

# Predicting **yelp** Review Quality

Jeff Johannsen



# Jeff Johannsen

- Illinois to Colorado in 2016
- Finance > Supply Chain > Process Improvement > Data Science
- I love to explore and support local businesses



# Quality Reviews Are Important

Review Sites



Better user  
engagement and  
satisfaction

Users



Better choices and  
easier decision  
making

Small Businesses



More effective  
advertising and  
customer acquisition

# Central Questions

Can the quality of a review be determined by the review text or the data surrounding the review?

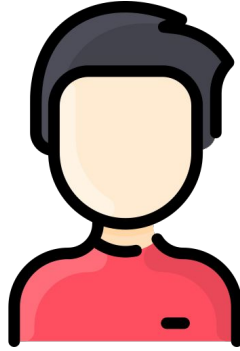
What types of data are most useful for predicting review quality?



# Yelp Open Dataset



Review Text



User Data



Business Data



Review Data

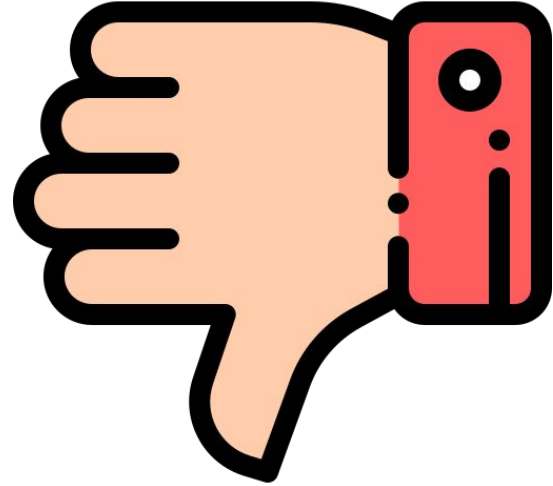


# Target - Binary Classification of Reviews



**Quality**

One or more votes



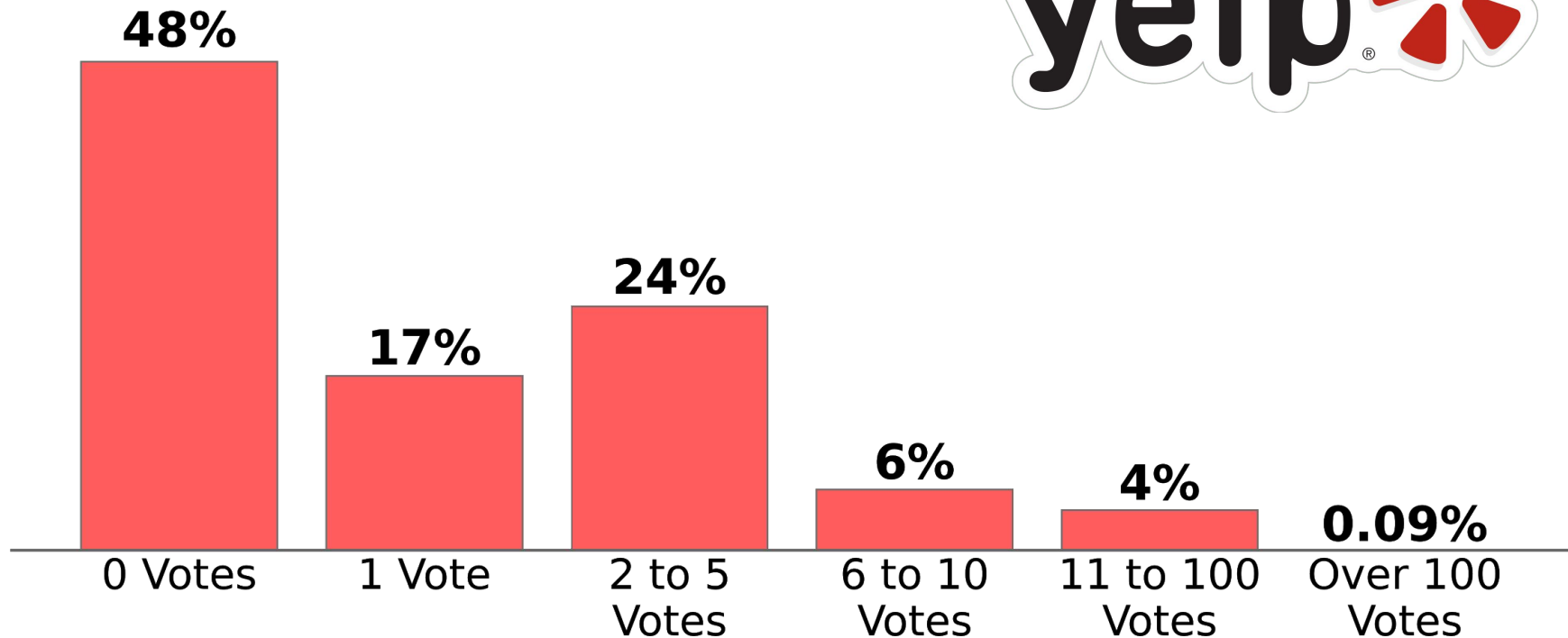
**Not Quality**

Zero votes

# Votes Per Review



Percentage of Reviews



Number of Useful, Funny, or Cool Votes



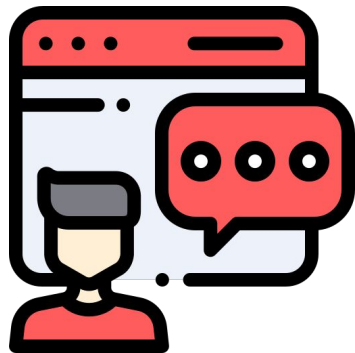
# Feature Engineering

Examples:

Review star rating vs  
average star rating

User's average votes  
per review

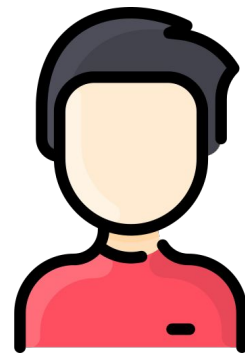
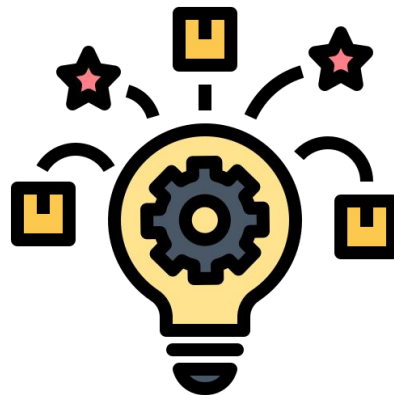
User's votes per year



Review Data



Date and Time



User Data



Business Data



# NLP Feature Engineering

## Basic Text Features

- Review Length, Word Count, etc.



# NLP Feature Engineering

## Basic Text Features

- Review Length, Word Count, etc.

## Readability

- Flesch–Kincaid Grade Level



# NLP Feature Engineering

## Basic Text Features

- Review Length, Word Count, etc.

## Readability

- Flesch–Kincaid Grade Level

## Parts of Speech

- Noun, Verb, Adjective, etc.



# NLP Feature Engineering

## Basic Text Features

- Review Length, Word Count, etc.

## Readability

- Flesch–Kincaid Grade Level

## Parts of Speech

- Noun, Verb, Adjective, etc.

## Syntactic Dependency Relations

- Sentence Structure

## Named Entities

- Person, Place, Event, etc.



# NLP Feature Engineering

## Basic Text Features

- Review Length, Word Count, etc.

## Readability

- Flesch–Kincaid Grade Level

## Parts of Speech

- Noun, Verb, Adjective, etc.

## Syntactic Dependency Relations

- Sentence Structure

## Named Entities

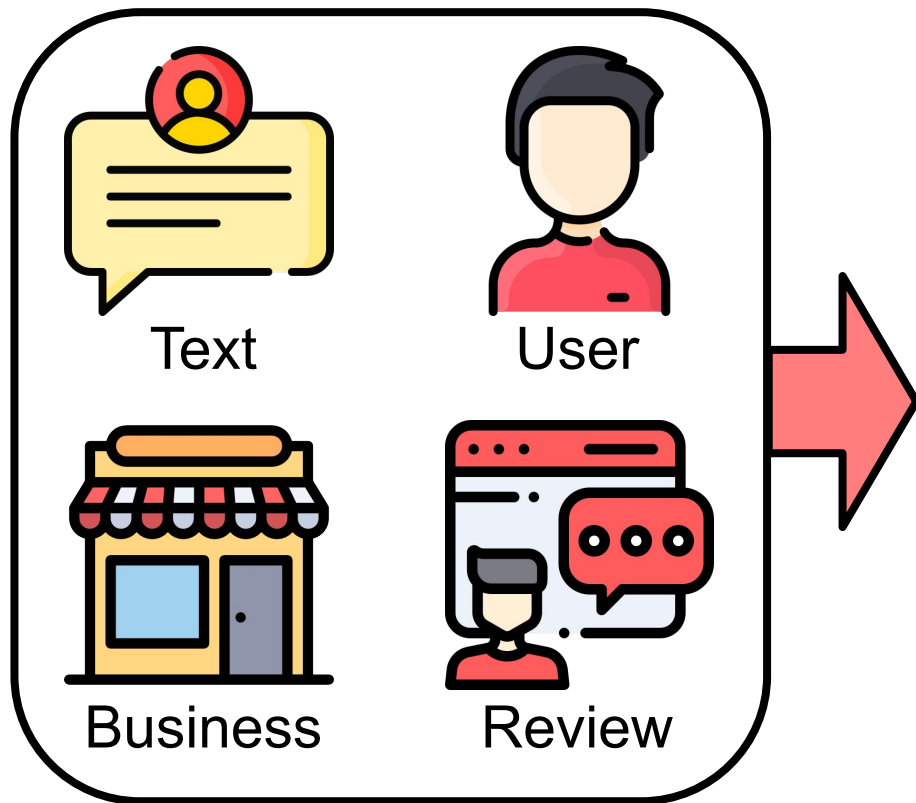
- Person, Place, Event, etc.

## ML Model Predictions

- **SVM and Naive Bayes using TF-IDF**

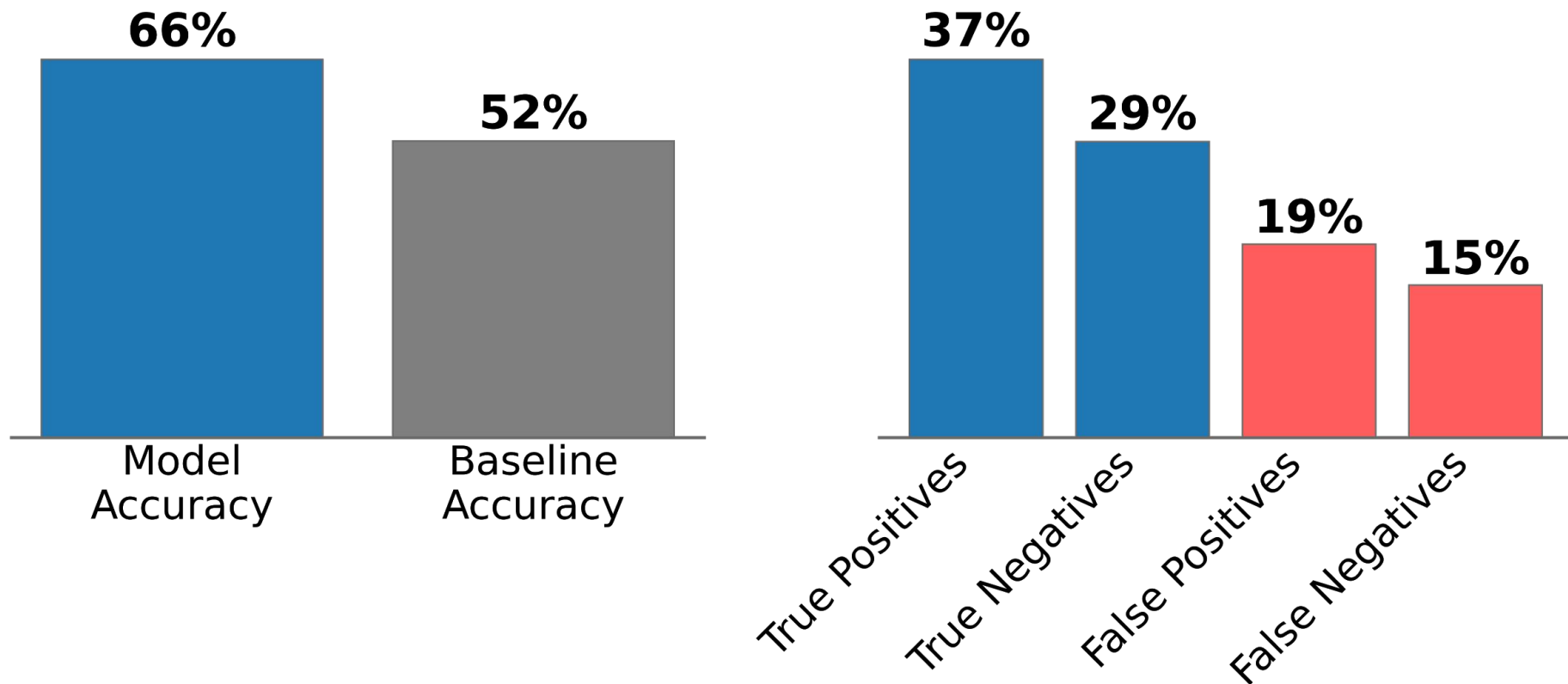


# Making Predictions



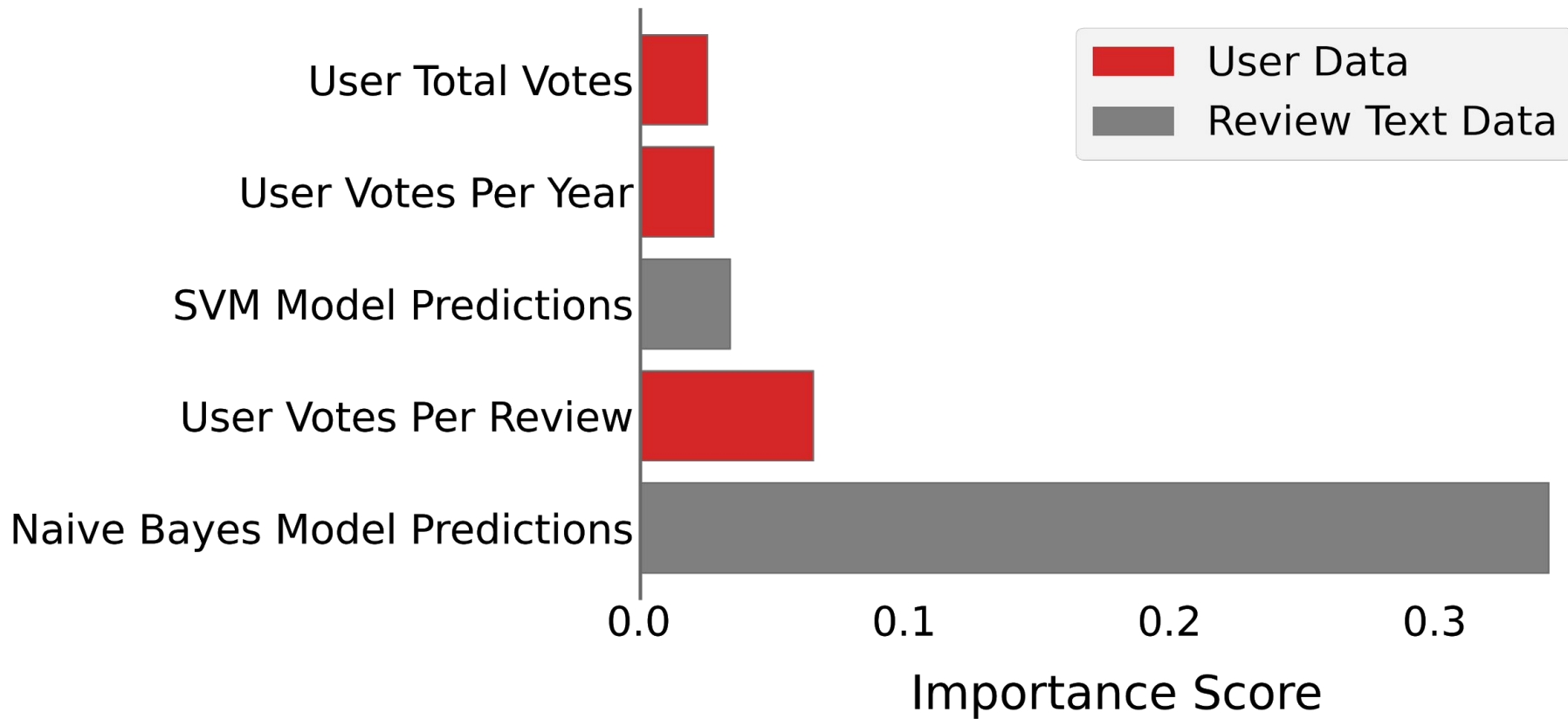
**Random Forest**

# Model Prediction Results





# Important Features



# Conclusions

The quality of reviews can be determined.

Data about the review text and the user are the top predictors of review quality.

## Next Steps

- Advanced NLP
- Scale up using Apache Spark and AWS



# Jeff Johannsen



Gmail

- jeffjohannsen7@gmail.com




GitHub

- jeffjohannsen

LinkedIn

- jeffjohannsen

 pandas

  
docker

  
jupyter



spaCy

  
scikit  
learn

  
matplotlib

python™



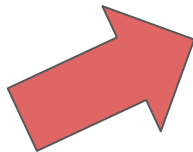
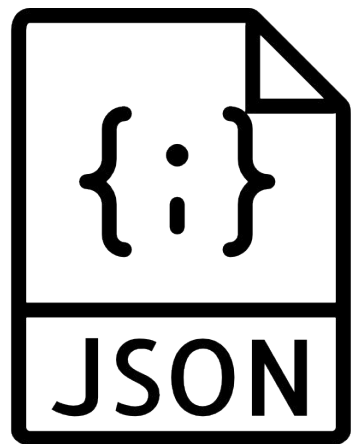
PostgreSQL

# Credits

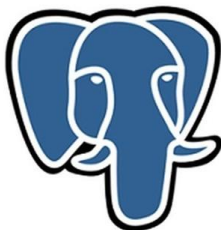
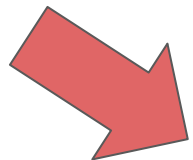
- Yelp - Dataset and logo images
- Great open source data science tools like Pandas and SkLearn
- All of my teachers and fellow cohort members at Galvanize-Denver



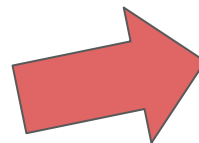
# Data Storage



mongoDB



PostgreSQL



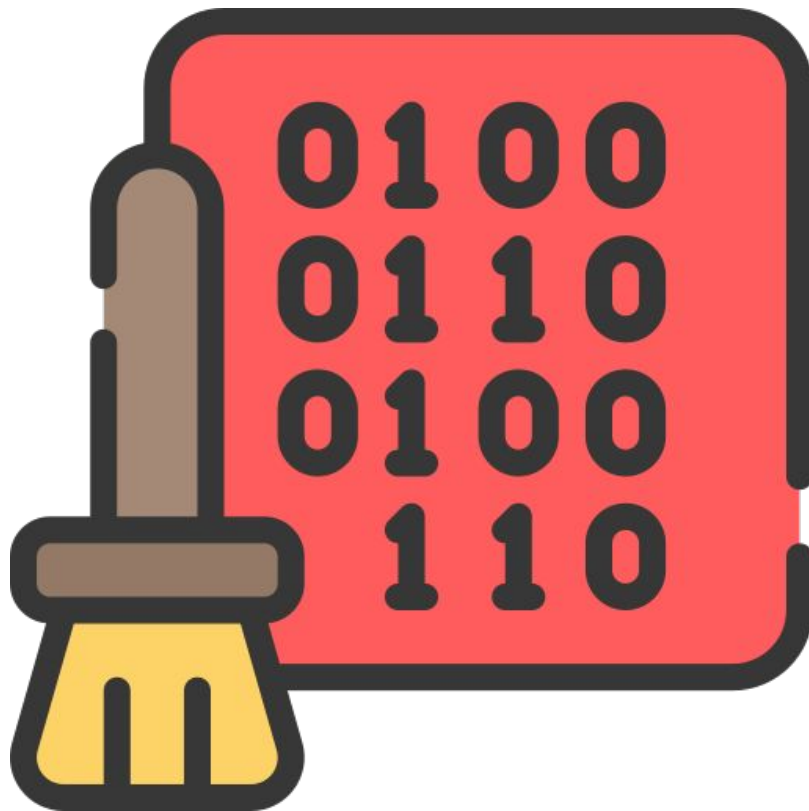
aws



Amazon RDS

# Data Cleaning

- Dropped Nan/Null Values
- Removed Duplicate Records
- Deleted Unnecessary Features
- Converted Data-types
- Organized Features



# Time Discounting

## Targets

- Votes Per Year instead of Total Vote Count

## Features

- Actual or estimated counts at the time of the review instead of at the time the dataset was released

