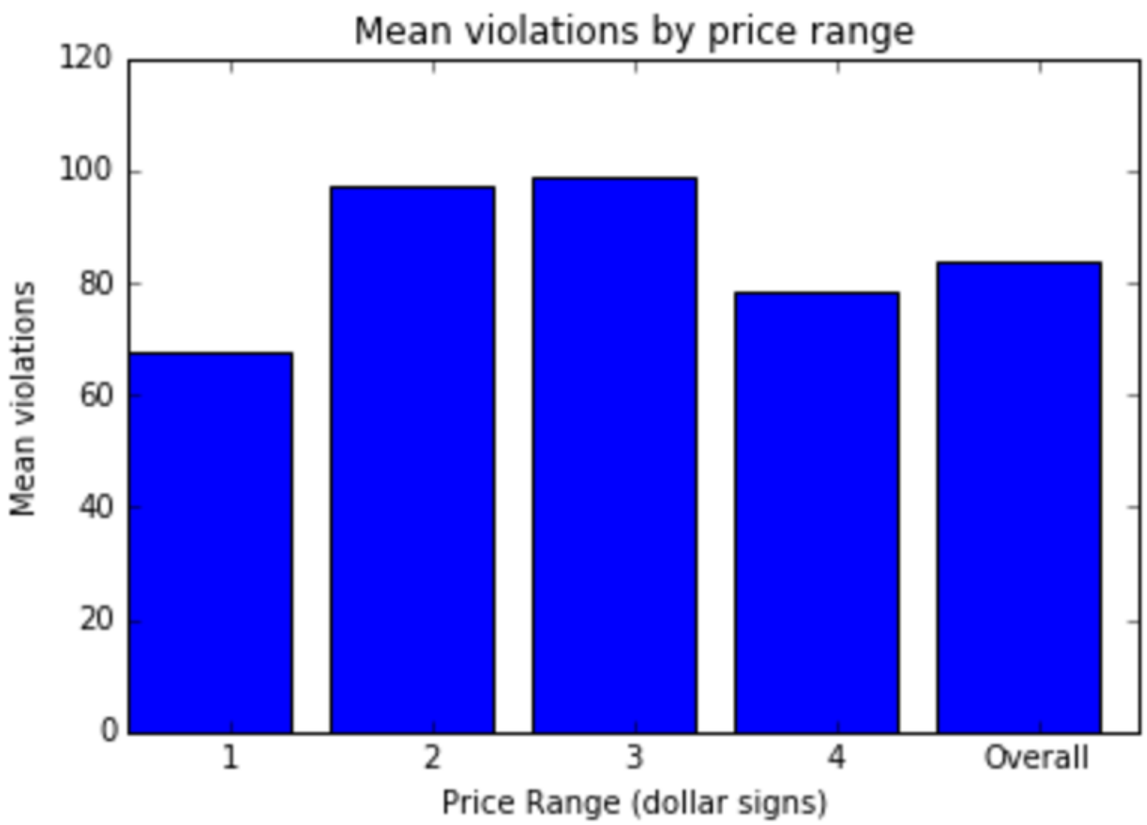
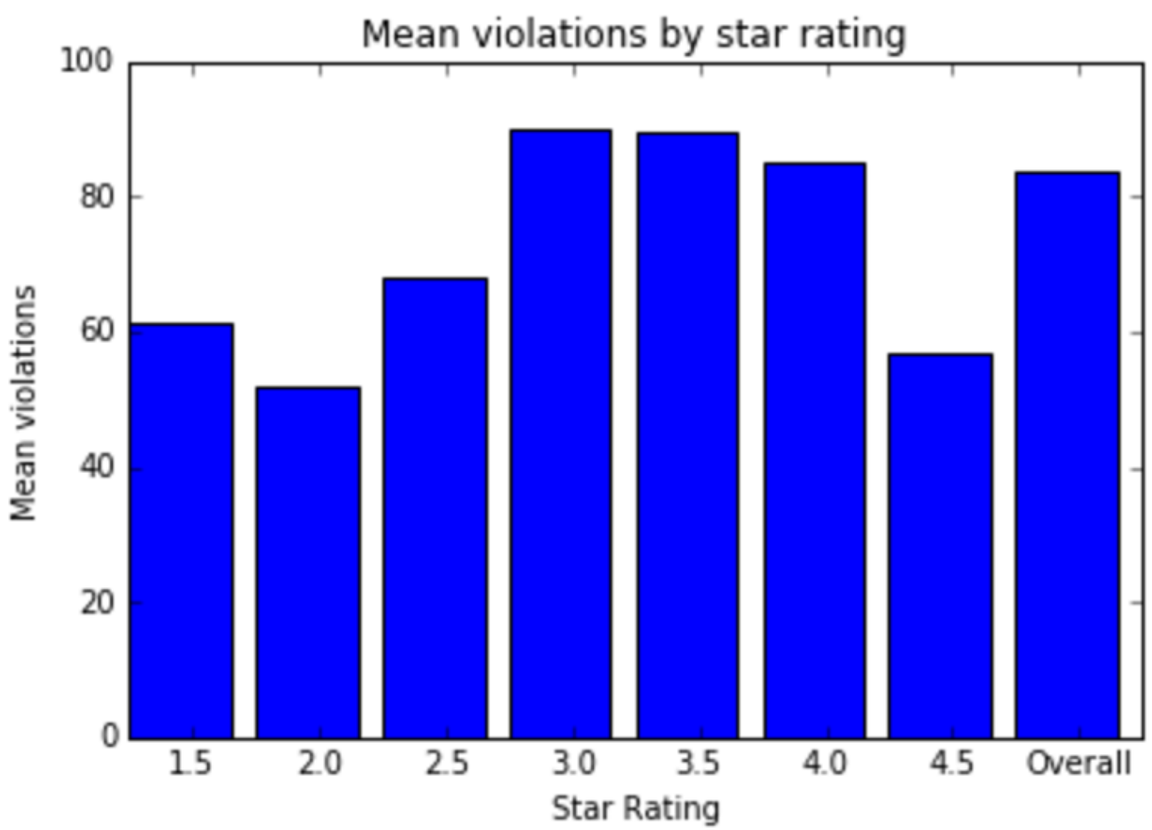
HEALTH VIOLATIONS FROM YELP DATA

Derek Updegraff

ISSUE

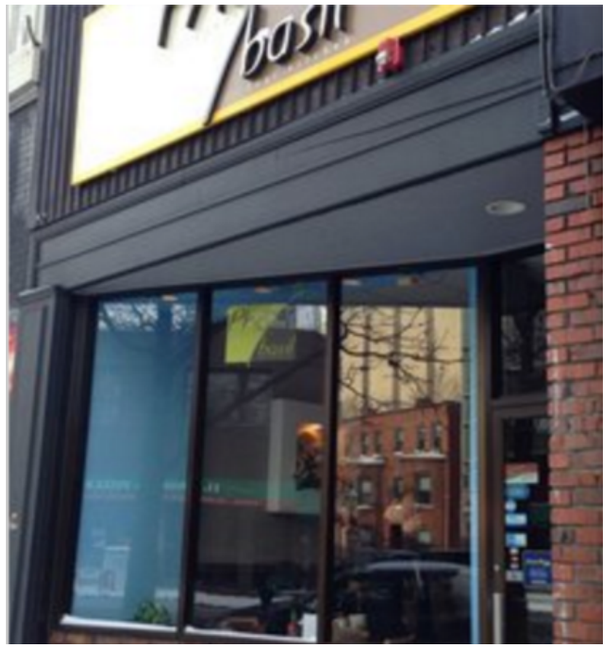
- 48 Million Americans sick with foodborne illness annually
- $15.6 Billion in economic cost
- 75% of these from restaurants vs. in-home





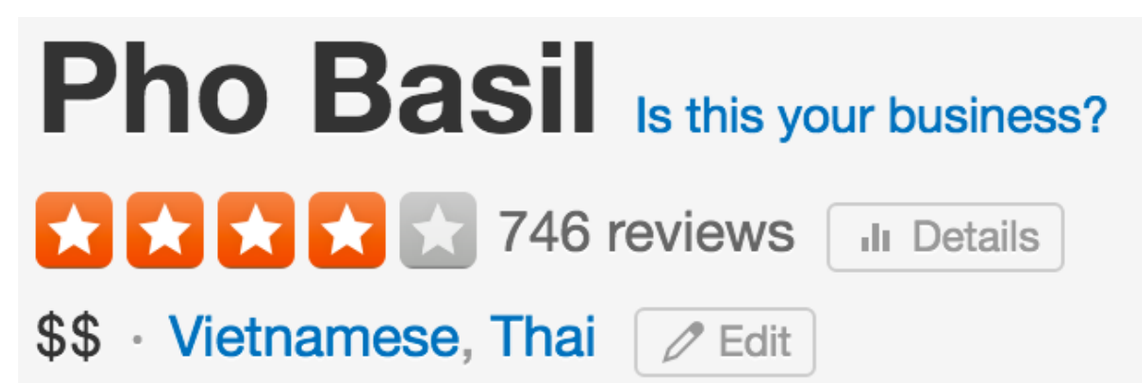
ISSUE

- Health inspection data not integrated with reviewer sites/apps
- Not simple to predict from other data



ISSUE

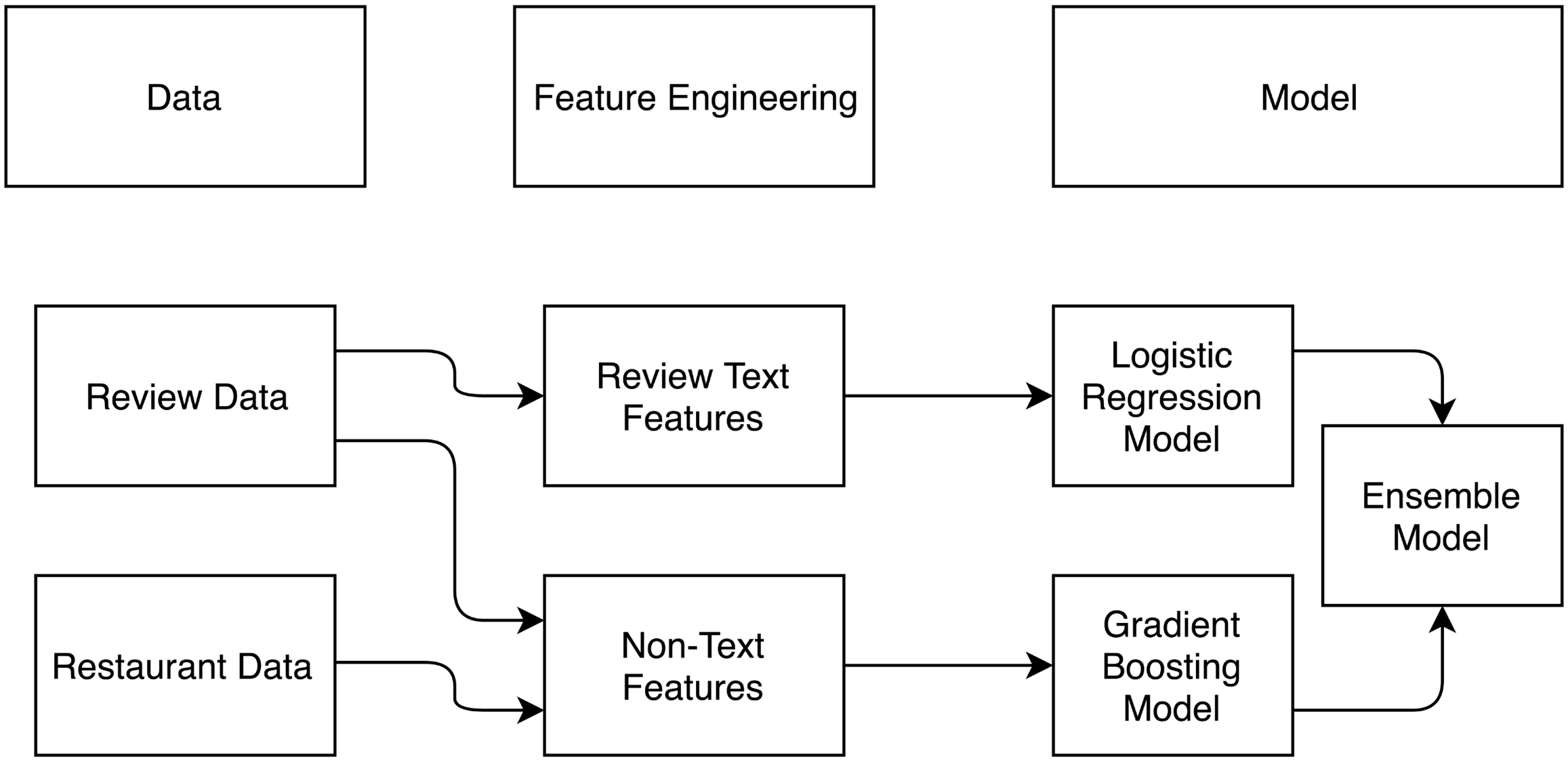
- Good looking establishments can be high risk
- But the information is in there if you look…



DATA/SCOPE

- Classify 10% of restaurants with most violations from review aggregator data
- Data from Boston, 2012-2015
  - 1567 restaurants, all yelp data and review text
  - City health inspection outcomes
    - 98% of restaurants have at least one violation

PROCESS/MODEL



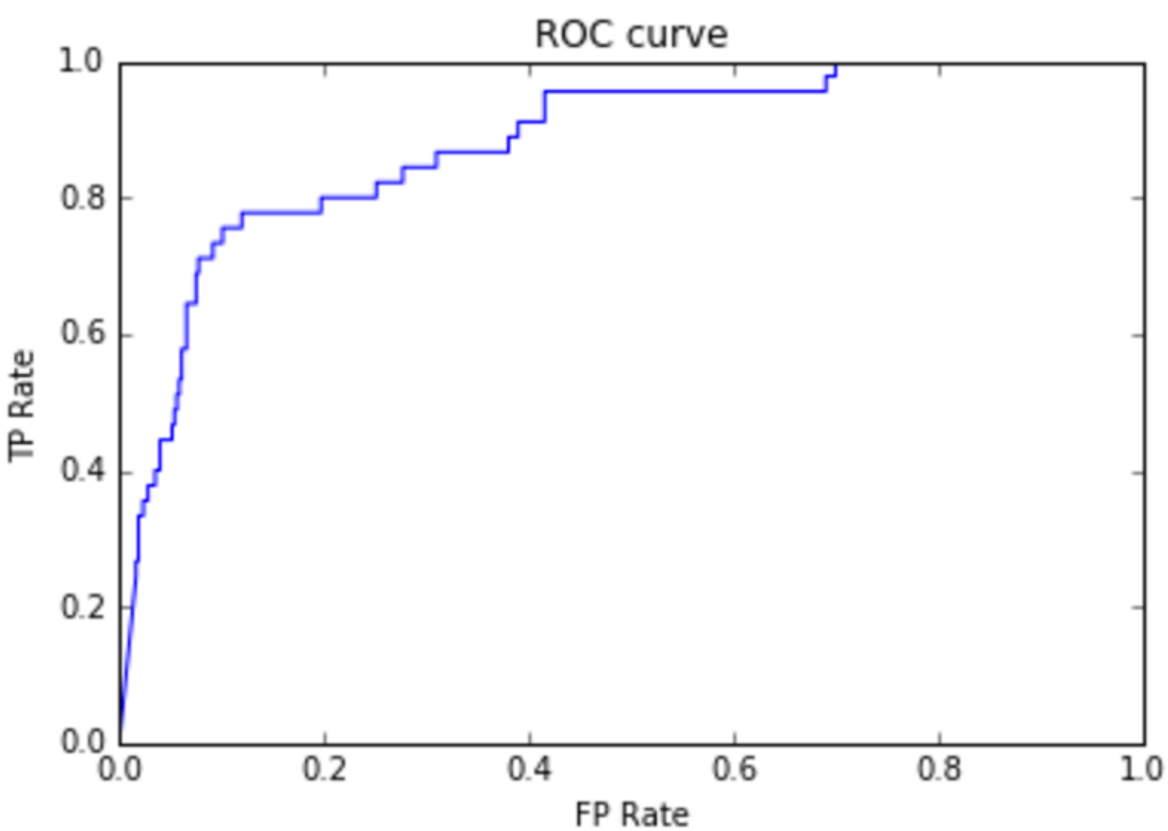
RESULTS/CONCLUSION

beef: 0.607
shrimp: 0.473

bread: -0.244
pasta: -0.474

gross: 0.065
dirty: 0.066

smoothie: -0.104
coffee: -0.577



- Model has significant predictive power
  - ROC AUC score 0.88
- Largely about food types
- Hygiene words less important
  - Don’t come up as much

FUTURE WORK

- Validate on other cities
- Inspection-level predictions
- Specific types of violations
- App development

QUESTIONS, COMMENTS, INQUIRIES?

Derek Updegraff

B.S. Physics, University of Maryland
wupdegra@gmail.com
*linkedin.com/in/**derekupdegraff***
*github.com/**WDUpdegraff***