Yelp Review Usefulness Prediction

Leilei Liu

Introduction

- Yelp
- Yelp Review
- Goal of the project: presenting high quality reviews to the users

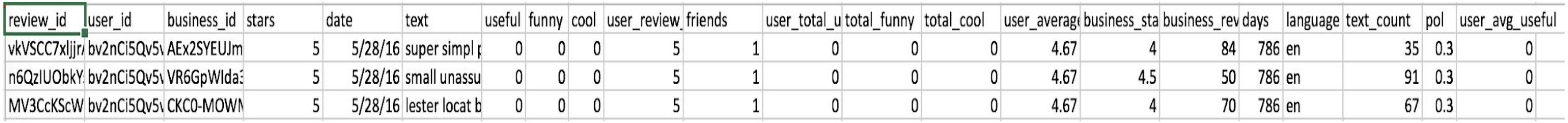
Dataset & Data Fact

Dataset:

- Kaggle Dataset

Data Fact:

- 22 columns * 5201136 rows after combining and cleaning



review_id	user_id	business_id	stars	date	text	useful	funny	cool	user_review_	friends	user_total_u	total_funny	total_cool	user_averag	business_sta	business_rev	days	language	text_count	pol	user_avg_useful
vkVSCC7xljjr/	bv2nCi5Qv5v	AEx2SYEUJm	5	5/28/16	super simpl p	0	0	0	5	1	0	0	0	4.67	4	84	786	en	35	0.3	0
n6QzIUObkY:	bv2nCi5Qv5v	VR6GpWIda3	5	5/28/16	small unassu	0	0	0	5	1	0	0	0	4.67	4.5	50	786	en	91	0.3	0
MV3CcKScW	bv2nCi5Qv5v	CKC0-MOWN	5	5/28/16	lester locat b	0	0	0	5	1	0	0	0	4.67	4	70	786	en	67	0.3	0

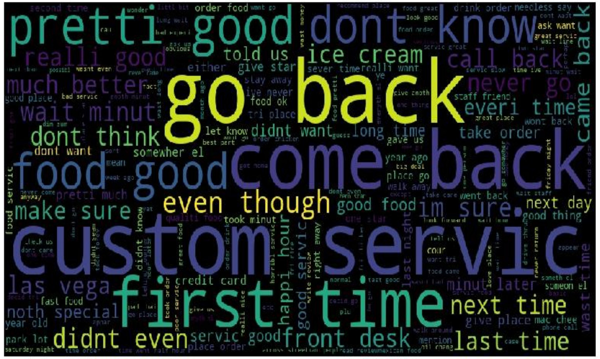
Exploratory Data Analysis

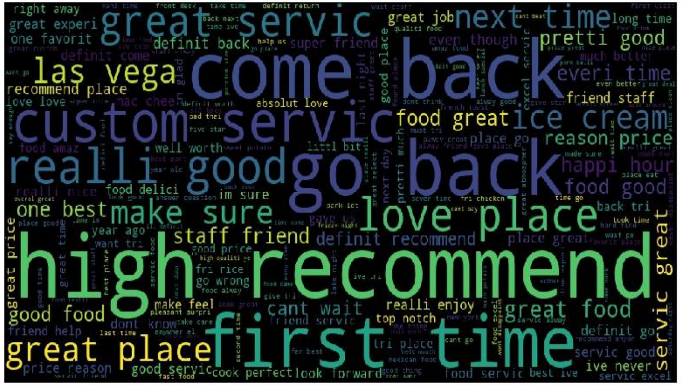
- Review word cloud

definit go
wait staff
im sure
servic good
fast food
good food
wait minut
food great
right away
happi hour
reason price
ice cream
even though
look forward
second time
told us
go back
year old
staff friend
make sure
mac chee
friend help
much better
dont think
place go
food good
great place
great experi
great price
pretti good
realli enjoy
great servic
friend staff
next time
recommend
place
come back
definit recommend
high recommend
realli good
came back
dont know
food servic
last time
las vega
first time
everi time
ive never
great food
one best
front desk
custom servic
long time
love place
servic great
definit come
one favorit
good thing
tri place

Exploratory Data Analysis

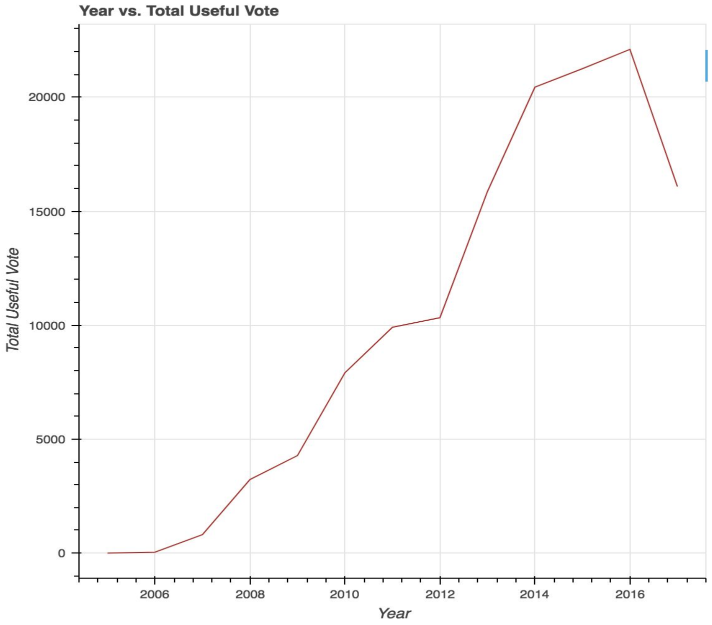
- Positive reviews word cloud
- Negative reviews word cloud

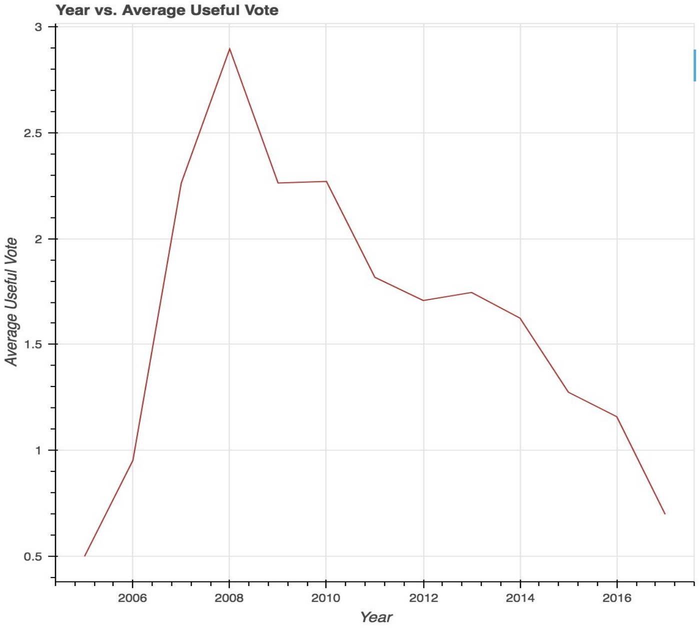




Exploratory Data Analysis

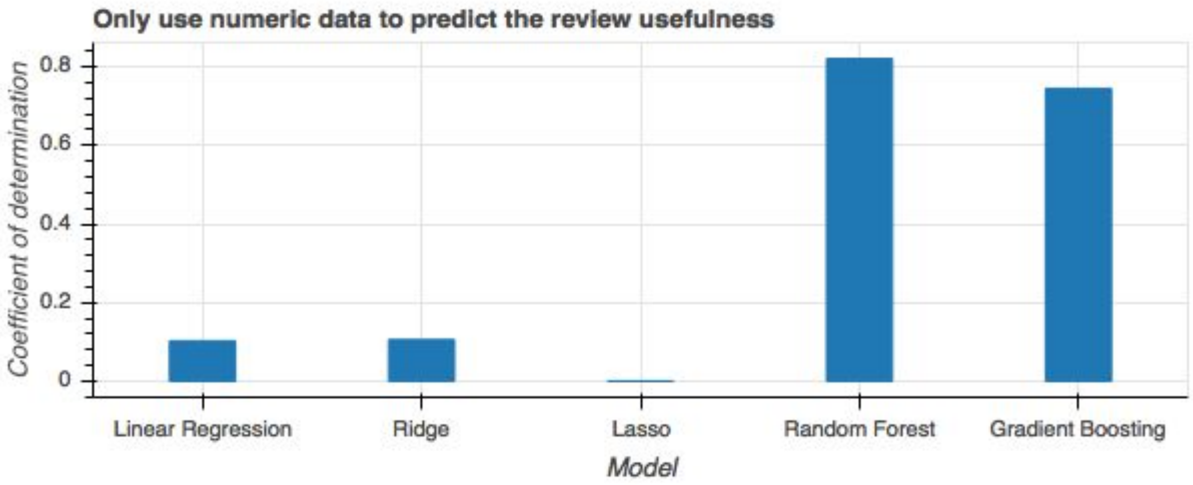
- Time influence on total useful vote and average useful vote





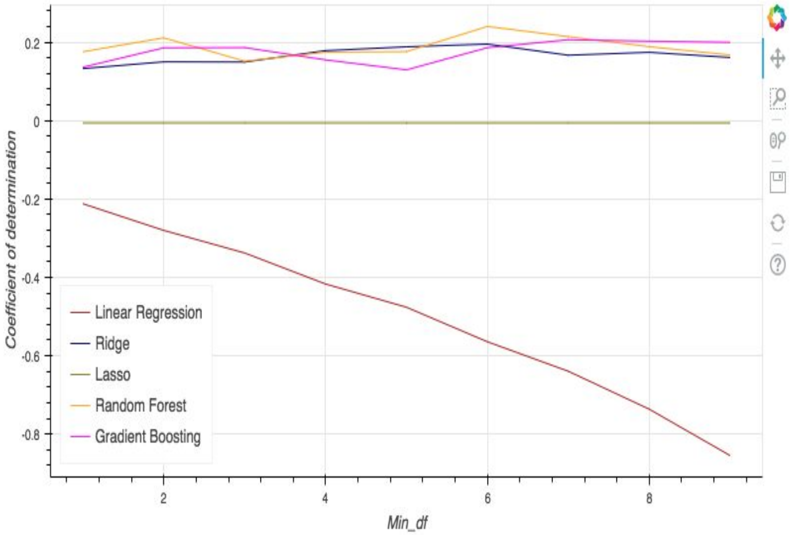
Machine Learning

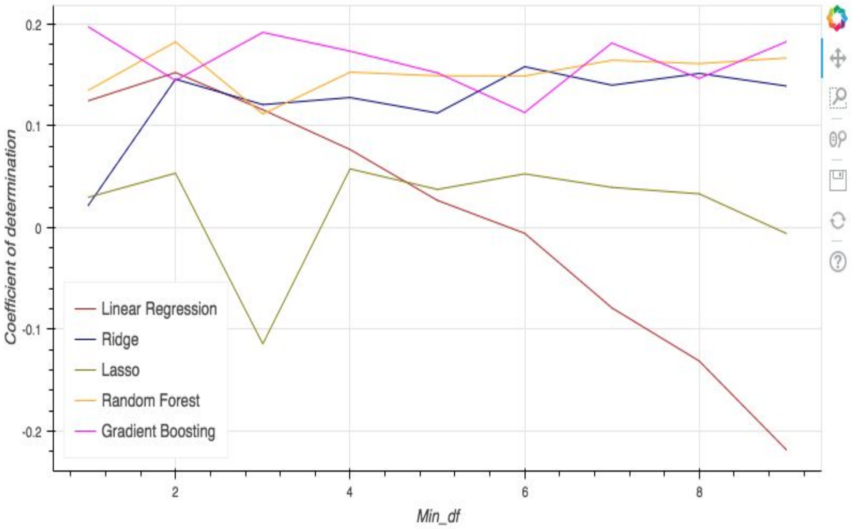
- Five machine learning models on numeric data



Machine Learning

- Five machine learning models on text data using CountVectorizer & TfidfVectorizer





Machine Learning

- Two machine learning models on numeric and text data using CountVectorizer & TfidfVectorizer:

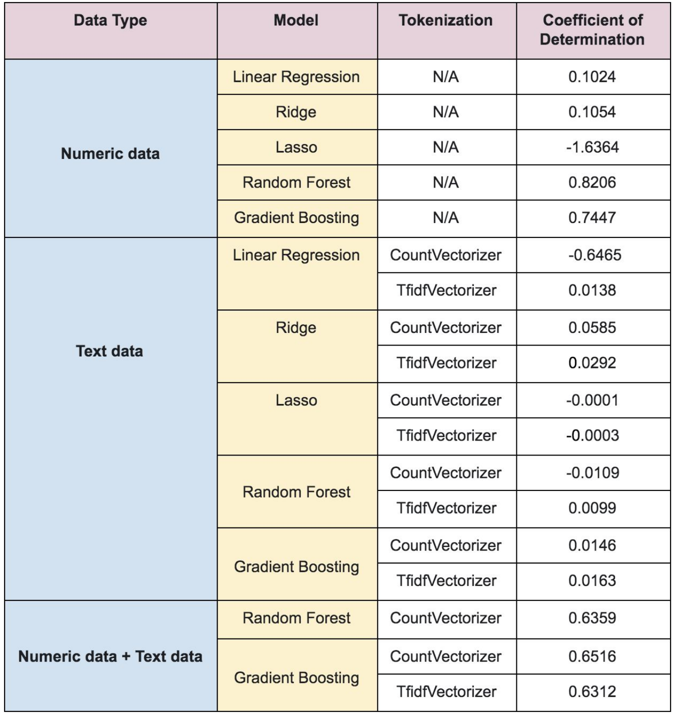
  Random Forest + CountVectorizer

  Gradient Boosting + CountVectorizer

  Gradient Boosting + TfidfVectorizer

Machine Learning

- Random Forest on numeric data performs best



Data Type	Model	Tokenization	Coefficient of Determination
Numeric data	Linear Regression	N/A	0.1024
	Ridge	N/A	0.1054
	Lasso	N/A	-1.6364
	Random Forest	N/A	0.8206
	Gradient Boosting	N/A	0.7447
Text data	Linear Regression	CountVectorizer	-0.6465
		TfidfVectorizer	0.0138
	Ridge	CountVectorizer	0.0585
		TfidfVectorizer	0.0292
	Lasso	CountVectorizer	-0.0001
		TfidfVectorizer	-0.0003
	Random Forest	CountVectorizer	-0.0109
		TfidfVectorizer	0.0099
	Gradient Boosting	CountVectorizer	0.0146
		TfidfVectorizer	0.0163
Numeric data + Text data	Random Forest	CountVectorizer	0.6359
	Gradient Boosting	CountVectorizer	0.6516
		TfidfVectorizer	0.6312

Conclusion

- The model with highest accuracy is Random Forest Regression on numeric data, the top 3 most important features are user_total_useful, total_cool and total_funny, represent the total useful vote, total cool vote and total funny vote the user got from his/her other reviews