

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

Exploratory Data Analysis

- Review word cloud



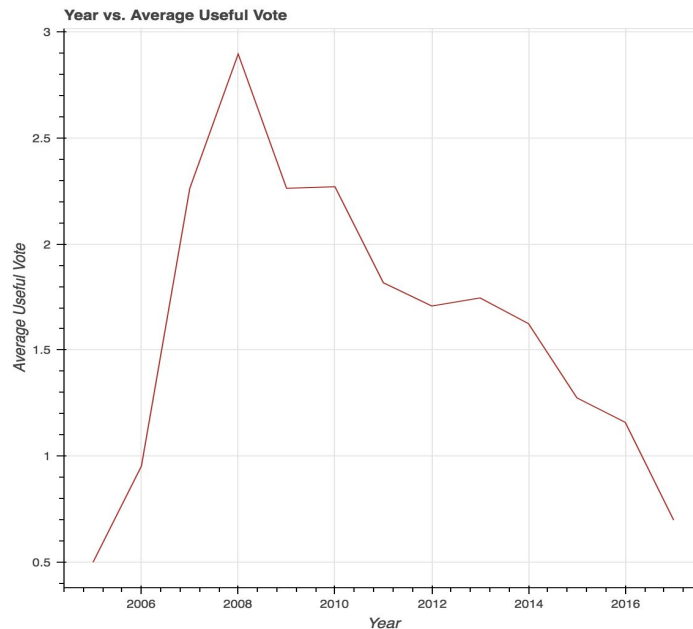
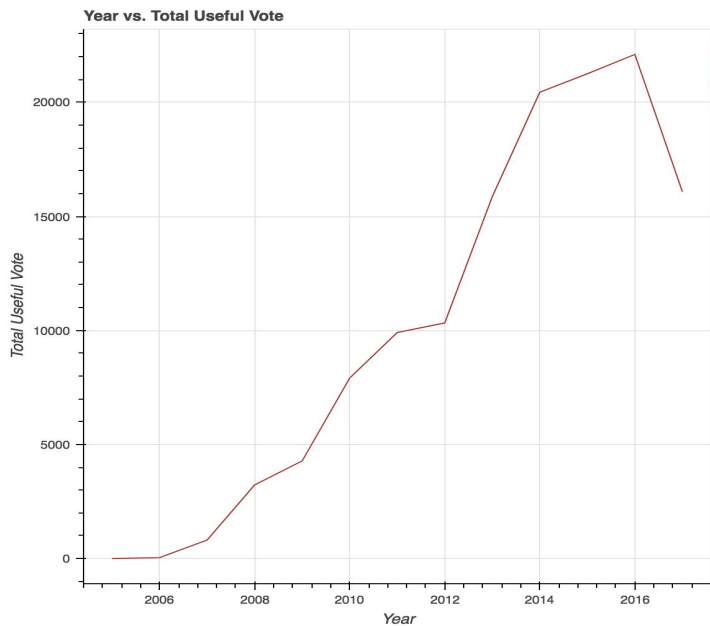
Exploratory Data Analysis

- Positive reviews word cloud
- Negative reviews word cloud



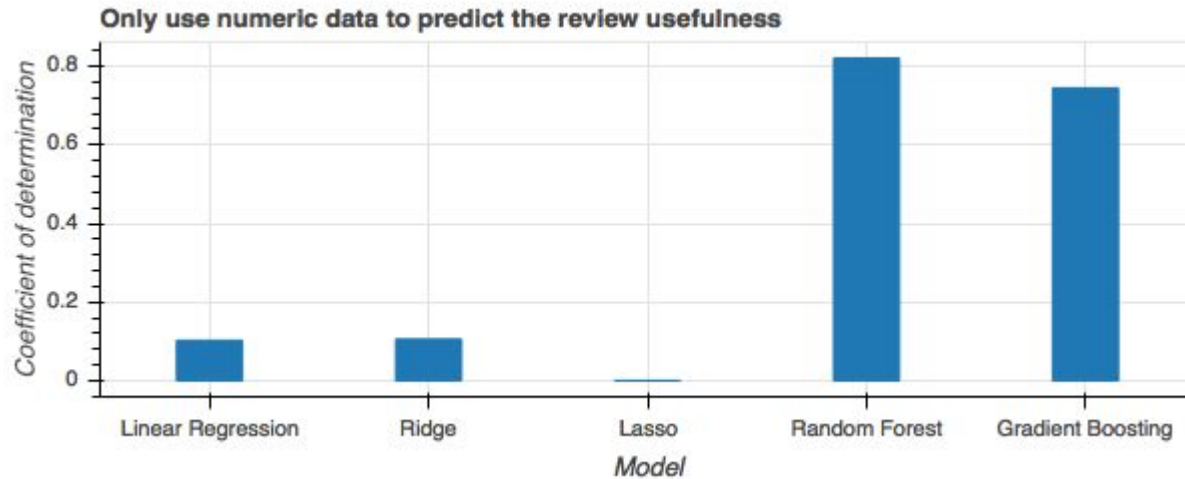
Exploratory Data Analysis

- Time influence on total useful vote and average useful vote



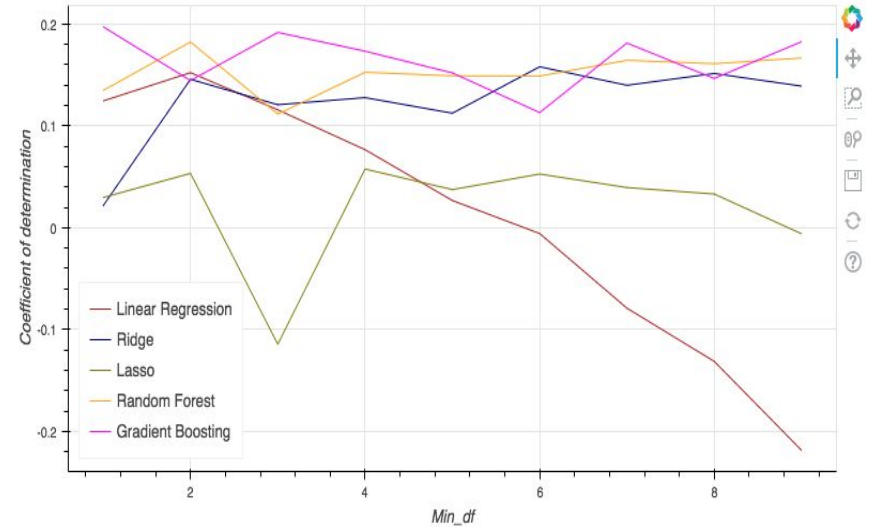
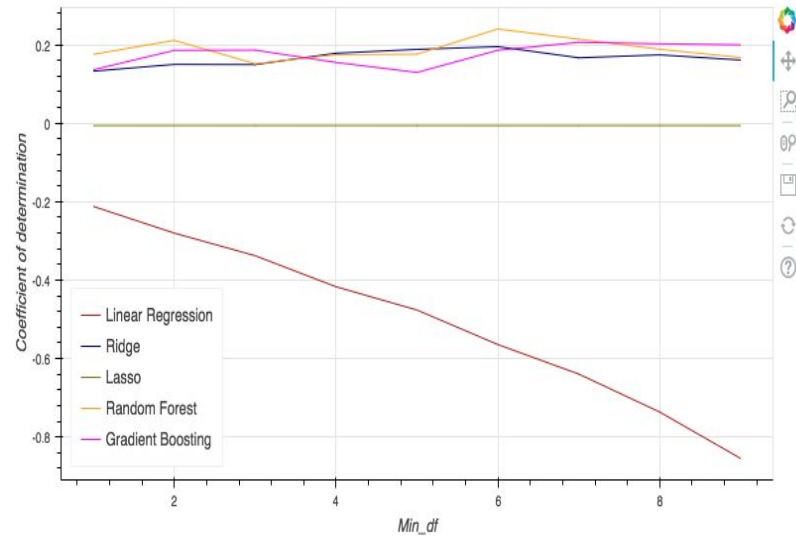
Machine Learning

- 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