Tripadvisor Reviews

w/ Natural Language Processing

Anisha Malhotra
Ryan Lewis

Table of contents

01
Business Problem & Overview

02
Exploratory Data Analysis

03
Preprocessing & Vectorizing

04
Modeling

05
Conclusion & Next Steps

01

Business Problem

- Categorize reviews as “poor”, “average”, and “excellent”
- Helps consumers and business owners get a better summary of each product or service
- Helpful for systems that do not already have preset review or rating system

Tripadvisor

- World's largest travel guidance platform
- Helps travelers plan, book & take trips
- Assists travelers discover where to stay, eat & sleep
- 884M+ reviews of ~8M businesses globally

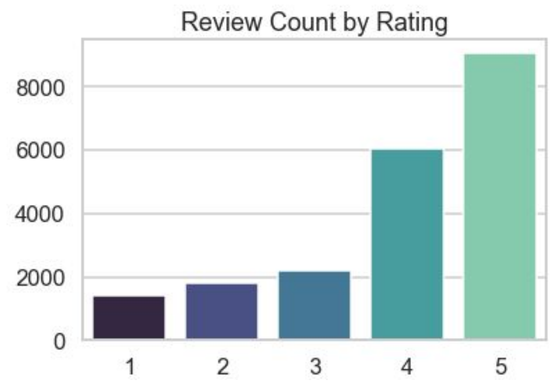
Dataset

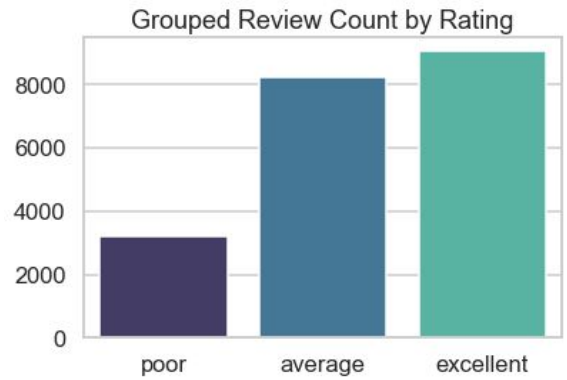
- Pre-scraped Kaggle dataset
- 20K+ hotel reviews
- Rating based on 1-5 scale

02

Exploratory Data Analysis

- Due to class imbalance in original data set, we grouped rating scores together
  - ➢ 1 & 2 -- 'poor'
  - ➢ 3 & 4 -- 'average'
  - ➢ 5 -- 'excellent'





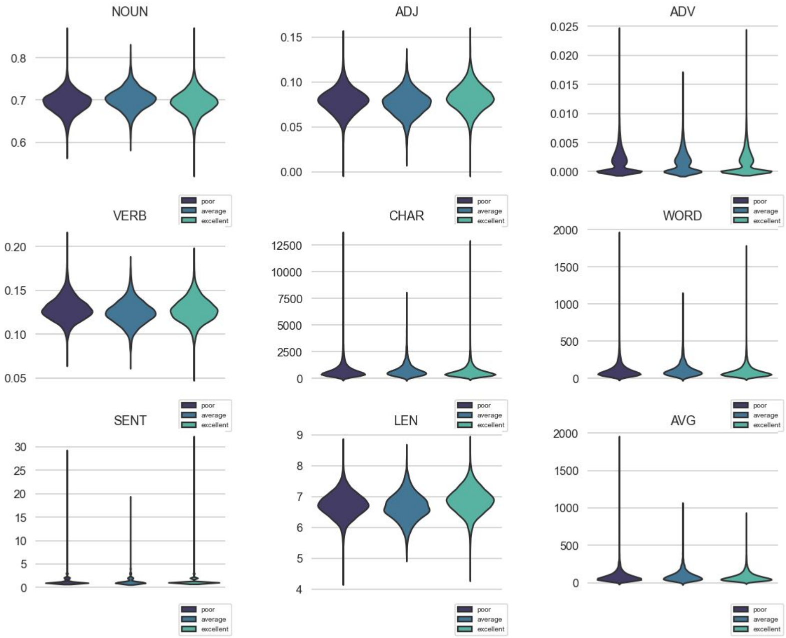
Word Clouds

- ❖ Shows how common certain words are for each category
  - ➢ “best”, “nice”, “deceptive”

excellent

average

poor



Violin Plot

- Frequency of nouns, adjectives, verbs are fairly even across all three review types
- Word count, character count and average length of sentences tend to be higher for 'poor' reviews

03

Preprocessing & Vectorizing

- Preprocessing
  - removing punctuation
  - lower-cased words
  - removing stop-words
  - assigned part of speech tags
  - lemmatizing words
  - tokenized remaining words

- TF-IDF Vectorizer
  - Better results than Countvectorizer
  - Min_df = .10
  - Max_df = .80
  - uni-grams

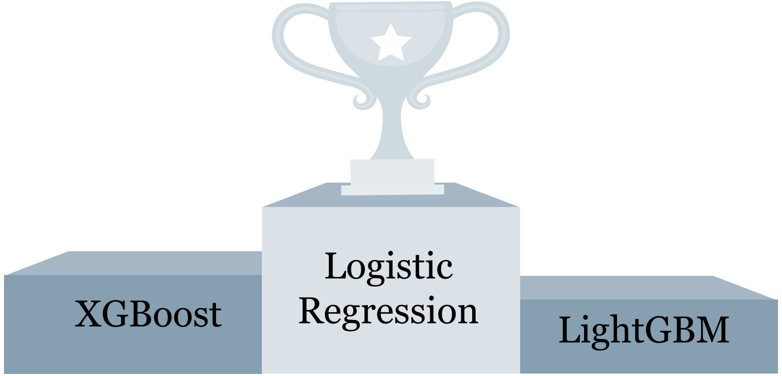
04

Modeling

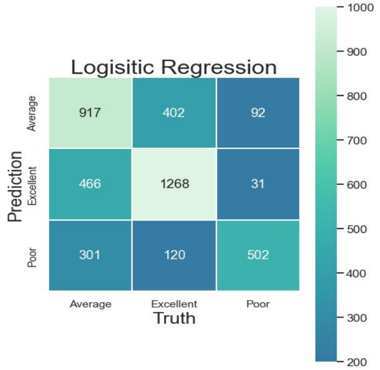
name	accuracy score	f1 score
Naive Bayes	0.624787	0.604537
Logistic Regression	0.655526	0.653858
Logistic Regression (PCA)	0.650159	0.653858
Decision Tree	0.572091	0.571000
Decision Tree (PCA)	0.595755	0.593609
Random Forest	0.616004	0.599506
XGBoost	0.650646	0.648684
Light GBM	0.647963	0.645438
KNN	0.601610	0.592157

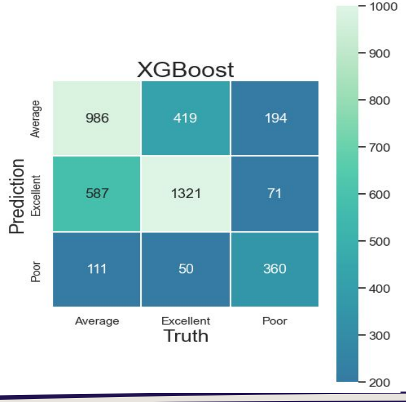
- All models were grid-searched to determine optimal hyperparameters
- Focused on accuracy score
- Top model: Logistic Regression
- Accuracy score: .655
- F1 score: .653

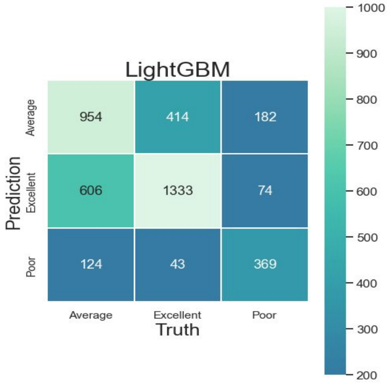
- Top 3 Models: Logistic Regression, XGBoost, & LightGBM
- Logistic Regression model did the best at identifying the 'poor' reviews
- 'average' and 'excellent' were relatively similar across the three models



Confusion Matrices

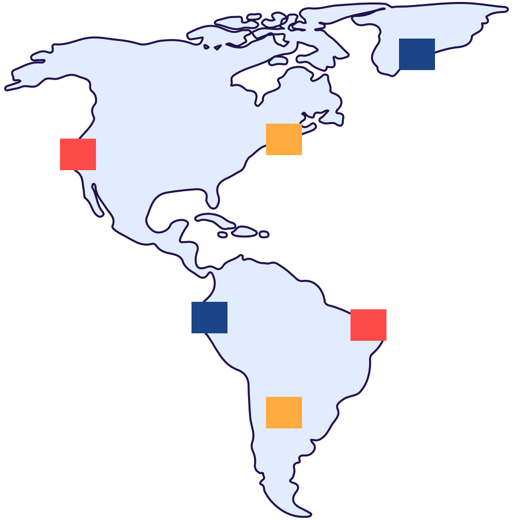






05

Conclusion & Next Steps



- Per our business problem, Logistic Regression model is best model for taking in text data and accurately categorizing the user review sentiment

- Next Steps:
  - Utilize deep-learning to create potentially more accurate models
  - Incorporate n-grams
  - Try different preprocessing techniques
  - Bring in new unseen data to assess our model performance
  - Resample “poor” reviews to even out classes
  - Include exogenous features

Thank you!

Anisha Malhotra - https://github.com/anisha732

Ryan Lewis - https://github.com/rylewww