Using Social Data to Predict If a Restaurant Will Fail

Alick Xu and Grant Apodaca

Abstract— **Restaurants are an extremely fickle business. They can never take off, explode in popularity, or run out of business despite years of success. As students of University of California, Santa Barbara, living in Isla Vista, we have noticed the number of restaurants that have come and gone - many restaurants standing today were not here when we first came.**

We wanted to see if we could find out what makes restaurants fail. H. G. Parsa was one of the first to attempt to truly identify the failure rates of the restaurant industry within the United States. However, his study was more empirical and did not delve into the precise reasons why a restaurant would fail. We propose a method to quantitatively evaluate restaurants to determine whether they succeed or fail. To further analyze the results of his statistical analysis, we attempt to use a Support Vector Machine with linear classification to model and classify restaurants and determine if restaurant's success could be identified based on their characteristics. We successfully determined a specific subset of 168 identifiable characteristics from gathering data using Yelp and other local databases. Using 13 distinct features, we could achieve a 69% accuracy in determining the success of a restaurant.

While there are other factors involved that we could not gather data for, such as geographic location and local demographics, we are still able to show some promising results and provide the framework for future research.

I. Introduction

The restaurant market has always had a stereotype of being lucrative, and risky, due to the high perceived rates of failure. As many people try to open up restaurants, they rely on either sound business advice, or a substantial amount of capital to stay in business until profits can outweigh the running costs. Studies have been done in the past attempting to quantify the actual failure rates among the restaurant industry, but little has been actually done to explain why, or provide factual evidence supporting what makes or causes failures in a successful restaurant. One study in particular by H. G. Parsa [1] from the Ohio State University has lead the business world on the statistic that restaurants have a 60% cumulative probability that they will close (or ownership will turnover) within the first three years of business. Their study had some statistical analysis attempting to examine contributing factors, but due to the type of their data, they had very little depth in this facet of their study. In other words, they identified the trend of restaurants closing in 3 years - but they could not explain why.

Our study attempts to use large-scale data mining and machine learning to classify successful restaurants in the Isla Vista, CA region, and identify what about restaurants are most important to their success or failure. We then attempt to use this feature-based classification model to attempt to predict newly opened restaurants in a local area to study the accuracy or viability of using these feature analysis techniques as a tool for guiding the restaurant industry. Using Social Media and Web-scraping, we collect 168 features, verify using several classification methods, and eventually use recursive-feature elimination to create a model with moderate verification accuracy. We then analyze the effectiveness of this model, and attempt to understand empirically the highest rank features.

II. Related Work

H. G. Parsa, John T. Self, David Njite, and Tiffany King performed a study from Ohio State University attempting to quantify the actual failure rates among restaurants, and compare them to other industries. Thier study examined 2,439 restaurants within the greater Chicago area. They studied the rates of the restaurants in the areas failures or turnovers, and found that the total cumulative turnover rate over three years was 59.74% between the years of 1996 through 1999. Parsa then compared the differences in these rates between franchised and independent restaurants, which only had a 4% cumulative difference over the three year period. The study then attempted to examine properties of the restaurants that failed, including differences in cuisine types, and several empirical hypothesis. The rest of the study concludes with comparing the failure rates in the restaurant industry to that of other industries, showing their main goal that restaurant failure rates werent as high as the previous 90% myth. This work, however, didnt spend too much time attempting to find factual aspects of the businesses that failed, and instead studied their true percentages over the previously unidentified estimates.

Our work attempts to further this research by using machine-learning classifiers to model the features and aspects of these failing business in attempt to truly identify distinct differences between restaurants which succeed, and those which fail.

In addition, we did not find any previous studies where business features were used to evaluate the success of restaurants. This could be due to a lack of features available, interest in the problem, or other factors. In any case, this was an exciting area to dive into, being the first to do such a quantitative study.

III. Methodology

A. Data Gathering

To gather data efficiently, we bug by using existing databases which have recorded factual features about restaurants Several of these existing data sets include the Yelp website, OpenTable, Google Pages, and several others. Given the API resources available to us, the Yelp API and Google

Pages API were our most viable sources to begin scraping social data. The Yelp network holds data about thousands of restaurants among other businesses in over 50 countries, seeing near 102 million comments from users to date. [3] A majority of this data is opinion based (ratings, comments, reviews), but there is also a large set of data which are factual features about these restaurants. Using the Yelp API, we could gather restaurants from a specific geographic location, and request further specific information using a webscraper on each business's specific Yelp website. The initial features set includes over 168 distinct features, with approximately 110 features being binary in value. An example of the feature-set can be seen in Table 1, but the features themselves described factual aspects about the business, ranging from average meal cost, to the inclusion of waiter service, to the various cuisine types that were offered.

These business features can be seen on every Yelp page. These features are determined both by the business owners who run the Yelp page as well as patrons of the business. When writing a review, patrons are given the opportunity to mark which features the business has with a simple Yes/No answer choice. This makes it so that the business owner cannot falsely advertise features that they do not have.

Due to the difficulty of acquiring a restaurant's true age without individually researching each establishments records in a public city record, we also used the delta between the timestamps of Yelps oldest and newest reviews to estimate each locations age of business. It is important to note that while we were able to obtain overall rating information, we did not include this in our feature set. This is because opinion-based data would not be helpful in establishing a true descriptor of each restaurant, since these can be biased, and fluctuate, as well as be altered by fake-reviews and comments. [2] While our data is focused on the Isla Vista, CA region, we have developed a framework to obtain similar data from any geographic area.

One issue that arose in the initial gathering of data was the isclosed data value returned in the Yelp search API. This feature is used to determine if a restaraunt was permanently closed or not, which was required for the establishment of a ground truth. However, after initial data gathering was underway, we discovered that all restaurants which were permanently closed were actually filtered out of the search results, meaning the only way to find each yelp business page of these closed restaurants was to individually search for them. Using records in Santa Barbara [4], we were then able to acquire a list of 25 permanently closed restaurants from the past 10 years, and used this to guide our web-scraper to establish a ground truth.

We attempted to use the Google API for the same purpose as the Yelp API, but we found that Google Pages currently lacks the same type of factual descriptors of restaurants, and mostly contains usage statistics, and reviews. Because only factual features of restaurants were needed, we deemed that the data provided from the Google API would not be necessary.

Category	Options
Bike Parking	No, Yes
Good for Kids	No, Yes
Attire	No, Casual, Dressy, Formal (Jacket Required)
Noise Level	No, Quiet, Average, Loud, Very Loud
Alcohol	No, Beer & Wine Only, Full Bar
Afghani	No, yes
Argentine	No, Yes
Burgers	No, Yes
Hot Dogs	No, Yes
Mexican	No, Yes
Good For	No, Breakfast, Brunch, Lunch, Dinner, Late Night, Dessert
Price Range Per Person	No, $, $$, $$$, $$$$
Ambience	No, Divey, Hipster, Casual, Touristy, Trendy, Intimate, Romantic, Classy, Upscale
Accepts Credit Cards	No, Yes

TABLE I

A SAMPLE OF THE 168 FEATURES THAT WE USED. MOST FEATURES WERE BINARY FEATURES, WITH A MAJORITY BEING CUISINE TYPE

B. Establishing Classification

Using our 168 features, we decided to use linear classification to attempt to classify restaurants. The binary classification was whether they WOULD or WOULD NOT permanently close within the first three years of their business. Our training set included 25 businesses whom were currently permanently closed, and 20 restaurants which were currently open and have been open for longer than 3 years. The restaurants were then classified as have been closing in less than 3 years of their operation, meaning some of the permanently closed restaurants in our training set were still classified as WOULD NOT permanently close in 3 years. This was important, because this allowed us to compare results to H. G. Parsas study, as well as kept the Binary distinction to a fairly empirically average timeline of when restaurants are usually posed to fail.

C. Prediction Attempts

We decided to use a Support Vector Machine as our classifier, as most literature points to SVM to having the best accuracy. Using the SciKit-learn library, we were then able to perform different experiments using various kernels of classification, such as linear and quadratic. Through experimentation, we found that the linear classification model for the SVM gave us the greatest accuracy. The following results reflect the accuracy of the SVM using the linear kernel. We first tried 5 fold cross validation to evaluate the accuracy, but the accuracy was very low, as we show in Table 2. Therefore, we chose to do an 8 fold cross validation, because our limited training set needed a many restaurants as possible, and with fewer folds we were receiving varying and non-consistent results.

Due to the limitations of the dataset, it was very difficult to get more results. Yelp was founded in 2004, so data only goes back that far. It was also hard to get information about restaurants that have closed past 10 years ago. It was also

	Accuracy	# of Useful Features
W/out Feature Removal:		
5-fold CV SVM	50.5%	N/A
8-fold CV SVM	52.5%	N/A
With Feature Removal:		
5-fold CV SVM	54.5%	24
8-fold CV SVM	66.9%	13

TABLE II

RESULTS OF THE CLASSIFIER WITH AND WITHOUT FEATURE REMOVAL

hard to establish a ground truth - all of the restaurants that we found were used to train the classifier, and the raw number of such restaurants was very low.

D. Classifier Evaluation

When we used 8-fold cross verification, we found that the model with the original 168 features yielded an accuracy of 52.5%, which is no better than flipping a coin. Upon receiving such inaccurate results, we began to analyze the features to determine if our feature set was too large for our given limited data set. In order to trim relatively useless features, we used Recursive feature elimination, which recursively removes common features from the data set (features in which every item or high percentage of items in the set has the same value for), and re-evaluates their validity. The scikit-learn tools determined that there were 3 different feature sets that were important, numbering in 13, 14, or 19 of the original 168 features. The graph showing the results of all of the recursive attempts can be seen in Figure 1. The highest ranked 13 features (based on how decisive they are in determining the classification) can be seen in Table 3. The most important features, which were all ranked as equally important are in the table, as well as the 7 next important features which were also tied. The results of using the classification model using the 13 features in Table 2, while the 8 fold validation gave us the highest accuracy of 66.9%.

Unfortunately, we were unable to get results such as the accuracy and f-score of the classifier, simply due to a lack of data. Truth values are needed to compare the classifier's predictions to, and there were simply not enough restaurants to break off of the training set and use as our truth value, as the classifier performance suffered substantially after excluding even a couple of training restaurants. As future work, more data needs to be gathered (perhaps in a larger area than Isla Vista) so that a truth set can be created that does not need to be included in training.

IV. ANALYSIS

Although the classification model is not as accurate as we would have hoped, the features it determined to be most important did make a great deal of sense, empirically. Based on the Isla Vista area, which is dominated heavily by a certain culture, the features can be viewed as attributes of these culture phenomena. For example, the Bike Parking feature would be important to restaurants near a University

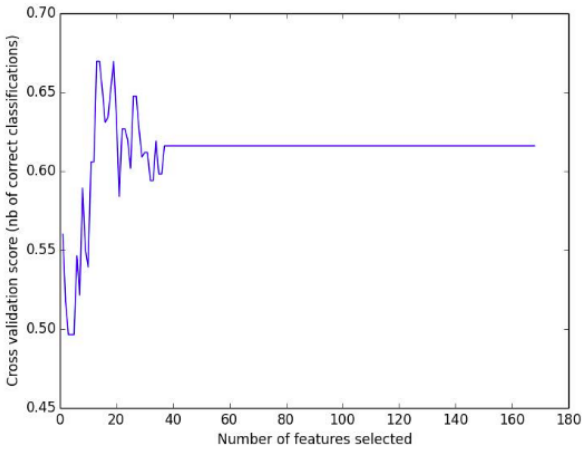

Fig. 1. Graph detailing the number of features used and the resulting accuracy of the validation

First (Most Important Features)	alcohol, bike parking, cafes, chinese, hotdogs (fast food), italian, mediterranean, outdoor seating, pizza, price range per person, vietnamese, waiter service, wi-fi
Second	caters, fast food*, asian fusion, parking, accepts credit card, breakfast and brunch, has a tv

TABLE III

LIST OF THE MOST IMPORTANT FEATURES.

campus with one of the largest known bicycle infrastructures. Another is the cuisines which were chosen to be dominant, specifically fast-food, pizza, Italian, and cafes, all of which are common food-types present near college campuses, mostly due to their low cost.

One downside to the empirical relevance of these features is that the classification model cannot be used nationally, but then must be trained to specific geographic locations. To fix this, use of all of the features would be needed (not just the 13 we determined to be most accurate). To make better use of all 168 features, our training set would need to be significantly larger, and also reflect establishments from various geographic locations, not just one area. A caveat to this is that for more personal results to a specific geographic area, this model could be useful then, and would require training the model specific to restaurants in a specific area, and not from a diverse location set.

For the purposes of predicting the failure of a restaurant of not, our classification model needs more training data to be of much use. Its current accuracy is ok, but still not within a trustworthy range to be used widespread. The process of creating the model did show that using linear classification, factual descriptions can be determined to identify what makes a closed restaurant tick over a successful restaurant. This amount of analysis shows that not only can the failures of restaurant be predicted using probabilities (as in the study by H.G. Parsa), but that identifiable characteristics can be

determined among these failed and successful restaurants. The fact that using machine learning was partially successful at all shows that there is a correlation between attributes in restaurants and their ability to stay open longer than three years.

It would be very exciting to see our framework applied to other geographic areas, along with an analysis of the demographics of those geographic areas and having those demographics applied as features. One dataset we were unable to find was the demographics of the population of Isla Vista, which we presume would've helped us a lot with classification. A bigger geographic area would also give more data, which means more accurate results and the ability to quantify the accuracy of classifier.

V. Conclusions

We predicted the failure of restaurants given certain features that they provide. Attempting to further the research by H. G. Parsa, we attempted to use linear classification to model and classify restaurants and determine if restaurant's success could be identified based on their characteristics. We successfully determined a specific subset of 168 identifiable characteristics and determined them to be most useful in characterizing a restaurant as either prone to fail, or succeed. Using 13 distinct features, we could achieve a 69% accuracy in determining the success of a restaraunt. Using further data gathering to achieve a larger ground truth, this method could be expanded from our currently geographically biased model to a more universal model. Further study merits obtaining a larger dataset of restaurants to expand our training and testing sets in our classification model.

References

[1] Parsa, H. G. (2005). Why Restaurants Fail. The Cornell hotel and restaurant administration quarterly, 46(3), 304-322.

[2] Santosh KC and Arjun Mukherjee. 2016. On the Temporal Dynamics of Opinion Spamming: Case Studies on Yelp. In Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 369-379. DOI=http://dx.doi.org/10.1145/2872427.2883087

[3] http://www.yelp.com/factsheet

[4] http://www.santabarbara.com/dining/restaurant_list.asp?show=closed