Yelp
Predicting Useful Votes of Reviews

Sinem Yekbun Polat
18/12/2017

Overview

- Problem Definition
- Dataset Overview
- Exploratory Visualization
- The ML Algorithms & Evaluation
- Conclusions

Are reviews important?

→ Limited budget
→ Best experiences
→ Ask friends, ask others
→ Limited time to read all reviews

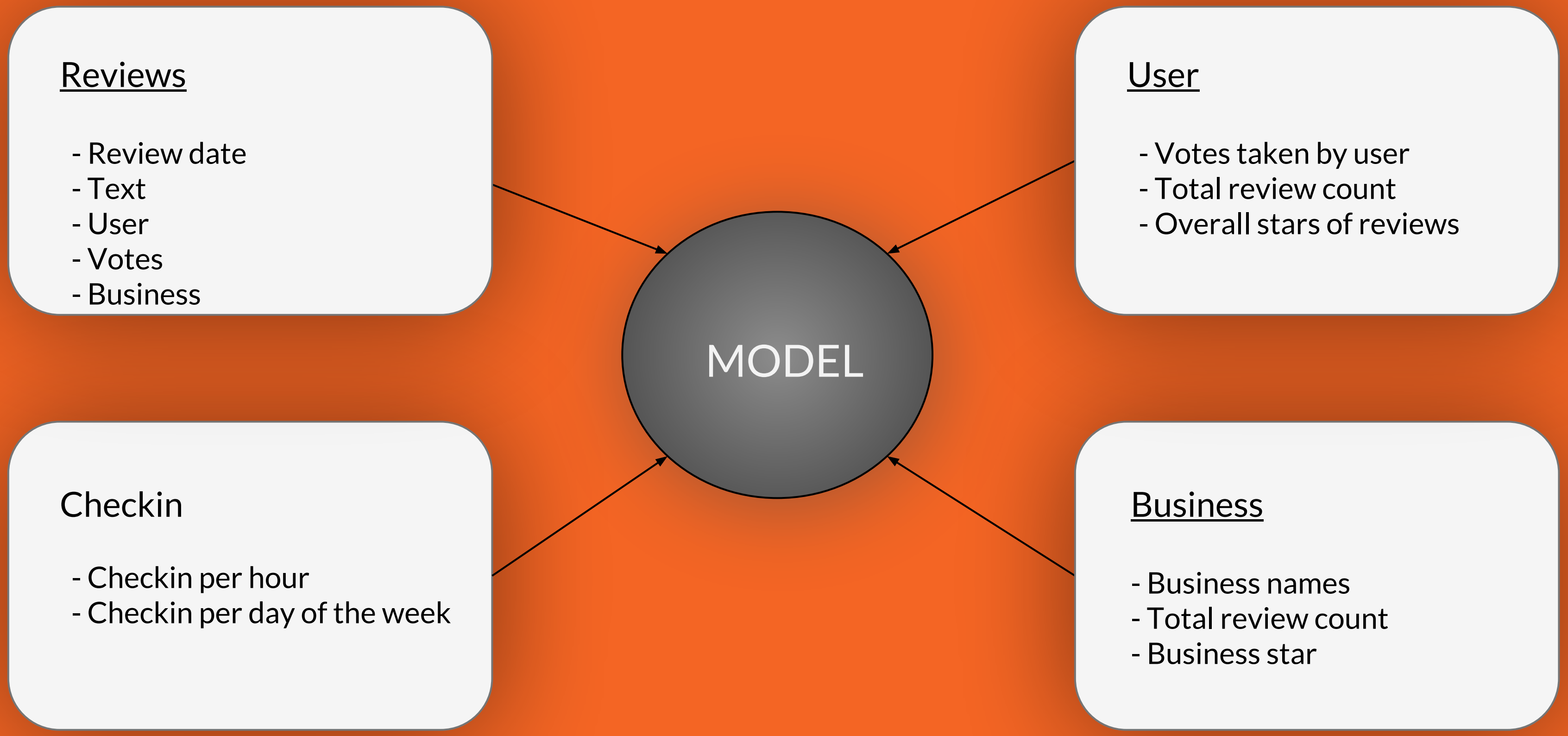
Is it worth to pay my money?

Are reviews important?

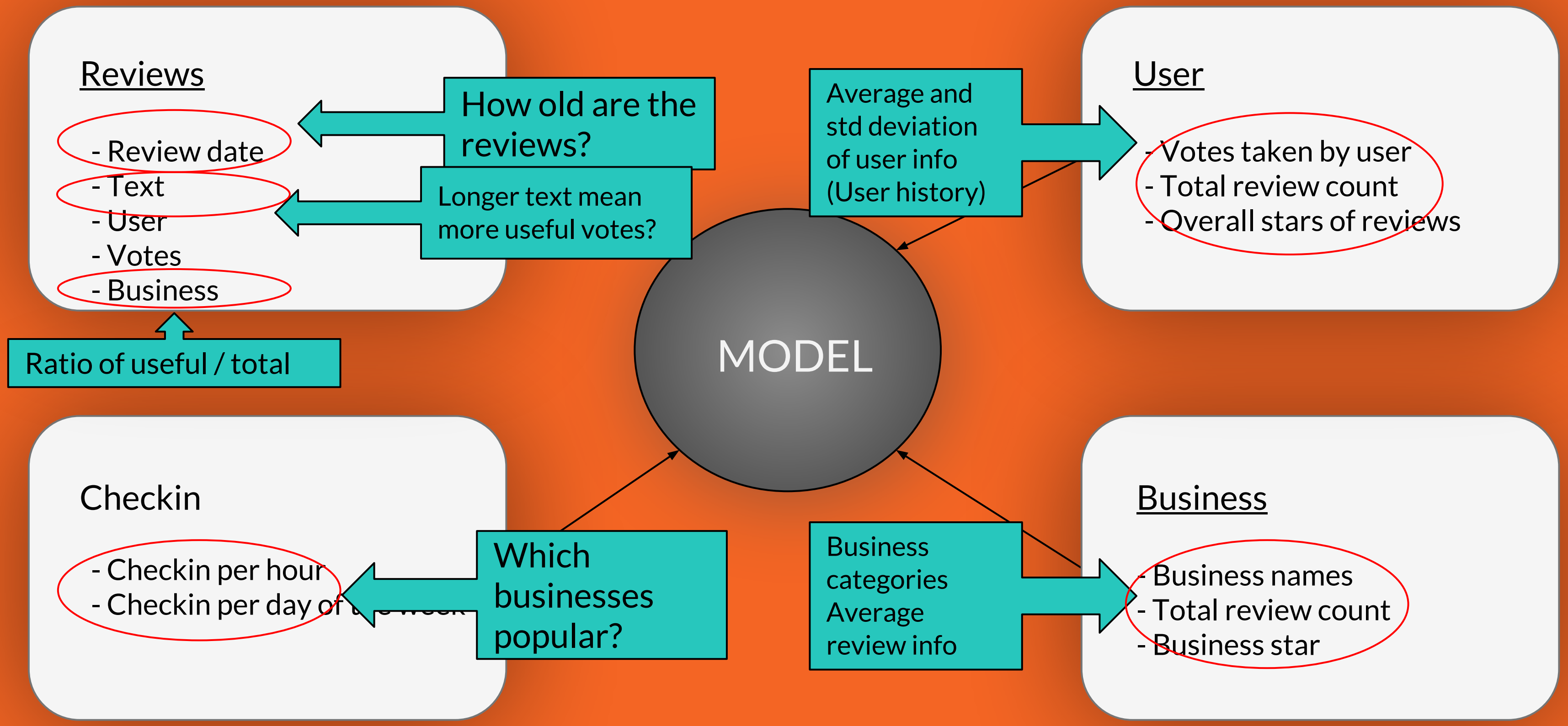
- Limited budget
- Best experiences
- Ask friends, ask others
- Limited time to read all reviews

Is it worth to pay my money?

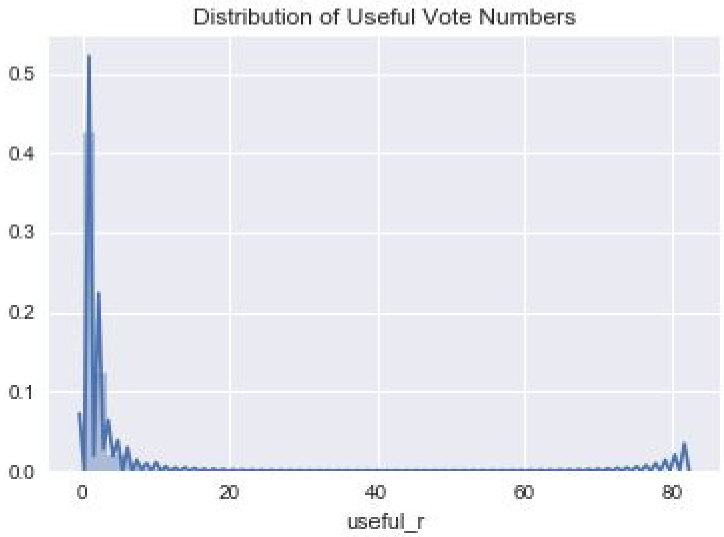
Dataset Overview

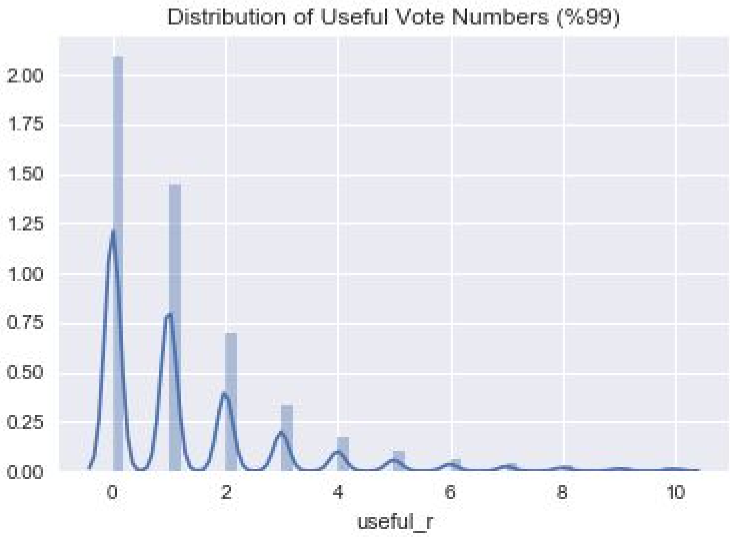


Extra Features from Datasets



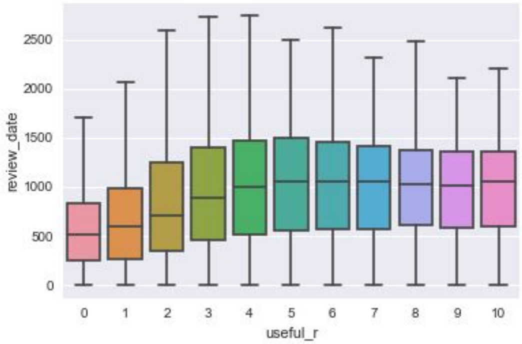
Distribution of Useful Votes



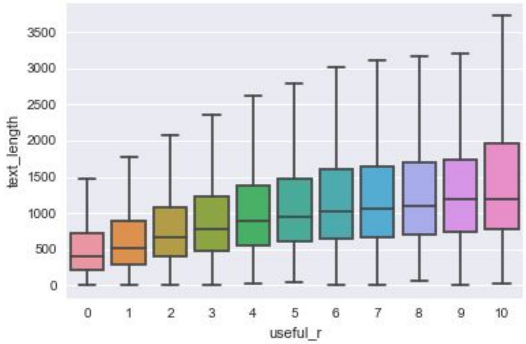


- Useful votes are distributed between 0 and 80.
- Most of them are between 0 and 10.
- There is a huge accumulation in number of 0 useful reviews.

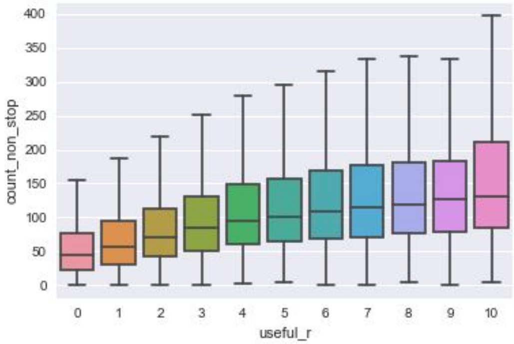
Text-related Features



Old reviews vs Useful votes

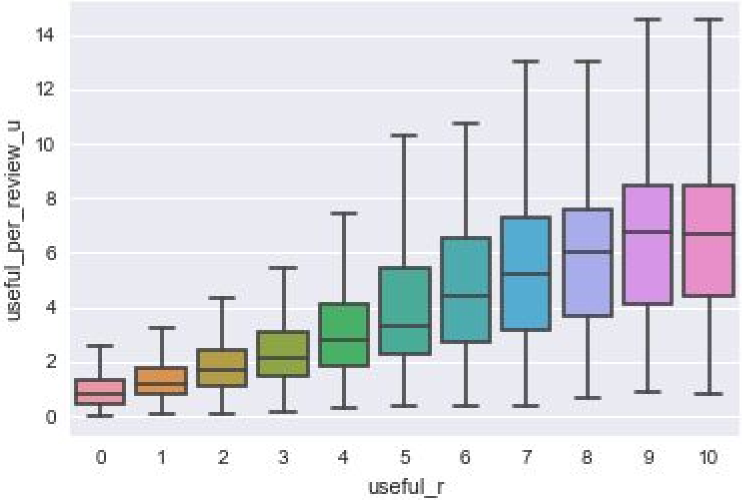


Review length vs Useful votes

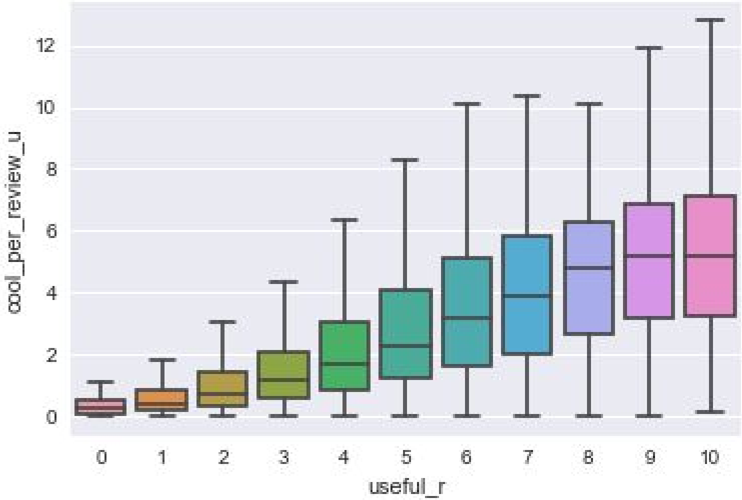


Non-stop words vs Useful votes

User-related Features



Users' average useful votes per reviews



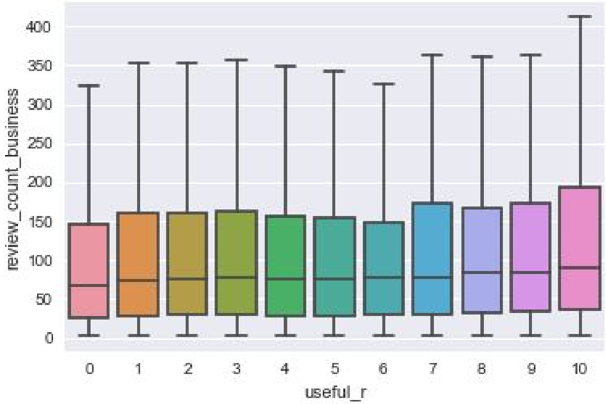
Users' average cool votes per reviews

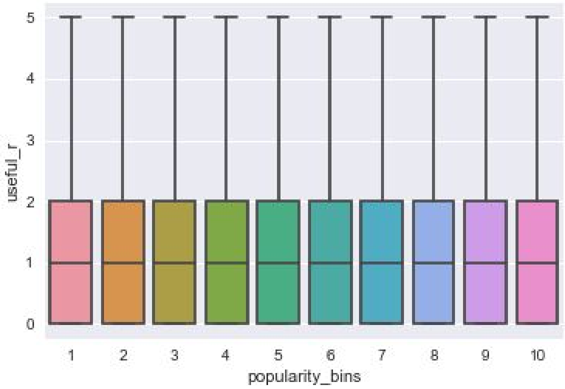
Business-related Features

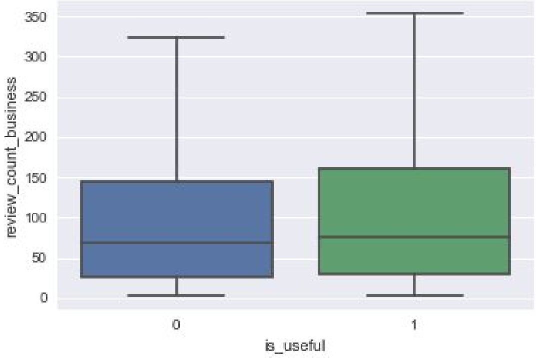
Business-related features such as:

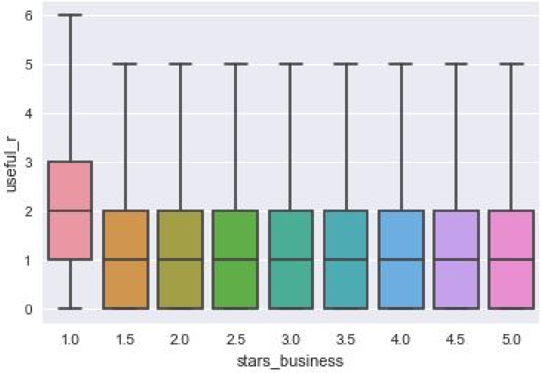
- total review count,
- popularity,
- average stars

don’t have huge effects on useful votes counts.









Algorithms

	Decision Tree Classifier	Random Forest Classifier
Accuracy Score	0.63	0.66
After optimizing modeling parameters (GridSearchCV)		
Accuracy Score	0.69	0.66
Precision	0.69	0.66
Recall	0.69	0.66
F-1 Score	0.68	0.66

Confusion Matrices

Decision Tree

	0	1	2	3	4	5
0	7696	4726	4			
1	3635	12471	240	3		
2	6	726	469	2		
3		20	51	17		
4		2	1			
5		1	1			

Random Forest

	0	1	2	3	4	5
0	8235	4183	8			
1	4922	11223	204			
2	21	732	446	4		
3	1	19	63	5		
4		2	1			
5		2				

0 = [0]
1 = (0, 5]
2 = (5, 15]
3 = (15, 30]
4 = (30, 50]
5 = (50, 100]

Most Important Features

Decision Tree

1	useful_per_review_u	0.524248
2	cool_per_review_u	0.209570
3	review_date	0.121847
4	review_user_std_dev	0.051389
5	count_non_stop	0.036831
6	review_count_user	0.016238
7	stars_review	0.014313
8	text_length	0.010739
9	review_count_business	0.006591
10	funny_per_review_u	0.002356

Random Forest

1	useful_per_review_u	0.108358
2	cool_per_review_u	0.080562
3	funny_per_review_u	0.072615
4	review_user_std_dev	0.071686
5	review_date	0.071396
6	text_length	0.061717
7	count_non_stop	0.059481
8	review_count_user	0.052696
9	stars_user	0.047257
10	review_count_business	0.046735

Conclusions

- Whenever a new review is written by a user having history, we can estimate how many 'useful' votes will be given to this review *with a 0.69 accuracy score.*

- Keep your users that love your product, and have usage history.

- Do improvements to encourage users to write more reviews. (no matter if it's useful or not)

- And write more 'useful' reviews.

Further Improvements:

- More data can be gathered.

- With help of NLP methods, different aspects of a review can be examined. (Which words are used to describe place, quality, food etc.)