Predicting Success of Restaurants In Las Vegas

Viet Vo, Sang Goo Kang
vtvo@stanford.edu, sanggookang@stanford.edu
Mentor: Nihit Desai

Motivation

The well-being of many businesses today heavily rely on the positive ratings given by their customers. With the founding of Yelp in 2004, the relationship between businesses and their customers has become more dynamic. In this project, we studied the success of restaurants in Las Vegas by predicting the star ratings of restaurants and finding the most useful traits in determining their success.

Data and Features

Data

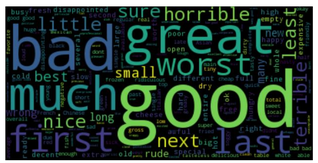
The Yelp Dataset Challenge was used for the project.

- JSON business file: location, restaurant category, and opening and closing hours
- JSON reviews file: customer reviews
- 6,764 restaurants from Las Vegas w/ star rating labels
  a) 5,764 for training set
  b) 1,000 for test set

Features

106 Restaurant Characteristics	200 Adjective Unigrams	150 Bigrams w/ Adjective
Price Range	“great”	“the best”
Review Count	“delicious”	“my favorite”
Categories (Food Truck, Japanese, Seafood, etc.)	“bad”	“the worst”
Drive-Through	“attentive”	“nothing special”
…	“horrible”	“too bad”
	…	…

Unigram 5-star review adjectives



Unigram 1-star review adjectives

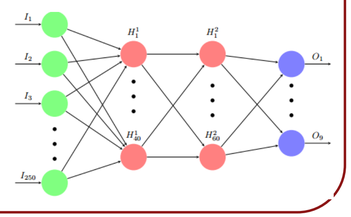
Models

Classification

- Perceptron Neural Networks
- SVM

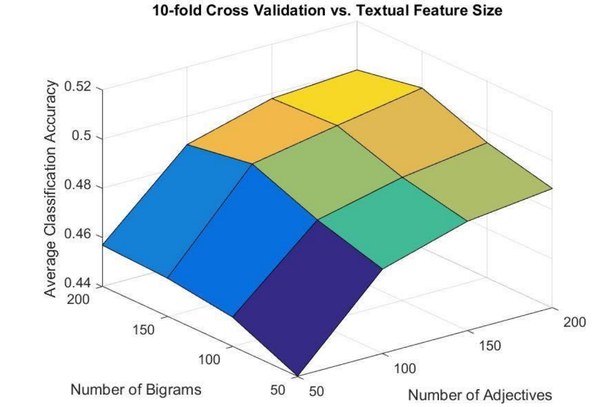
Regression

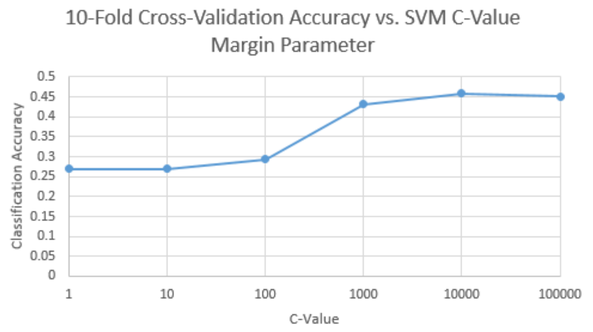
- Linear Regression
- SVR



Results and Analysis

- Business star ratings can be predicted using regression or as a 9-class classification problem since ratings range between (1, 1.5, …5)





Predicting Star Ratings using Textual Unigram + Bigram Features

Feature: Unigrams + Bigrams	Train Classification Error	Train RMS Error	Test Classification Error	Test RMS Error
Linear Regression*	0.251344786	0.31993053	0.594	0.563249501
SVR*	**0.247961131**	**0.332725456**	**0.597**	**0.538052042**
SVM	0.174041298	0.281022463	0.583	0.562138773
Neural Networks	**0.154606976**	**0.232256881**	**0.521**	**0.521296461**

Predicting Star Ratings using Restaurant Characteristics

Restaurant Characteristics	Training Classification Error	Training RMS Error	Test Classification Error	Test RMS Error
Linear Regression*	**0.339927121**	**0.469367941**	**0.814**	**0.839642781**
SVR*	0.314766615	0.42980753	0.800	0.844393273
SVM	0.257331251	0.478498287	0.814	0.856446145
Neural Networks	**0.307305223**	**0.473737672**	**0.768**	**0.849117189**

*We rounded regression results to nearest 0.5 to calculate the classification accuracy.

Discussion & Future Steps

- Using unigrams and bigrams that contain adjectives provide better results for both classification and regression than restaurant characteristics
- Neural networks was the best algorithm for classification, while linear regression and SVR were the best for regression

- In the future, we can try other techniques like regression trees and random forest classification
- Look at the number of upvotes in a given review
- Take into account how influential a reviewer is in the Yelp community

References

[1] Public data: http://www.yelp.com/dataset_challenge
[2] Fan Mingming, Khademi Maryam, “Predicting a Business’ Star in Yelp from Its Reviews’ Text Alone.” University of North Carolina, Charlotte, University of California, Irvine, 2014.
[3] Farhan Wael, “Predicting Yelp Restaurant Reviews.” *USCD, 2015.*
[4] Yu Mengqi, Xue Meng, Ouyang Wenjia. “Restaurants Review Star Prediction.” *UCSD, 2015.*