Predicting the most profitable area a new/existing business can expand to

Aditya Chavan
aschavan@usc.edu

Likhit Dharmapuri
ldharmap@usc.edu

Venkata Sai Praneeth Nallamalli
nallamal@usc.edu

Tanushree Joshi
tanushrj@usc.edu

ABSTRACT

Business location plays a key role in attracting the right customers and therefore affects the business's profits and popularity directly. In this paper, we introduce a recommendation model that will help new and existing restaurant owners to identify the neighborhood that their business can expand to. Hence, our solution focuses on empowering the business owners to make informed decisions for the same. Our model will take city, restaurant categories and its price rating as an input and predict the neighborhood where the business is most likely to become successful.

KEYWORDS

Naïve Bayes classifier, clustering

1 INTRODUCTION

The US restaurant industry generated a revenue of around $799 billion in 2017 and employs around 10% of the overall US workforce [1]. According to a study by Ohio State University [2], around 60% restaurants fail in a span of three years. Further, their research also suggested that the restaurant location can directly affect its survival [3]. Given the huge size of this industry and the business failure statistics, we decided to build a recommendation model which can give business owners and investors an idea about a restaurant's success in a neighborhood and help them make informed decisions. For this purpose, we used the data from the Yelp Dataset challenge. Yelp business owners may provide check-in offers to their customers which promotes customer loyalty. It also increases the business's social media presence and hence brings in new customers. Therefore, check-ins give a fair idea of the restaurant's popularity. Thus, our model focuses on combining results from three classifiers which predict the check-ins, star ratings and the probability of it shutting down respectively, to decide the most favorable neighborhood for a restaurant.

2 DATASET

The Yelp Dataset consisted of 6 json files out of which we have used the business and checkin data for this model. The Checkin.json file consist of 157075 records and the business.json data consists of 188593 businesses. Out of these, we filtered out restaurants using 'categories' attribute (that involve only 'restaurant', 'food', 'bar') and got a total of 72624 businesses (~40% original). In the check-ins dataset, each business has a "time" column. This column consists of check-ins based on hour of the day, and day of the week. Due to lack of clarity in the dataset documentation and for purpose of simplification, we assumed these entries are based on all time. We calculate the total number of check-ins per restaurant to use as a feature.

The following features were used as input to all models:

- Neighborhood
- The restaurants categories
- The restaurants price rating (between 1-4)

The following features behaved as the output to our models:

- Check-ins to a restaurant
- If the restaurant is open or has been shutdown
- Restaurant's rating

3 METHOD

3.1 Preparing the data

The Yelp business data consists of 188593 businesses which span across more than 500 categories. We filtered the data using 'Food', 'Restaurants' and 'Bar' categories, which resulted in a total of 72624 businesses. On analyzing this filtered data, we found out that a lot of unnecessary data of businesses like Caterers, Shopping Markets, Gas stations etc. got retained. So, we made a list of 91 of the most frequent categories which cover all the restaurant data and

filtered it further. We targeted 4 of the top cities, by restaurant count, to build the predictors for:

1. Toronto: 8888 restaurants (6150 still open)
2. Las Vegas: 7872 restaurants (5467 still open)
3. Phoenix: 4666 restaurants (3345 still open)
4. Montréal: 3741 restaurants (2901 still open)

3.2 Location Clustering

A lot of restaurants lack neighborhood information. For example, in Toronto, ~20% of the restaurants don't have any neighborhood information. On the other hand, Phoenix doesn't have a single restaurant with neighborhood information. Since neighborhood prediction is the main goal of this effort, we need to create neighborhood labels for each restaurant.

Neighborhoods were created for restaurants without any neighborhood by using k-means clustering. If a restaurant already had a neighborhood, it wasn't reassigned a to a different neighborhood. Instead, the k-means algo was seeded with initial clusters comprising of the original neighborhoods.

3.3 Predicting the best neighborhood

We model the problem of predicting the best neighborhood to open a restaurant by first training three Naïve Bayes classifiers to predict each of the follow attributes per restaurant:

- Check-ins to a restaurant
- If the restaurant is open or has been shutdown
- Restaurant's rating

The problem of predicting the best neighborhood is then modeled as finding the neighborhood – with the same set of category, and price rating features – which provides the highest probability for each of the classifiers. Therefore, we provide three signals back to the user.

Naïve Bayes is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting [4]:

$$y = \underset{c_i}{argmax}\ P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

Figure 1

It uses the Bayes rule

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

and the naïve assumption that the features are conditionally independent given the class for classification. Below are the types of Naïve Bayes classifier [5] that we explored:

- *Gaussian Naïve Bayes:* Deals with continuous data using the assumption that the continuous values associated with each class are distributed according to a Gaussian distribution.
- *Multinomial Naïve Bayes*: With a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial $(p_1,\ldots,p_n)$ where p_i is the probability that the event i occurs.
- *Bernoulli Naïve Bayes:* In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs.

3.3.1 Predicting check-ins

The check-ins data follows an exponential distribution, as can be seen in Figure 2. In order to convert a continuous variable like check-in into a discrete variable to use as a output label with the Naïve Bayes classifier, we binned the check-ins into 5 different classes using the quantile function for the exponential distribution. This gives us 5 balanced classes.

We found that the Bernoulli Naïve Bayes Classifier using 5-fold cross validation to predict the check-ins rating gave us the best results.

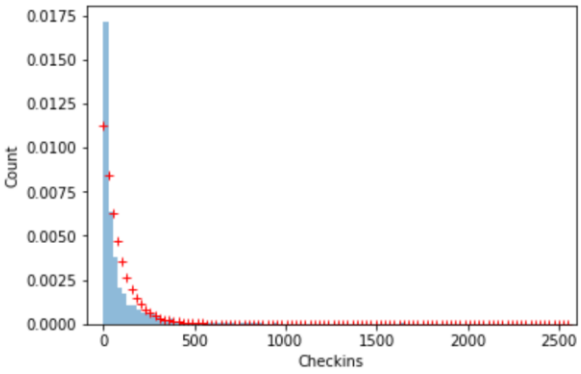

Figure 2: Toronto check-ins

3.3.2 Predicting star rating

The star rating is discrete with value within (1,2,3,4,5). We built a Multinomial Naïve Bayes Classifier using 5-fold cross validation to predict the star rating given the categories, price rating and neighborhood of a restaurant. We also experimented with using the price rating as a categorical feature instead of as a numeric feature but got similar results.

3.3.3 Predicting survival

The 'is_open' attribute in the dataset was used as a binary variable to predict the probability of a restaurant surviving in a neighborhood. We built a Bernoulli Naïve Bayes classifier and used the 5-fold cross validation technique, to predict the survival of a business.

4 RESULTS AND DISCUSSION

In order to evaluate the classifiers, we calculated the prediction accuracy, and the Kolmogorov–Smirnov test on a hold out set (20% the size of the number of restaurants per city). Table 1, 2, 3, and 4 shows the results for each city, for each classifier.

Table 1 Toronto

Classifier	Accuracy	KS-Statistic	KS-p value
Check-in	38%	0.29	1.7e-65
Star Rating	29%	0.18	2.07e-27
Survival	68%	0.22	4.97e-40

Table 2 Las Vegas

Classifier	Accuracy	KS-Statistic	KS-p value
Check-in	26%	0.28	2.89e-55
Star Rating	27%	0.10	1.82e-08
Survival	68%	0.19	1.11e-27

Table 3 Phoenix

Classifier	Accuracy	KS-Statistic	KS-p value
Check-in	44%	0.31	3.38e-42
Star Rating	27%	0.10	7.5e-5
Survival	69%	0.22	1.39e-20

Table 4 Montréal

Classifier	Accuracy	KS-Statistic	KS-p value
Check-in	36%	0.30	6.73e-30
Star Rating	31%	0.20	1.54e-14
Survival	74%	0.22	8.3e-17

On an average, over all cities, we have an accuracy of 38% for the check-ins predictor, 28.5% for the star rating predictor, and 70% for the survival predictor.

The accuracy for the check-ins, and star rating classifiers was found to be low because accuracy penalizes predictions into any wrong class equally. So even if a restaurant was classified one star less than the actual rating, it was equally penalized as if the prediction was off by multiple stars.

In order to perform a softer check on the predicted labels, we also performed the KS-test. The output of the test shows that the predicted distributions are not very far off from the actual distributions (a KS-statistic of 0 means that the distributions are same, and infinity means the distributions are completely different).

Unfortunately, the KS-test gives us a very small p-value which indicates that the we cannot reject the null hypothesis of the two sided KS-test that the distribution predicted by our classifiers and the actual distribution followed by the data are the same.

ACKNOWLEDGMENTS

We would like to thank Professor Rafael Ferreira da Silva and Professor Anoop Kumar for their guidance and support.

APPENDIX

Github link to code: https://github.com/saip009/yelp-dataset-challenge

Categories selected to use as feature:

coffee & tea, specialty food, sandwiches, breakfast & brunch, chinese, cafes, canadian (new), bakeries, fast food, pizza, desserts, italian, japanese, burgers, pubs, american (traditional), sushi bars, indian, juice bars & smoothies, asian fusion, korean, mexican, middle eastern, thai, mediterranean, salad, chicken wings, ice cream & frozen yogurt, seafood, beer, wine & spirits, vegetarian, comfort food, vegan, greek, barbeque, vietnamese, diners, caribbean, french, american (new), halal, ethnic food, gluten-free, delis, tea rooms, gastropubs, tapas/small plates, soup, steakhouses, bubble tea, dim sum, noodles, donuts, chicken shop, portuguese, chocolatiers & shops, ramen, tapas bars, latin american, bagels, pakistani, fish & chips, taiwanese, modern european, tex-mex, british, creperies, southern, filipino, african, hot dogs, irish, poke, ethiopian, afghan, turkish, falafel, hot pot, spanish, local flavor, himalayan/nepalese, hawaiian, lebanese, persian/iranian, polish, waffles, soul food, malaysian, sri lankan, live/raw food

REFERENCES

[1] National Restaurant Association: https://www.restaurant.org/News-Research/Research/Facts-at-a-Glance

[2] H.G.Parsa. Why restaurants fail https://journals.sagepub.com/doi/abs/10.1177/0010880405275598

[3]H.G.Parsa. Why restaurants fail II https://www.researchgate.net/publication/254362606_Why_Restaurants_Fail_Part_II_-_The_Impact_of_Affiliation_Location_and_Size_on_Restaurant_Failures_Results_from_a_Survival_Analysis

[4] Naïve Bayes classification: https://towardsdatascience.com/introduction-to-naive bayes-classification-4cffabb1ae54

[5] Types of Naïve Bayes: https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Parameter_estimation_and_event_models