# CSE 255 Assignment 1
# Predicting price range of a restaurant based on previous visit

Dev Agarwal
devagarwal@eng.ucsd.edu

***Abstract*—In this project, patterns related to the price range of a restaurant are explored. A model has been designed which would like to predict the next restaurant to be visited's price range given information about the current restaurant (it's price range, location, cuisine and day of the week).**

***Keywords*—Price range, Yelp, Multiclass Classification**

## I. Introduction

With the rise of online review websites, there has been an abundance of data to analyze. This has enabled us to better study the human behavior. Yelp, one of the most popular online business ratings and reviews website, has a lot of interesting data in the restaurant domain.

Majority of restaurants on Yelp have a price range (varying from $ to $$$$) associated with them. The hypothesis here is that there might be a general trend people seem to follow. There might be some $$ restaurants which are visited frequently after some type of $ restaurant (e.g. Italian restaurant visited during first week of the month). There could be a variety of temporal factors (e.g. day of the week/month), social factors (e.g. city) and other factors (e.g. type of food) which could help us find such a pattern. This problem is interesting because this can help us differentiate between restaurants having same price range, and it can help Yelp in it's ranking problem. For example, if Yelp had to decide between a $ and $$ restaurant, it could use the information about the user's previous visit and recommend more accurately.

## II. Dataset

The Yelp dataset is used for this assignment. This dataset is from the fifth round of the Yelp Dataset Challenge. The dataset provides information about the following:

1) Business (61,184 businesses)
2) Review (1,569,264 reviews)
3) User (366,715 users)
4) Check-in (Check-in information for 45,166 businesses)
5) Tips (495,107 tips)

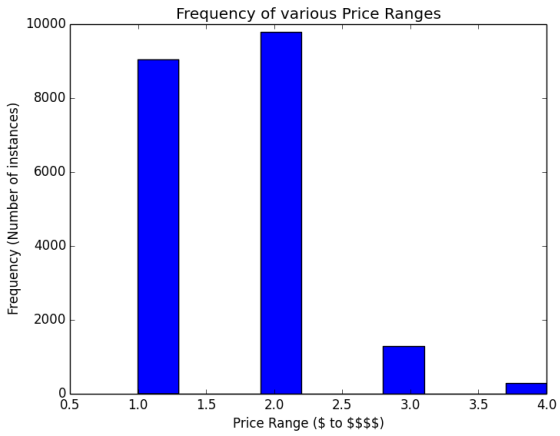
For this assignment, only those businesses were considered which were Restaurants and had their Price Range information given. This reduced the number of businesses from 61,184 to 20,430.

The Review information is most relevant as it gives the time the review was written, for which business (hence the price range) it was written and which user wrote it. With the review information, we can calculate the sequence of restaurants visited by a user and the price range of each restaurant. This information is crucial for our analysis.

Note: I have assumed that the reviews must have been written in the same chronological order as the restaurants the user visited and the review must have been written soon after the visit. The Check-in information gets rid of this assumption, but it doesn't mention the order by which the restaurants were visited.

## III. Exploratory Analysis

### A. Distribution of Price Ranges



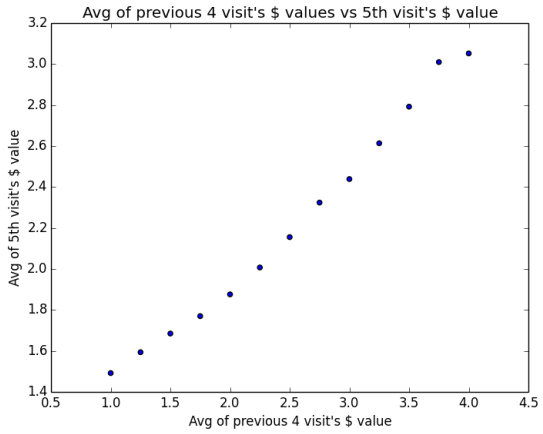

The plot above shows the distribution of the price range of restaurants in the dataset. We can see that $$ and $ restaurants are much more abundant in number as compared to $$$ and $$$$ restaurants.

### B. Relationship between two subsequent visited restaurants' price ranges

1) Average price range of past 4 visits vs 5th visit
   The average price range of the past 4 visited restaurants is plotted against the price range of the 5th visited restaurant. As there are different 5th visited restaurant's $ value for the same average of past 4

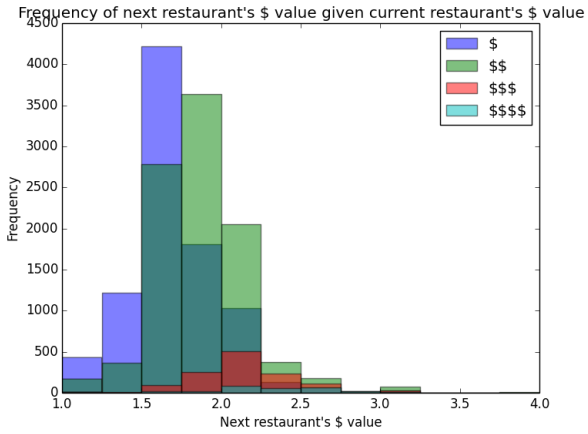
visited restaurants' $ value, the average of all the various 5th visited restaurant's $ value is plotted instead.

For this, only those users who have given more than 4 reviews are considered. The whole "user - sequence of restaurant visits" data is windowed with window size = 5. By doing this, 454,098 windows are generated.




We can see that there is a linear correlation between the average of previous 4 visits and the 5th visit.

2) Price range of current visit vs Price range of next visit




The purple histogram above ($ price range currently) shows that most frequently the subsequent restaurant's price range has an average of 1.5-1.75. The green histogram ($$ price range currently) shows that most frequently the subsequent restaurant's price range has an average of 1.75-2. The red histogram ($$$ price range currently) shows a similar trend. Therefore, there exists a relationship between previous restaurant's $ value and next restaurant's $ value.

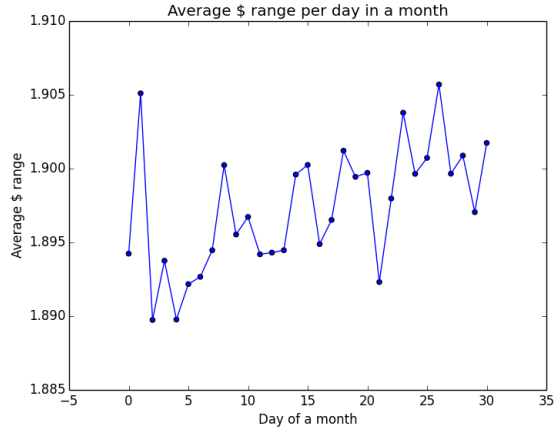
### C. Temporal factor (monthly)

The effect a day in a month could make in the average price range of the restaurant is analyzed here. For example, it could be possible that people prefer going to expensive restaurants around the time they get their salaries.

1) Frequency of reviews in a month


First, we see how the reviews frequency varies as a function of days in a month. There doesn't seem a very obvious pattern, but it signals that there might be a weekly pattern.

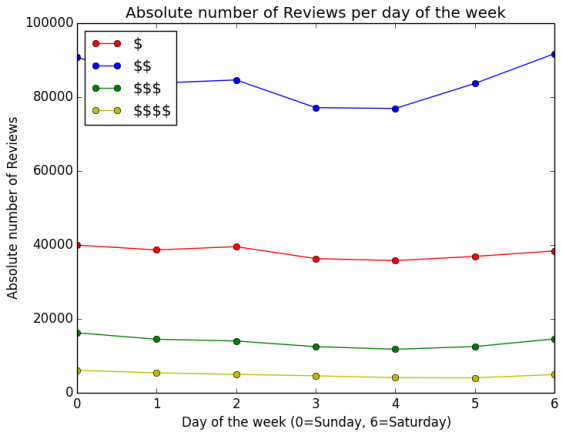
2) Average price range of restaurants in a month based on reviews




Here too we can't see a very obvious trend, but this too signals that there might be some weekly trend.

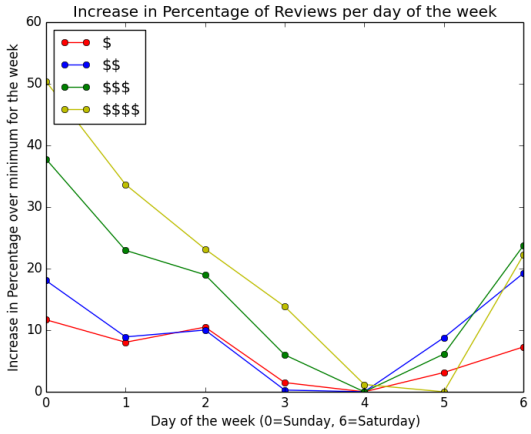
### D. Temporal factor (weekly)

1) Absolute number of reviews as a function of day of the week




There is no helpful information regarding the price range of a restaurant.

2) Increase in percentage of reviews as a function of day of the week

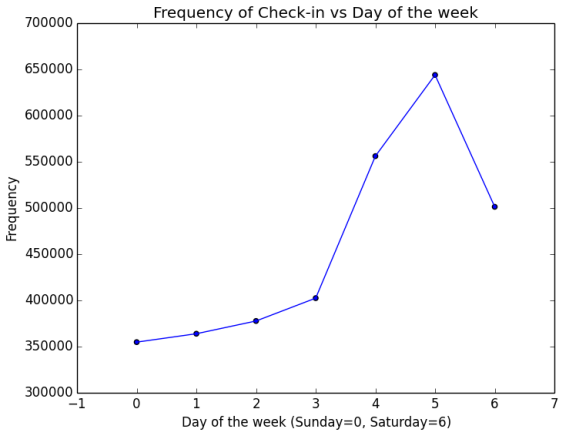



We can conclude that people prefer reviewing more during the weekends as compared to weekdays.

*E. Temporal factor (weekly) based on Check-ins*

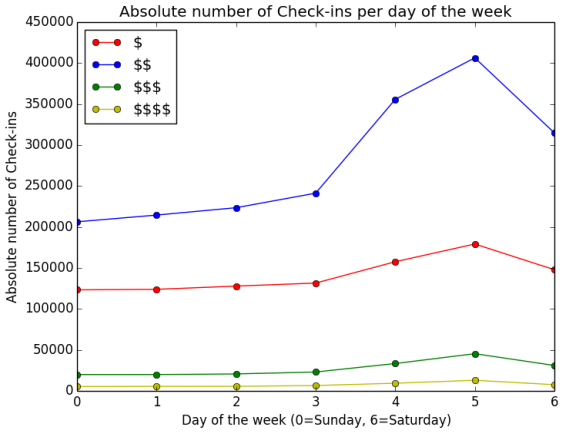
A better alternative to review data is the check-in data. The Check-in data gives us the information about which day of the week each restaurant was checked-in and by how many people.

1) Frequency of Check-ins vs Day of the week



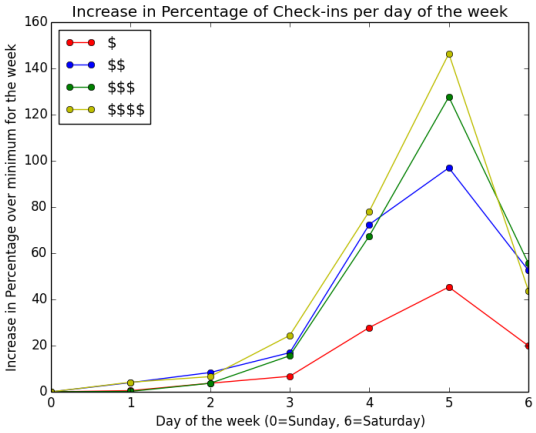

As expected, Friday seem to be the busiest days for restaurants.

2) Frequency of Check-ins (for various price ranges) vs Day of the week




No interesting pattern is found here.

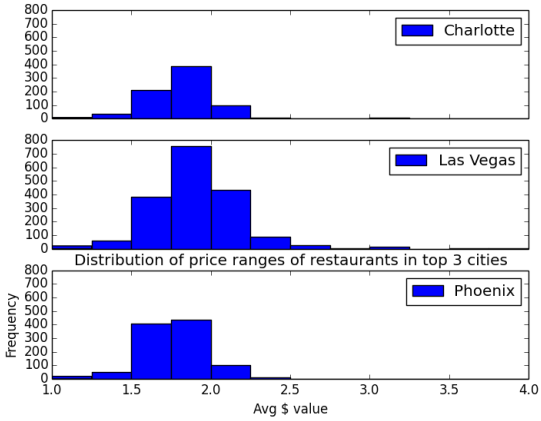
3) Increase in percentage of check-ins vs Day of the week




This is an interesting plot. We see that there is greater increase in $$$$'s popularity as compared to others during the weekend. This signifies that there is a

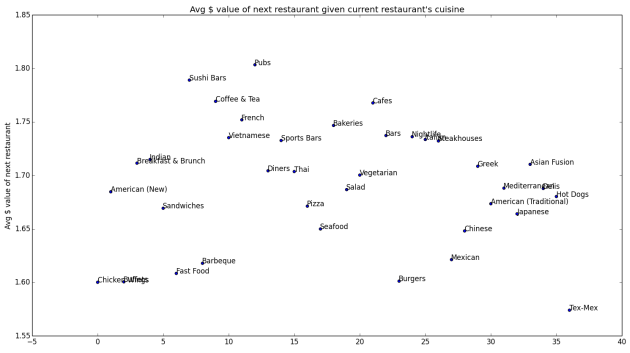
slightly higher chance to visit a $$$$ restaurant on a weekend as compared to a weekday.

### F. Effect of location




The histograms above denote the distribution of average price ranges across 3 different cities. We can see that the distribution is not the same. Therefore, the location of the restaurant might play a role in the price range of a restaurant.

### G. Effect of cuisine




The scatter plot above shows the average price range of the next visited restaurant with the current restaurant belonging to the given category. We see that there is a high variance, probably implying that the cuisine of a restaurant might be a factor in the price range of the next visited restaurant.

## IV. Predictive Task

The Prediction Task chosen is to predict the next restaurant to be visited's Price Range ($ value) given the current restaurant's information and the visit information.

One baseline would be predicting next restaurant's price range randomly from the price range distribution. Second baseline would be predicting next restaurant's price range solely based on previous restaurant's price range.

This is a multiclass classification problem. The input features are the current restaurant's details and the visit details (further discussed in Section VI: Features). The output can take one of the four possible price range values (1 to 4).

The dataset is split into training (75%) and testing (25%) set. The prediction model is built on the training set. The validity of the model's prediction is checked by running on the unseen testing set. The training and the testing set is randomly picked to remove any bias.

## V. Literature

The dataset being used is Yelp's dataset. The dataset is available online for research purpose $(https : //www.yelp.com/dataset\_challenge/dataset)$. This dataset has been used heavily to predict a user's rating of a business given the business and user's attributes [1].

I was inspired to work with price range after coming across an interesting study which explored the relationship between price range of a restaurant and the text used while writing a review for such a restaurant [2].

## VI. Features

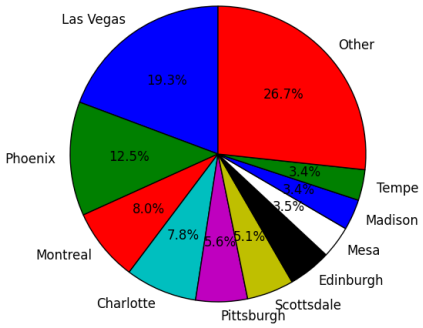
From the exploratory analysis done in Section III, there seems to be information in day of the week (Section III-E), $ value of previous visits (Section III-B), location of the restaurant (Section III-F) and cuisine (Section III-G) for the price range prediction of the next restaurant. Day of the month didn't seem to have useful information.

The input features can be divided into 2 categories:

1) Information about the current restaurant
   a) $ Value
      We can see from the exploratory analysis done above that the previously visited restaurant's $ value gives some indication towards the $ value of the next restaurant to be visited. This is a **single feature**, whose value can range from 1($) to 4($$$$).
   b) Cuisine
      The top 32 most popular cuisines are chosen. After considering only the top 32 cuisines, only 1575 instances out of 20430 instances (7.7%) didn't have any of these cuisines as a part of their categories. The top 32 cuisines are: 'Chicken Wings', 'American (New)', 'Buffets', 'Breakfast & Brunch', 'Indian', 'Sandwiches', 'Fast Food', 'Sushi Bars', 'Pizza', 'Coffee & Tea', 'Vietnamese', 'French', 'Pubs', 'Diners', 'Sports Bars', 'Thai', 'Barbeque', 'Vegetarian', 'Bakeries', 'Salad', 'Seafood', 'Cafes', 'Steakhouses', 'Burgers', 'Nightlife', 'Italian', 'Bars', 'Mexican', 'Chinese', 'Greek', 'American (Traditional)', 'Mediterranean', 'Japanese', 'Asian Fusion', 'Delis', 'Hot Dogs', 'Tex-Mex'.
      This is a categorical feature. Therefore, **32 binary features** are created, each representing one of the category. Note: A given restaurant can

belong to more than one category. In that case, all the corresponding features will be set to 1.

c) Location/City
There are totally 247 cities mentioned in the dataset. We take the top 10 cities which comprises 73% of the data. The top 10 cities are 'Las Vegas', 'Phoenix', 'Montreal', 'Charlotte', 'Pittsburgh', 'Scottsdale', 'Edinburgh', 'Mesa', 'Madison', 'Tempe'.




This is a categorical feature. Therefore, **10 binary features** are generated, each representing one of the cities. If the instance doesn't belong to any of the cities, then all the 10 features are left to be 0.

2) Information about the current visit

a) Day of the week
We know that there is a higher increase in the frequency of $$$$ restaurant during the weekend as compared to a $ restaurant. This makes sense as people are more willing to spend during the weekend and enjoy their meals. As the trend is clearly divided into weekend and weekday, I thought having **two binary features** (representing weekend and weekday) was a better idea than representing each day as an independent feature. Therefore, there are 2 binary features. One feature will represent if it is a weekend, and the other will represent if it is a weekday.

## VII. Model

The dataset looks like this,

| Information about current Restaurant | | | Information about the Visit | Output |
|---|---|---|---|---|
| $ Value | Cuisine | City | Weekday or Weekend | Next Restaurant's $ Value |
| | | | | |

"$ Value" is a single feature with values ranging from 1 to 4. "Cuisine" is a collection of 32 binary features. "City" is a collection of 10 binary features. "Weekend/Weekday" is a collection of 2 binary features. "Next Restaurant's $ Value" is a multiclass feature with values ranging from 1 to 4.

The various models tried are:

1) Baseline 1: Randomly predict from distribution
The plot in Section III-A shows that the frequency of $$ restaurants is the maximum. If we only know that, then the optimal approach would be to predict 2 ($$) every single time.

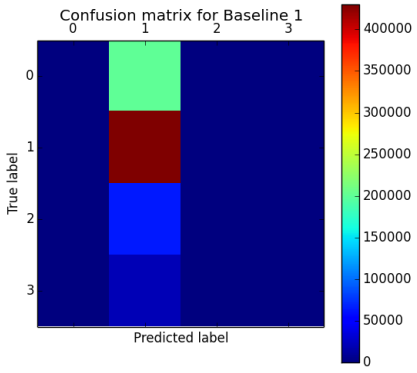
2) Baseline 2: Predict next price range solely based on current price range
As the plot in Section 2 seemed promising, we'll just predict the next restaurant's price range to be the same as current restaurant's price range.

3) Logistic Regression
Logistic Regression in it's original form is a binary classifier. To extend it to a multiclass classifier, we apply OneVsRest approach. In this approach, a logistic regression is trained for each class against all other classes combined. While predicting, the class for which the logistic regression has maximum probability/confidence is predicted.
The weakness of Logistic Regression is it's assumption for a linear decision boundary. To overcome this assumption, Support Vector Machines is tried.

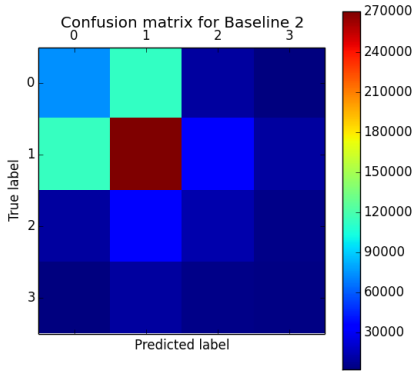
4) Support Vector Machine
Support Vector Machine is a maximum-margin classifier. With the help of Kernels, non-linear decision boundaries can be constructed easily. For this project, I chose the Gaussian Kernel as it proved to perform best (via 4-fold Cross Validation). Similar to Logistic Regression, SVM in it's original form is a binary classifier. To extend it to a multiclass classifier, we again apply OneVsRest approach. The hyperparameter value of C is also chosen via 4-fold Cross Validation.

## VIII. Results & Conclusion

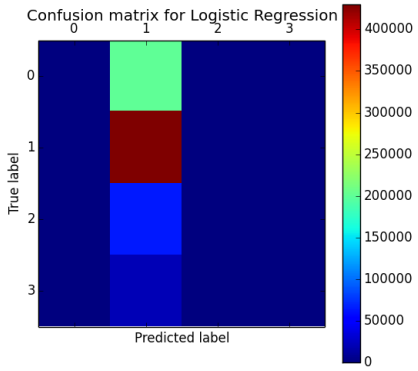
1) Baseline 1: Randomly predict from distribution
This method has an accuracy of 60.02%. This is because $$ price range restaurants make 60% of dataset.



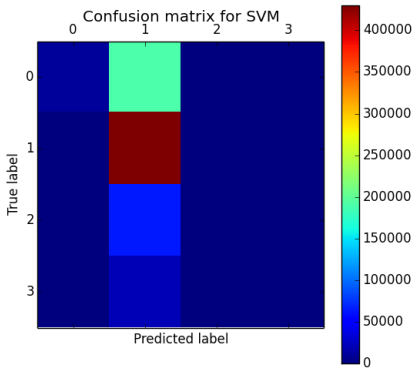
2) Baseline 2: Predict next price range solely based on current price range
This method has an accuracy of 50.44%. The distribution of predictions can be seen from the Confusion Matrix,




3) Logistic Regression
Logistic Regression doesn't perform well. The data overfits, and it always predicts $$. The accuracy of this model is 60.02%.




4) Support Vector Machine
Support Vector Machine performs slightly better than the other models. The accuracy of this model is 61.5%.




Overall, the data overfits to a good extent. This is probably because of the presence of noise in reviews (as compared to Check-ins) and lack of $$$ and $$$$ data. SVM performs better than Logistic Regression probably because of a non-linear separating hyperplane.

In conclusion, the model constructed does better than the two baselines, but by a very small margin.

## ACKNOWLEDGMENT

I would like to thank Prof McAuley for his guidance and suggestions.

## REFERENCES

[1] Kevin Reschke, Adam Vogel, and Dan Jurafsky. Generating recommendation dialogs by extracting information from user reviews. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 499-504, Sofia, Bulgaria, August 2013. Association for Computational Linguistics

[2] Chahuneau, Victor, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, and Noah A. Smith. "Word salad: Relating food prices and descriptions." In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1357-1367. Association for Computational Linguistics, 2012.