Predict Yelp Ratings Using Machine Learning*

1st Kai Luo
Department of Computer Science
Unniversity of Virginia
Charlottesville, VA USA
kl3pq@virginia.edu

2nd Jinpin Lin
Department of Computer Science
University of Virginia
Charlottesville, VA USA
jl4sh@virginia.edu

3rd Guan Lingjun
Department of Computer Science
University of Virginia
Charlottesville, VA, USA
lg3bv@virginia.edu

I. Abstract

Food delivery is a popular service and its industry is prosperous nowadays. Companies like Yelp have constructed some well-structured platform for people to order food and have it delivered. [1] Food restaurants are ranked depending on many features like reviews and rating. A high rating restaurant often indicates its popularity. Basing on its location, food kinds, facilities and other features, our model can predict the rating of a newly open restaurant or help an old restaurant to improve. By doing so, we help people make better strategies in their restaurant business.

II. Motivation

Our goal is to create a star rating prediction model that can provide accurate rating estimation on the restaurants. Suppose people want to open a new restaurant in VA, they would like to know if the restaurant will be a success. Therefore, our model is going to help them make decisions. Given the information of the restaurant, we are able to make predictions on its probable rating points. Moreover, our model also serves as a simulator. That is, if the restaurant owner wants to improve the situation, he or she can change the features of the restaurant and run the model to get better results. That will serve as a reference to restaurant modification.

III. Method

A. Data collection

In the proposal, we mentioned that the data set from Yelp dataset challenge (https://www.yelp.com/dataset/challenge) is available, however the data set contains only one restaurant in Virginia and most of the restaurants locates in other states. So, we built a tool to scrap the data from its website and get our own unique dataset. The tool is flexible and stable. You can collect any kind of data from any city as long as the information shows on the website.

We have collected about 7256 VA restaurants' information. Their location includes all the independent cities and part of the counties in VA.(about 450 restaurants from Charlottesville). Each restaurant contain about 30 or more features, which provides more information than Yelp's original data set. Each restaurant's information had the following format:

```
{
    "Name": "Cafe Kindred"
    "Ratings by Yelp" : 4.0,
    "Ratings by Calculation" : 4.173,
    "Number Of Reviewers" : 364,
    "Price" : "$",
    "Price Range" : "Under $10",
    "Category" : "American(Traditional)
    Breakfast&Brunch,CocktailBars",
    "Address" : "450 N Washington
    StSte FFalls Church, VA 22046",
    "Phone" : "(571) 327-2215",
    "Website" : "cafekindred.com",
    "Open Hours" : "Mon Closed
    Tue 7:00 am - 3:00 pm ...",
    "Business Info" : "Liked by
    Vegetarians,Good For Dancing,
    Smoking..."
}
```

Listing 1: Restaurant Information Example

Yelp calculated the restaurants' ratings by averaging all reviewers' rating and replace them by the nearest 0.5 level, we calculate the rating by ourselves in order to get a more precise label. The attribute called ”Rating by Calculation” did improve our model's accuracy. Each restaurant has about 25 different attributes in ”Business Info” and it varies with different restaurants, which provides rich information.

We build model and write code based on the useful software named Octoparse (https://www.octoparse.com/). It runs slow in order not to be detected by Yelp. If you request the web pages too often, Yelp may record your IP and ban any request from it. It takes about 8 hours to collect 1000 restaurants' information.

In order to speed up, we open 5 EC2 instances to run the software at the same time. After getting the data in xlsx format(the software can also output json format), we also wrote a macro in Excel to clean the data(like deleting the useless information or combing the same features.)

B. Data cleaning

We eventually get a data set of around 140000 rows in a form of CSV file eventually after crawling down data from the Yelp website. However, due to the restriction of the crawling software, we need to merge the data of the same restaurant. After the merging process, we finally have around 7200 restaurants. After that, we dropped all the null data.

C. Feature expanding

As we are considering which feature is important, we first notice some features that are definitely unrelated to the star rating, which are ”Website” and ”Phone”. So we drop them. We’ve read some related papers and they claim that the ”Category” has very little to do with the rating. Besides, this attribute is hard to be transformed to some usable values. So we decide not to consider it. Then we consider to transform the ”Address” to some usable data, like the coordinate of the restaurant. Therefore we use Google Map API to fetch the coordinates and drop the original address. After that, we try to make the best use of the ”Openhour” attribute, and we decide to expand it. We disintegrate it into three parts: First part is the total open hour of a day; Second part is the open time of that day; Third part is the close time of that day;(If the restaurant is closed at that day, it’s marked as closed, so is open all day). And we have these features for every day in a week, which means we have 21 new features totally. Then we take a look at the ”BusinessInfo”. For convenience, we only consider the binary class in it, such as ”WiFi: Yes/No”. And when a restaurant doesn’t have that particular feature, we set it to ”No”.

D. Training and Predicting

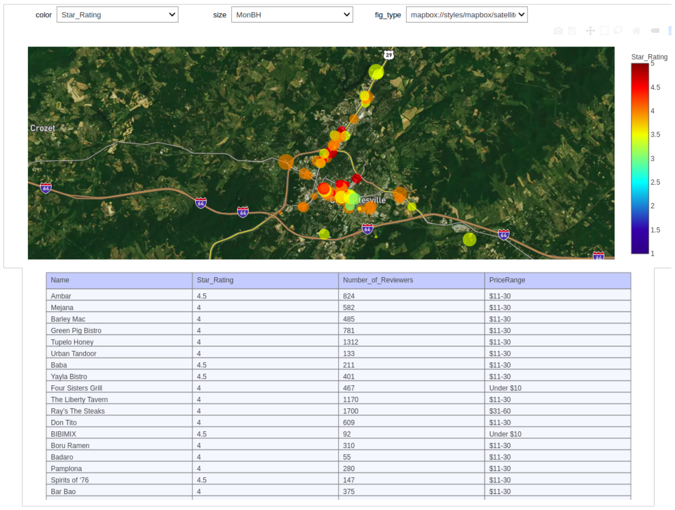
We set the ratio between train set and test set to 4:1. Since we don’t want the separation to be biased, we use stratified split to make sure the two sets have almost the same ratio of samples of a certain rating value. We select 6 regression models, and the results are listed in the table below. After we’ve finished regression task. We want to know about the results of classification too. So we transform the rating back to the original form and label it. The results are listed in the table below.

E. Data visualization and interpretation

Geographic Information Technologies (GIT) is applied for the data visualization, and Principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are employed for the interpretation. More data has been collected with much more features since the checkpoint. The data contains an array of 48 features. Apart from the name of the restaurant (string), all other 47 features are numerics (all categorical features has been transformed to numerical features with the OneHotEncoder). A sample of 2159 restaurants at 8 cities (Virginia beach, Richmond, Charlottesville, Alexandria, Fairfax, Ashbum, Leesburg and Dumfries) in Virginia is adopted for the sake of running time, which can be enlarged to the whole dataset.

For the data presentation, each restaurant is represented as a circle with different size and color in the Google map. Fig. 1 shows a sample representation of the restaurants in Charlottesville, where the color represents the star ratings, and the circle area represents the Monday business hours. Different map styles can also be selected (street, satellite, etc.). The detailed information of selected restaurants will also be displayed in the table below the plot, which can help with exploring localized relation between restaurants and the difference between various districts. The figure can be used to provide intuitive relation between different features together with the geographical information. It is seen that most restaurants have a star rating at of 3.5 to 5.0, and the restaurants with longer opening hours on Monday doesn’t necessarily mean the restaurant will get higher ranks. The restaurant with the highest ranks are in the downtown of Charlottesville and along the Route 29.

PCA and t-SNE are used for data interpretation. The result of PCA with 3 components for the features shows the explained variance ratio as [9.93971946e-01, 3.54340243e-03, 9.38501702e-04], which means there is only one main component in the selected features. A plot of t-SNE with 3 components and perplexity of 20 is shown as below in Fig. 2. The number of iterations is set to be 20000 for convergence. In the obtained 3D space, it is observed that the restaurants are clustered along a non-linear 3D band. The restaurants with highest ranks are mostly narrowly clustered at one end of the band, while most ones with lower ranks are sparsely distributed at the other end. In the middle is a curved band with a complicated shape.


color: Star_Rating
size: MonBH
fig_type: mapbox://styles/mapbox/satellit

Name	Star_Rating	Number_of_Reviewers	PriceRange
Ambar	4.5	824	$11-30
Mejana	4	582	$11-30
Barley Mac	4	485	$11-30
Green Pig Bistro	4	781	$11-30
Tupelo Honey	4	1312	$11-30
Urban Tandoor	4	133	$11-30
Baba	4.5	211	$11-30
Yayla Bistro	4.5	401	$11-30
Four Sisters Grill	4	467	Under $10
The Liberty Tavern	4	1170	$11-30
Ray's The Steaks	4	1700	$31-60
Don Tito	4	609	$11-30
BIBIMIX	4.5	92	Under $10
Boru Ramen	4	310	$11-30
Badaro	4	55	$11-30
Pamplona	4	280	$11-30
Spirits of ‘76	4.5	147	$11-30
Bar Bao	4	375	$11-30

Figure 1. Geographic plot for restaurants in Charlottesville

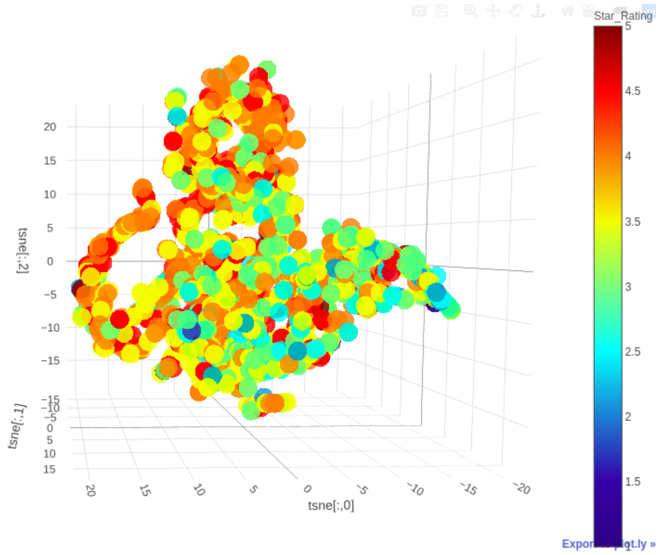


Figure 2. t-SNE plot for the sample data set

IV. Preliminary Experiments

We test the two model mentioned above with the ratio between train set and test set being 4:1. Since we don't want the separation to be biased, we use stratified split to make sure the two sets have almost the same ratio of samples of a certain rating value. Although the range of the rating is between 2.0 to 5.0 and has the interval of 0.5, we perform regression other than classification to be more precise. We use two different regressors–SVR and RandomForestRegressor and measure the results by mean squared error. It turns out that these two models perform similarly with a relatively high error of about 0.46, which indicates that our model is underfitting.

V. DISCUSSION AND RESULTS

1.feature selection to give advice.
We've found that the most important features that affect the rating results are: Whether it's liked by Vegetarians; Whether it closes on Monday; Whether it has wheelchair access; Whether it has gender neutral restrooms; Whether it's liked by vegans; Whether it has waiter service.

2. show and compare our result with Stanford
We have done both classification and regression using many different models. The Table I has shown the RMSE with different models. As we can see, the Adaboost and Random Forest has the best performance. The RMSE within 0.40 gives a good prediction. The regression result is shown in Table II.

Since Yelp's dataset challenge has draw many people's attention, Standford students have done the similar work based on this dataset. In the Kyle Carbon's paper[], they floored the business' rating to integer, thus make "improvement" to their training accuracy and test accuracy.
Our models have higher accuracy because we have more features by collecting our own dataset, which provides more information than the Yelp dataset.

Model	**Hyperparameter**	RMSE
RandomForest	max_features=200,n_estimators=1000	0.40
Adaboost	base_estimator=rbf,n_estimators=100	0.398
SVR	kernel rbf	0.458
Decision Tree	default	0.569
Gradient Boosting	n_estimator==200	0.463
Bagging	default	0.441

Table I
REGRESSION RESULT WITH DIFFERENT MODEL

Model	**Our Test Accuracy**	**Their Test Accuracy**
SVC	0.44	0.44
RandomForest	0.49	043
Decision Tree	0.52	0.43
AdaBoost	0.60	None
MLP	0.54	None
Gradient Boosting	0.45	None
Bagging	0.58	None
Voting	0.55	None

Table II
CLASSIFICATION RESULT WITH DIFFERENT MODEL

VI. Future work

The prediction accuracy can be further improved with the following methods:

1.Collecting more data There are 95 counties in VA, right now we have just collect data from 20 counties in it. More data will help with achieving better models.

2. Obtaining more appropriate features Since our model is under-fitting and we are clear about the reason(we are not using the most important features of "category" and "business info"). The next step we are going to take is to generate those features and use more models. Basically, we are going to use the seven models that we mentioned above and use AdaBoost or stacking to boost the performance. Moreover, we are going to take into consideration of the distribution of the rating throughout the whole dataset. It seems that most of the restaurants have the rating between 3.5 and 4.5 inclusively. This might affect the prediction results and need to be paid attention to. Last but not least, we will use cross validation and grid search to fine-tune the models.

3.Using more prediction model We just applied some simple models right now, we'll use 7 other models[1] to train the data and find the best one.

4. Using more data and models for visualization and dimension reduction Future work for visualization includes generating the geographic plots with the whole dataset and more features, changing parameters (perplexity, etc.) in t-SNE for better visualization of high dimensional data, adopting more methods to find the most important components in the data for dimension reduction, etc.

VII. Contribution

The contributions of team members are as follows: Kai Luo: Data collection and data cleaning.
Lingjun Guan: Data processing, model building and evaluation
Jinpin Lin: Data visualization and interpretation

References

[1] Gingerich, Travis, and Yevhen Bochkov. ”Predicting Business Ratings on Yelp.” Stanford University. 2015.