**Multifactor Analysis of Yelp! Ratings**

**1 Introduction**

Many restaurant entrepreneurs are faced with questions about where to open a restaurant, pricing, ambience, demographic background of the chosen location etc. These are factors that contribute to the success (star rating) of a restaurant. Existing work focuses on studying the impact of a single factor such as location or price on the rating.[7] Our objective is to analyze the Yelp dataset and identify the top factors that contribute to restaurant success, visualize them and allow users to experiment with different values for the identified factors to observe the impact on the predicted rating so they can make informed decisions about which factors to focus on for success.

**2 Proposed Methods**

The overview of the proposed method is:

* Obtain Yelp data
* Analyze data to determine which factors contribute to ratings and by how much
* Visualize the contributions of the top factors
* Build interactive UI for the user to experiment with the values for the contributing factors and reflect change in rating

The dataset is restaurant data from the Yelp Dataset Challenge. To augment this data and consider external factors, socioeconomic and demographic data about age, ethnicity and income levels was collected from the U.S. Census Bureau [1],[8]. Currently, information for Phoenix, AZ (Maricopa County) has been used to build a minimum working system.

**2.1 Data Cleaning, Representation and Storage**

The Census bureau divides US territory into blocks [9]. Socioeconomic data was available at the block group level. We appended the Census data to the Yelp data by identifying the block group in which a restaurant is located (using geo-blockgroupid API to convert latitude-longitude into block group ids) and cross-referenced the age and income data by block group ID.

Cleaning was required to integrate the data sources. Census data had inconsistent size of age and income group buckets. Such fields were aggregated to have uniform bucket size using Python. These fields were added to Yelp dataset, creating a list of JSON objects, each object representing a restaurant. The JSON had nested fields that couldn’t directly be analysed. Flattening of the nesting was performed to expand lists and dictionaries. The flattened JSON was inserted into an SQLITE database.

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

0 – 2.5 Pretty Bad

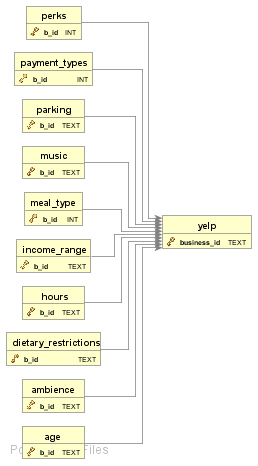
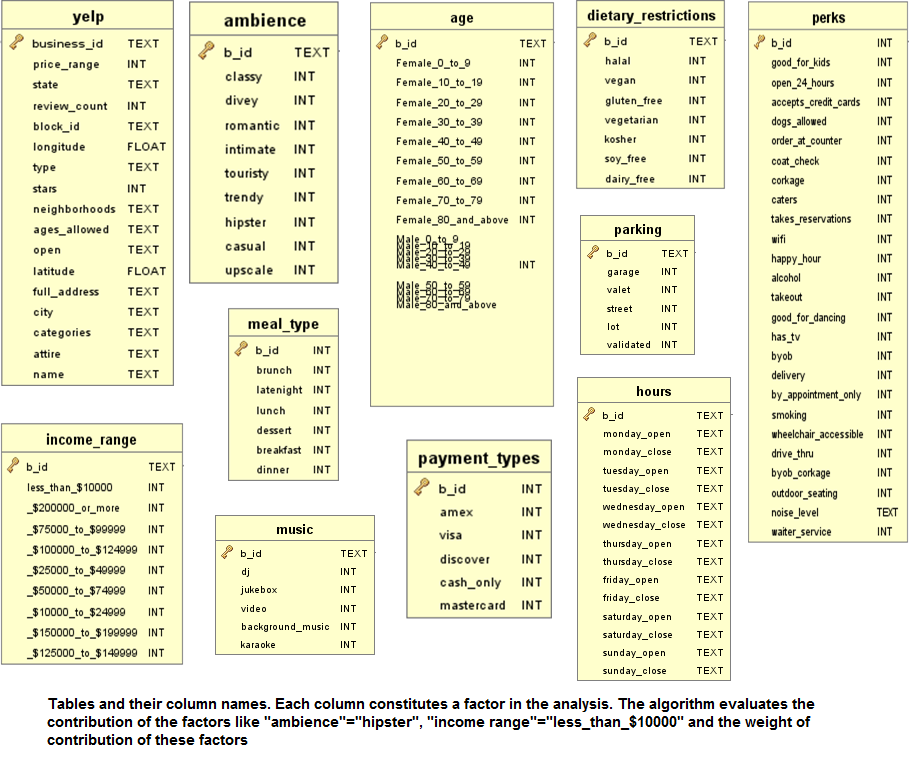
>=3.0 Not Bad

3.0 to 4.0 Good

=3.0 Fair

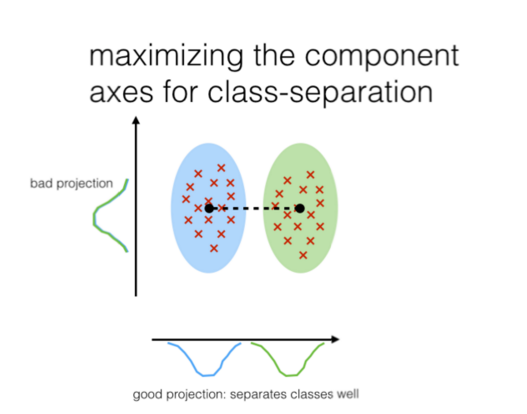
3.5 – 4.0 Above Average

4.5 – 5.0 Extremely good



**2.2 Class separation Analysis**

The data was analyzed using multiple classifiers to find the linear combination of variables which best separated the data into classes. The initial approach for finding the important factors used PCA and then Linear Discriminant Analysis on the components of PCA and then found correlation co-efficients between the components and the features to determine importance. Analysis and experimentation found LDA unsuitable because it failed the Shapiro test[2]. Therefore, we tried Quadratic Discriminant Analysis (QDA) but the accuracy wasn’t good enough to have confidence in our factor weights. We experimented with AdaBoost, GradientBoost, PolynomialSVM and finally choose RandomForests which gave best accuracy. We used the Feature importances property of classifiers in scikit-learn to rank the weights of features [3]. We chose features which consistently ranked high across all algorithms. We intend to use feature selection methods like MRMR to independently verify that our analysis is correct [4]. The results returned are the top ‘k’ features contributing to ratings and their weights (contributions) on a subset of the data filtered by user criteria. ‘k’ can be varied to obtain desired number of factors. The user can choose any filters and analyse the factors contributing to success of these filtered restaurants.



**2.3 Testing:**

Testing was done by dividing the data into training and testing sets using randomized stratified shuffling. Accuracy of prediction was calculated using the model learned.

**2.4 Expected outcome and Prototype Interface:**

The final objective is to provide the user with the weights that each factor contributes to the ratings and also to allow the user to tweak the values of weights to see the how the predicted rating changes with new values. At the midterm stage, the data cleaning, storage and implementation of classifier are complete and the system outputs the weights for top k contributing factors. An initial web interface has been setup using Django and visualizations have been designed. Filter menu is also available. The complex visualizations and interactive graphs remain to be completed.

**Mock-Ups:**

Layouts of the screens that we will be working on and some basic visualizations.

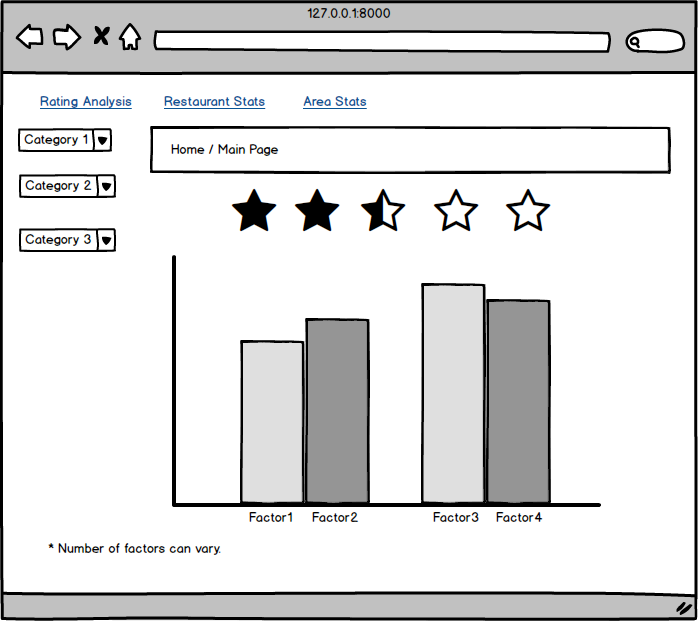


Fig 1: Prediction of ratings

The bar charts in figure 1 here plots the weights of the factors (top k) that contribute to the particular star rating group. 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. The user can then drag the bars to set them to different values. For e.g., he can change “Cuisine” to have a higher weight and “Price range” to be lower. The star rating will then change to a predicted value based on the new weights.

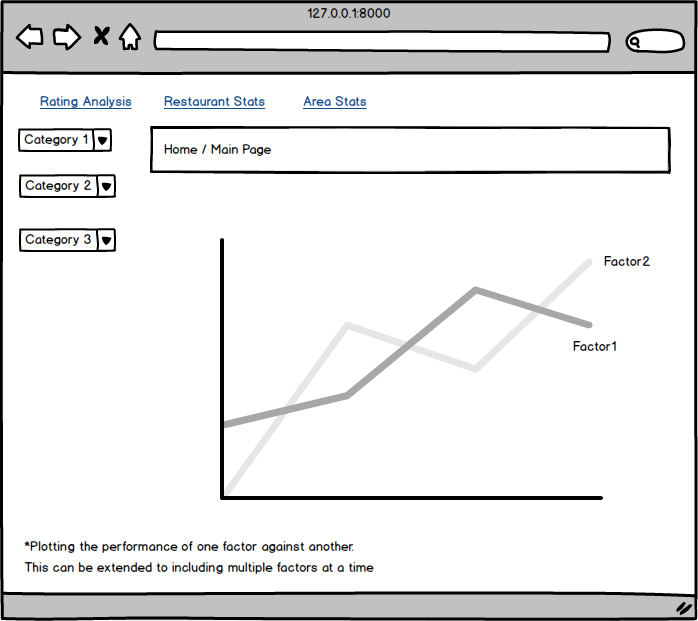
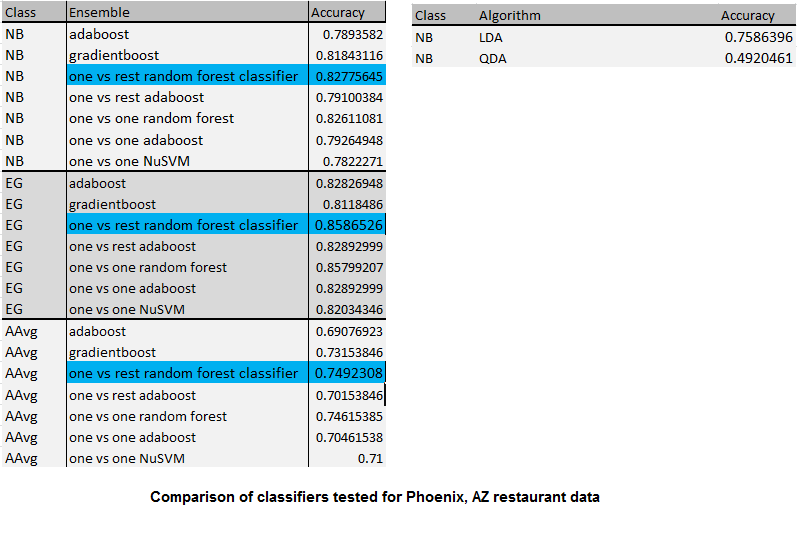


Fig 2: Comparison between factors that contribute

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. This will help analyze which factor should be given more importance while making the decision about opening the restaurant.

**3. Experimental Results and Initial Interface**



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

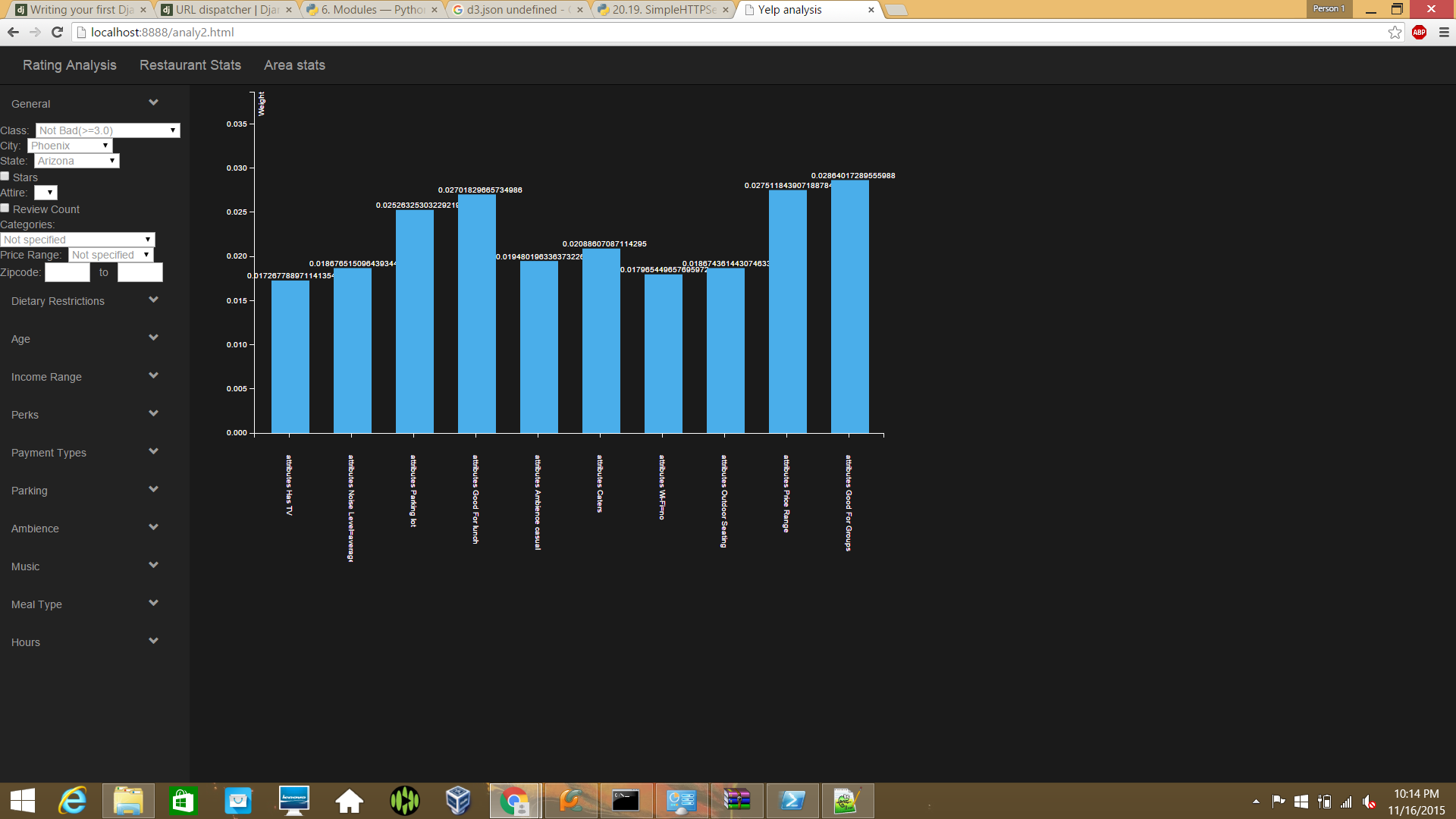
**EG(Extremely Good)** category-



The above graph plots the following values:

[(‘Wheelchair Accessible', 0.016708343155191818), (‘Price Range', 0.015882610276886186), (‘Caters', 0.015183111314425108), (‘Outdoor Seating', 0.014394343906883151), (‘Wi-Fi=no', 0.014198985318470461), ('Noise Level=average', 0.012722890371051039), ('Has TV', 0.012441531068174517), ('Waiter Service', 0.012377714113338302), ('open', 0.012143786619022717), ('Alcohol=none', 0.012055734230117387)]

**AAVG(Above Average)** category-



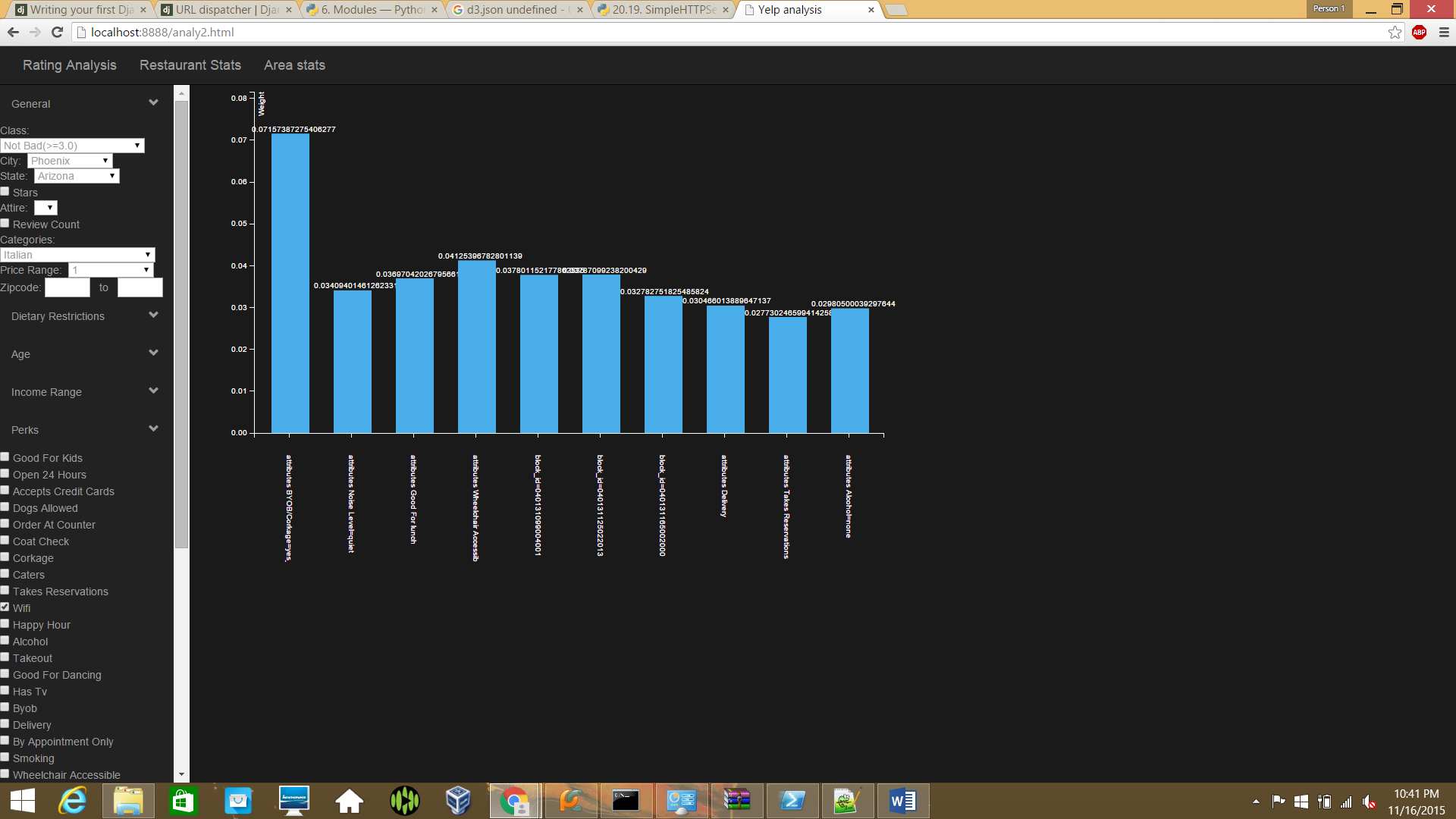
[('Caters', 0.019432465441215505), ('Price Range', 0.018251319490358915), ('Ambience casual', 0.017265075471278982), ('Outdoor Seating', 0.01662760420613299), ('Wheelchair Accessible', 0.016099825655570429), ('Wi-Fi=no', 0.015799907071278974), ('Has TV', 0.015648666419009055), ('categories Buffets', 0.015218934968452044), ('Noise Level=average', 0.014621606916995454), ('open', 0.0145569981824247)]

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%.**

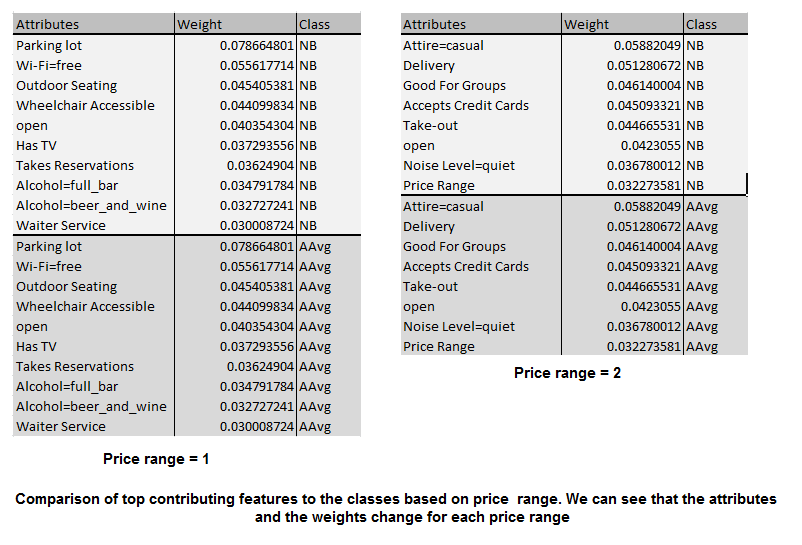
2. For the filter conditions, the features considered are: **Categories: Chinese, Pizza Italian** and ' Wifi, Reservations, Ambience(casual, touristy). The experiment is repeated for Price range values: 1 and 2. By running the algorithm for k=10 we get the top 10 features and their weights. It is seen that different factors contribute with different weights to the two price ranges and the cuisines do not contribute.



Price range: 2



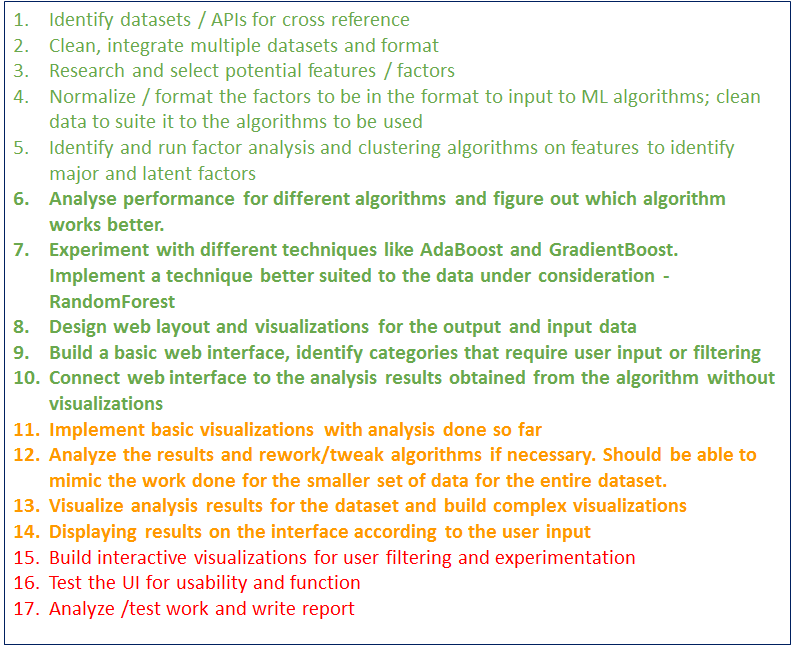
Price Range : 1



The weights for each attribute are the same however since restaurants between rating 3-4 have similar contributing factors.

**4. Current progress**





**All members contributed equally to the work thus far.**

**5. Improvements/Future Work**

1. Consider reviews for restaurants among the factors. [5],[6]

2. Correlation between tips and reviews.

3. Analyzing reviewer-restaurant relationships

**6. Innovations**

1. Applying algorithms to most relevant features to determine the contribution to star ratings.

2. Designing visualizations for the users to understand demographic and existing restaurant statistics and parameters

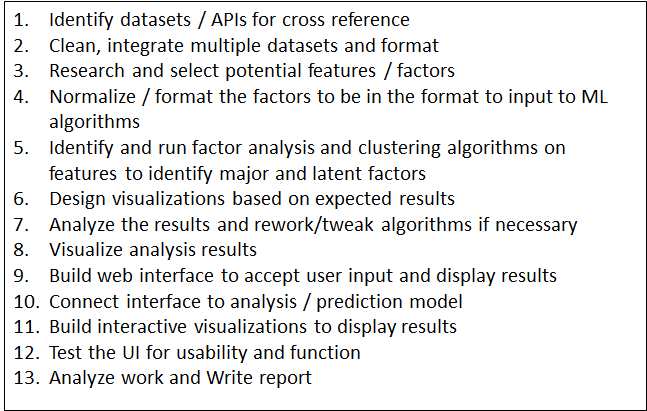
3. Allowing users to experiment with weightage for each factor and analyze impact on rating.

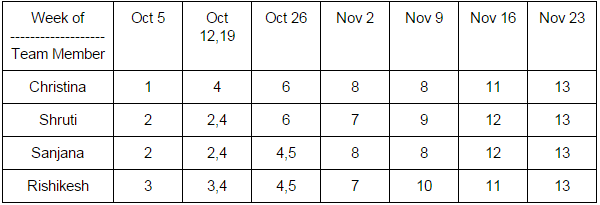
**7. Conclusions**

We have a fully working and validated algorithm, basic visualization designs and web interface setup that a user can use to filter restaurant parameters and analyze what factors contribute most to success with a high degree of accuracy. We intend to improve and add visualizations.

**Appendix**

Original task breakdown and schedule:





**Bibliography**

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

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

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

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

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

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

[7]*”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

[8]Muller, Christopher C., and Crist Inman. "The geodemographics of restaurant development." *The Cornell Hotel and Restaurant Administration Quarterly* 35.3 (1994): 88-95.  
[9] Census Block Groups (https://www.census.gov/geo/reference/gtc/gtc\_bg.html)