Predicting Restaurants Rating

Cristián Vildósola Michell

June 2020

1 Background

There are thousands of restaurants across the United States, if we observe the industry of dinning out, we can realize that total number of restaurants in the United States in 2018 were 660.755, and it keep increasing at 1-2% rate, also "The National Restaurant Association projects overall industry sales will hit a high of $863 billion in 2019, up 3.6% from last year." Based in the channel CNBC, the number of restaurants that fails each year its about 60%, and nearly 80% close before their 5 year anniversary. Also, in this news channel, the most common cause of failure its location and user reactions in social platforms.

2 Problem

In this project we are going to use machine learning algorithms with the intention of predict the rating of a restaurant based in some variables we are going to get from Yelp. The goal of this, its identifies where its more possible a business to succeed or if its more important the reviews that people make in order to evaluate the food and the place.

3 Methodology

For the realization of this project, we are going to assume that a business succeed depends of the rating of the users, also we are going to evaluate only the places that have a minimum amount of reviews.

The first part of the project it's connecting to Yelp API, then we make series of request in order to get data from the state of California, then we proceed to clean this data and make a exploratory analysis of it. For most of the analysis, we are going to first construct an unique model for the whole state, and then we are going to explore every city by itself. The first analysis will be a Geo referenced map of each restaurant. Then we are going to try to identifies clusters outsides the location, and then for each city.

In the last part of the project, we are going to predict a value between 1 and 5 for a new group of restaurants, first we are going to use our state model and then our city model. From this we are going to generate conclusion about data quality and the importance of how to compare elements inside a data set.

4 Sources and description of the data

In order to obtain the data, we used the API offered by Yelp, specifically the business search. Then we proceed to make a few requests and pass the data to pandas data frame. The fields acquired where: business_id, name, address, city, state, postal_code, latitude, longitude, stars, review_count, open, categories of food and a series of attributes. We are going to mainly use the locations coordinates, state, review_count and price.

5 Reference

1. **CNBC - Why restaurants fails?**

   https://www.cnbc.com/2016/01/20/heres-the-real-reason-why-most-restaurants-fail.html

2. **CNBC - Budget towars eating out**

   https://www.cnbc.com/2019/08/19/americans-putting-more-of-their-budget-toward-eating-out.html

3. **Data Acquisition**

   https://www.yelp.com/dataset/

4. **Troubleshooting**

   https://github.com/Yelp/yelp-fusion/issues/307