Data Science Specialization - Capstone Project

Predicting Frequently Reviewed Businesses on Yelp

November 21, 2015

Abstract

When users browse Yelp's website, they tend to look at businesses with a high review count. In this project, we explore different attributes of businesses in the Yelp dataset and develop a series of statistical models to predict if a business is likely to receive more than 100 user reviews. We combined business information provided by Yelp with maps, census data, and calculations of review timing and frequency to conduct exploratory analysis and create statistical models. While many of our predictors showed some promise in the exploratory phase, we could only successfully classify businesses as highly reviewed when the number of reviews generated per day was considered. Our most successful model had a precision rate of 79% and accurately predicted 217 of 465 restaurants with more than 100 reviews. Future research could improve on these results by incorporating sentiment and topic analysis of user reviews and merging Yelp's dataset with local traffic and tourism information.

Introduction

Business owners use Yelp to attract more customers and their goal is to get as many positive reviews as possible. Generally speaking, having a high number of reviews can help boost a business's profile on Yelp. If a given number of people have rated and reviewed a particular business, others are likely to patronize that business over a similar business with a low review count.

In this report, we consider the possible factors that account for the high number of reviews for a relatively small number of businesses on Yelp, and why these businesses seem to generate many more reviews in comparison with the majority of businesses. Does the discrepancy in the high number of reviews for a small fraction of businesses indicate that these select businesses are of higher overall quality than their Yelp competitors, and can high review counts be used as a proxy for success? Furthermore, could a business's success be predicted based on certain attributes? These questions led to the formation of our research question: **Can we predict frequently reviewed businesses?**

Methods and Data

Dataset

The dataset, provided by Yelp, consists of five JSON files: businesses, checkins, reviews, tips, and users. We utilized two of the files, businesses and reviews, which provide over 1,500,000 reviews from more than 61,000 businesses. Important identifiers considered include business ids, locations, ratings, dates, and categories.

Methods

Framework

For this capstone project, highly reviewed businesses were defined as businesses with 100 reviews or more, and we limited our research to the 6 cities in the United States.

Exploratory Analysis

The number of businesses with more than 100 reviews concern less than five percent of all businesses in the dataset (see Figure 1).

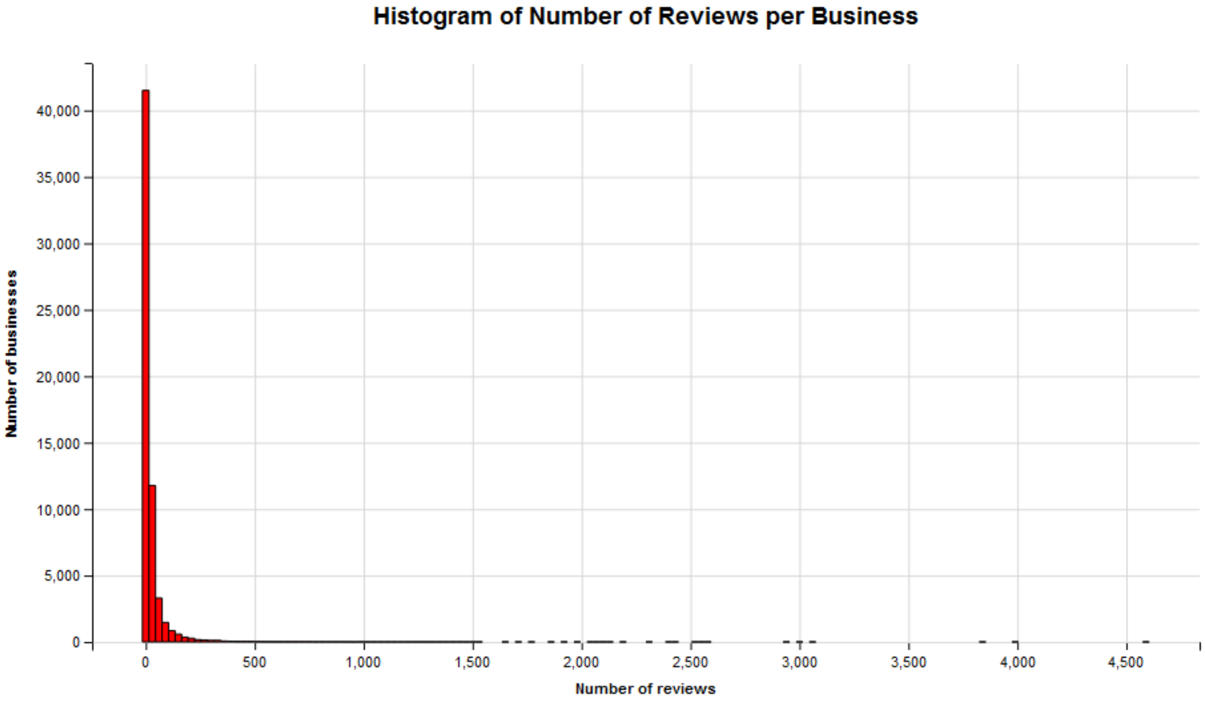


Figure 1: Histogram of Number of Reviews per Business

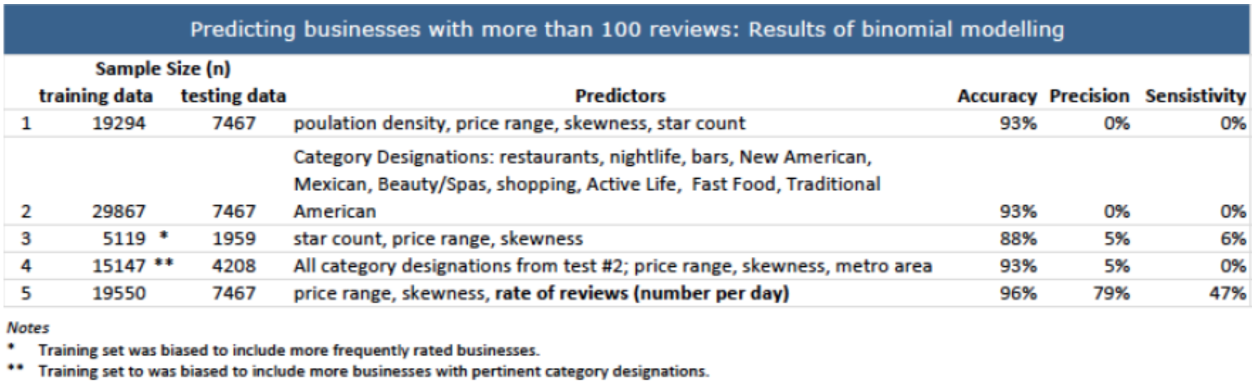
To answer our research question, we explored a number of predictors from the original Yelp business and review datasets, including, most commonly, review count (the number of total reviews a business has received), stars (the star rating a business received), and business price range. Business categories were also included in our analysis in order to determine what types of businesses received a disproportionately high number of reviews. Four additional predictor variables were created from a combination of sources. Population per square mile was calculated using U.S. Census data. Review timing refers to how the businesses accumulate reviews over the time of the business's existence. The variable 'open' indicates the number of days a business has been open, and 'rate' refers to the average number of reviews a business receives per day.

Prediction Modeling

We attempted to predict the likelihood that a business would receive more than 100 reviews on Yelp with two statistical models: the recursive partitioning and regression trees (rpart) package and the binomial family of models within R's glm function (logistic regression models). We incorporated a selection of both scaled and binary predictors, predicting the most effective independent variables from our exploratory analysis (see previous section). Our models were only successful, however, when they incorporated the number of reviews generated per day for each business.

Results

The results of five separate logistic regression models are shown in Figure 2. We trained the first two models with 80% of the business data, treating categorical and noncategorical predictors separately. However, neither of these models were able to predict the presence of any frequently (>100) reviewed businesses. We tried two additional models with biased training data: our third model was trained with a higher proportion of businesses with >100 reviews, and the fourth model was trained with a higher proportion of category designations. The precision of these models only improved to 5%. Incorporating the number of reviews generated per day improved the precision of our model to 79%, although the model was able to locate less than half of the restaurants in our test set with more than 100 reviews (217 of 465 for a sensitivity of 47%). Four decision tree models were created with R's rpart package based on a mix of predictor variables. The models varied by including or excluding the business attributes (price range, romantic, touristy, etc.) in our predictor list and the frequency of reviews per day (rate).



Predicting businesses with more than 100 reviews: Results of binomial modelling

	Sample Size (n)					
	training data	testing data	Predictors	Accuracy	Precision	Sensitivity
1	19294	7467	poulation density, price range, skewness, star count	93%	0%	0%
2	29867	7467	Category Designations: restaurants, nightlife, bars, New American, Mexican, Beauty/Spas, shopping, Active Life, Fast Food, Traditional American	93%	0%	0%
3	5119 *	1959	star count, price range, skewness	88%	5%	6%
4	15147 **	4208	All category designations from test #2; price range, skewness, metro area	93%	5%	0%
5	19550	7467	price range, skewness, **rate of reviews (number per day)**	96%	79%	47%

Notes

* Training set was biased to include more frequently rated businesses.

** Training set to was biased to include more businesses with pertinent category designations.

Figure 2: Evaluating logistic regression models.

The most successful rpart model included our entire set of non-categorical predictor variables: price range, skewness and frequency, star rating, metropolitan area, and nine ambience attributes. Using three ambiance attributes (casual, divey, hipster) as the most significant variables in tree construction, this model was able to correctly predict 1,596 of the 1,981 businesses with over 100 reviews, a sensitivity of 80%. However, when plotting this tree, we noticed that a large percentage (around 70%) of businesses were easily filtered from the first decision branch which hinged on the rate of review to be less than 0.044 per day (or about 16 reviews per year). Re-running our model but removing our rate variable gave only 807 out of 1,981 (40%) correct predictions that a business will get over 100 reviews.

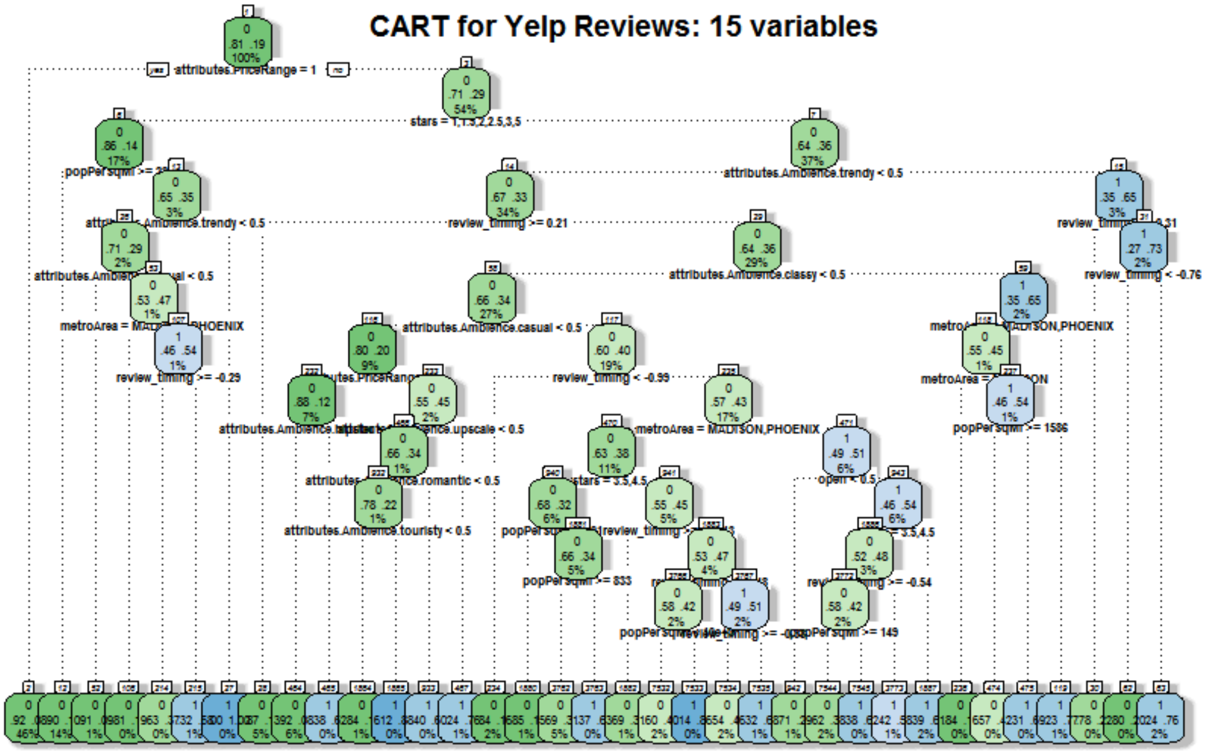

CART for Yelp Reviews: 15 variables
yes
attributes.PriceRange = 1
no
stars = 1,1.5,2,2.5,3,3.5
attributes.Ambience.trendy < 0.5
review_timing >= 0.21
attributes.Ambience.classy < 0.5
attributes.Ambience.casual < 0.5
review_timing < -0.99
metroArea = MADISON,PHOENIX
review_timing < -0.76
review_timing >= -0.29
attributes.Ambience.upscale < 0.5
attributes.Ambience.romantic < 0.5
attributes.Ambience.touristy < 0.5
popPersqmi >= 833
popPersqmi >= 1586
popPersqmi >= 149
review_timing >= -0.54
open < 0.5

Figure 3: Decision Tree Model

Discussion

While many of our predictors – category, population density, and skewness, for example – showed some promise in the exploratory phase, we were only able to successfully classify businesses as highly reviewed when the number of reviews generated per day was considered. This supports the results of Hood, Hwang, and King in the first Yelp dataset challenge, where indicators such as attention time and the maximum number of reviews in a day most accurately predicted the total number of reviews a business would receive.

The rate of reviews can be a useful predictor of emerging "hot spots". A cursory survey of frequently reviewed businesses in Las Vegas shows an exponential increase in review counts between July and December 2014. However, temporal indicators only show predictive power for a limited set of businesses that are "highvisibility." An ideal predictive model would find emerging businesses that open to little fanfare, but develop into prominent fixtures in their communities due to the excellence of their products and services.

While we have not yet found the optimal mix of variables that will predict these less obvious "up and coming businesses," we have identified some future directions for this research which could improve the sensitivity and precision of our model. Both sentiment and topic analysis of initial reviews of a business could highlight unique characteristics of those likely to receive more than 100 reviews. Local traffic and tourism data would improve the model, as well. We are particularly interested in stratifying business characteristics based on the level of tourism that a geographic area receives. In Las Vegas, for example, tourism accounts for roughly 20% of local GDP. The effects of tourism need to be considered more closely when evaluating the way that reviews are generated for individual businesses.

Conclusion

A variety of variables were taken into account in our attempts to predict successful businesses, defined as those having 100 or more reviews on Yelp. However, when excluding the predictor variable of rate of reviews generated per day, our generalized linear models and recursive partitioning and regression tree models proved unsuccessful in predicting the presence of frequently reviewed businesses with high precision. Further research is required to determine more effective predictor variables in order to identify emerging businesses new to Yelp or those aiming to accumulate a high review count. Sentiment and topic analysis of user reviews as well as merging Yelp's dataset with local traffic and tourism data might also be considered in formulating a better predictor model.