**Business Analysis of pubs in Wisconsin**

Yinqiu Xu, Xiaofeng Wang, Mi

**Introduction**

Data analysis could be a really powerful tool to extract information and draw conclusions based on scientific methods. Business owners usually try to improve their operation by analyzing their financial statements or consulting customers. Benefited by development of the internet, they upload their information to websites to expose more. Yelp could be helpful in promotion while the data could be informative as well. In this project we analyze the reviews customers give and other attributes of a restaurant, trying to extract useful information and give suggestions to businesses. Our analysis focuses on all pubs that have a full bar in Wisconsin. Among these pubs our specific goals are to analyze the most important hardware facilities and services that customers prefers and then provide suggestions to pubs with low ratings. Rating of a business is measured by the stars score of a business. We hope to answer the following several questions:

1. What facilities will affect businesses’ rating? How will they influence?
2. Will opening hour influence their rating?
3. What extra services could be beneficial to pubs rating? For example, business-parking, valet, garage?

Yelp is an Internet company that provide a platform for users to write reviews of businesses. We use a real dataset from Yelp and conduct some exploratory data analysis. A multiple regression model is built by stepwise selection to predict the rating of a business. In the following sections, we will show details about our analysis and model.

**Data preprocessing**

1. **Data that we use and sample size:**

Yelp receives a very large amount of data every day. Yelp dataset is released by Yelp encouraging students to do research on it. All the data are stored in four json files which contains information about business, review, users, and tips. We mainly used stars, open hours and facilities data in business json file, and content of review in review json file.

After filtering all the open restaurants with a full bar in Wisconsin, we got 466 pubs with 69 attributes and its stars, 50569 reviews with its content, and corresponding business id.

1. **Clean attributes variables of business**

To obtain attributes of each pub, we separate BusinessParking, GoodForMeal, and Ambience into several binary variables. Some redundant characters are deleted from levels of attributes factor. For example, “u’free’” is the same as “free” in attributes Wifi. We deleted the redundant “u’”. Missing values of both nominal and ordinal variables are interpolated by their mode.

1. **Create new predictors**

Opening hours of each day are new predictors that calculated by hours.Monday-hours.Sunday attributes.

To get the information about reviews’ sentiment of each pub, we create a new predictor called positive review ratio by dividing the number of positive reviews by the number of reviews of each pub. Sentiment analysis is used to evaluate the positiveness and negativeness of a review. First, we parsed each review into single words and deleted stopwords like he, she, the. Then, we count the number of positive words and negative words in each review. Finally, we calculate the sentiment of each review by substracting the number of negative words from the number of positive words. The sentiments of each word are defined and stored in a lexicon which is provided by the tidytext library in R. After conducting sentiment analysis of each review, we group them by business id to get the number of positive reviews. Review data and business data are merged by business id column.

**EDA/ statistics tests:**

1. **Plots showing proportion of attributes by rating of business**

Note that most of attributes are discrete. We plot the histogram of businesses’ stars under each level of nominal variable.

图表, 直方图

描述已自动生成图表, 直方图

描述已自动生成

Figure1: Delivery Figure2: TakeOut

Depicted by the Figure1, businesses offering delivery service are mostly rated between 2.5 and 3.0 and businesses without delivery service rated high. This indicates that delivery service may not make a big improvement of business rating.

By the Figure2, businesses offering takeout service are mostly rated between 2.5 to 4.5 and most of businesses without takeout services rated very high. This indicates that takeout service could not be a competitive power of a business. Businesses without takeout services can still have high rating.

图表, 直方图

描述已自动生成图表, 直方图

描述已自动生成

Figure3:GoodForGroups Figure4: HasTV

Depicted by the Figure3, businesses that are good for groups rated close to uniformly while

Depitcted by the Figure4, more businesses with TV have lower ratings and more business without TV have greater ratings. This may indicate a negative influence of placing TV in a pub.

1. **Plots showing frequency of words by rating of business**

**图表, 条形图, 直方图

描述已自动生成图表, 条形图, 直方图

描述已自动生成图表, 条形图, 直方图

描述已自动生成图表

描述已自动生成**

1. Distribution of ratings; mean ratings; mean length of reviews

图表, 直方图

描述已自动生成图表, 直方图

描述已自动生成

* + - EDA plots of nouns, and adjectives
    - see if nouns like atmosphere are in top 20/30
    - see if abjectives like friendly, fresh, busy are in top 20/30
    - Tests to draw some conclusions that can answer our questions

**Analysis of pubs market**

We construct a binary variable stars\_level which equals 1 if a business’s star is greater than its median and equals 0 if a business’s star is less than its median.

conduct chi square test of test the independency between attributes and the … variable.

A multiple regression:

Output: stars

Explanatory variables: some key attributes of business(discrete) + opening hours + positive review ratio

**Recommendations for business:**

According to our analysis above.

**Limitations:**

Since we use bing, nrc, Afinn sentiment lexicons which are unigram based, our sentiment analysis cannot identify sarcasm and negated text.

A text the size of many paragraphs can often have positive and negative sentiment averaged out to about zero, while sentence-sized or paragraph-sized text often works better.

**Introduction:**

1. Will the atmosphere, music, light and temperature affect their ratings? If so, what kind of atmosphere are good for their business?
2. Do different kinds of alcohol drinks and brands of drinks affect rating? If so, what kind of drink or brand of drink do customer prefer?

**Data preprocessing:**

Discuss any relevant aspects about the data

Process of data cleaning

* 1. Using R to handle the following cases
     + n’t to not; ‘d to would
     + he’s to he is, she’s to she is , i’m to i am，it’s to it is
     + change to lower case

output is review\_text.json file

* 1. using R to filter all pubs in Wisconsin by the following steps
     + business json file, filter with attribute alcohol==”full\_bar”and is.open==TRUE
     + retain all the reviews of pubs
     + replace reviews json file’s text with cleaned review\_text
     + retain all the tips of pubs
     + retain all the users of pubs
  2. tokenization
     + using tidyverse and tidytext library to tokenize review\_text
     + get the most frequent nouns and abjectives in review\_text

select the top 100 or top 0.01 nouns and adjectives

* 1. repeat steps 1~3 to tips text
  2. Sentiment analysis to create some new predictors
     + Analysis the sentiment of a review text, positive, neutral, or negative
     + Analysis the sentiment of a tip text, positive, neutral, or negative
     + Get sentiment grade of each review and each tip