Analysis of Yelp Dataset

Multifactor Analysis of Yelp! Ratings

1 Introduction

Many restaurant entrepreneurs are faced with the question about what makes a restaurant successful. The objective is to analyze the Yelp dataset and identify success factors and visualize them and also to allow tweaking values of the factors to see the impact on the predicted rating based on new parameters.

2 Proposed Methods

From the methods considered in the literature survey presented in the proposal, it was found that neither PCA nor Linear Discriminant Analysis provided good results to accomplish the task we outlined so we experimented with many algorithms and settled on OneVsRest Ensemble classifier [1] with RandomForests [2] as the building block. The idea is to build a classifier with a good accuracy and then analyze the features which separate the classes the most and present them to the user along with their importances.

The dataset is restaurant data from the Yelp Dataset Challenge. In order to augment this data with location information and consider external factors that contribute to the ratings, socioeconomic and demographic data about age, ethnicity and income levels at the Census block group level was collected from the U.S. Census Bureau [3]. Currently, information for Phoenix, AZ (Maricopa County) has been added to get a working system up and running. This will be extended to all the cities in the dataset for the final submission. The additional data has been appended to the Yelp dataset.

2.1 Data Cleaning, Representation and Storage

The Census bureau divides US territory hierarchically to carry out census, the smallest of which is a census block(size). But socioeconomic and other data is only at the block group level(size). The first task was to append the additional Census data to the Yelp data by the block group id in which a restaurant is located. We used the geo-blovkgroupid api to convert latitude-longitude into block group ids. These were then cross-referenced with the census data age and income data by the block group level.

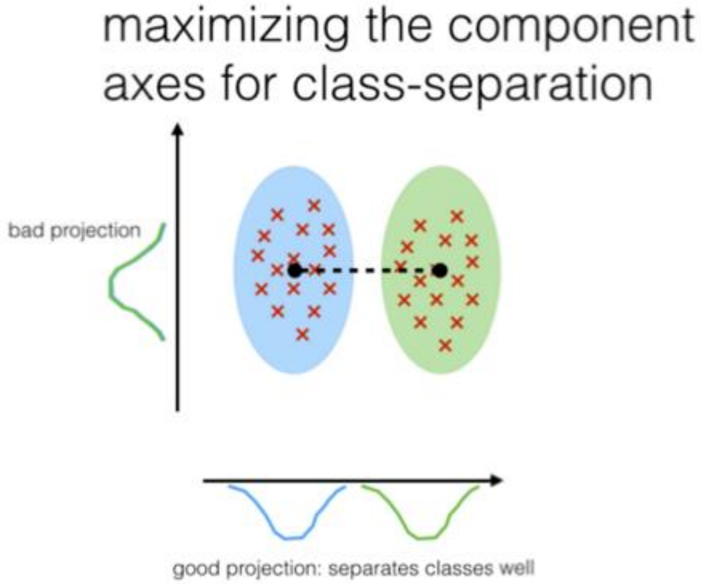
The age data from the Census also had inconsistent size of the age group buckets. For e.g., fields were given as “Male:10-15”, “Male:16-18”, “Male:18-20”, “Male:21-25” etc. Such fields were combined to have a uniform split of ten years per age group using Python scripts and Excel, given the smaller size of the Arizona dataset. Income data was also similarly grouped. These fields were then added to the Yelp dataset. This was now in the form of a set of JSON object, each object representing a restaurant. The JSON also had nested fields that could not directly be used in the analysis or to store in the database. Further flattening of the nesting was performed using Python scripts to expand lists and dictionaries. Once the flattened JSON was ready, the data was inserted into an SQLITE database.

A hierarchical classing strategy is used since Yelp ratings (classes from 0.0 to 5.0 with 0.5 point increments) proved to be too fine grained and the points identified had the clearest class separation. The new classes are:

0 – 2.5 Pretty Bad
>=3.0 Not Bad
    3.0 to 4.0 Good
        =3.0 Fair
        3.5 – 4.0 Above Average
    4.5 – 5.0 Extremely good

2.2 Class separation Analysis

The data was analyzed using multiple classifiers. The idea was basically to find a linear combination of variables which best separated the data into classes. The initial approach for finding factors was to do PCA and then Linear Discriminant Analysis on the components of PCA to find the most important features and then find correlation co-efficients between the components and the features to determine importance. But with a little more analysisand experimentation, we found LDA isn't suited to the data because it failed the Shapiro test[4]. Therefore we decided to try Quadratic Discriminant Analysis (QDA) but found that the accuracy wasn't good enough to have confidence in our factor weights. So we experimented with AdaBoost, GradientBoost, PolynomialSVM and finally choose RandomForests as the best accuracy was obtained with this classifier. We looked at the Feature importances property of classifiers in scikit learn to rank the weights of features [5]. We choose features which consistently rank high across all algorithms. We intend to use feature selection methods like MRMR to independently verify that our analyses is correct[6]. For our algorithm we have considered the features from the Yelp+Census datasets. The results returned are a set of top k features contributing to ratings and their weights (contributions) on a subset of the data filtered by user defined criteria. The parameter k can be varied to obtain the desired number of factors. The user can choose any kind(s) of restaurant with filters and analyse the factors contributing to success of these restaurants.


maximizing the component
axes for class-separation
bad projection
good projection: separates classes well

2.3 Testing:

We tested our methods by dividing the data into training and testing sets using randomized stratified shuffling and calculated the accuracy of prediction using the model learned. After obtaining a high enough accuracy we analyzed which features most contributed to the model to select the top k features to visualize.

2.6 Expected outcome:

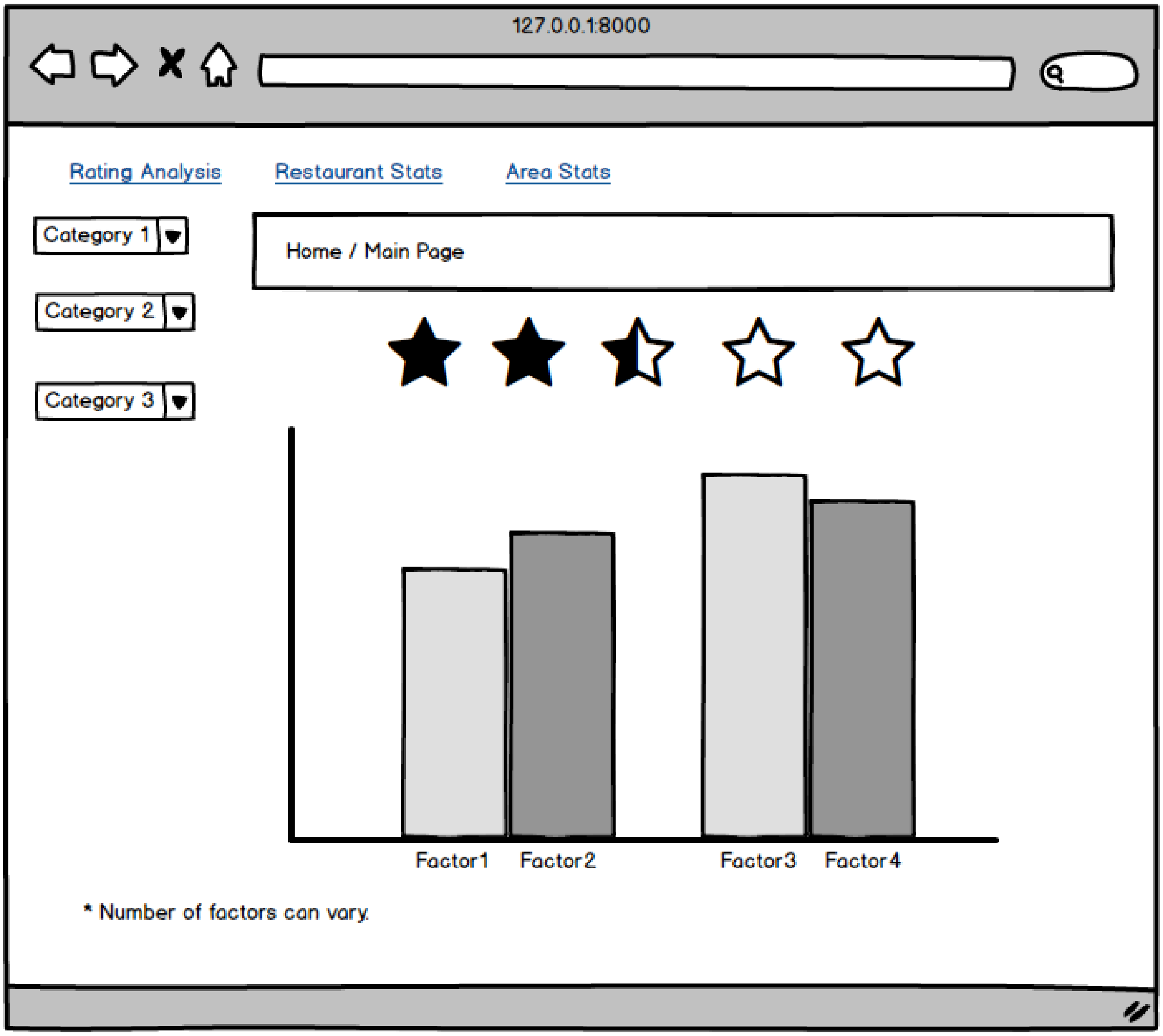
The final objective is to provide the user with the weights that each factor contributes to the star ratings and also to allow the user to tweak the values of features to see the how the predicted rating changes with new restaurant parameters. At the midterm stage, the data cleaning, storage and implementation of classifier are complete and the system outputs the weights for top k factors contributing to success. An initial webapp interface has been setup using Django and basic statistical visualizations are ready. Users can select the specific type of restaurants they are interested in and then run the analysis on this filtered set to see decisive features and their weights.

Prototype Interface

An interactive web interface wherein the user can select the factors and parameters to be visualized or analyzed will be implemented. The weights for the factors are computed and represented as a bar graph for a given class of ratings. A calculation of the star ratings for the user-specified factor values, with the weight known or the chosen parameters can be made.

Mock-Ups:

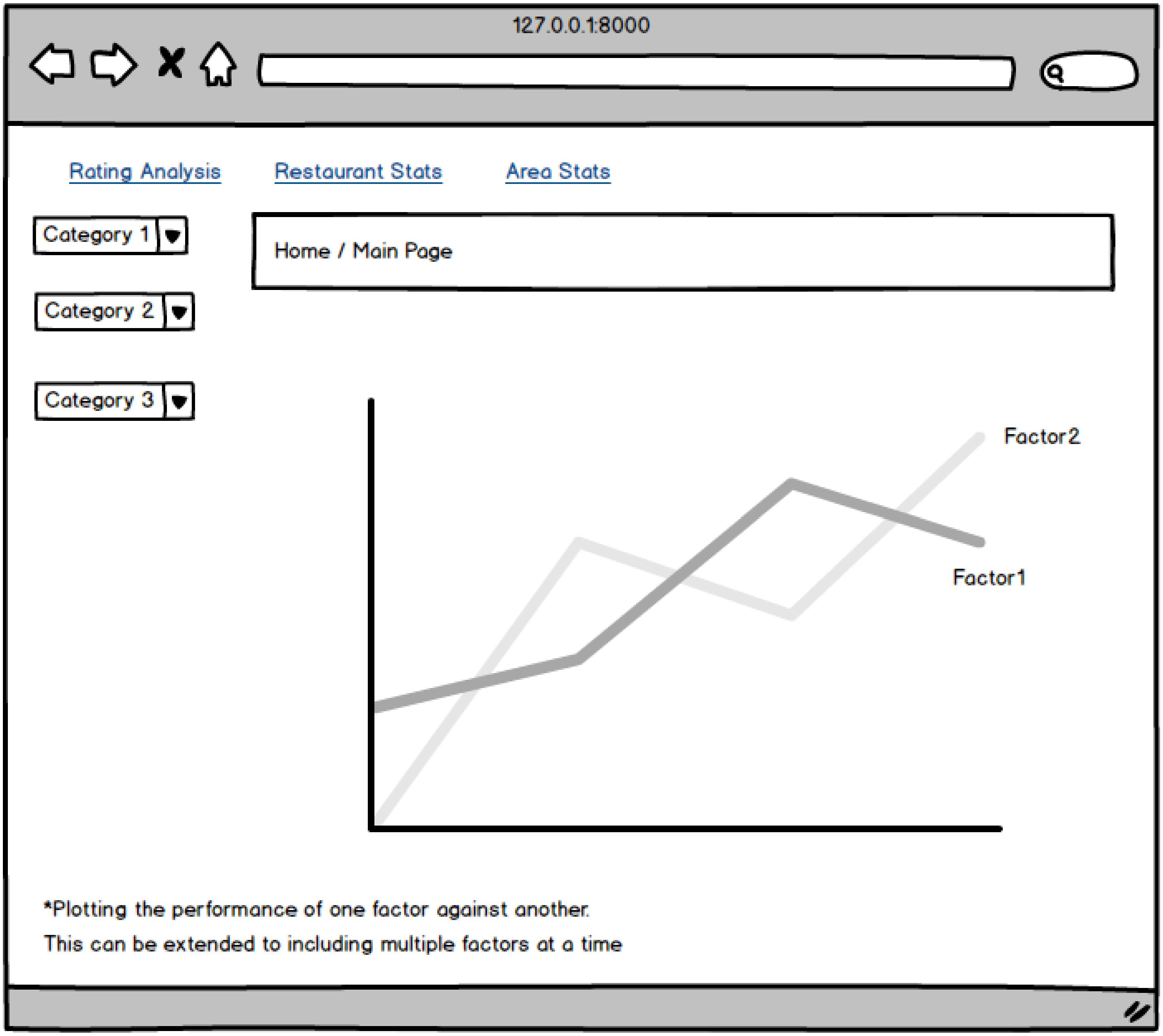
Layouts of the screens that we will be working on and some basic visualizations. The below represent the mock ups that we started off with and attempt to achieve by the end of the project:



* Number of factors can vary.

Fig 1: Prediction of ratings

The bar charts in figure 1 here plots the weights of the factors (top k) that contribute to the success of the restaurant. The star ratings are predicted based on the contribution these factors make. The user can select categories like “Cuisine: Italian, Indian, Chinese”, “Age-groups: 20-30, Female”, “Price range: $ to $$” etc. and view the contributions of the category to the rating.



*Plotting the performance of one factor against another.
This can be extended to including multiple factors at a time

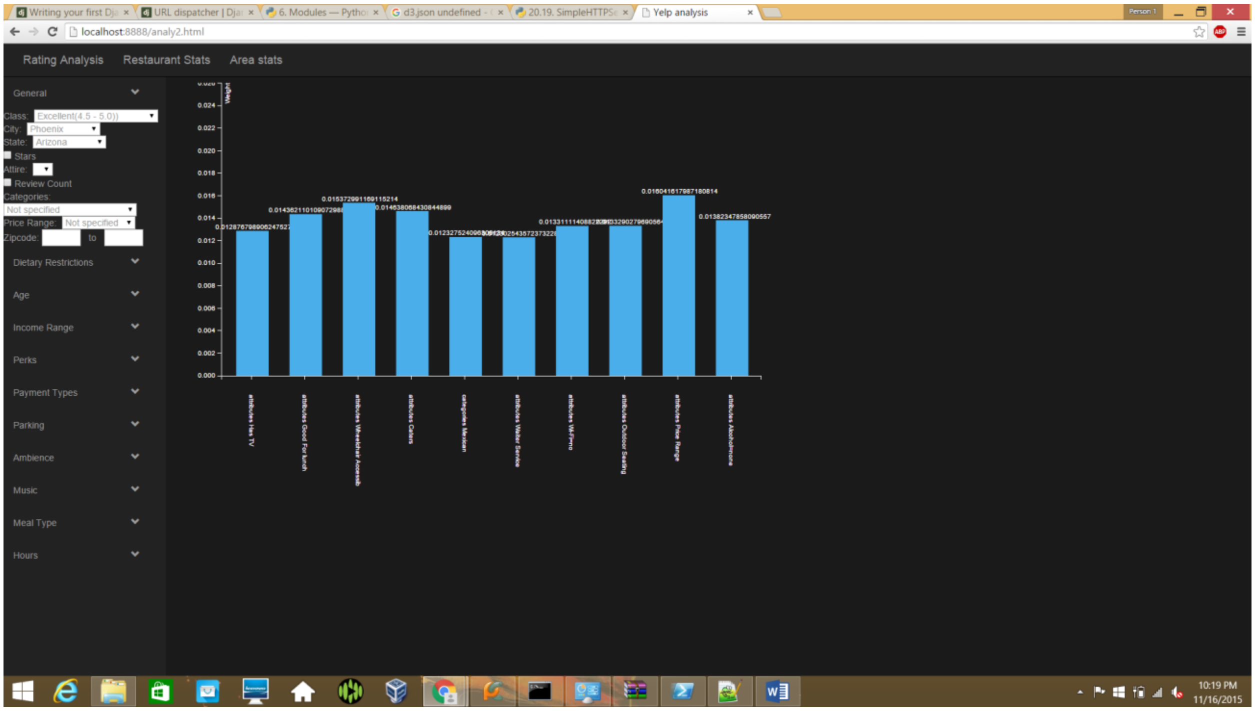
Fig 2: Comparison between factors that contribute

The line charts in figure 2 represent how one contributing factor performs with respect to another. For example, the average values of Restaurant Price vs. Area Income for a chosen location can be easily compared. Other categories can also be selected to view statistics about the area or rating group. For example, the user can view graphs on how many Chinese restaurants priced at $$ in a particular neighborhood are present. This will help analyze which factor should be given more importance while making the decision about opening the restaurant.

3. Experimental Results

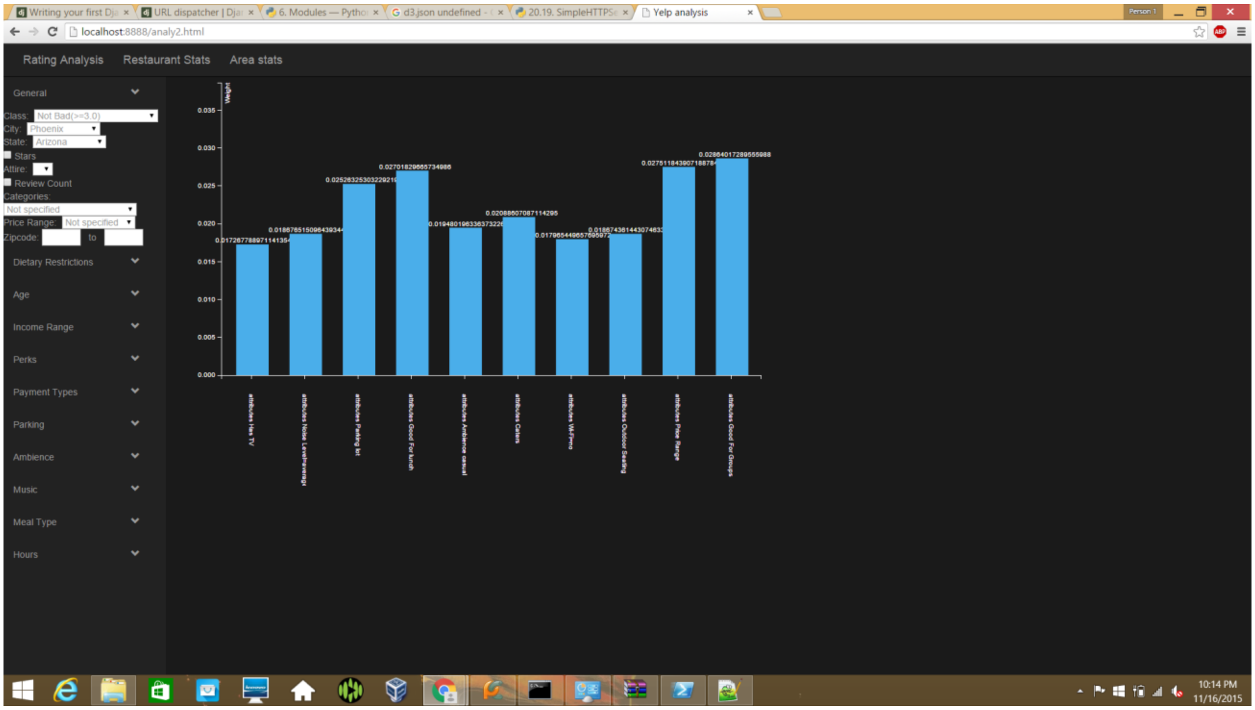
1. The output that we obtain for the city ‘phoenix’, state ‘Arizona’ and category ‘restaurants’ are for all categories/ attributes / features of restaurants.

These are for the **EG(Extremely Good)** category-



Class: Excellent(4.5 - 5.0))
City: Phoenix
State: Arizona
□ Stars
Attire:
□ Review Count
Categories: Not specified
Price Range: Not specified
Zipcode: to

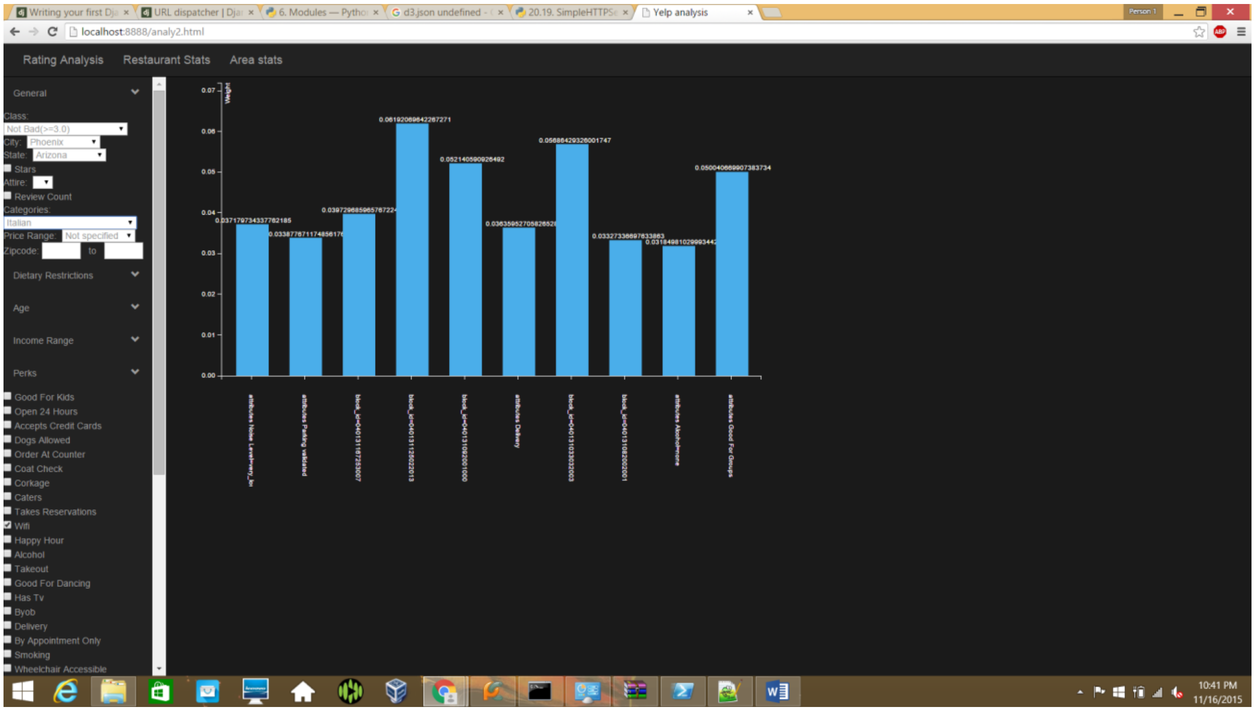
These are for the **AAVG(Above Average)** category-



Class: Not Bad(>=3.0)
City: Phoenix
State: Arizona
□ Stars
Attire:
□ Review Count
Categories: Not specified
Price Range: Not specified
Zipcode: to

We finally get three results. Result set 1 has weights of features for restaurants with star ratings greater than 2.5 with an **accuracy of 83%**. Result set 2 has weights of features for restaurants with star ratings greater than 4.5 with an **accuracy of 85%**. Result set 3 has weights of features for restaurants with star ratings between 3 and 4.5 with an **accuracy of 73%.**

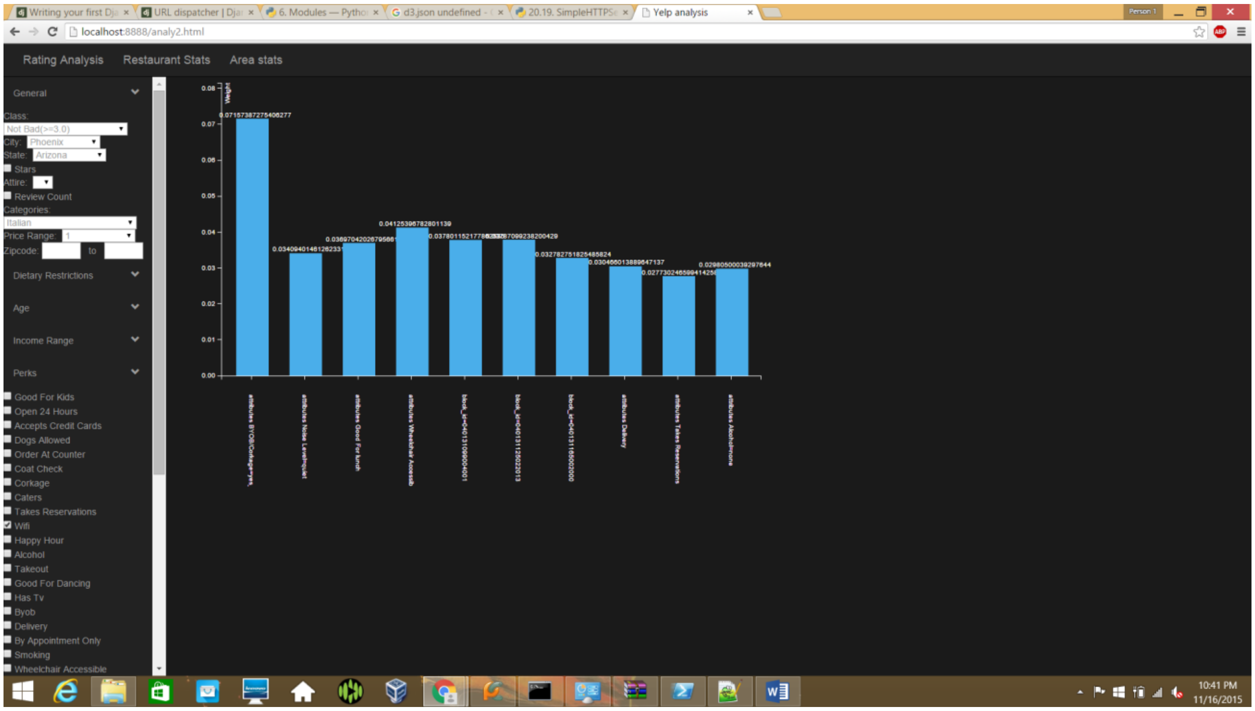
2. For the filter conditions, the features considered are: **Categories: Chinese, Pizza Italian** and attribute Wi – fi, Reservations, Ambience(casual, touristy). The experiment is repeated for Price range values: 1 and 2. By running the algorithm for k=10 we get the top 10 features and their weights. These do not contain the categories considered as a part of our filter conditions.



Class: Not Bad(>=3.0)
City: Phoenix
State: Arizona
☐ Stars
Attire:
☐ Review Count
Categories: Italian
Price Range: Not specified
Zipcode: to

☐ Good For Kids
☐ Open 24 Hours
☐ Accepts Credit Cards
☐ Dogs Allowed
☐ Order At Counter
☐ Coat Check
☐ Corkage
☐ Caters
☐ Takes Reservations
☑ Wifi
☐ Happy Hour
☐ Alcohol
☐ Takeout
☐ Good For Dancing
☐ Has Tv
☐ Byob
☐ Delivery
☐ By Appointment Only
☐ Smoking
☐ Wheelchair Accessible

Price range: 2



Class: Not Bad(>=3.0)
City: Phoenix
State: Arizona
☐ Stars
Attire:
☐ Review Count
Categories: Italian
Price Range: 1
Zipcode: to

☐ Good For Kids
☐ Open 24 Hours
☐ Accepts Credit Cards
☐ Dogs Allowed
☐ Order At Counter
☐ Coat Check
☐ Corkage
☐ Caters
☐ Takes Reservations
☑ Wifi
☐ Happy Hour
☐ Alcohol
☐ Takeout
☐ Good For Dancing
☐ Has Tv
☐ Byob
☐ Delivery
☐ By Appointment Only
☐ Smoking
☐ Wheelchair Accessible

Price Range : 1

4. Improvements/Future Work

Additional features to the existing work:

1. Consider reviews for restaurants. Sentiment analysis based on the top 5 reviews.
2. Finding out correlation between tips and reviews.
3. Analyzing review count and user reviewing

Feature selection to improve ranking:

1. Mutual information based ranking
2. Pure correlation
3. Max relevance minimum redundancy
4. Out of bag error

5. Summary of innovations

1. Applying algorithms to features that matter the most to determine the top factors contributing to the star ratings.
2. Designing visualizations for the users to understand and easily pick out categories by which they want to make their decisions about opening a restaurant in a certain location.

6. Conclusions

We have a fully working and validated algorithm and a basic visualization scheme that a user can use to filter restaurant groups and analyze what factors contribute most to success with a high degree of accuracy. We intend to improve the visualizations and do external validation for the finals.

Bibliography

[1] *Wikipedia Page on George H. Heilmeier*. Wikimedia Foundation, n.d. Web. 04 Oct. 2015. <https://en.wikipedia.org/wiki/George_H._Heilmeier>.d

[2] Lee, Ryong, and Kazutoshi Sumiya. "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection." *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*. ACM, 2010.

[3] Sawant, Sumedh. "Collaborative filtering using weighted bipartite graph projection: a recommendation system for yelp." *CS224W: Social and Information Network Analysis (December 10, 2013)* (2013).

[4]Pileggi, Hannah, et al. "Snapshot: Visualization to propel ice hockey analytics." *Visualization and Computer Graphics, IEEE Transactions on* 18.12 (2012): 2819-2828.

[5]Flood, Mark D., et al. "The application of visual analytics to financial stability monitoring." *Office of Financial Research Working Paper* 2 (2014).

[6]Sitaram Asur,Bernardo A. Huberman.”Predicting the future with Social Media.”http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf

[7]Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis Foundations and Trends in Information Retrieval, 2(1-2), pp. 1135, 2008.

[8]James Huang, Stephanie Rogers, Eunkwang Joo.Improving Restaurants by Extracting Subtopics from Yelp Reviews.

[9]Zhang, Jason Q., Georgiana Craciun, and Dongwoo Shin. "When does electronic word-of-mouth matter? A study of consumer product reviews."*Journal of Business Research* 63.12 (2010): 1336-1341.

[10]Hu, Longke, Aixin Sun, and Yong Liu. "Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014.

[11]Lyu, Yan, et al. "Using multi-criteria decision making for personalized point-of-interest recommendations." *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2014.

[12]*”Evaluation of factors affecting customer loyalty in the restaurant industry”*Mohammad Haghighi, Ali Dorosti et al in African Journal of Business Management Vol. 6(14), pp. 5039-5046, 11 April, 2012

[13]Muller, Christopher C., and Crist Inman. "The geodemographics of restaurant development." *The Cornell Hotel and Restaurant Administration Quarterly* 35.3 (1994): 88-95.

[14] The Yelp Dataset Challenge (http://www.yelp.com/dataset_challenge)