Yelping Around

Group 2

Ishan Shah, Svetak Sundhar, Katherine Wang,
Jessica Ho, Alex Kan

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

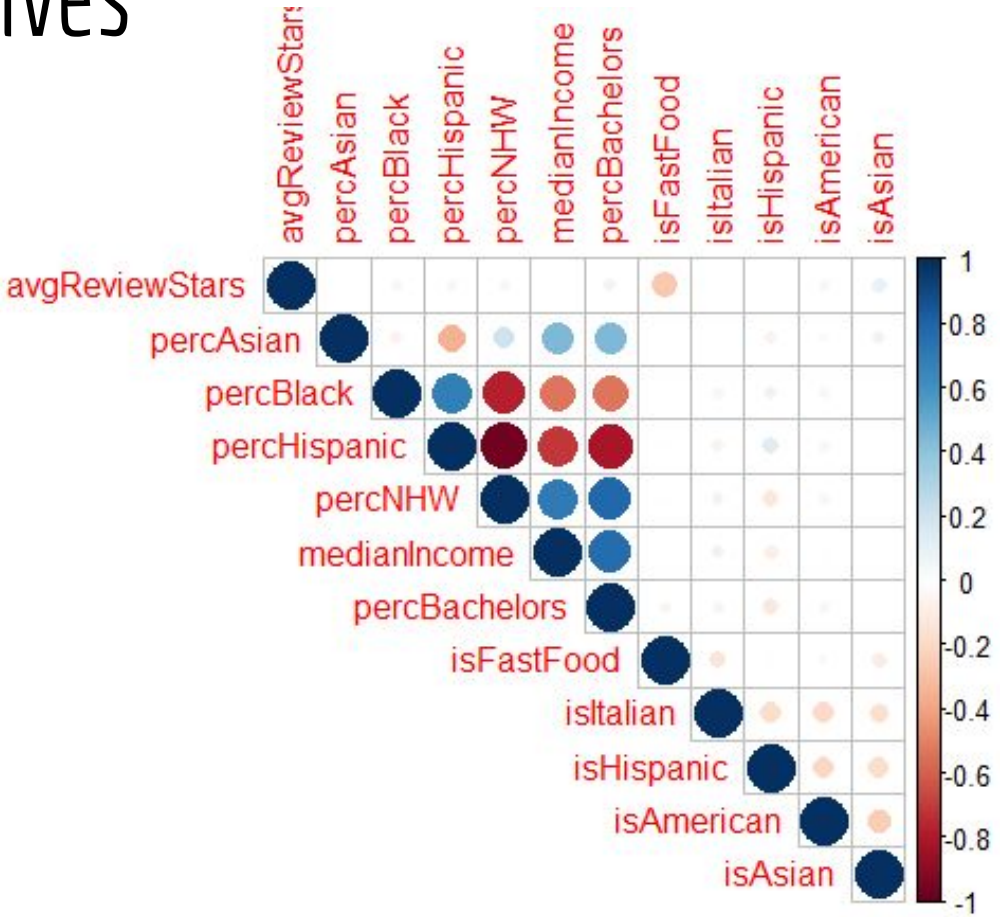
Yelp?

- Tons of places to build a restaurant – why not use Yelp to help?
- What attributes would you need to be successful in a specific geographic location?
  - What demographics should you target?
  - How would you expect to be rated by customers?
  - How could you predict your number and quality of review??

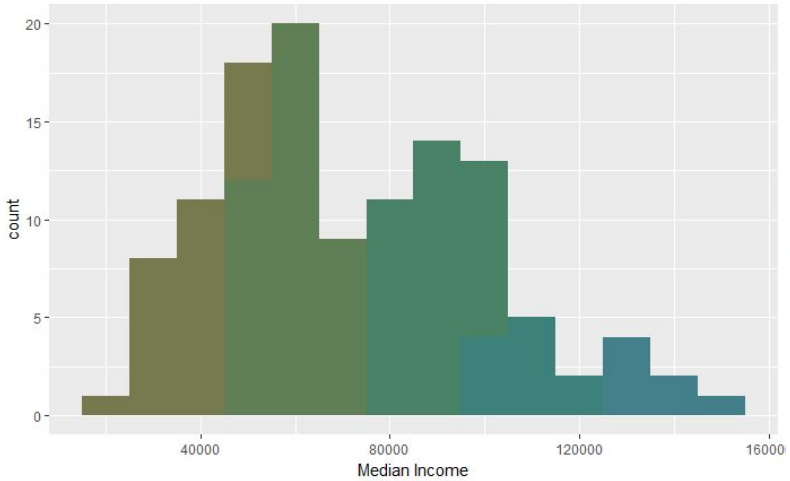
The Data

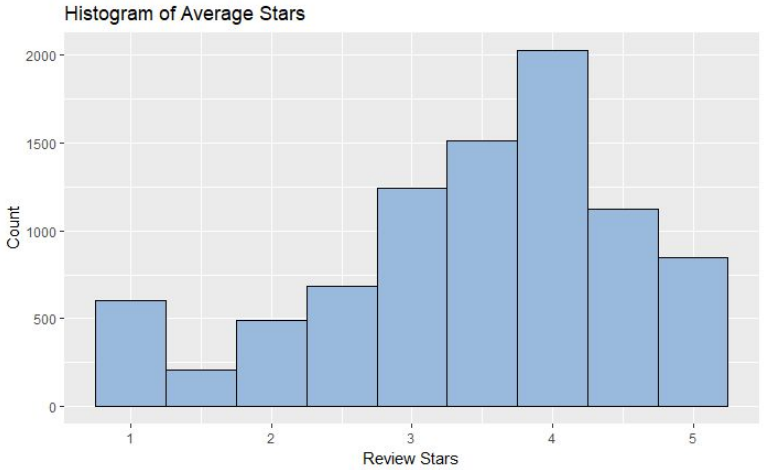
- Yelp Challenge Dataset
  - Tables for all businesses, users, and reviews
- Census Demographic Data
  - Racial distribution, median income, age distribution, and education level by ZIP Code
- Pre-Processing
  - Filtered out restaurants within a 30 mile radius of Phoenix downtown
  - Aggregated by business ID
  - Focused on average star review, average Yelp age of reviewers, characteristics of the reviews, and dummy variables for types of food served
  - Combined both datatables – 8,700 rows, 49 variables

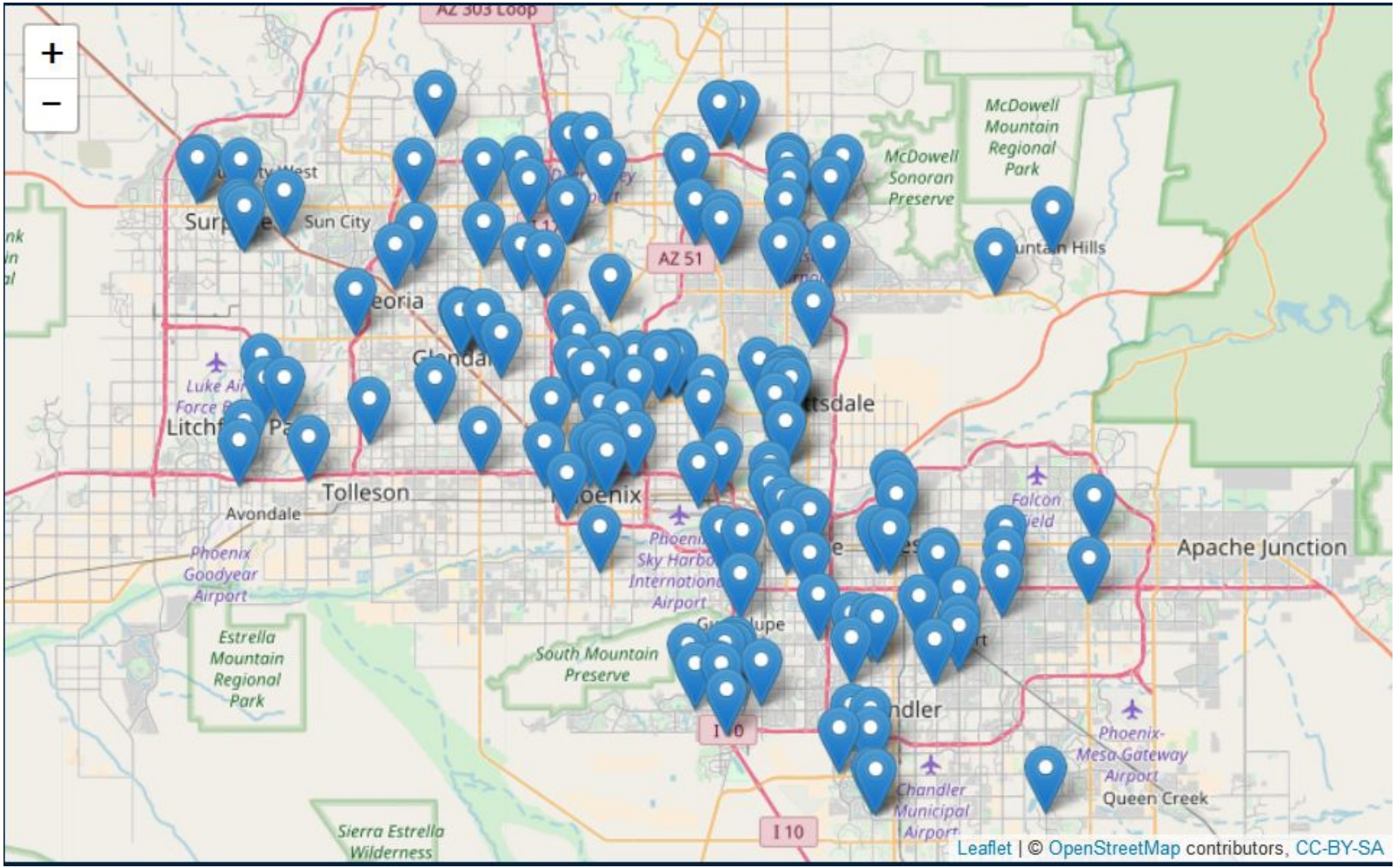
Descriptives

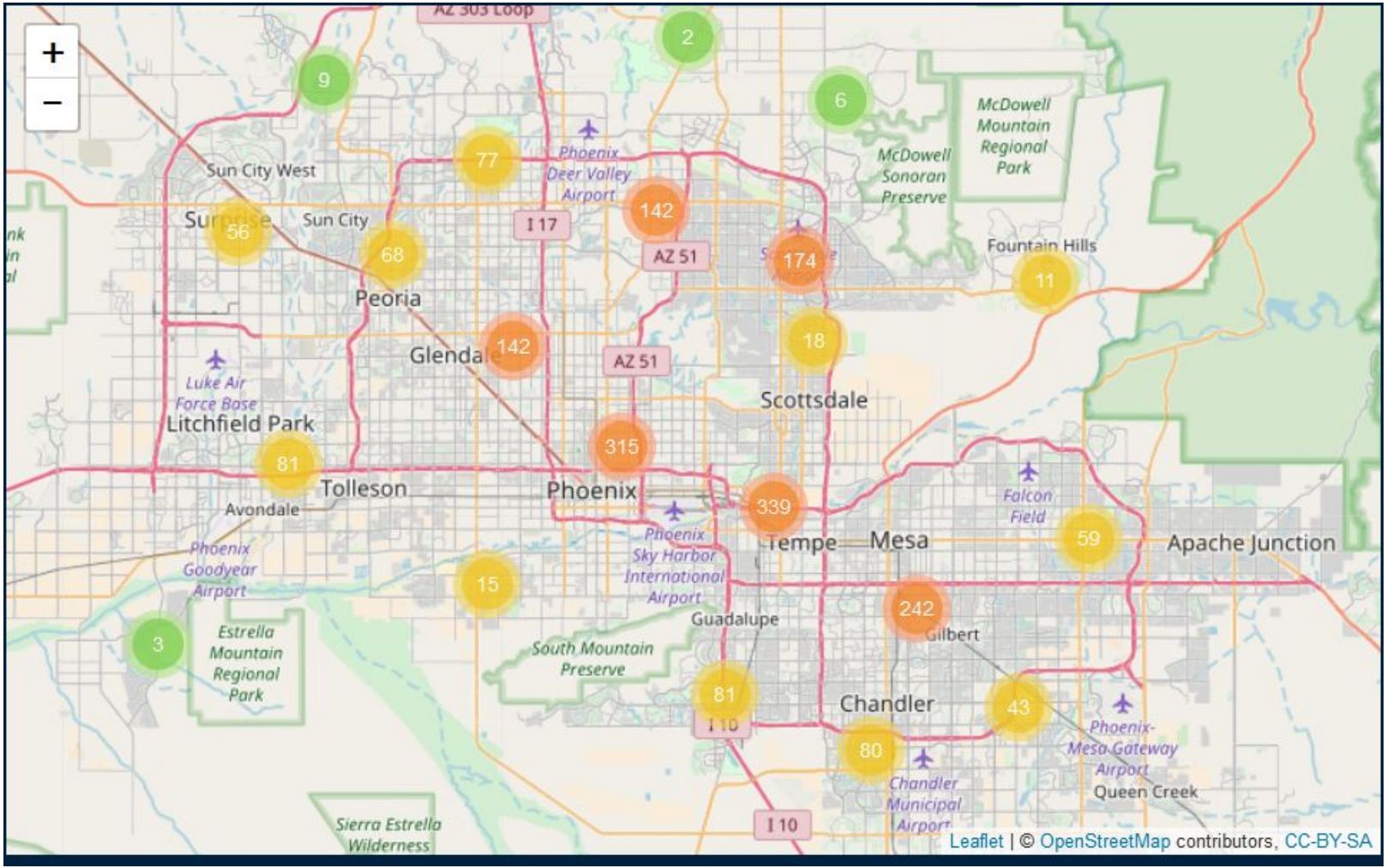


Descriptives






AZ 303 Loop
McDowell Mountain Regional Park
McDowell Sonoran Preserve
Sun City
AZ 51
Peoria
Luke Air Force Base
Litchfield Park
Tolleson
Avondale
Phoenix Goodyear Airport
Phoenix Sky Harbor International Airport
South Mountain Preserve
Estrella Mountain Regional Park
Sierra Estrella Wilderness
Apache Junction
Falcon Field
Phoenix-Mesa Gateway Airport
Chandler Municipal Airport
Queen Creek
I 10
Leaflet | © OpenStreetMap contributors, CC-BY-SA

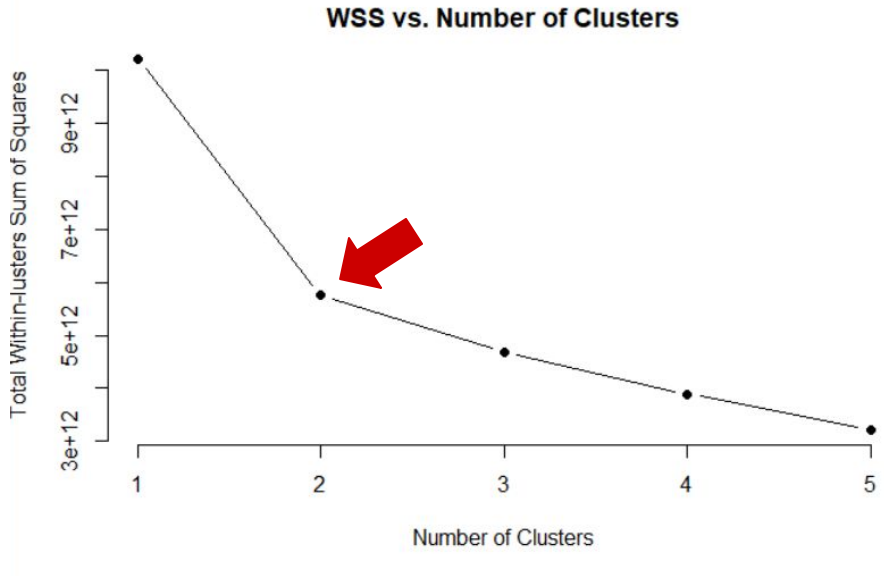

AZ 303 Loop
2
9
6
McDowell Mountain Regional Park
McDowell Sonoran Preserve
77
Phoenix Deer Valley Airport
Sun City West
Sun City
142
I 17
AZ 51
56
68
174
Fountain Hills
11
Peoria
18
Glendale
142
AZ 51
Luke Air Force Base
Litchfield Park
Scottsdale
315
81
Tolleson
Phoenix
339
Falcon Field
Avondale
Phoenix Goodyear Airport
Phoenix Sky Harbor International Airport
Tempe
Mesa
59
Apache Junction
15
242
Guadalupe
Gilbert
3
Estrella Mountain Regional Park
South Mountain Preserve
81
I 10
Chandler
43
Phoenix-Mesa Gateway Airport
80
Chandler Municipal Airport
Queen Creek
Sierra Estrella Wilderness
I 10
Leaflet | © OpenStreetMap contributors, CC-BY-SA

So now...the Machine Learning

- K-means Clustering
- Classification – two classes (≤ 3.5 stars vs. > 3.5 stars)
  - Logistic Regression
  - Quadratic Discriminant Analysis
  - Naive Bayes Classification
- Decision Trees

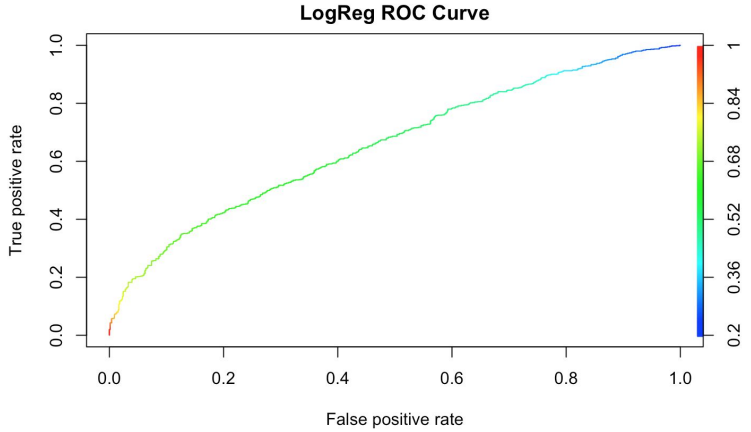
K-means Clustering

- Optimal number of clusters: $K = 2$
- Reinforces decision to use two classes for classification model

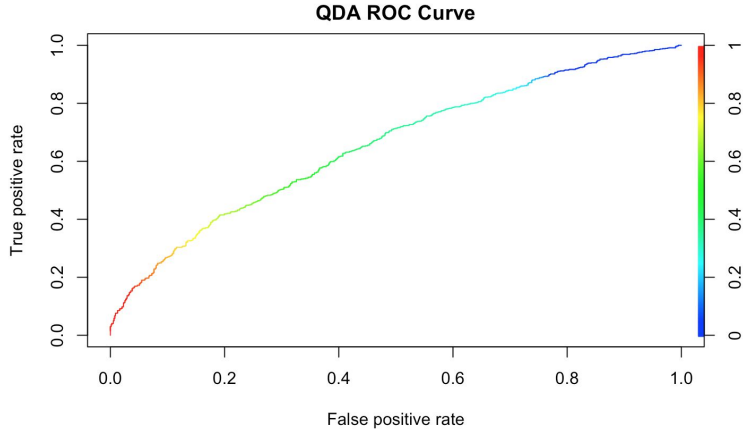


Logistic Regression & QDA

- Logistic Regression
  - Overall lower test error
  - Only correctly predicted **38.8%** of poorly-rated restaurants, but correctly predicted **79.1%** of well-rated restaurants

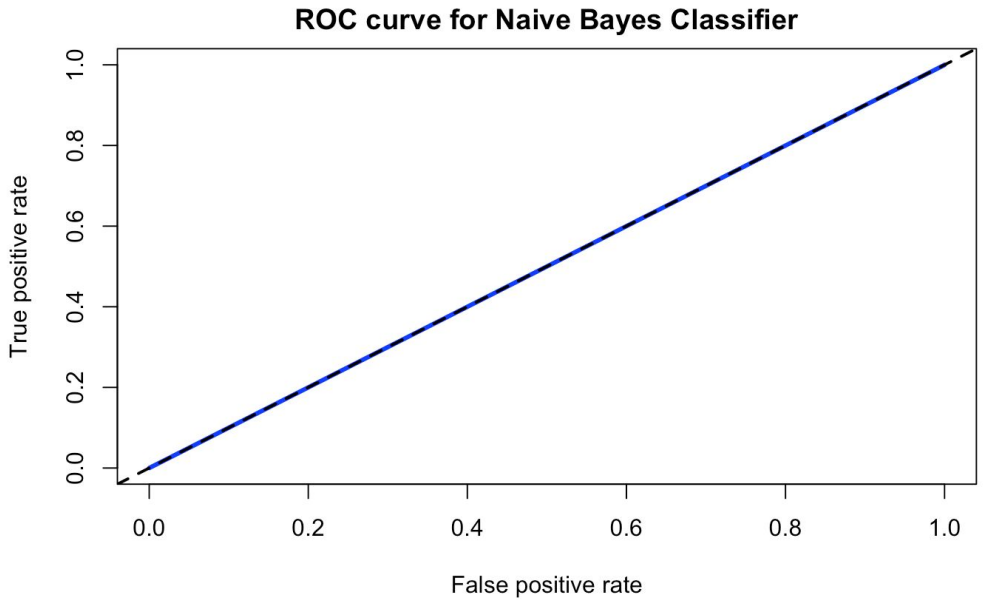


- QDA
  - Slightly higher test error
  - Better at predicting poorly-rate restaurants (**65.8%**), but not well-rated restaurants (**54.2%**)



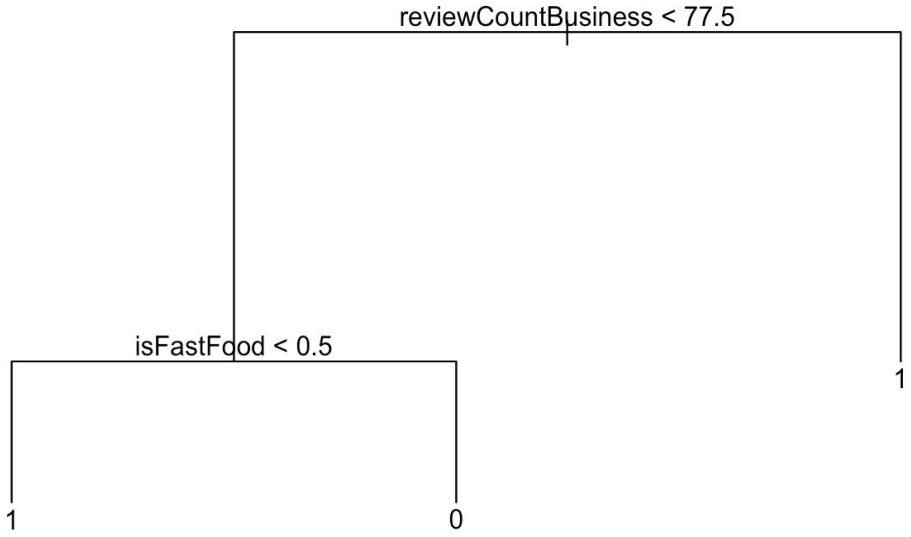
Naive Bayes for Classification

- Error rate is a little better than random guessing
- Incorrectly predicted poorly-rated
- AUC value: 0.5
- Picks majority class



Decision Trees

- Tree splits into only 3 nodes without the need for pruning
- Easy to predict rating based upon review count and whether or not the business is fast food
- Random forest model also clustered to majority class



Results

Technique	Test Error	Comments
Logistic Regression	0.3899371	• Predicted highly-rated restaurants better
QDA	0.4059463	• Predicted poorly-rated restaurants better
Naive Bayes	0.4894225	• Predicted all as majority classifier

Conclusion & Future Work

- Every city is different
  - Would need to run models on every city to get most accurate results
- Naive Bayes did not perform well
  - Sentiment analysis could yield more insight into classification
- Important variables:
  - *reviewCountBusiness, avgDaysSinceJoined, isFastFood, isAsian, isAmerican, isBreakfast*
- More specific business attributes & demographics
- Define a better metric of success rather than star rating

Got questions?