Yelp Review Rating Prediction

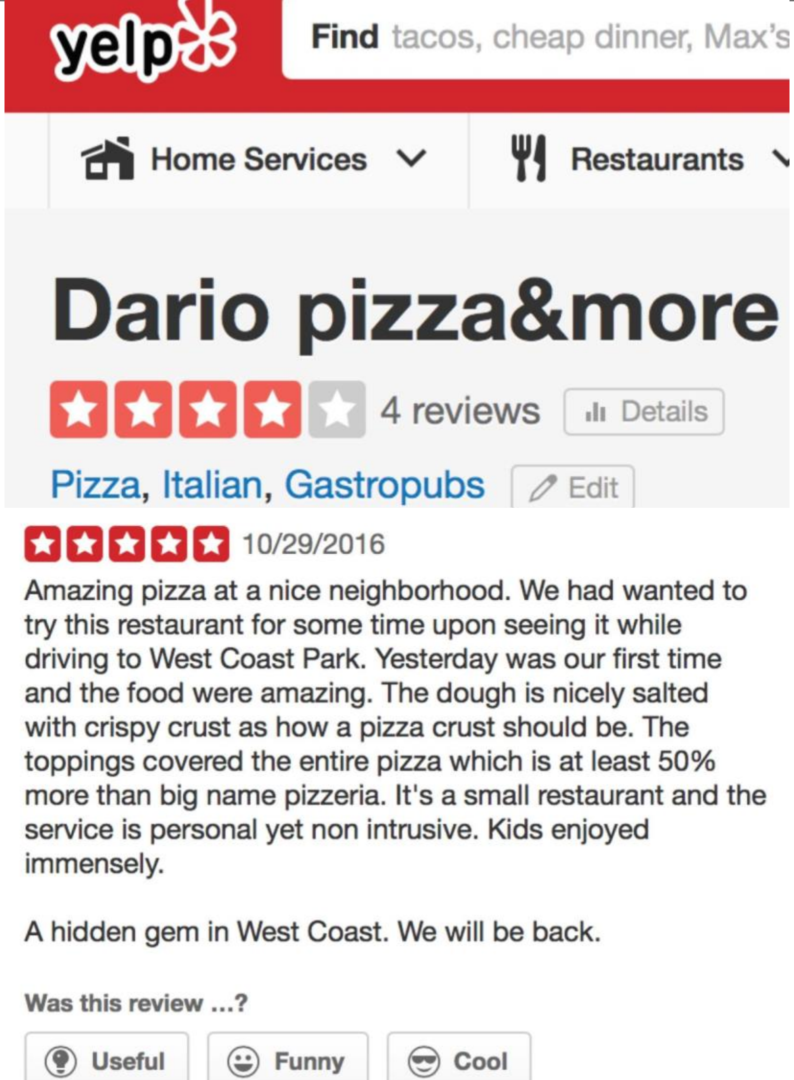
PRESENTED BY:
LI LIPING (A0186040M)
REN JIEWEN (A0186102N)
WANG XINRUI (A0186103M)
XIAO RUI (A0186000W)

Overview

1. **Overview of Problem**
2. **Data Pre-processing**
3. **Model Results**
4. **Conclusion**

Section 1

Overview of Problem



Try to answer…

- What are potential variables correlated with ratings?
- How well can we predict the rating by a customer?

Business value…

- Restaurants
- Yelp

Section 2

Data Pre-processing

Yelp Dataset 2013 Sample

31,693

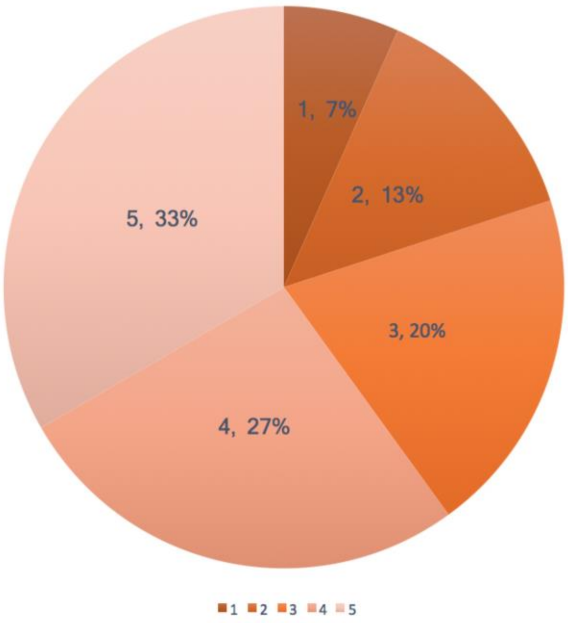
samples

4,362

Restaurants

14,893

Customers



Distribution of review stars

Feature Selection

- Related to the review:
- text, vote.funny, vote.cool, vote.useful

- Related to the customer:
- funny, cool, useful, review.count

- Related to the restaurant:
- city, category, review number

Feature Construction

cuisine

25 cuisines summarize 204 categories of restaurants in original dataset

city

10 levels summarize 46 cities, most of which are in Arizona

wcount

Total number of words in each review text

qmark & emark

Total number of question/exclamation marks in each review text

Section 3

Model Results

Baseline Model Results

Baseline model RMSE		
linear regression	stepAIC	glmnet-lasso
1.11721	1.11369	1.11618
glmnet-ridge	glmnet-elastic net	knn
1.11815	1.11624	1.15793

Find DSC 5103 Statistics   Near NUS Business School, Singapore

Baseline Model Interpretation

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.07488    0.03187 127.841  < 2e-16 ***
## city.Mesa            0.08226    0.04405   1.867 0.061858 .
## city.Others         -0.05550    0.03464  -1.602 0.109121
## city.Scottsdale     -0.06736    0.02382  -2.828 0.004689 **
## city.Tempe          -0.13378    0.02913  -4.593 4.41e-06 ***
## review_count.x       0.80059    0.05174  15.473  < 2e-16 ***
## review_count.y      -0.47936    0.13151  -3.645 0.000268 ***
## yelp.votes.funny    -4.37729    0.43566 -10.048  < 2e-16 ***
## yelp.votes.useful   -9.80270    0.73517 -13.334  < 2e-16 ***
## yelp.votes.cool     17.20662    0.82707  20.804  < 2e-16 ***
## cuisine.American    -0.22691    0.05800  -3.912 9.18e-05 ***
## cuisine.Breakfast   -0.53685    0.05169 -10.386  < 2e-16 ***
## cuisine.Chinese     -0.41505    0.10482  -3.960 7.54e-05 ***
## cuisine.Fast.Food   -0.79075    0.09625  -8.216 2.27e-16 ***
## cuisine.Fusion      -0.23313    0.09001  -2.590 0.009602 **
## cuisine.Italian     -0.17478    0.04626  -3.778 0.000159 ***
## cuisine.Japanese    -0.26415    0.06622  -3.989 6.66e-05 ***
## cuisine.Korean      -0.35243    0.16162  -2.181 0.029231 *
## cuisine.Mexican     -0.35836    0.03389 -10.575  < 2e-16 ***
## cuisine.Nightlife   -0.18784    0.03470  -5.413 6.30e-08 ***
## cuisine.Snacks      -0.12123    0.03164  -3.832 0.000128 ***
## cuisine.SouthAsia   -0.29339    0.20077  -1.461 0.143935
## cuisine.Thai        -0.17899    0.04996  -3.583 0.000341 ***
## cuisine.Vietnamese -0.14743    0.05876  -2.509 0.012120 *
## wcount              -1.41087    0.08349 -16.898  < 2e-16 ***
## qmark               -4.04347    0.30227 -13.377  < 2e-16 ***
## emark                5.45072    0.25356  21.497  < 2e-16 ***
```

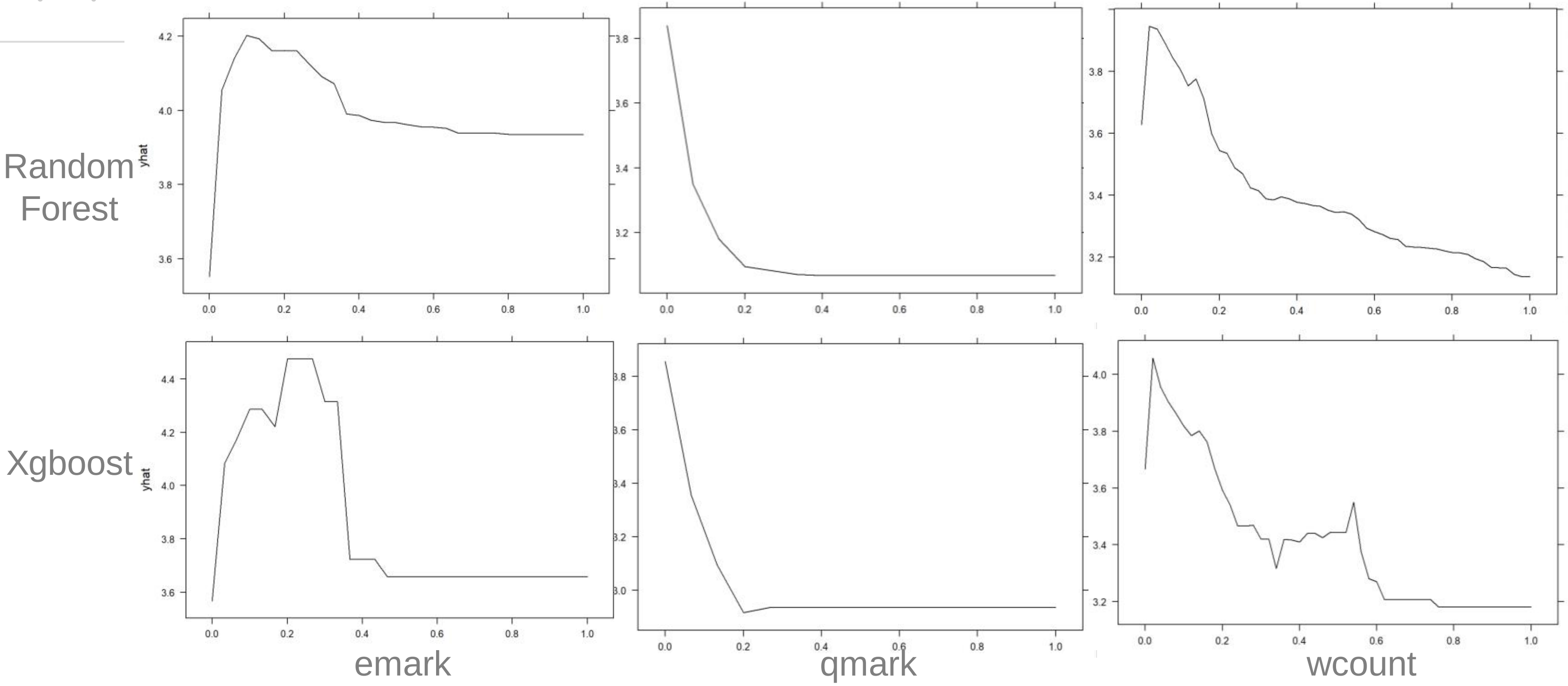
- Coefficient of 6 variables
- Positive or negative?
- Statistically significant

Ensemble Learning Results

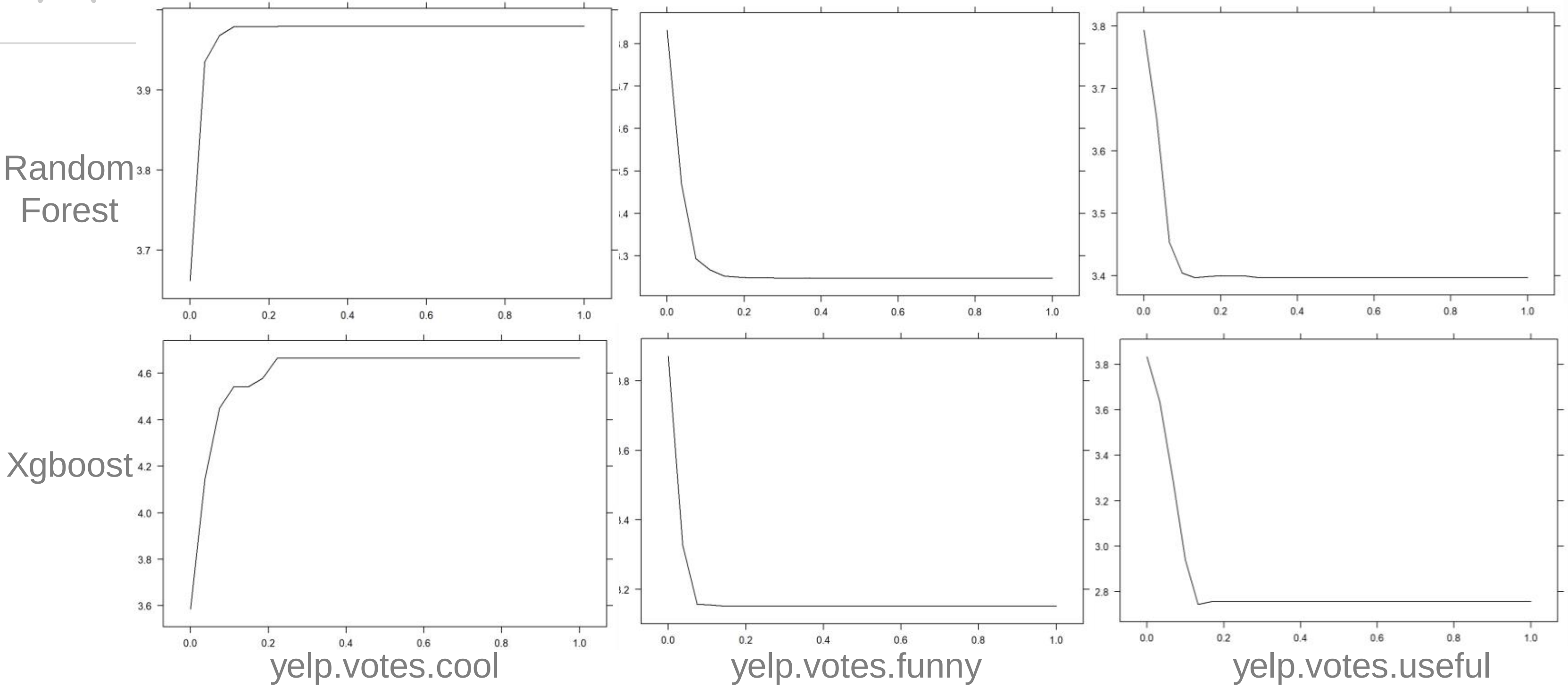
Random Forest	xgboost	Stacking
1.10221	1.09036	1.11108

- **Stacking:**
- level 0 - Random Forest, Xgboost, Step_AIC and Elastic net
- Meta-learner - Lasso regression

Variable Importance – Partial Plot I



Variable Importance – Partial Plot II



Feature Engineering Results

Model RMSE					
	linear regression	glmnet-elastic net	Random Forest	xgboost	Stacking
Before	1.11721	1.11624	1.10221	1.09036	1.11108
Feature Engineering	**1.10672**	**1.11657**	**1.0987**	**1.0906**	**1.1051**
Change	-0.01050	0.0003	-0.0035	0.0002	-0.0060

Section 4

Conclusion

1.0906

Our model’s lowest RMSE

From Xgboost after feature engineering

Find DSC 5103 Statistics Near NUS Business School, Singapore

“

- Algorithm selection: Xgboost performs the best
- Text mining techniques
- Extend prediction models to other business categories, such as shopping, hotels, etc.

THANK YOU!