### **Project Problem and Hypothesis**

* My project involves finding out if I can predict the ratings for a Yelp restaurant based on the features provided in the Yelp dataset. Are there certain attributes which are more associated with a successful rating?
  + For this, the value I’m predicting is a categorical number (from 1 to 5, in 0.5 increments). I am essentially solving a classification problem.
* Secondly, I’d want to find out if people in different cities within a state value different things in their eating establishments. For example, does a 5-star restaurant in certain parts of Phoenix have different features on average than a 5-star restaurant in certain Scottsdale?
* I think this kind of analysis would be interesting for restaurants deciding where to open their businesses or what features to include when opening a business in a particular location.

### **Datasets**

* My data has 72,742 rows (each a different business) and 89 different columns (various characteristics of the business as provided by different Yelp fields).
* Of these rows, only 22,264 are restaurants, which is the subset I want to focus in on.
* This set includes information about local businesses in 11 cities across 4 countries.
* A number of fields have a great deal of null values (for example, not every restaurant has the field “Smoking” filled out). One of the things I will have to think about is how to deal with these null values.
* Data table on last page.

### **Domain knowledge**

* I do not have much experience in analyzing this kind of data (or thinking about the Yelp business model). However, the Yelp dataset challenge has been around for a while, and many projects have been done using this data.
* For example, below are a couple projects/conclusions related to this dataset that I found online already:
  + Ratings for food quality and service tend to be positively correlated
  + Restaurants are rated lower by 2/5ths of a star when the reviewer visited during busiest time period

### **Project Concerns**

* I am still unsure about what the best model is to use for my classification problem: whether it’s a logistic regression or random forest model. It seems that random forest models are better than logistic regressions when dealing with non-linearity, and in this case I have a lot of binary independent features, so a random forest model might be best.
* So far, I’m not planning on using the actual text reviews that users leave on Yelp pages, which could also be an interesting area of study. For example, it’d be interesting to see what words/phrases are most associated with good/bad reviews. If I end up having extra time, this would be an interesting additional avenue to pursue.
* Risks to the project: If my model is incorrect about what features are associated with positive/negative ratings and how important these features are, then restaurant owners would be led astray in thinking about their business models.
* Assumptions/Caveats: I’m also not entirely sure how complete the dataset I am working with (unclear exactly over what time period it’s been collected and if it’s every Yelp review that is in that time period or just a sample).

### **Outcomes**

* I expect the output to show me either an R^2 (if I end up using a logistic regression) or an Area Under the Curve (if I use a random forest method) to show me how good of a fit the model is. More importantly, I want to see which features are the most important in predicting the rating, which is the part that will be most interesting for my target audience.
* I doubt there will be just one feature will very large predictive power on its own, but rather I expect a combination of features to be important.
* If my project is a bust, then I’ll try to either ask a different question or approach the problem in a different way. There is a lot of data in this dataset, so I believe there can be some interesting conclusions drawn, even if it happens to not be with my current questions.

**Data Table:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| business\_id | integer (0 or 1) | business\_id |
| city | string | city |
| full\_address | string | full\_address |
| latitude | double | latitude |
| longitude | double | longitude |
| name | string | name of restaurant |
| review\_count | integer | number of reviews |
| stars | double (rounded to nearest 0.5) | rating of the restaurant |
| state | string | state |
| type | string | all of these are labeled as business |
| Sunday\_Open | double (time) | Sunday\_Open |
| Sunday\_Close | double (time) | Sunday\_Close |
| Monday\_Open | double (time) | Monday\_Open |
| Monday\_Close | double (time) | Monday\_Close |
| Tuesday\_Open | double (time) | Tuesday\_Open |
| Tuesday\_Close | double (time) | Tuesday\_Close |
| Wednesday\_Open | double (time) | Wednesday\_Open |
| Wednesday\_Close | double (time) | Wednesday\_Close |
| Thursday\_Open | double (time) | Thursday\_Open |
| Thursday\_Close | double (time) | Thursday\_Close |
| Friday\_Open | double (time) | Friday\_Open |
| Friday\_Close | double (time) | Friday\_Close |
| Saturday\_Open | double (time) | Saturday\_Open |
| Saturday\_Close | double (time) | Saturday\_Close |
| Noise Level | string | Average, Quiet, Loud, Very Loud |
| Attire | string | Casual, Dressy, Formal |
| Alcohol | string | Full Bar, Beer and Wine, None |
| Price\_Range | integer (1 to 4) | 1 (cheap) to 4 (expensive) |
| Delivery | integer (0 or 1) | 0 = N, 1 = Y |
| Outdoor\_Seating | integer (0 or 1) | 0 = N, 1 = Y |
| Drive-Thru | integer (0 or 1) | 0 = N, 1 = Y |
| Good\_for\_Groups | integer (0 or 1) | 0 = N, 1 = Y |
| Has\_TV | integer (0 or 1) | 0 = N, 1 = Y |
| Caters | integer (0 or 1) | 0 = N, 1 = Y |
| Waiter\_Service | integer (0 or 1) | 0 = N, 1 = Y |
| Good\_for\_Kids | integer (0 or 1) | 0 = N, 1 = Y |
| Accepts\_Credit\_Cards | integer (0 or 1) | 0 = N, 1 = Y |
| Takes\_Reservations | integer (0 or 1) | 0 = N, 1 = Y |
| Wi\_Fi | string | No, Free, Paid |
| Happy\_Hour | integer (0 or 1) | 0 = N, 1 = Y |
| Good\_for\_Dancing | integer (0 or 1) | 0 = N, 1 = Y |
| Smoking | integer (0 or 1) | Outdoor, No, Yes |
| BYOB | integer (0 or 1) | 0 = N, 1 = Y |
| Corkage | integer (0 or 1) | 0 = N, 1 = Y |
| Take\_Out | integer (0 or 1) | 0 = N, 1 = Y |
| Coat\_Check | integer (0 or 1) | 0 = N, 1 = Y |
| Parking\_Street | integer (0 or 1) | 0 = N, 1 = Y |
| Parking\_Valet | integer (0 or 1) | 0 = N, 1 = Y |
| Parking\_Lot | integer (0 or 1) | 0 = N, 1 = Y |
| Parking\_Garage | integer (0 or 1) | 0 = N, 1 = Y |
| Parking\_Validated | integer (0 or 1) | 0 = N, 1 = Y |
| Music\_DJ | integer (0 or 1) | 0 = N, 1 = Y |
| Music\_Karaoke | integer (0 or 1) | 0 = N, 1 = Y |
| Music\_Video | integer (0 or 1) | 0 = N, 1 = Y |
| Music\_Live | integer (0 or 1) | 0 = N, 1 = Y |
| Music\_Jukebox | integer (0 or 1) | 0 = N, 1 = Y |
| Music\_Background\_Music | integer (0 or 1) | 0 = N, 1 = Y |
| Is\_Restaurants | integer (0 or 1) | 0 = N, 1 = Y (is it a restaurant) |
| Sandwiches | integer (0 or 1) | 0 = N, 1 = Y |
| Fast Food | integer (0 or 1) | 0 = N, 1 = Y |
| Nightlife | integer (0 or 1) | 0 = N, 1 = Y |
| Pizza | integer (0 or 1) | 0 = N, 1 = Y |
| Bars | integer (0 or 1) | 0 = N, 1 = Y |
| Mexican | integer (0 or 1) | 0 = N, 1 = Y |
| Food | integer (0 or 1) | 0 = N, 1 = Y |
| American (Traditional) | integer (0 or 1) | 0 = N, 1 = Y |
| Burgers | integer (0 or 1) | 0 = N, 1 = Y |
| Chinese | integer (0 or 1) | 0 = N, 1 = Y |
| Italian | integer (0 or 1) | 0 = N, 1 = Y |
| American (New) | integer (0 or 1) | 0 = N, 1 = Y |
| Breakfast & Brunch | integer (0 or 1) | 0 = N, 1 = Y |
| Thai | integer (0 or 1) | 0 = N, 1 = Y |
| Indian | integer (0 or 1) | 0 = N, 1 = Y |
| Sushi Bars | integer (0 or 1) | 0 = N, 1 = Y |
| Korean | integer (0 or 1) | 0 = N, 1 = Y |
| Mediterranean | integer (0 or 1) | 0 = N, 1 = Y |
| Japanese | integer (0 or 1) | 0 = N, 1 = Y |
| Seafood | integer (0 or 1) | 0 = N, 1 = Y |
| Middle Eastern | integer (0 or 1) | 0 = N, 1 = Y |
| Pakistani | integer (0 or 1) | 0 = N, 1 = Y |
| Barbeque | integer (0 or 1) | 0 = N, 1 = Y |
| Vietnamese | integer (0 or 1) | 0 = N, 1 = Y |
| Asian Fusion | integer (0 or 1) | 0 = N, 1 = Y |
| Diners | integer (0 or 1) | 0 = N, 1 = Y |
| Greek | integer (0 or 1) | 0 = N, 1 = Y |
| Vegetarian | integer (0 or 1) | 0 = N, 1 = Y |
| Number\_of\_Checkins | integer | Number\_of\_Checkins |
| Number\_of\_Tips | integer | Number\_of\_Tips |
| Number\_of\_Tip\_Like | integer (0 to 18) | Number\_of\_Tip\_Like |