# Predicting Yelp Rating

A guide by Josh, Ian, Ryan and Douglas

# Topic: Yelp Review Predictor

- Testing our predictive model on  Yelp reviews. Can we predict a Yelp star rating based on various attributes specific to the restaurant?

- What attributes would customer (reviewers) consider most important to their rating? Does it depend on the type of restaurant?

# Steps in the Process

- Gather Data
- Import Json into Jupyter Notebook
- Drop unwanted columns/features, filtered data by category, refined to OH
- Factorized attribute columns to break object into separate individual feature strings
- Convert data to binary via out of the box get dummies function to prep for machine learning
- Trained data, and ran multiple different models to see best fit
- Random Forest ftw
- Compared Actual vs. Predicted Yelp stars
- Took a look at feature importance

# Cleaning the Dataset

Step 1: remove all businesses that are not restaurants

```
df_new = data[data["categories"].str.contains("Restaurants", na = False)]
df_new.dropna(subset = ['attributes'], inplace = True)
df_new.head()
```

Step 2: include only restaurants in Ohio

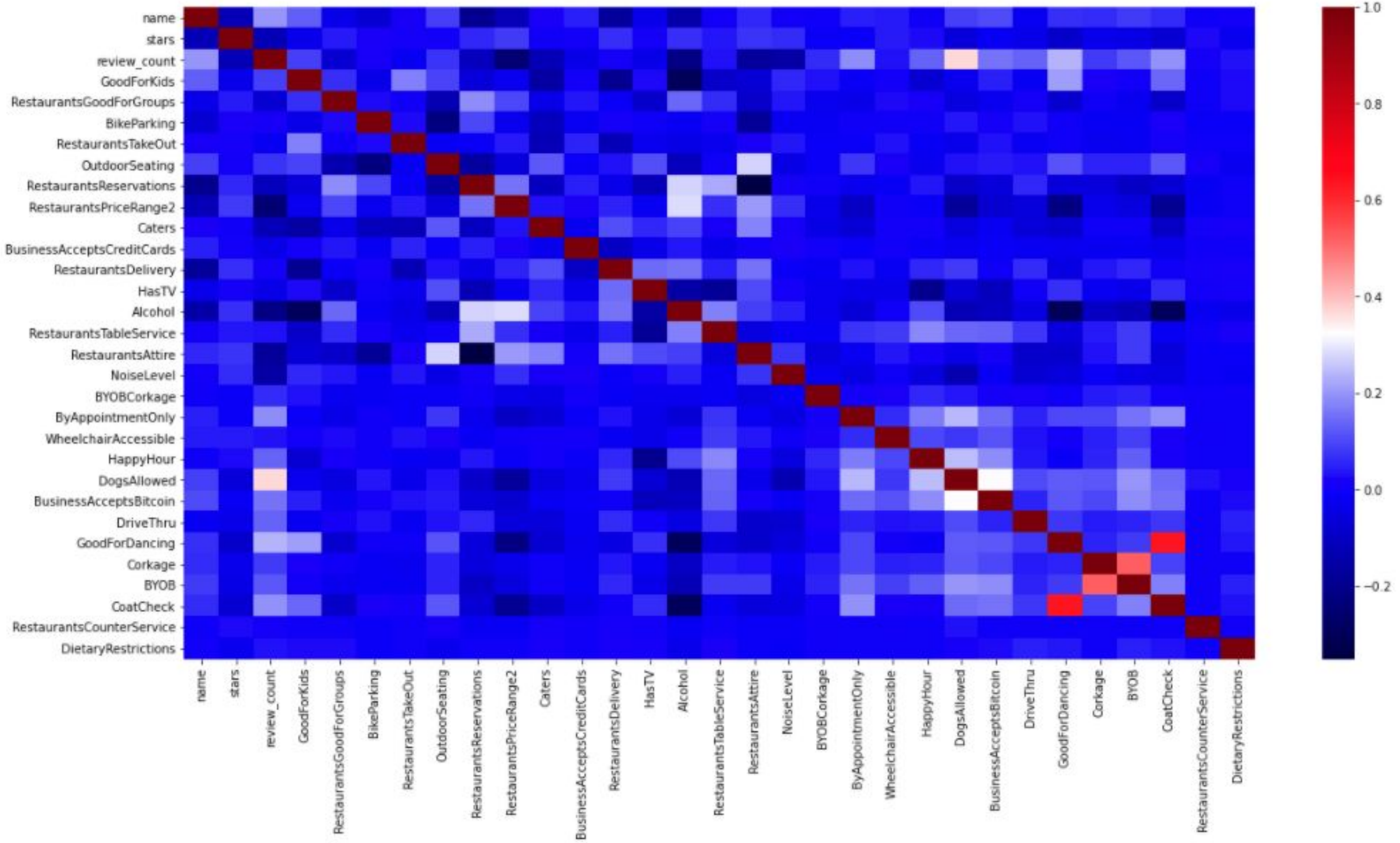
```
df_ohio = df_new[df_new['state']== 'OH']
df_ohio
```

Step 3: cleaning and restructuring columns

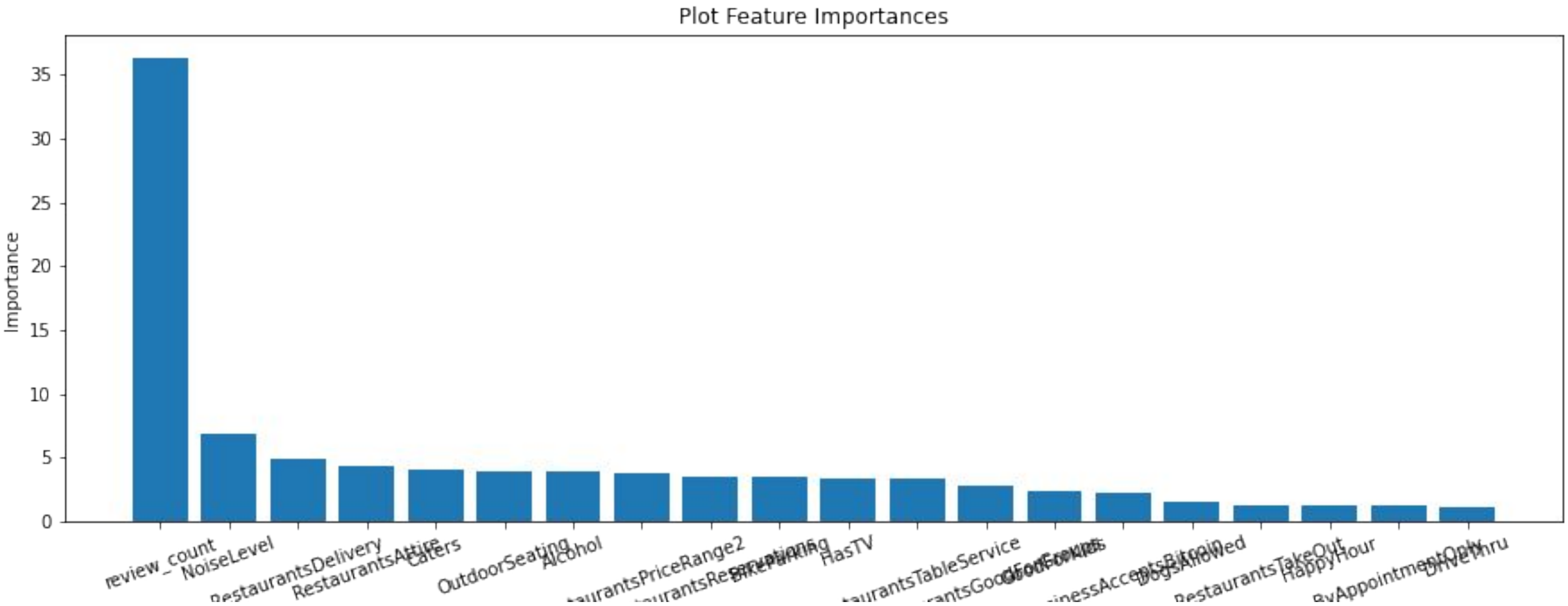
```
df_test = df_ohio
df_test['attributes'] = df_test['attributes'].astype('str')
df_test['attributes'] = df_test['attributes'].apply(lambda x : dict(eval(x)) )
df_test2 = df_test["attributes"].apply(pd.Series)

df_test2
```

```
df_test3['GoodForKids'] = pd.get_dummies(df_test2['GoodForKids'])
df_test3['RestaurantsGoodForGroups'] = pd.get_dummies(df_test2['RestaurantsGoodForGroups'])
```

# Feature Importance Heatmap



# Feature Importance Plot

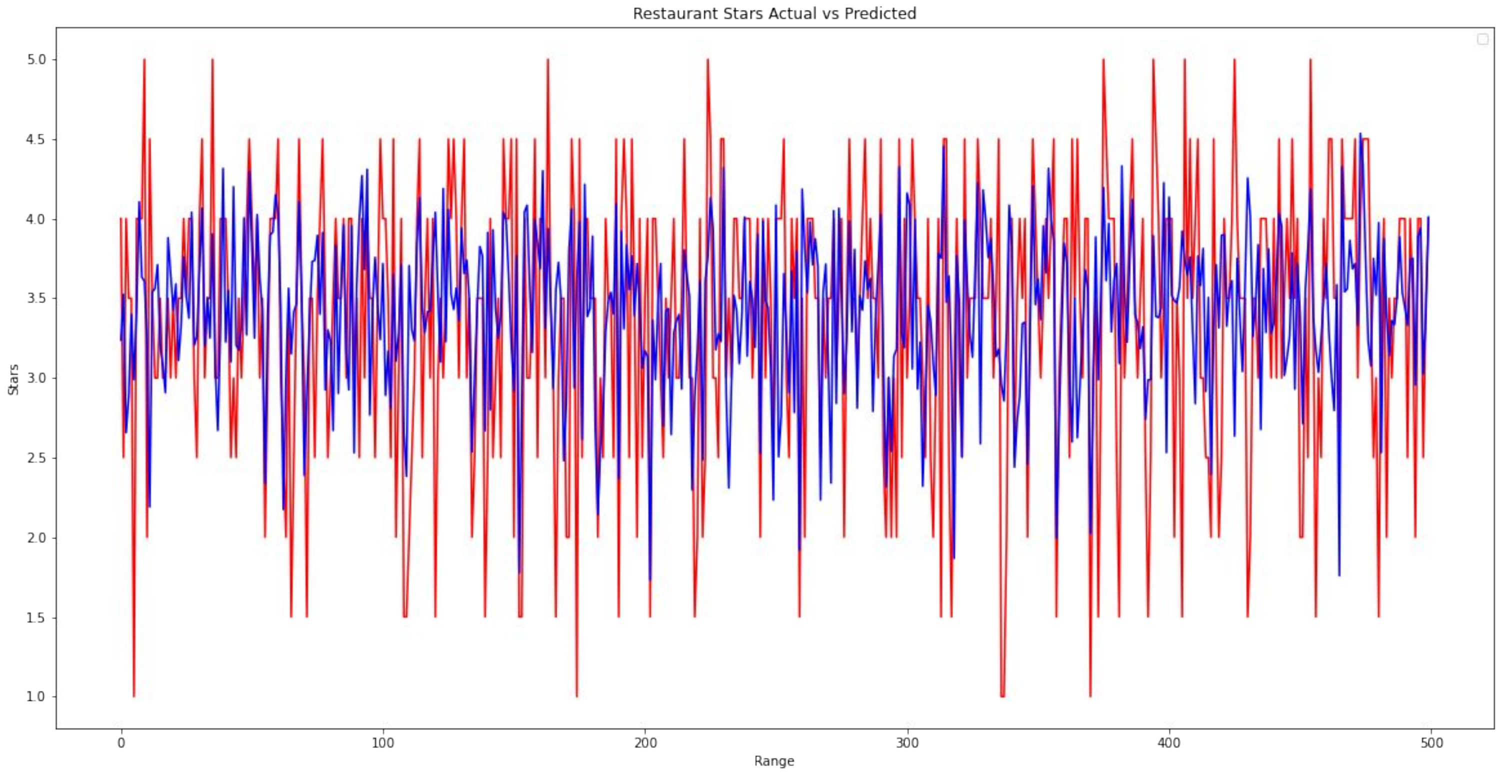


# Feature Importance Ratings

| Feature | Importance |
|---|---|
| review_count | 0.362741 |
| NoiseLevel | 0.068062 |
| RestaurantsDelivery | 0.049246 |
| RestaurantsAttire | 0.043898 |
| Caters | 0.040941 |
| OutdoorSeating | 0.039555 |
| Alcohol | 0.039482 |
| RestaurantsPriceRange2 | 0.037685 |
| RestaurantsReservations | 0.035099 |
| BikeParking | 0.034689 |
| HasTV | 0.034134 |
| RestaurantsTableService | 0.033454 |
| RestaurantsGoodForGroups | 0.028092 |
| GoodForKids | 0.023478 |

| | |
|---|---|
| BusinessAcceptsBitcoin | 0.022882 |
| DogsAllowed | 0.015573 |
| RestaurantsTakeOut | 0.013263 |
| HappyHour | 0.013102 |
| ByAppointmentOnly | 0.013047 |
| DriveThru | 0.011705 |
| CoatCheck | 0.010056 |
| BYOB | 0.008278 |
| WheelchairAccessible | 0.007128 |
| GoodForDancing | 0.004966 |
| BYOBCorkage | 0.004684 |
| BusinessAcceptsCreditCards | 0.002469 |
| Corkage | 0.002293 |
| RestaurantsCounterService | 0.000000 |
| DietaryRestrictions | 0.000000 |

# Yelp Stars, Actual (red) vs Predicted (blue)

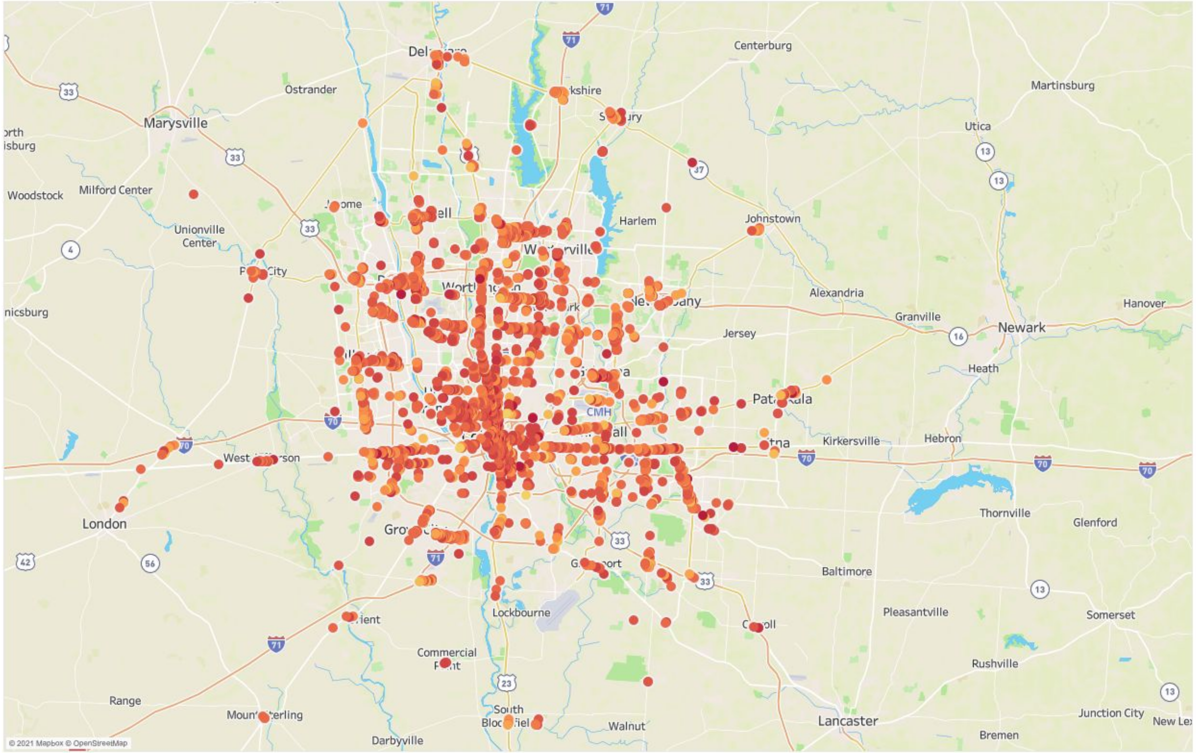


# Actual vs. Predicted Ratings

| Actual | Predicted |
|---|---|
| 4.0 | 3.237500 |
| 2.5 | 3.525000 |
| 4.0 | 2.656167 |
| 3.5 | 2.885000 |
| 3.5 | 3.399167 |
| 1.0 | 2.988085 |
| 4.0 | 3.422500 |
| 4.0 | 4.105000 |
| 4.0 | 3.633119 |
| 5.0 | 3.606628 |
| 2.0 | 3.237229 |
| 4.5 | 2.189929 |
| 3.5 | 3.540000 |
| 3.0 | 3.560000 |
| 3.0 | 3.710917 |
| 3.5 | 3.200000 |
| 3.0 | 3.080000 |
| 3.0 | 2.907500 |
| 3.5 | 3.880000 |
| 3.0 | 3.655000 |
| 3.5 | 3.426167 |
| 3.0 | 3.590000 |
| 3.5 | 3.110000 |
| 3.5 | 3.330000 |
| 4.0 | 3.760000 |
| 3.5 | 3.510000 |
| 4.0 | 3.375000 |
| 4.0 | 4.040000 |
| 3.0 | 3.205000 |
| 2.5 | 3.260000 |
| 4.0 | 3.678750 |
| 4.5 | 4.065000 |
| 3.0 | 3.200000 |
| 3.5 | 3.505000 |
| 3.5 | 3.250667 |
| 5.0 | 3.905000 |
| 3.0 | 3.075000 |
| 3.0 | 2.670000 |
| 4.0 | 3.310000 |
| 4.0 | 4.315000 |

# Yelp Stars, Ohio Dataset

Map - Star Ratings



# 1. Challenges/Limitations

➔ **Finding small business dataset**
Obtaining actual restaurant data.

➔ **A  Lack of Restaurant Data**
The industry is very inept when it comes to data collection.

➔ **Breaking up feature object into separate strings**
Out of the box factorize pandas function

➔ **Dealing with null / naan values**
Treating them as false, rather than dropping or true.

# 2. If we had more time...

- ➔ **Utilize web scraping and NLP**
  Adding further dimensions and accuracy to predictions
- ➔ **Take a zoomed look at each feature**
  Histograms per feature, determine actual feature importance
- ➔ **Expand analysis to identify biases**
  Include intangible features and expand data set.

# Conclusion

# Questions ?

LET'S GO BROWNS!