and services offered to patrons such as WiFi, home delivery, valet parking, and wheelchair accessibility. Ordinarily, a new restaurant owner would have to invest time into building their menu, setting the price range, and scoping out their competition to determine how best to attract more customers and improve their rating on Yelp. With the Yelp Dataset, one can automate the third step in this process using relatively simple models to positive effect. In this paper, we describe a novel system that predicts the success rate of restaurants and offers suggestions for these restaurants to improve upon based on their competition.

The model works in two phases. First, we employ several different Machine Learning techniques including Linear Regression, Multi-Layer Perceptron (MLP), and XGBoost to estimate the rate of success of a restaurant. Next, the model identifies and suggests improvements by sampling restaurants within the same category and price range from the Yelp Dataset that have higher rating than our model's rating.

Additionally, restaurants can be assigned one or more categories, and there are over a thousand possible categories in the Yelp Dataset. This paper explores an unsupervised learning technique based on Singular Value Decomposition (SVD) and Hierarchical Agglomerative Clustering (HAC) to identify a much smaller number of higher-level categories and assign these unique labels to restaurants.

A demo of this system can be found at *https://bit.ly/2QapL3s*.

Keywords

Yelp, success prediction, machine learning, Linear Regression, XGBoost, Multi-Layer Preceptron, unsupervised learning, Singular Value Decomposition, clustering, Hierarchical Agglomerative Clustering

1 INTRODUCTION

Yelp is a currently one of the most popular platforms for crowdsourced searching for businesses. For restaurants especially, it provides a holistic view of businesses based on the information published on their site, the menu, prices, user reviews, pictures, and so on. A restaurant's rating on Yelp is also a very important indicator of whether that restaurant is actually popular or successful in reality.

The ability to identify business features that are most indicative of restaurant success can help business owners devise sensible strategies to improve their restaurant's ratings. A higher rating on Yelp indicates that the restaurant provides quality food and service, but it is not as straightforward. The manner in which "success" is quantified for a restaurant would depend on its targeted audience, category, price range, and so on. Our model aims to capture these relationships between categories and attributes, and accurately estimate a restaurant's rate of success.

2 DATA PREPROCESSING

We use the data from Round 12 of the Yelp Dataset Challenge[1] to train our models. In particular, we make use of the data in *business.json* and *review.json*[2]. The dataset is initially filtered to retain only the restaurant data. Additionally, only restaurants having categories occurring 100 times or more in the dataset are retained, to allow edge cases to be handled more easily.

2.1 Feature Engineering

Three types of features are used to train the model: (1) business attributes, (2) restaurant category, and (3) height above sea level of the business' location.

2.1.1 Business Attributes: Restaurants in the Yelp Dataset have, on an average, **11** non-null attributes and **4** categories. With over **40** possible attributes, the dataset has a lot of missing data and is sparse. Attributes are then featurized along the following lines:

(1) Boolean attributes with "True" and "False" values are converted to 0 and 1, respectively.
(2) Categorical attributes are converted to real-valued features using Mean Frequency Encoding, given by equation 1:

$$p(x_{nd} == c) = \frac{|\{n : x_{nd} == c\}|}{N} \quad (1)$$

Where $c \in [C]$ is a specific feature value that can be associated with a feature d of a sample x_n and N is the total number of training samples.

Using Principal Component Analysis (PCA) on the featurized data, it is possible to determine the features with the best signals. We chose the following **19** attributes to train our success prediction model:

```
Alcohol, BikeParking, BusinessAcceptsCreditCards,
Caters, GoodForKids, NoiseLevel, RestaurantsDelivery,
WiFi, OutdoorSeating, HasTV, RestaurantsReservations,
RestaurantsTableService, WheelchairAccessible,
RestaurantsPriceRange2, RestaurantsTakeOut,
RestaurantsGoodForGroups, elevation,
{catRegion_label OR cat152_label}
```

Note that the label for the category is either *catRegion_label* or *cat152_label* depending on the category assignment scheme used, as described in Section 2.1.2.

2.1.2 Categories: Restaurant categories in the Yelp Dataset are organized messily. There are over **850** categories relating to restaurants alone, and restaurants may have one or more categories (with the most being 37 categories assigned to a single restaurant), and this increases the complexity of assigning a unique category label to restaurants.

Many Yelp categories are closely related to each other, and can be modeled as a hierarchy. We can represent these 850 categories as subclasses of more abstract, higher-level entities. There can either be done using a simple approach of manually identifying higher-level entities from the category pool, or an unsupervised approach with Singular Value Decomposition (SVD) and Hierarchical Agglomerative Clustering (HAC) to model them.

2.1.2.1 Manual Assignment: One method of modeling high-level categories for restaurants is to select those categories corresponding to regional cuisine this includes, but is not limited to Chinese, Thai, Indian, Mediterranean, American, and so on. In this way, **32** higher level categories can be identified from the Yelp Dataset, and only those restaurants belonging to one of these categories are used for the suggestion phase.

2.1.2.2 Unsupervised Approach: This approach uses SVD to assign vectors to the categories, and then runs the vectors through a HAC subroutine to assign these vectors to clusters representing more abstract, higher-level categories. For consistency, the number of clusters here is the same as the number of regions identified during manual assignment.

If a restaurant can be assigned to possibly two (or more) higher level categories, then then one with the highest frequency of occurrence is considered.

2.1.3 Elevation: An observation on the dataset is that restaurants are geographically distributed across various countries and continents. To enhance the richness of dataset and model location-related data, we extracted contours from digital elevation datasets and combined altitude (measured in meters above sea level) as a feature with the dataset. QGIS has been used to perform point sampling on the lat-long coordinates[3].

Figure 1 shows where restaurants in the Yelp dataset are located on the world map.

Figure 1: Yelp businesses on the world map

2.2 Modeling Business Ratings as Ground Truth

To accurately predict the success of the restaurant using user rating, we must understand the quality of a restaurant as perceived by each individual and how it changes over time. A restaurant which has seen an increase in its rating year after year can be perceived as a successful restaurant. Only the last ten years of restaurant review data have been considered. User ratings are mean-shifted to account for user-bias, and then weighted to give preference to more recent ratings. The final rating of a restaurant is then the mean of these review ratings for that restaurant.

$$s(b) = \frac{1}{N}\sum_{u=0}^{N} \mu[t] * [r(u,b) - r_{avg}(u)] \tag{2}$$

where s(b) is the predicted success for business b,$r_{avg}(u)$ is the average rating for user u, r(n,b) is the rating user u have given for restaurant b. $\mu(t)$ is given by

$$\mu(t) = \begin{cases} 1 - \gamma * t, & \text{if } t < 10 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where t is the number of years since the rating was posted to Yelp, and $\gamma = 0.1$, is the discounting factor. Finally, the ratings are scaled to lie between 1 and 10. Table 1 shows the distribution of success ratings in our dataset.

Table 1: Distribution of Success rating of restaurants

Range	Count
(10% - 20%]	19
(20% - 40%]	2358
(40% - 60%]	43045
(60% - 80%]	66507
(80% - 100%]	1236

3 METHODS AND APPROACHES

Figure 2 shows a high-level overview of our system architecture.

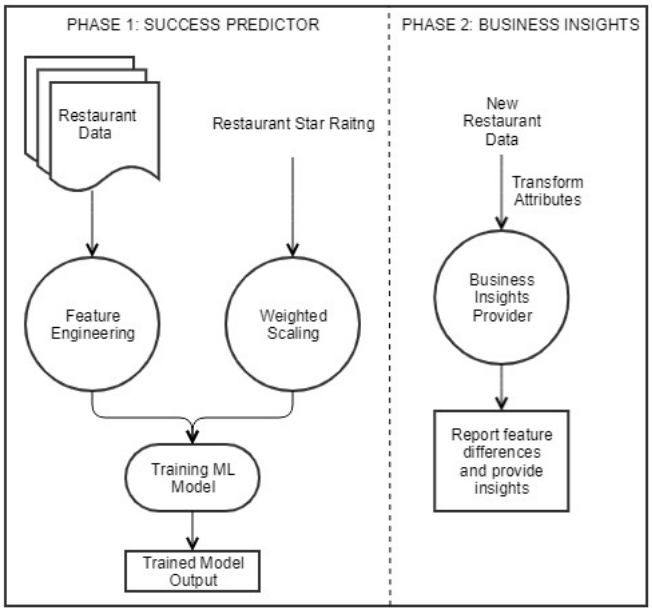

Figure 2: Flow chart for the success predictor model

The system is divided into two phases. In the first phase (labeled "SUCCESS PREDICTOR" in Figure 2), a Machine Learning (ML)

model is trained on the features from the filtered, featurized, and augmented dataset. In the second phase, (labeled "BUSINESS INSIGHTS" in Figure 2), the trained model from the first phase is used to predict the success rate for a new restaurant, and improvements are suggested by sampling the top **10%** of restaurants that have the same category and similar price range.

3.1 Phase 1: Predicting Success

Various different ML models have been experimented with (as described in Section 4), but only the two best models are described in this section.

3.1.1 Multi-Layer Perceptron: Multi-Layer Perceptron (MLP) Regressor is a supervised learning algorithm that learns a function f(·) : $R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features X=x1, x2, ..., xm and a target y, MLPRegressor[6] has capability to learn non-linear models in real-time and trains using backpropagation. Here we construct the neural network with 3 hidden layers of sizes 128, 32, and 10, respectively. Rectified Linear Unit (ReLU) is used as the activation function.

3.1.2 XGBoost: eXtreme Gradient Boosting, or XGBoost, is an ensemble method that trains a large number of weak estimators on subsets of the data to learn a non-linear decision boundary[5]. An XGBoost model is trained with different combinations of parameters using scikit-learn's GridSearchCV API, and the best model and the corresponding set of parameters is returned. 5-Fold cross validation is used to evaluate the model. The advantage of XGBoost over other ML methods experimented with in this paper is that it trains very fast, and can handle missing data without the need for imputation.

3.2 Phase 2: Providing Business Insights

The second phase provides valuable insights to the user, specifically calling out the areas in which the business can improve upon. It also suggests the extent to which a business can improve, assuming these suggestions are incorporated by the business owner.

A new business is mapped to a category during prediction. If the user defined category cannot be directly mapped to one of the predefined categories that we have extracted from the business dataset, we make use of the SpaCy library's word2vec model to determine the closest category to which the business belongs using cosine distance as a similarity measure.

Once the category for the new business is determined, its success rating is predicted using the model trained in section 3. Next, all businesses in the dataset having same category and with an absolute difference of 1 or less for the price range are extracted. From these businesses, only the top 10% of restaurants with a higher rating than our model prediction are considered, denoted by top_k. Now for each attribute d in the list of attributes from section 2.1.1, we determine the feature value v that should be present in the improved version of the new business, denoted by $suggested_d$, by grouping each feature by its feature values and summing the weighted ratings. The weighted rating is obtained by dividing the success rating by the rank of the restaurant. The feature value v that has the highest aggregated weighted rating is assigned to $suggested_d$. Mathematically, a categorical feature d, can be grouped into multiple values v. On the other hand, for a boolean feature $v \in \{0, 1\}$. Let

$$\gamma_{ndv} = \mathbb{1}[v == x_{nd}] \tag{4}$$

Be an indicator that groups together businesses that have same feature value v for a feature d. Then,

$$suggested_d = \underset{v \in V}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{v \in V} \gamma_{ndv} * r'_n, \forall d \in D \tag{5}$$

$$r'_n = \frac{r_n}{rank_{x_n}} \tag{6}$$

Equation 5 Represents the suggested attributes values v that should be assigned to each feature d learned from the top_k restaurants in a category. The attributes in $suggested_d$ are then compared with attributes originally given during prediction, and the differences are reported. An example of this can be seen in Table 2.

Table 2: Insights: Business Improvement Areas

Attributes	Current	Suggested
Alcohol	none	beer_and_wine
BikeParking	absent	present
HasTV	nan	present
NoiseLevel	loud	average
OutdoorSeating	present	absent
RestaurantsAttire	formal	casual
RestaurantsDelivery	present	absent
RestaurantsGoodForGroups	absent	present
RestaurantsTableService	absent	present
WheelchairAccessible	absent	present
WiFi	no	free

The "Current" column represents feature values for each attribute for a business taken from the dataset, and the "Suggested" column shows suggestions as given by our model.

4 RESULTS AND DISCUSSION

For the purpose of evaluating our system, the dataset was divided into train and test portions with 90% of the data for training and the remaining for testing. The ML models from the success prediction phase are evaluated using the Root Mean Squared Error (RMSE) measure. "SVD" and "Regional" refer to the category assignment schemes described in section 2.1.2.

Table 3: RMSE for Regression Models

Algorithm	SVD	Regional
XGBoost	0.85	0.87
Gradient Boosting	0.86	0.87
MLP Neural Nets	0.88	0.88
Extra Trees	0.89	0.89
Linear Regression	0.87	0.86
Decision Trees	0.93	0.92

Some of the regression models described above can also be used for classification tasks, with some changes. The success ratings are rounded off to nearest integer to represent 1 of 10 possible

classes. Table 3 gives a description of the results for the regression models. Table 4 describes the results of the classification models used. Accuracy measure is used to evaluate these models.

Table 4: Accuracy Values on Classification Models

Algorithm	No of Classes	SVD	Regional
Logistic Regression	10	40%	39.9%
MLP Classifier	10	42.5%	43%
XGBoost Classifier	10	62.1%	64.2%

The influence of category on the success rating prediction model was also investigated. Table 5 shows the RMSE for various classifiers with and without using the category as an additional feature.

Table 5: RMSE for Regression Models using Category as a Feature versus not

Algorithm	Without Category	With Category
Linear Regression	1.08	0.84
Decision Trees	0.98	0.92
MLP Neural Nets	0.97	0.883
Gradient Boosting	0.88	0.86
Extra Trees	0.89	0.89

To gauge the efficacy of the business suggestion phase, 1000 restaurants from the test dataset were sampled at random, and run through our system. The ML regression model predicts the success of some restaurant x_n, given by r'_n. The feature differences are then computed as described in section 3.2. Using the new feature values v for the corresponding feature d, we estimate the new rating r''_n for the improved business. The accuracy of the model is then determined by

$$accuracy_{business\_insights} = \frac{\sum_{n=1}^{N} \mathbb{1}[r''_n > r'_n]}{N} \tag{7}$$

$$average\_improvement_{business\_insights} = \frac{\sum_{n:r''_n > r'_n} | r''_n - r'_n |}{| n : r''_n > r'_n |} \tag{8}$$

Where, N is the total number of samples used for evaluation. The results of evaluation for different random states are shown in Table 6.

Table 6: Prediction Accuracy for Phase 2

	Accuracy	Average Improvement
Seed = 0	83.5%	8%
Seed = 42	86.8%	8.10%
Seed = 1337	85.6%	8.12%

5 FUTURE WORK

In addition to knowing how successful a business would be and the insights to incorporate to improve the success rating of a business, one might be interested in knowing the geographical location where the investment should be made to reap maximum benefit. For instance, a Mexican-themed restaurant may become more popular in the southern or eastern states. To incorporate this change, it would be necessary to assign each restaurant to a region label (such as "US West", "Asia Middle-East", and so on), based on its lat-long coordinates. Then this feature would automatically be incorporated while providing business insights as an additional suggestion to the user.

References

[1] Yelp Dataset Challenge - Round 12 - https://www.yelp.com/dataset/challenge
[2] The documentation for the Yelp Dataset Challenge data can be found at https://www.yelp.com/dataset/documentation/main
[3] Ujaval Gandhi. *Sampling Raster Data*.
[4] Michel Goossens, Frank Mittelbach, and Alexander Samarin, *Predicting Business Ratings on Yelp*
[5] Tianqi Chen and Carlos Guestrin. [*XGBoost: A Scalable Tree Boosting System*]. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining XG Boost Documentation*
[6] Machine Learning Classification and Regression Algorithms *scikit-learn*.