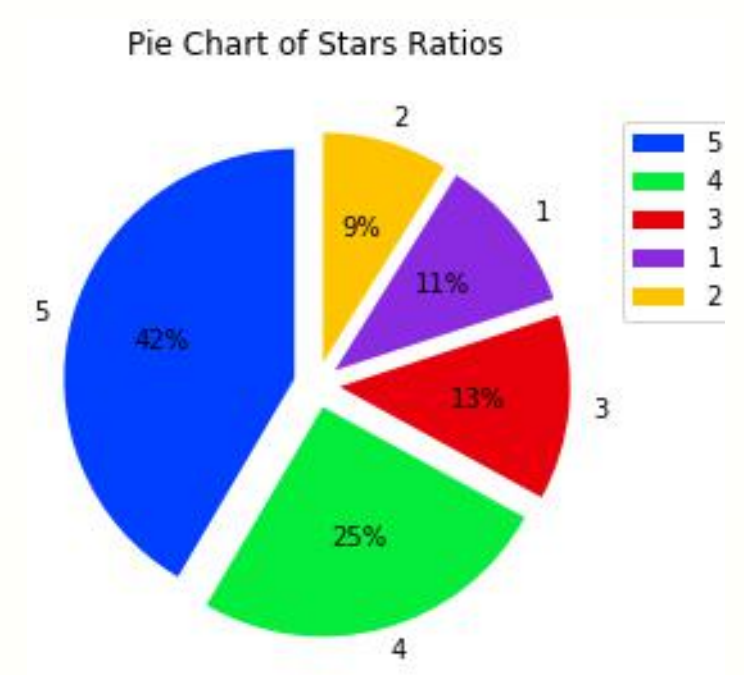
Modeling Customer Satisfaction from Yelp Data

Li Liu
M.A. Student in Computational Social Science, The University of Chicago
Email: liu431@uchicago.edu; GitHub: liu431; Date: 6/5/2019

Research Quesion

- Marketing research: measure the heterogeneity in customer satisfaction by surveys, focus groups, etc
- Problem: low-response rate; cost time and money; not scalable; sampling error; missing data...
- Opportunities: large data on customers' behavior and machine learning methods
- **Question: What are the determinants of customer satisfaction in the restaurant industry?**

Independent Variable



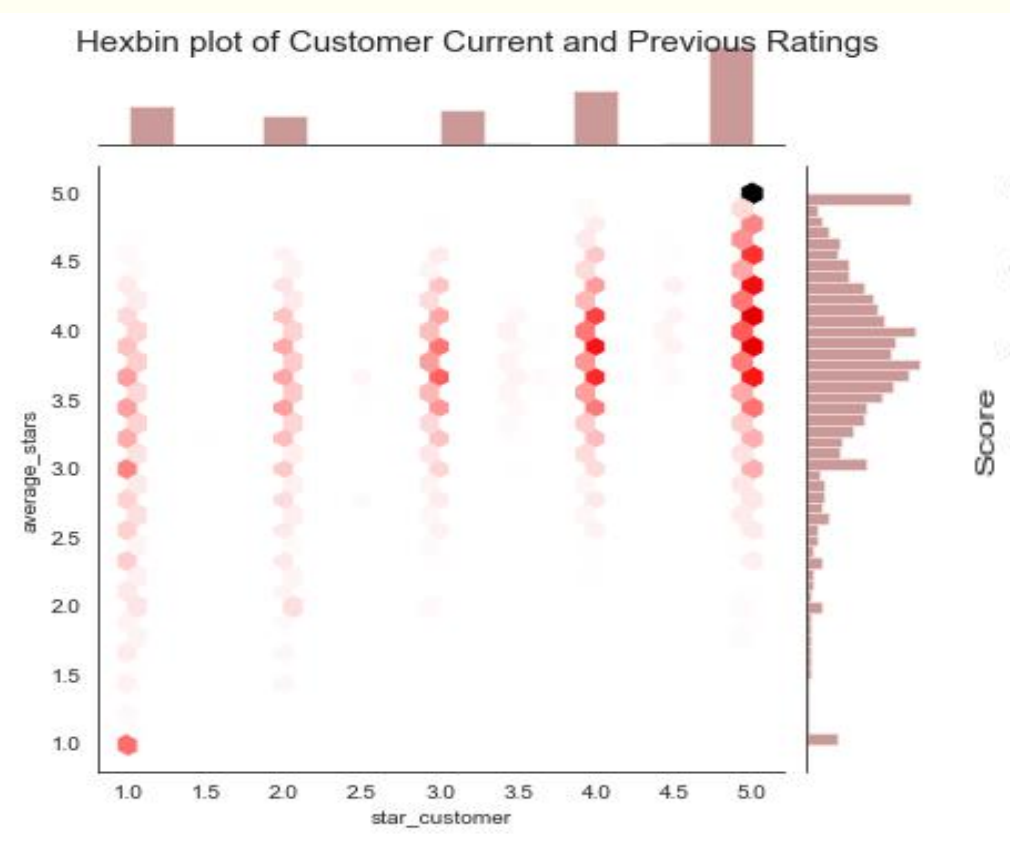
- Start with customsers' rating for the restaurants
- Level: 1~5 stars
- 67% are 'satisfied'

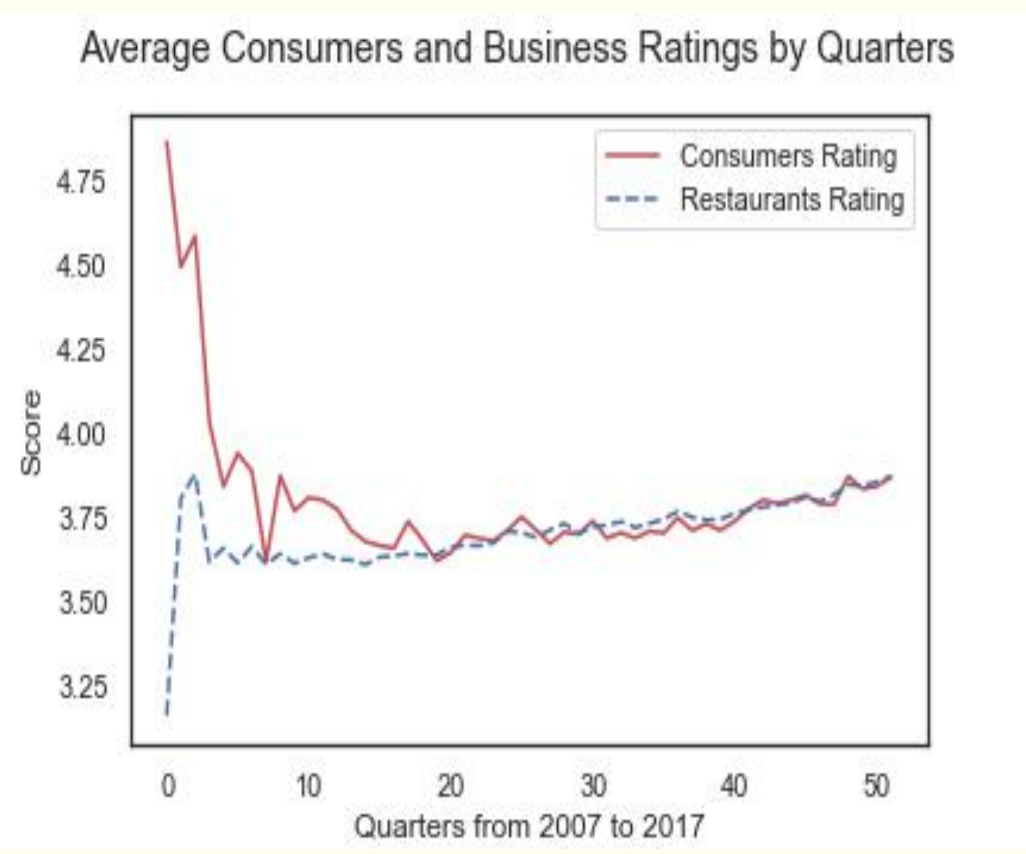
Pbm 1: Customers have different grading scale

Pbm 2: Data spans 10 years. Non-stationary ratings. Time trend in ratings.

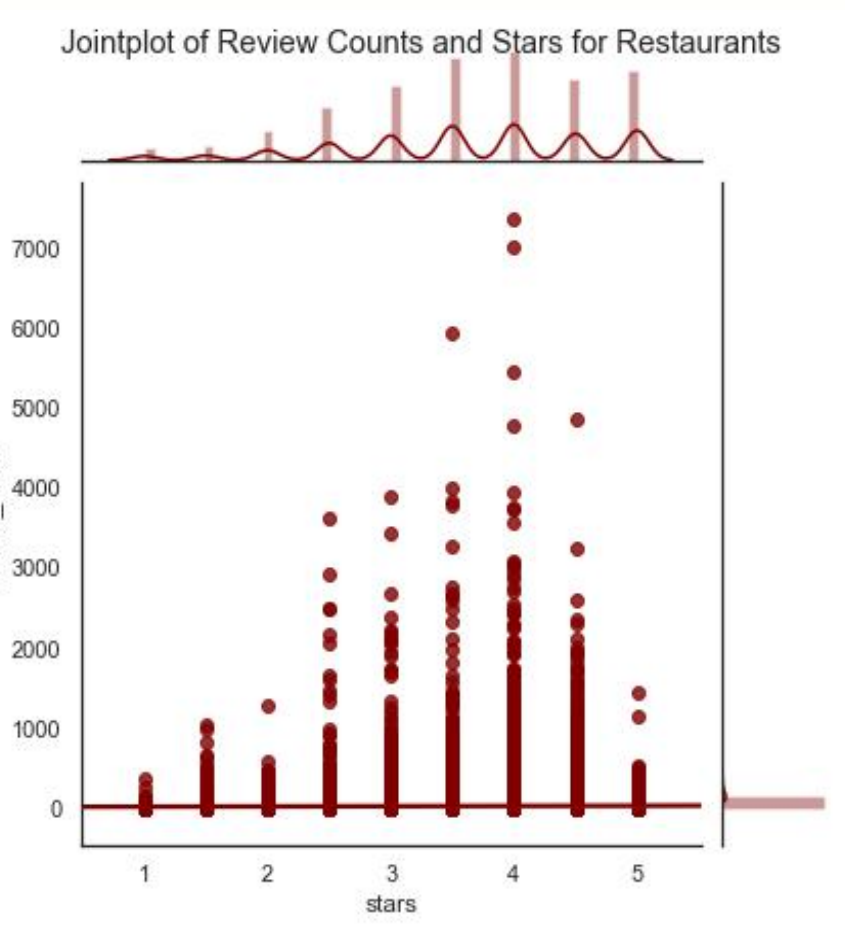
Sol to 1: Reweighed current ratings by accounting for the previous average ratings

Sol to 2: Detrending by accounting for the average growth rate

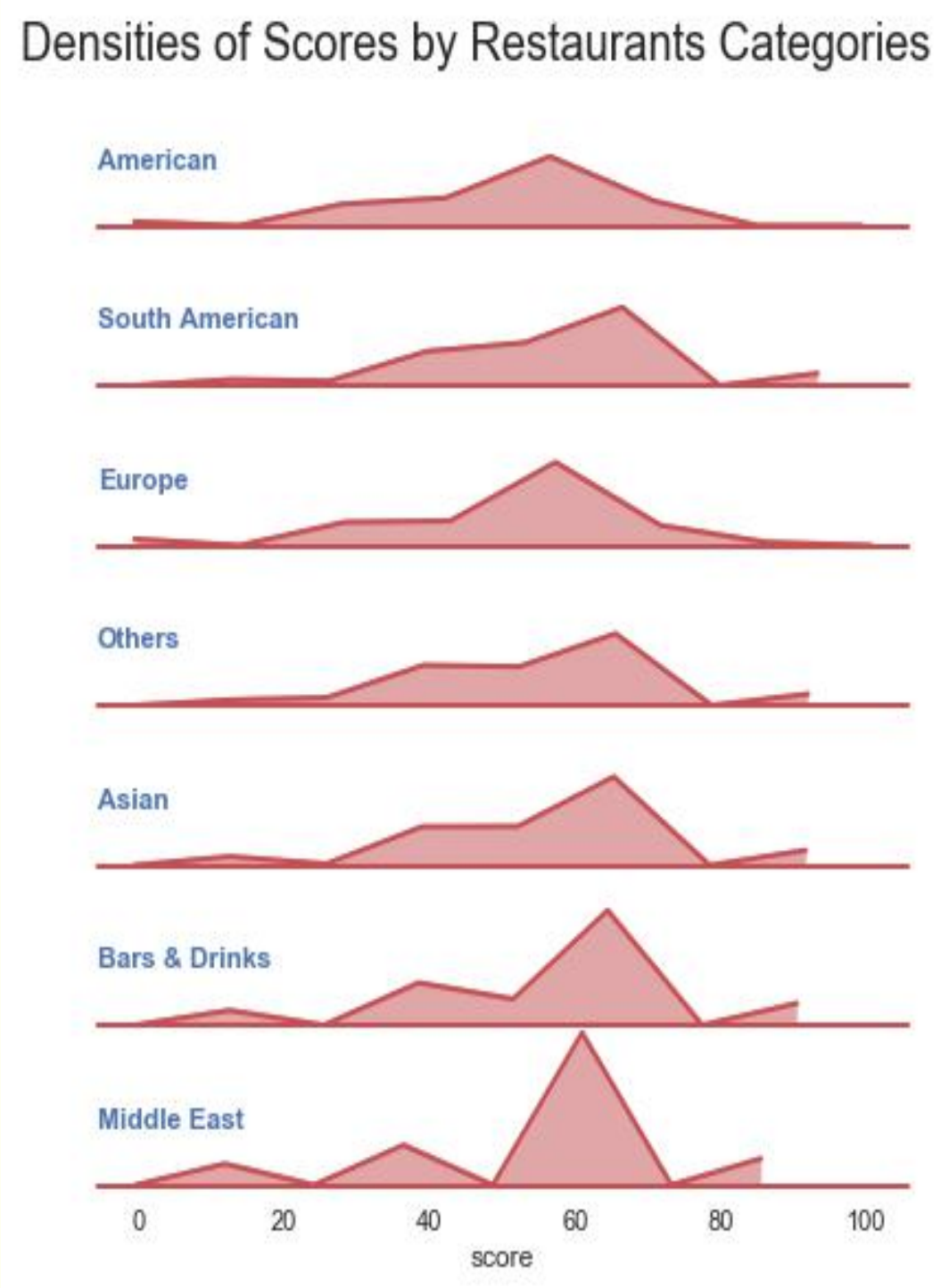




X-restaurants

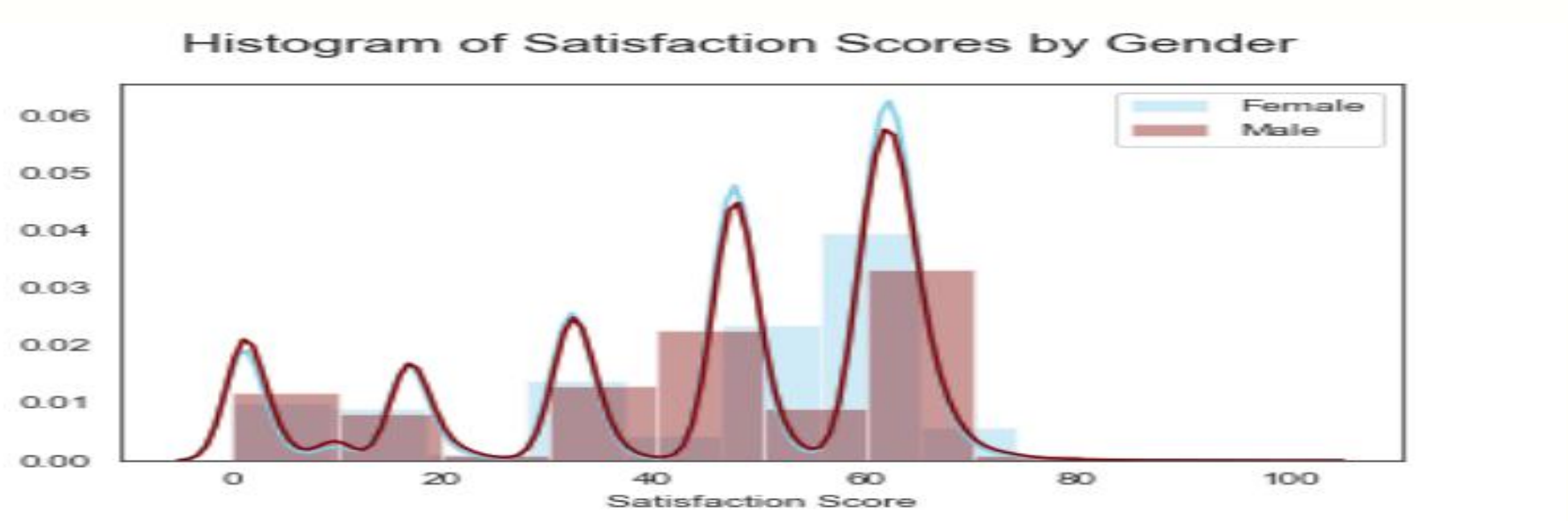


- Restaurants reputations form customers' prior belief
- Use stars ratings and review counts
- Categories (ex. cuisine type) influence people's expectation
- Classify restaurants into 7 categories by keywords searching
- Use dummy variables for the categories

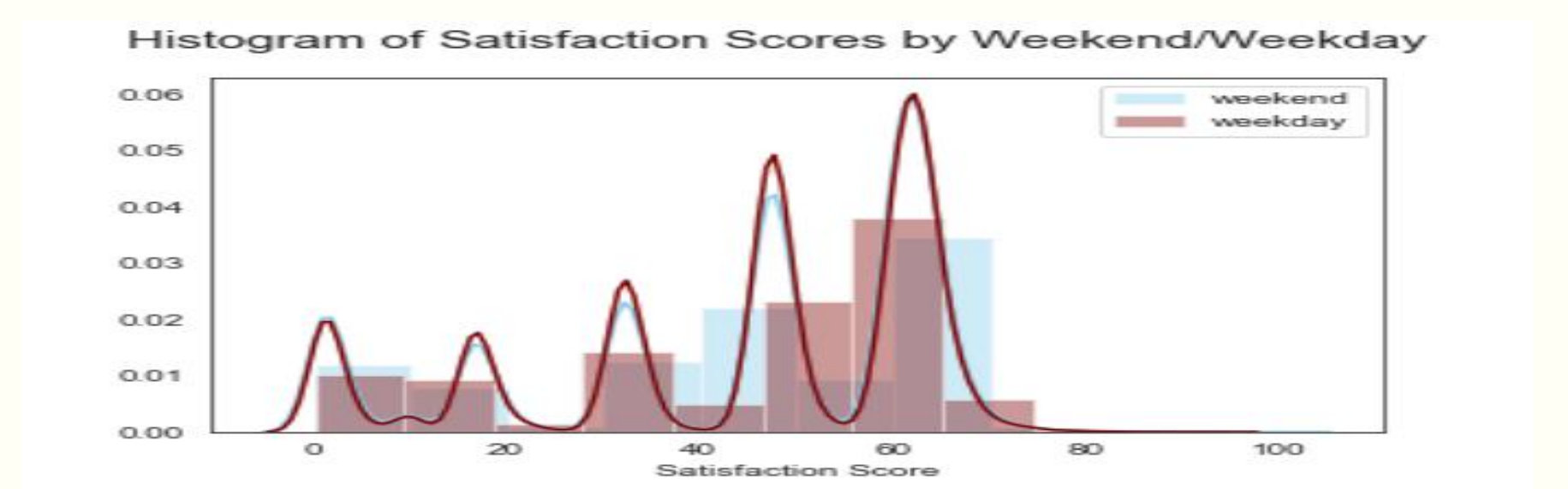


X-users

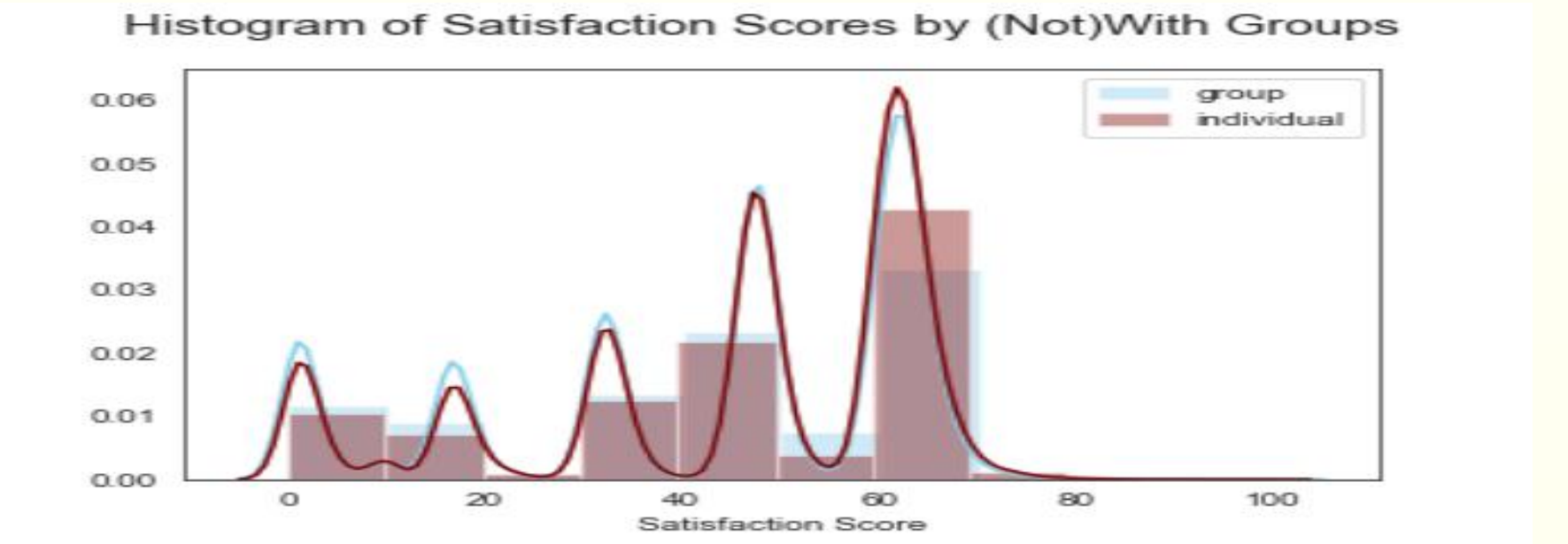
- Gender: inferred 42% are female users



- Weekend/Weekday: inferred 42% of the visits are during Friday to Sunday

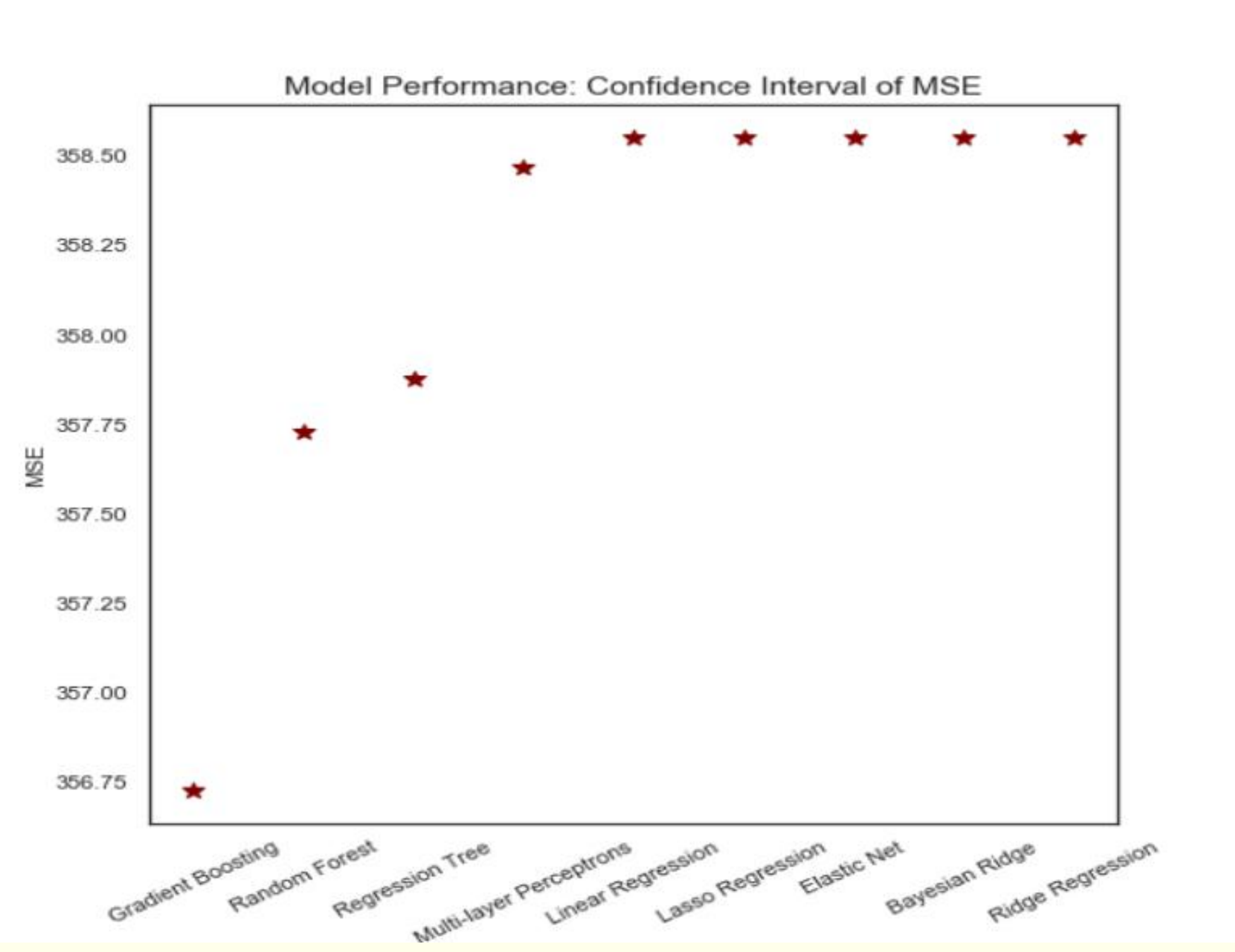


- Group/individual: inferred 51% of the visits are group outings



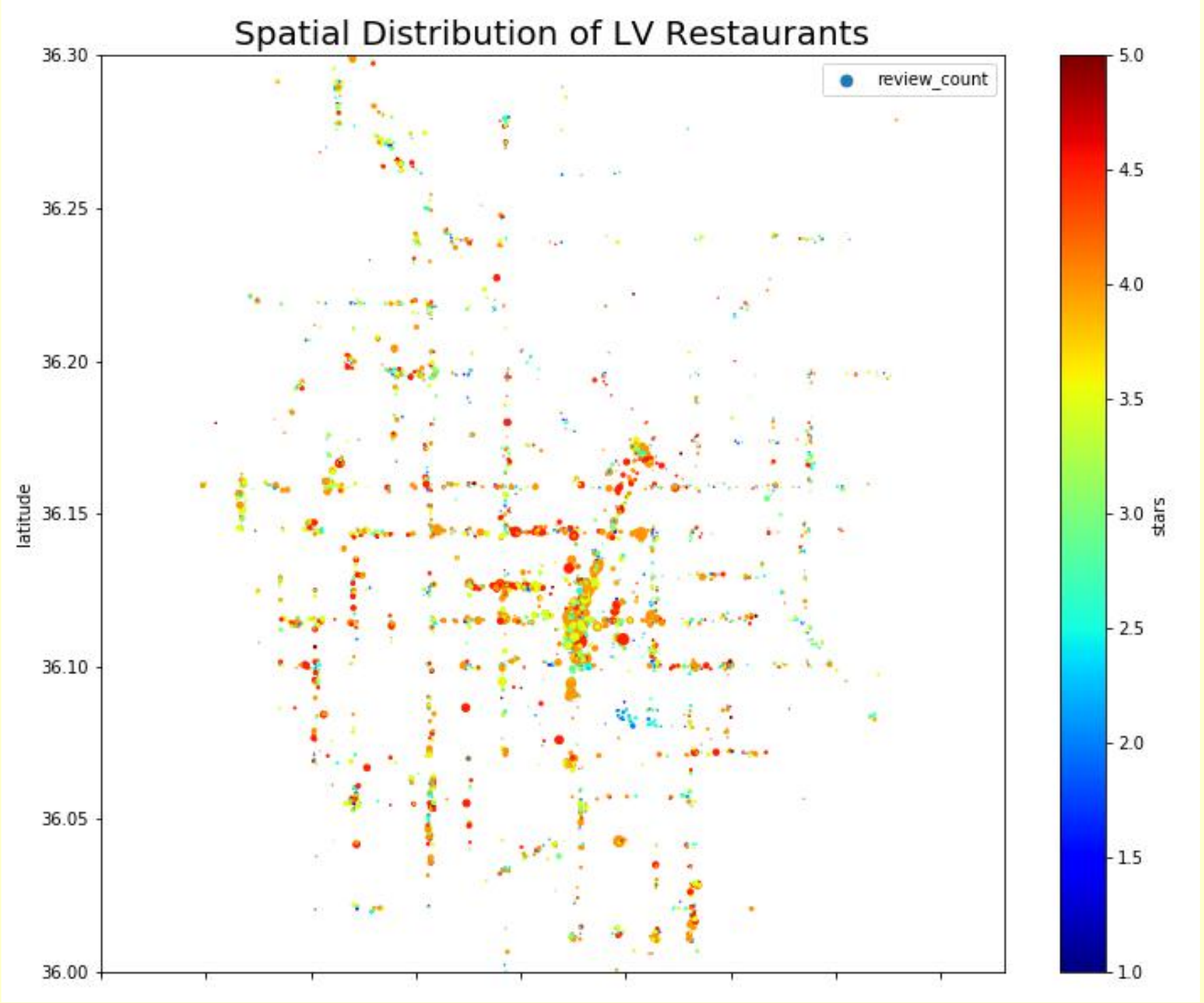
Conclusion

- Tree-based models perfomed slightly better than linear models
- **Restaurants reputation is important for modeling customer satisfaction**
- Customers heterogeneity (X-users) variables are relatively not predictive
- **Machine learning is good supplement to existing marketing research methods and models**

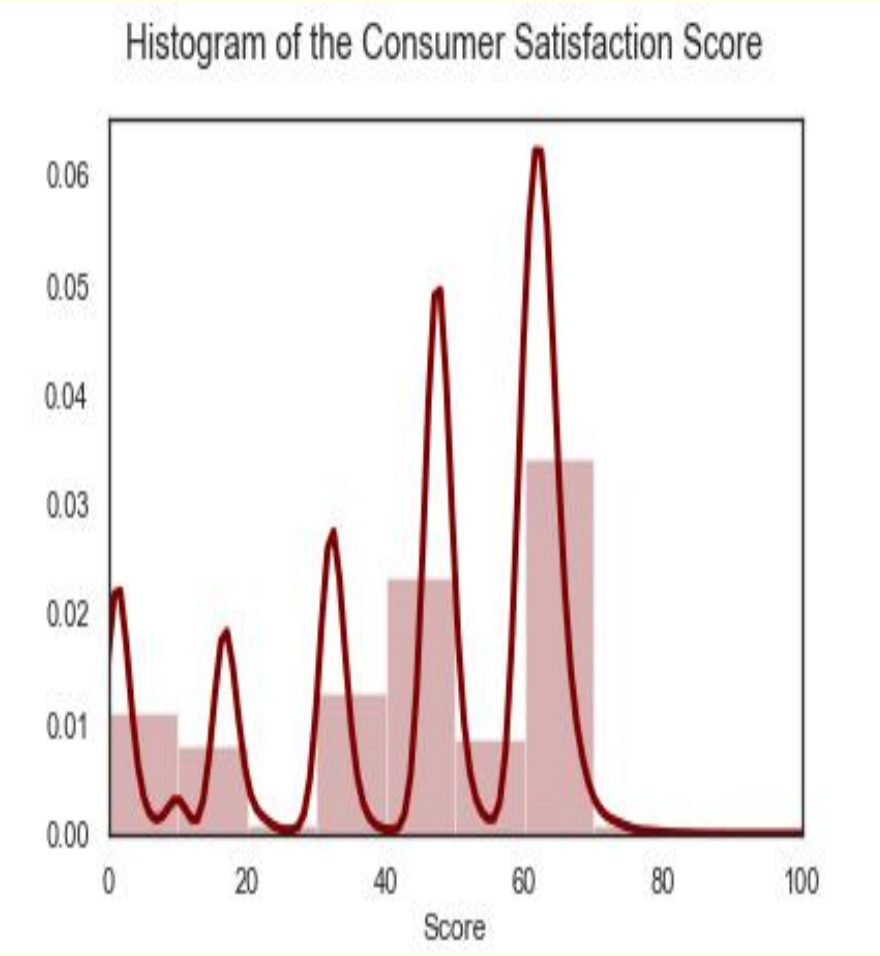


Data

- Online open Yelp data (8GB)
- Subset: Open Restaurants in Las Vegas, 2007-2017
- Merge reviews, business, and users data; each row indicates a visit
- 0.8 million rows; 36 columns



- y as the measure of customer satisfaction
- 0-100 scale
- Mean: 44; Std: 21; Left skewed



Dependent Variables

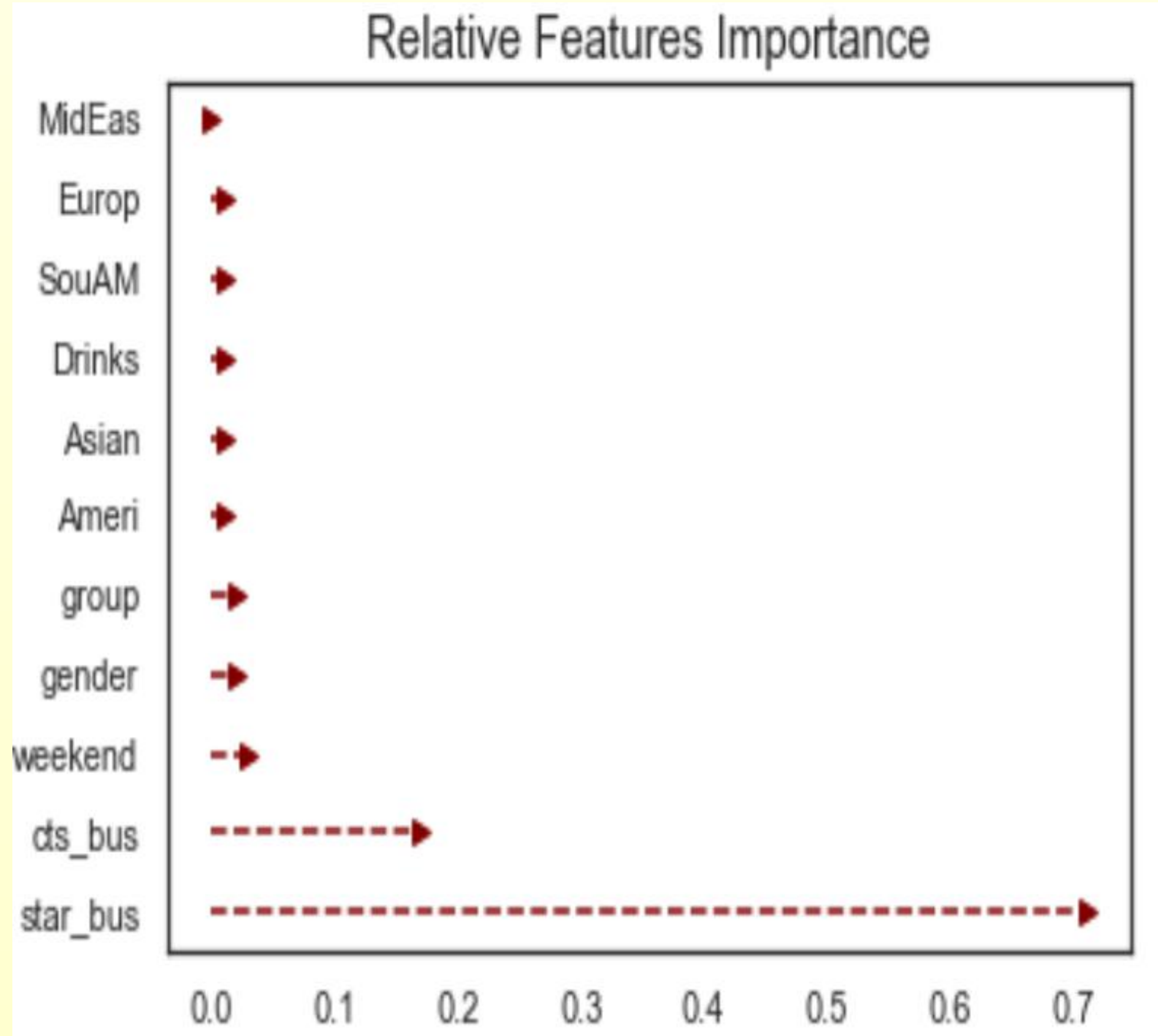
- Determinants of the satisfaction function
- Should be observable for non-Yelp users
- Two categories of the X:

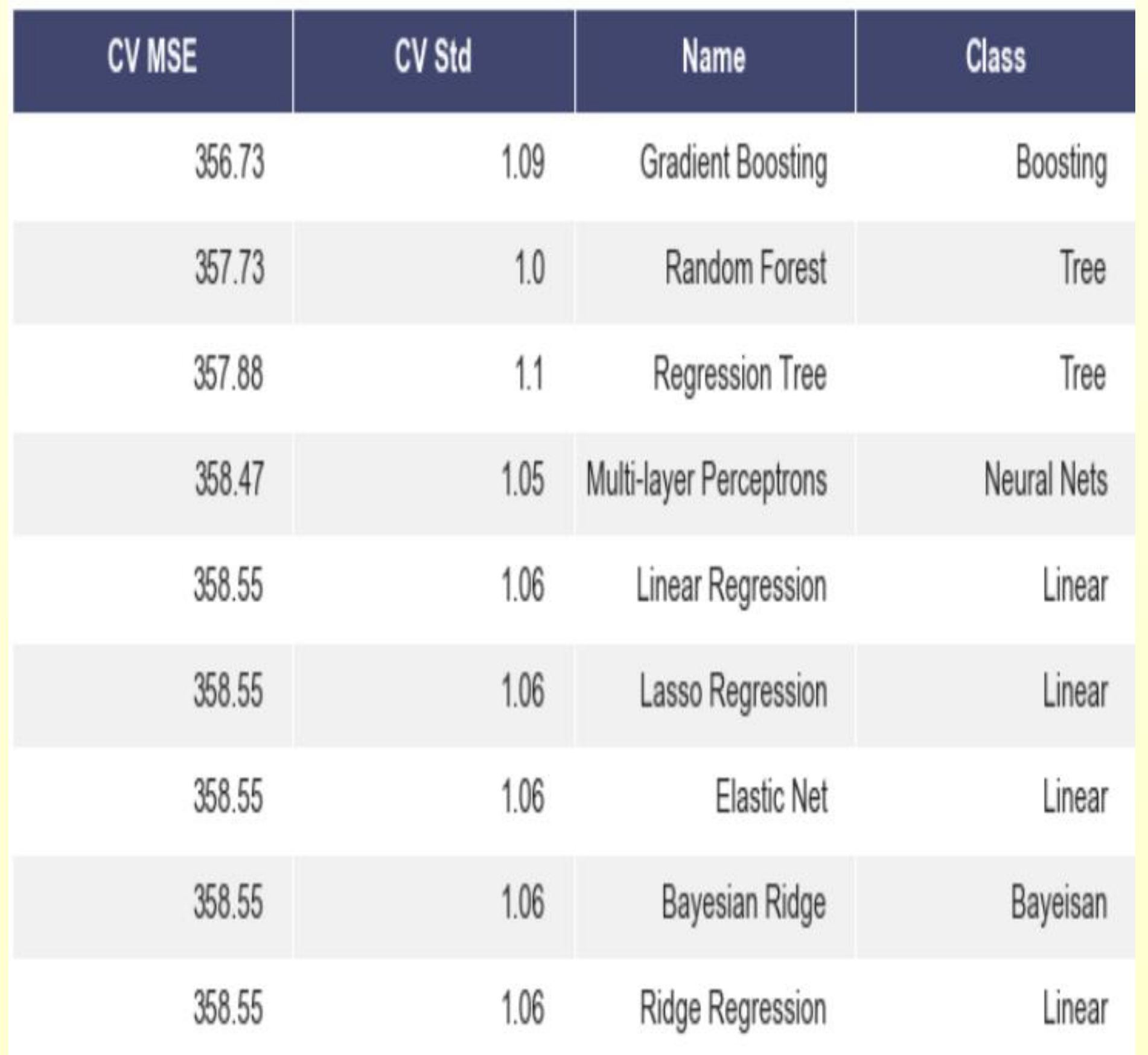
1. X about the restaurants: reputation (restaurants review ratings and counts)

2. X about the users: individual attributes (Not directly available from the data)

Methods & Results

- Features Importance by Random Forest: leave not the restaurants categories as they don't help prediction;
- Divided the data into training and set test, with ratio 3:1
- Trained the data on 14 supervised regression algorithms
- Measured accuracy by MSE with 5-fold cross validation
- Tuned hyperparameters with randomized or grid searches
- Left the worst 4 models out (KNN, SVM, XGBoost, Adaboost)





CV MSE	CV Std	Name	Class
356.73	1.09	Gradient Boosting	Boosting
357.73	1.0	Random Forest	Tree
357.88	1.1	Regression Tree	Tree
358.47	1.05	Multi-layer Perceptrons	Neural Nets
358.55	1.06	Linear Regression	Linear
358.55	1.06	Lasso Regression	Linear
358.55	1.06	Elastic Net	Linear
358.55	1.06	Bayesian Ridge	Bayeisan
358.55	1.06	Ridge Regression	Linear

Limitations

- X-users variables are based on “mining” and “guessing”
- X-restaurants might be endogeneous
- Omitting variable bias
- Yelp users != whole users population

Future Work

- **Bayesian framework** could better capture customers' prior expectation
- Model the **dynamic searching and learning** of customers
- Link Yelp data with **other sources**, such as Census, business revenue/tax data, etc