# *Prediction of Weekend Check-in Ratio Based on Yelp Dataset*

Wentao Guan
University of California,
San Diego

Yan Huang
University of California,
San Diego

Qianchen Zhao
University of California,
San Diego

## Abstract:

Yelp contains a lot of data from the businesses and users. A lot of people use it to decide which business service is the best to provide. In this report, we investigate the data provided by the Yelp Dataset Challenge 9. We analysis the business information based on the business dataset and the check-in information based on the check-in dataset and explore the relationship between the business and ratio of weekends check-in ratio the whole week with linear regression and k mean cluster.

## Introduction

During the past years, there is a quick development in the area of machine learning, many models come out and can be used for different tasks. There is no doubt that, machine learning provides us a better way to figure out the latent relationship between given features and the aim to predict. Meanwhile, it is reasonable to believe that, a good predicting model can help people make better decision and maximize their benefits.

It is well-known that, Yelp is a very popular and helpful website, people often use yelp to decide which business is the best choice. A lot of statistics and analysis has been done to predict the rating based on the user and business information. However, there are much fewer works that have been done to help with businesses instead of users. Some of these information can be extremely helpful for business owners. In this report, we are building and illustrating models for weekend/all-week check-in number ratio predicting, which can be helpful for some certain business to balance their workload and realize a better management. During the following work, we start from an intuitive baseline, and try to improve the model by choosing more reasonable features and divide the ratio prediction into several processes and merge these models. According to the result, significant improvement is realized and the MSE is squeezed 12.12% from 0.05297 to 0.04655.

## Dataset Characteristics

We use the dataset from the yelp dataset challenge round 9. It contains five specific datasets, including business, review, user, check-in and tip. The two dataset that we use include the business dataset and check-in dataset, which contains business information for 144k local business and aggregated check-ins over time for each of the 125k businesses. The business information includes the location, rating, review count, attributes, and categories and open hours. In the location, it contains the city, state, postal code, longitude and latitude. The check-in data is in the format “Day-Hour: check-in number”. For example:” Mon-0: 20” means that 20 people check in between 0 o’clock and 1 o’clock on Monday. All information is stored in JSON file format as the following:

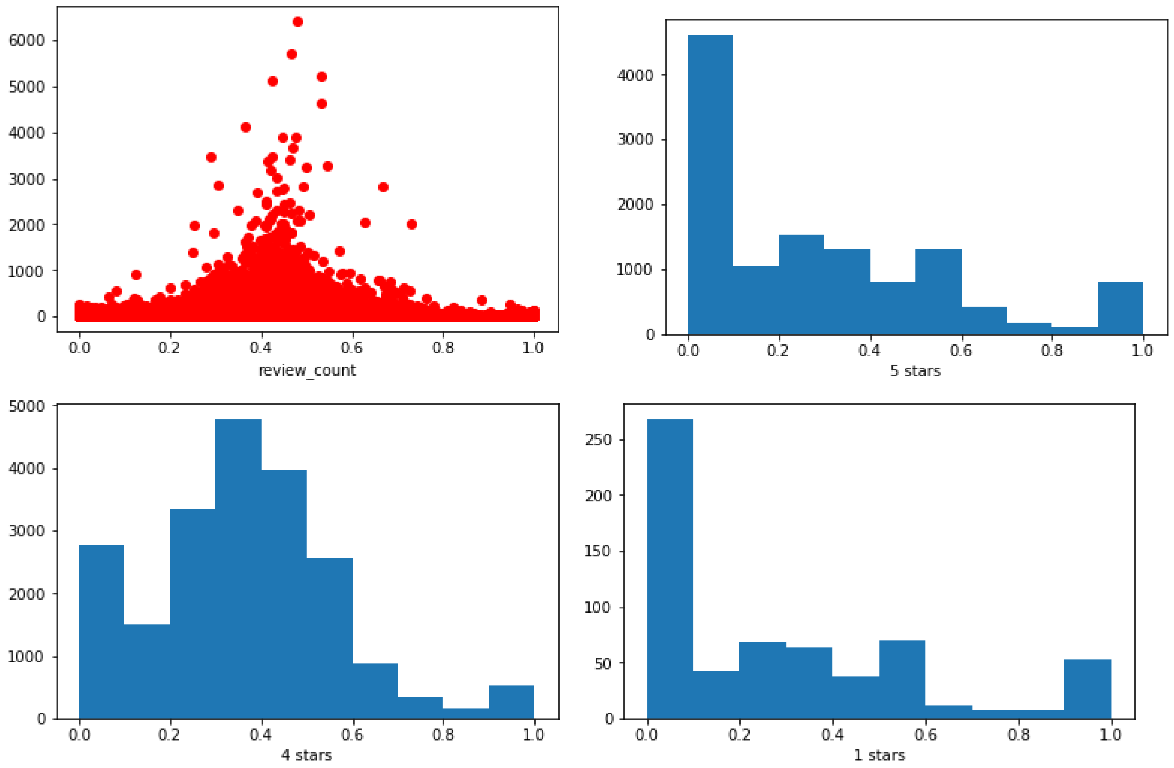

*Figure 1Distribution of weekend's ratio vs review_count and star*

yelp_academic_dataset_business.json

```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
"postal code":"postal code",
"latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```

yelp_academic_dataset_checkin.json

```
{
    "time":["an array of check ins with the format day-hour:number of check ins from hour to hour+1"],
    "business_id":"encrypted business id",
    "type":"check-in"
}
```

## Dataset Analysis

In order to extract useful information, we visualize some of possible important features and analyze them in this section.
According to figure 1, feature “stars” seems to have a very significant influence upon the ratio, and “review_count” can be helpful as well. As shown in the figure 1, businesses with review count above and below 30 have different weekend check-in ratio, it is reasonable to believe that, the feature could be chosen as “if the review_count greater than 30” and “review_count”, the training result on the validation set proved our thinking.

The dataset contains 144K number of businesses with more than 1000 categories. The businesses are from 29 states. There are total 32 attributes a business may have, and there are about 1171 categories that a business might belong to. A business may belong to multiple categories. Since there are too many attributes and categories, it is better to figure out which attribute and category are more important or occurs more frequently. To study which of these information are more important, we calculate the most frequent number of categories and attributes. The most frequent 10 categories and the corresponding count are: (8133, 'Local Services'), (8554, 'Automotive'), (9087, 'Bars'), (10476, 'Health & Medical'), (10524, 'Nightlife'), (11241, 'Home Services'), (13711, 'Beauty & Spas'), (21189, 'Food'), (22466, 'Shopping'), (48485, 'Restaurants'), and the most 10 frequent attributes and the corresponding count: (43485, 'RestaurantsReservations'), (43844, 'RestaurantsDelivery'), (46706, 'RestaurantsGoodForGroups'), (47657, 'OutdoorSeating'), (50514, 'RestaurantsTakeOut'), (54636, 'GoodForKids'), (66963, 'BikeParking'), (83173, 'BusinessParking'), (90300, 'RestaurantsPriceRange2'), (110105,'BusinessAcceptsCreditCards')
Similarity, location might be another important feature. However, there are too many different kinds of locations to represent by one-hot encoding, and using the magnitude of longitude or latitude does not make sense. Therefore, we decide to explore the distribution of businesses. To achieve this goal, we calculate the average longitude and latitude for the state, and then print it out in figure 2.

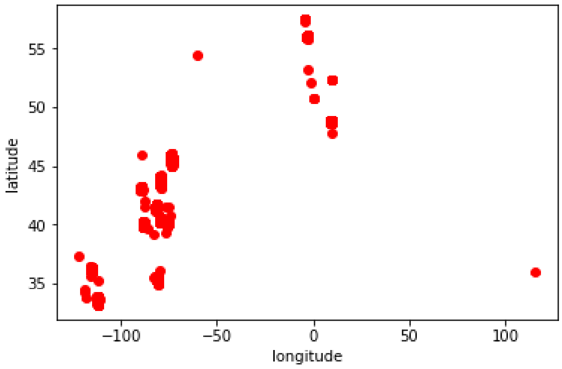

*Figure 2 Business location distribution*

Based on the figure 2, we find that the business are divided into several clusters. This indicates that we could classify the data by their location. We make an overall estimation that there are 5 to 15 clusters from the figure.

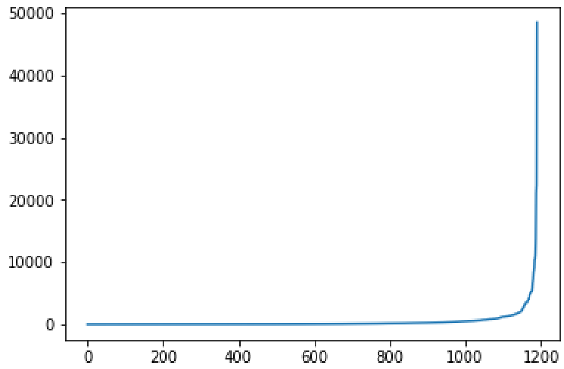

*Figure 3 Categories frequencies*

On the other hand, as the figure 2 shows, as for the category, only the most frequent 7 categories contains 97876 businesses out of 144k, and each other category contains less than 1000 businesses. In order to avoid the disturbance caused by the categories which contain only a small number of data, we decide to do our prediction based on the data which contains these 7 categories.

| | Restau rants | Shopping | Food | Home Services | Beauty & Spas | Nightlife | Health & Medical |
|---|---|---|---|---|---|---|---|
| Restaurants | 46417 | 283 | 8239 | 21 | 51 | 6203 | 31 |
| Shopping | 283 | 19781 | 1748 | 1187 | 1593 | 223 | 810 |
| Food | 8239 | 1748 | 20236 | 38 | 316 | 1283 | 103 |
| Home Services | 21 | 1187 | 38 | 5918 | 22 | 9 | 20 |
| Beauty & Spas | 51 | 1593 | 316 | 22 | 12011 | 23 | 1478 |
| Nightlife | 6203 | 223 | 1283 | 9 | 23 | 10126 | 8 |
| Health & Medical | 31 | 810 | 103 | 20 | 1478 | 8 | 8823 |

*Table 1 Business number of categories*

| | NFR | HH | BS |
|---|---|---|---|
| NFR | 61990 | 176 | 2125 |
| HH | 176 | 14721 | 3410 |
| BS | 2125 | 3410 | 30199 |

*Table 2 Business number of clusters*

NFR: “Nightlife”, “Food” and “Restaurants”;
HH: “Home Services” and “Health & Medical”;
BS: “Beauty & Spas” and “Shopping”

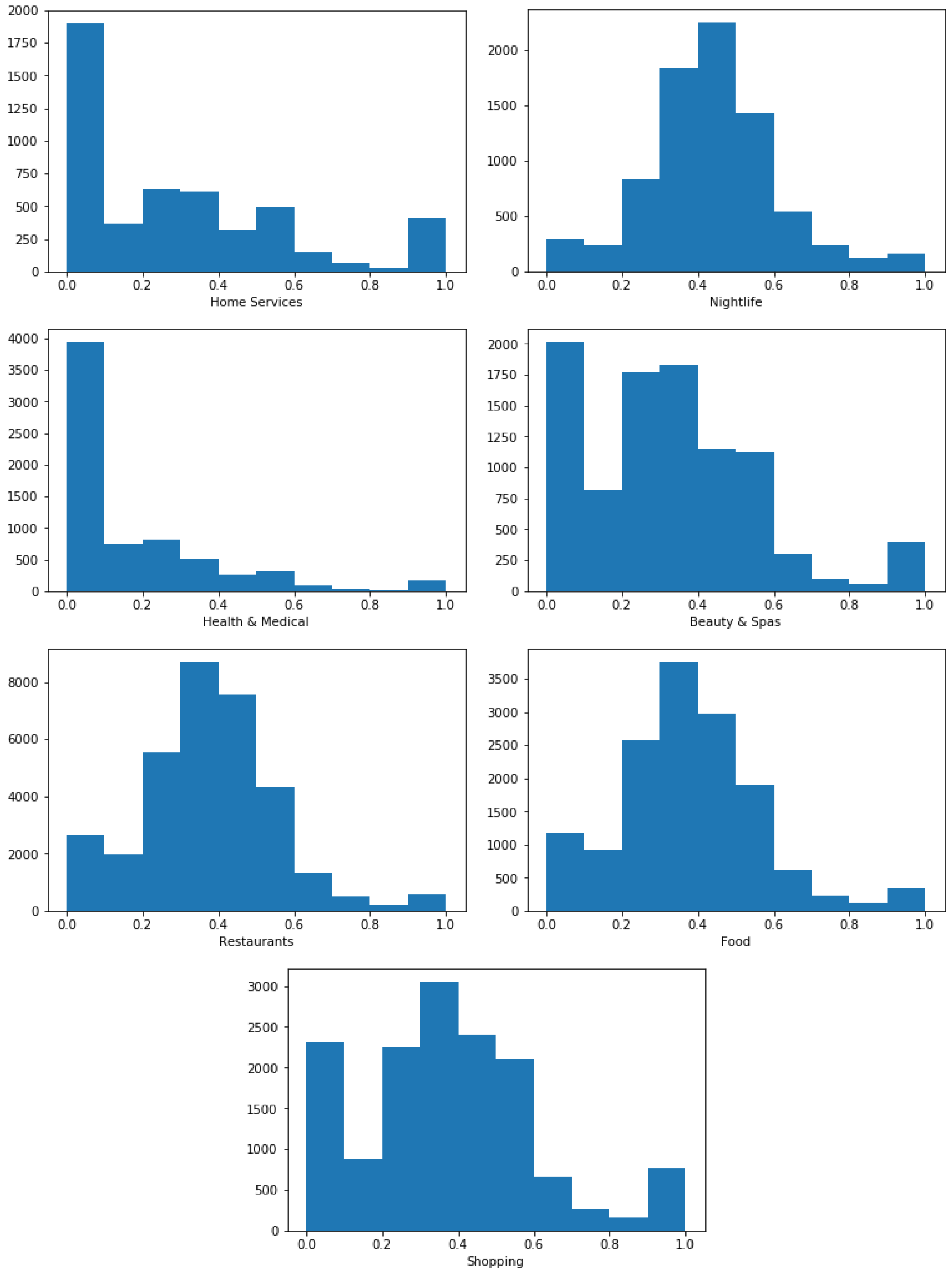

*Figure 4 Weekend ratio of different categories*

Meanwhile, the influence of categories is very interesting. There are over about 1171 categories, and some of the categories can be merged into one cluster. However, most of the categories seem to be useless. There are no clear hierarchy relation between categories; for example, “Restaurants” is not a sub-category under “Food”. Based on the weekend ratio distribution of business with certain categories, we find that, the top 7 most common categories can be merged into 3 clusters, “Home Services and Health & Medical”, “Nightlife, Food and Restaurants”, and “Beauty & Spas and Shopping”. The overlaps between clusters are not significant, so that three different models are trained for different

clusters. When a business belongs to multiple clusters, the most confident prediction is used as the final prediction. According to table 1, there is significant overlap between “Beauty & Spas” and “Health & Medical”, but the distributions of weekend ratio are very different, so it is not a good idea to merge these two categories, and the overlap samples are chosen to put into the “Health & Medical” cluster, based on the prediction confidence.
The attribute open hour should be important when predicting the check-in ratio, but open hour is not accurate in the dataset. We find several businesses which do not have open hour at weekends but still have check-in number. Therefore, it is not a good idea to use this attribute as the feature for our prediction.

## Predictive Task:

Our task is to predict the ratio between the check-in number for weekends and the check-in number of total week. There are two reasons why we choose the ratio to be our target. On the one hand, just as we mentioned in the Dataset Analysis section, some check-in numbers are really large, which may cause large error when we do the prediction. It should be noticed that, the check-in numbers of different business may not be very accurate since the time period is not clearly provided in the dataset. In another word, the dataset is so big and comes from different states and countries. We are not sure if Yelp has the accurate record of all check-in for all the area. However, it is reasonable to believe that, the ratio of weekend check-in and all week check-in is more accurate. On the other hand, the ration can be useful. Based on the prediction, we can provide the business owners suggestions on how to reallocate their recourse during weekends. For example, if the ratio is high for a restaurant, then the business owner may need to employ more waiters or waitresses and prepare more raw material for the weekends opening. If the ratio is low, they may want to employ less people and prepare less raw material. If the ratio is too low for the restaurant, then it is better for the owner to close it during the weekend.

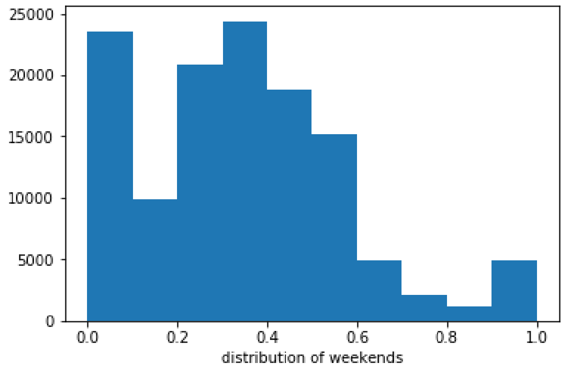

*Figure 5 Distribution of weekend ratio*

We divide the dataset into three parts, training data, validation data and test data, with the proportion 40%, 30%, 30%. We will evaluate the data by calculating the mean square error (MSE) of our prediction result and compare it to our baseline. Our baseline implementation is just calculating the average weekend check-in ratio for all the businesses and use it for prediction.

## Models Selection and Evaluation

### 1. K-mean Classification

Initially, we use states as the feature to represent location, but we did not get very good result. We decide to do the Kmean classification with longitude and latitude for the location. Then we use one-hot coding to achieve binary represent of clusters. For example. If the business belongs to cluster 2, then the third element in the location vector will be set to 1. Others remain 0.

Based on that, we add new features to the baseline. The model is:

$$ratio = \theta_0 + \theta_1[\,location\ vector\,] + \theta_2[\,rating\ stars\,] + \theta_3[\,review\ count\,]$$

**2. Multiple Linear Regression**

In order to get more information, the categories are taken into consideration. To add the categories information, a basic analysis is performed in previous section. From the previous analysis, we decide to use different linear repressor for different kind of business. We build three different linear model for "Home Services and Health & Medical", "Nightlife, Food and Restaurants", and "Beauty & Spas and Shopping" respectively.

**3. Add attributes features:**

According to the distribution, most of the attributes are very rare and several most common attributes are used as features, it should be noted that, some of the attributes share most of the information; for example, 'Noise Level' seems to be helpless if we already added the feature 'Ambience'. Different combination of attributes are used and according to the validation set result. Firstly, top 10 most common attributes are used; however, the correlation analysis shows that, some of the most common attributes have very little correlation with the ratio to predict. Therefore, we calculate the correlation between every attributes and the weekend ratio, and we use the top 10 most related attributes to predict the weekend ratio. In this work, covariance denotes the correlation.

$$Cov(x, y) = \frac{x \cdot y}{|x||y|}$$

**4. Other Models:**

We also tried other models. We tried K Nearest Neighbor Model. Since the dimension of the features and the total number of businesses are both very large. It takes long time to get a result for a certain K. So we finally give up this method.

## Models evaluation:

**1. K-mean Classification**

| K | Training MSE | Validation MSE |
|---|---|---|
| 3 | 0.051174 | 0.052339 |
| 5 | 0.051141 | 0.052303 |
| 7 | 0.051139 | 0.052340 |
| 9 | 0.051141 | 0.052303 |
| 11 | 0.051139 | 0.052301 |
| 13 | 0.051133 | 0.052302 |

*Table 3 K-mean MSE*

K = 11 is chosen because of the lowest validation MSE.

### 2. MSE of Different Models

| | Training MSE | Validation MSE | Test MSE |
|---|---|---|---|
| Baseline | 0.0517711266825 | 0.0529379115626 | 0.0520260085993 |
| K-mean Cluster | 0.0511389700674 | 0.0523013151357 | 0.0514774872786 |
| Categories | 0.04683415822 | 0.0479441364794 | 0.0470656355178 |
| Attributes | 0.0463228180031 | 0.0475619377472 | 0.0465487200801 |

*Table 4 MSE for different models*

The baseline model is very intuitive, and only the average of all the business are used for prediction.

In order to improve the model, “categories” are used to classify the data and separate them into three different linear models. The prediction result is improved a lot, clearly, different types of business tend to have different models.

Then we quantize and add attributes into the model. The attributes are chosen based on frequency and correlation with the target, and the choice based on correlation has a better performance.

In order to get the location information, K-mean cluster method is used upon the longitude and latitude. Compared to the state and postcode, the binary K-mean cluster is the most helpful, it can be easily understand, the distance between the locations contain most of the location information.

The final model can help us get a predicting MSE as small as 0.04655, and almost all the features can help deciding if a given business should pay more attention on the weekend business.

## Related Work:

Yelp public dataset challenge dataset [1] is a very popular public data set which contains local businesses in 11 cities across 4 countries. The dataset is released for Yelp Dataset Challenge 9. There are lots of researchers worked on the Yelp challenge dataset, and there are lots of good papers talking about this dataset. Most of previous papers only make the predictions on restaurants type business. Although some of previous researchers works on different type of business, the common method they use is just simply selecting 3 to 5 most common business types[3][4]. In addition, most of the previous works focus on the rating prediction based on a variety of information of business or review. However, there are not many works about check-in related predictions. One previous works was done by Tyler Daniel [5], who predicts the check-in numbers of all type of businesses based on their location information. He uses a linear regression model to predict the check-in numbers based on which city the business locates. The result has a huge mean square error since check-in numbers varies a lot based on the business type. For example, the check-in number of an airport is in about hundred-thousand level, but a common check-in number of a restaurant is in about several hundred, which can cause a big problem when we calculate the mean square error for the whole dataset. It implies that we need to either use business category as a feature in our model, or build different predictor for different business categories.

Another previous work which is related to check-in information is done by Davy Suvee [6]. He employ the Gephi graph visualization platform for interpreting the identified business communities/clusters. The dataset he used is a subset of the Yelp dataset in Phoenix metropolitan area which includes 11,000 businesses, 8,000 check-in sets, 43,000 users and 230,000 user reviews. He use Pearson Correlation Coefficient to identify the correlation between two businesses to determine whether two business can be classified to one category. His result shows that the business in Phoenix metropolitan area can be category to 8 clusters based on their check-in information, and each cluster is meaningful. For example, one cluster represents breakfast diner’s restaurants, and another cluster represents various department stores like Costco and IKEA. His results shows that business can be classified based on the check-in information, which implies that we can use business to predict check-in information.

## Conclusion:

During this work, we build a linear regression model for weekend/all-week check-in ratio prediction, based on the Yelp dataset. It could be helpful for business owners to manage their schedule and resource. According to the result, the classification based on the categories can help significantly improve the prediction. It can be understand, the different types of business tend to have different models. Actually, merging different models can sometimes be a great direction to improve the predicting performance, however, the merging sometimes can be trivial, since we can just use the limited information again and again. How to choose more important features, how to choose a proper model and how to combine different models will always be the most difficult part of task. From this work, we realize that, the visualization can be very helpful and we should pay more attention before we design and implement our model. Meanwhile, we should start with a very simply model, and a complex model can perform quite bad and hard to analyze.

## Further Work

For further work, review text and customers can be taken into consideration, since review text can contain more information and the hobbit of customers can also help with business type classification, for example, customer who only has time on weekends will help us figure out some of most weekend-busy businesses. Bag of word and n-grams model can be used to process review text, and Latent Factor model can be used for customer and business pairs as a criterion for classification. In addition, many other predictions can be implement based on the given Yelp dataset.

## Reference:

[1] Yelp dataset challenge: https://www.yelp.com/dataset_challenge
[2]"Topic Regularized Matrix Factorization for Review Based Rating Prediction" Jiachen Li, Yan Wang, Xiangyu Sun, Chengliang Lian, and Ming Yao, from the Language Technologies Institute, School of Computer Science, at Carnegie Mellon University.
[3] Naomi Carrillo, Idan Elmaleh, Rheanna Gallego,Zack Kloock, Irene Ng, Jocelyne Perez,Michael Schwinger, Ryan Shiroma. Recommender Systems Designed for Yelp.com; August 17, 2013
[4] Tingting Zhang, Yi Pan, from University of Washington. Yelp Challenge Project Report
[5] Tyler Daniel. Yelp DataSet–Calculating Popularity Zones CSE 190 Assignment 2
[6] Davy Suvee. Yelp Graph: Business Clustering Based on Check-In Data https://dzone.com/articles/yelp-graph-checkin-based