RATINGS OF YELP REVIEWS PREDICTION

Stat 628 Module 2 [2/2]

Kai Wang, Yichen Shi, Yanran Wang

Group 4, Mon

TABLE OF CONTENTS

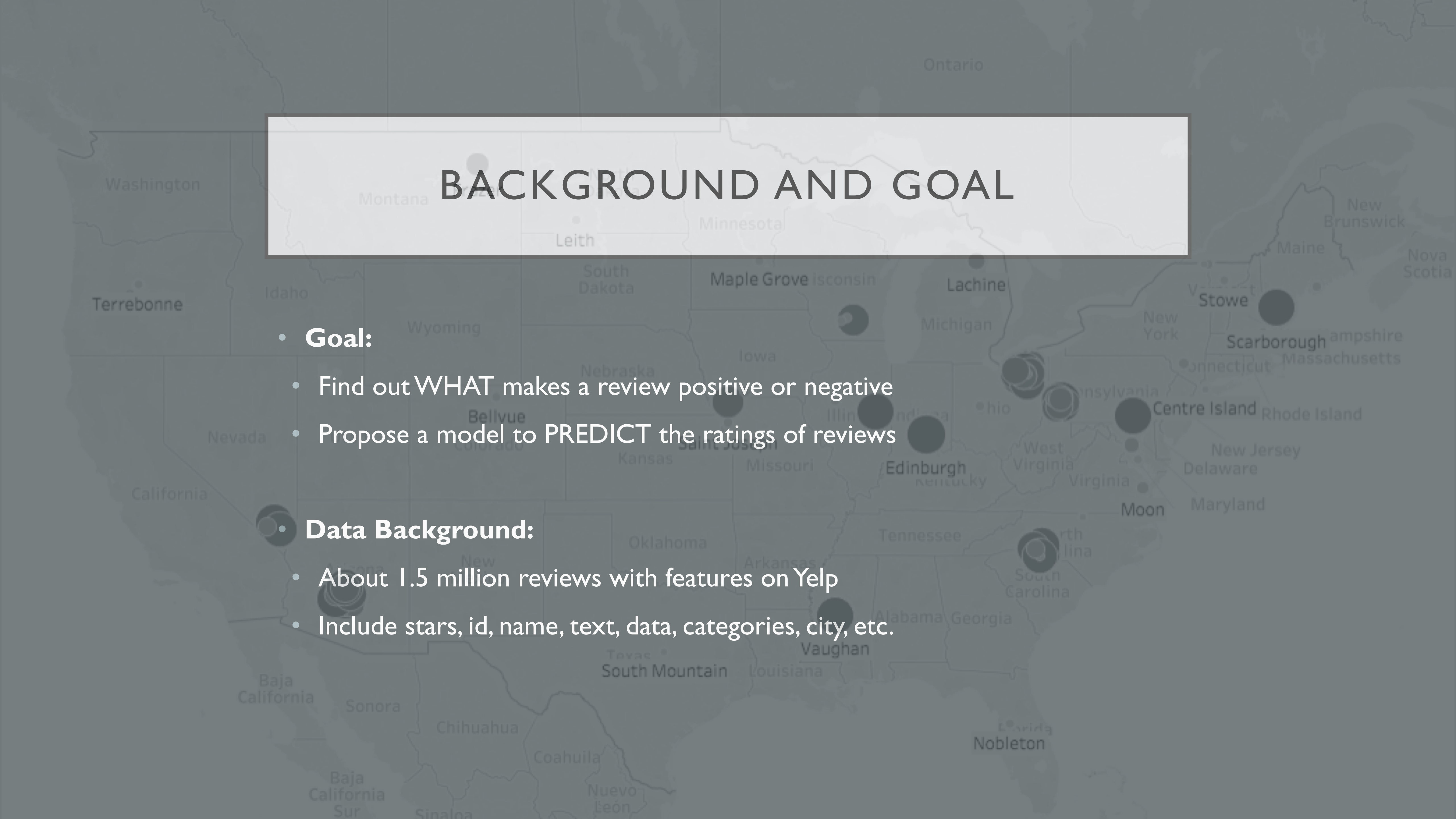
- **Background and Goal**
- **Features Processing**
  - Other feathers
  - Text
- **Model Fitting**
  - Prediction Model
  - Interpretable Model

BACKGROUND AND GOAL

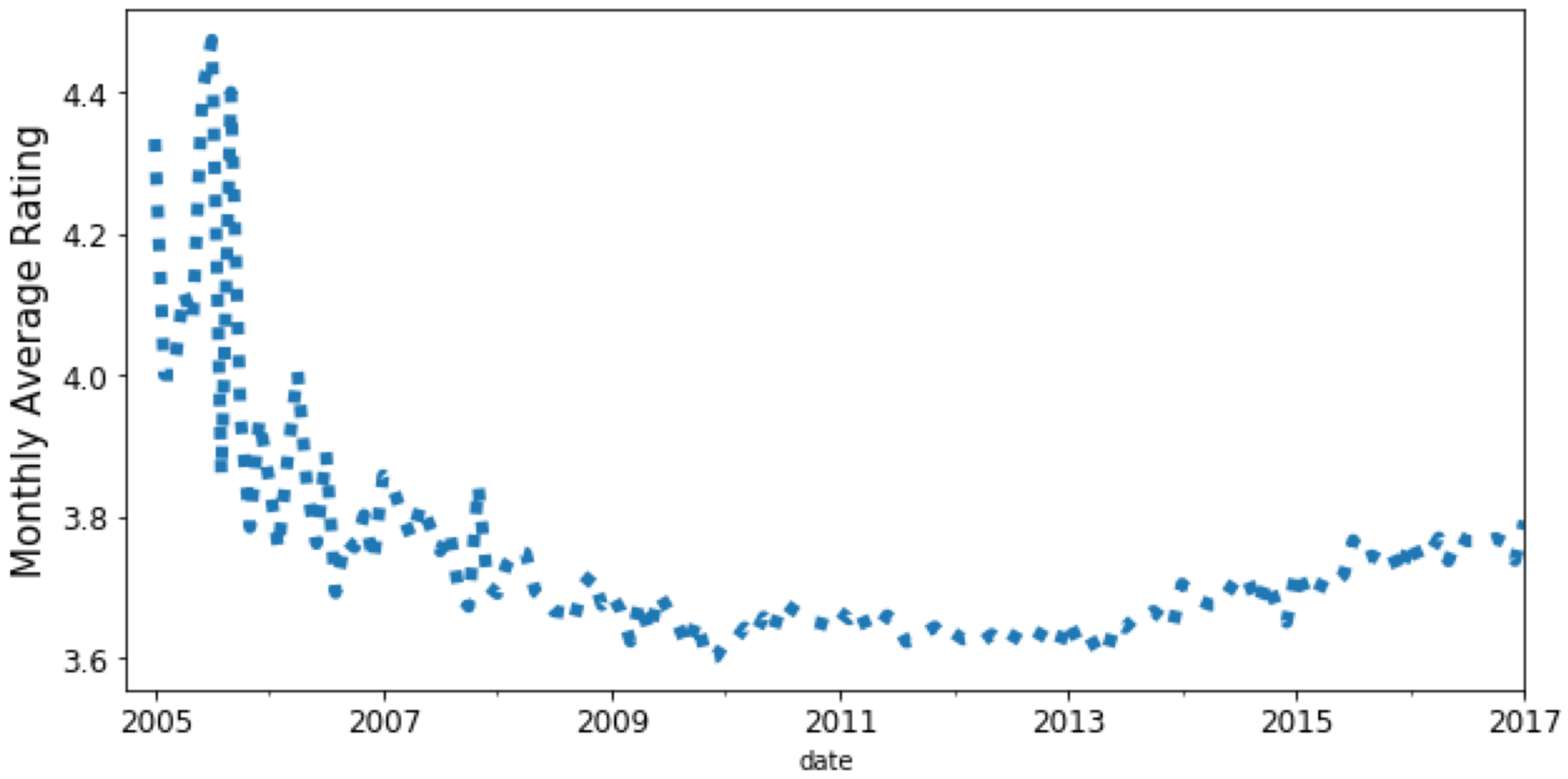
BACKGROUND AND GOAL

- **Goal:**
  - Find out WHAT makes a review positive or negative
  - Propose a model to PREDICT the ratings of reviews

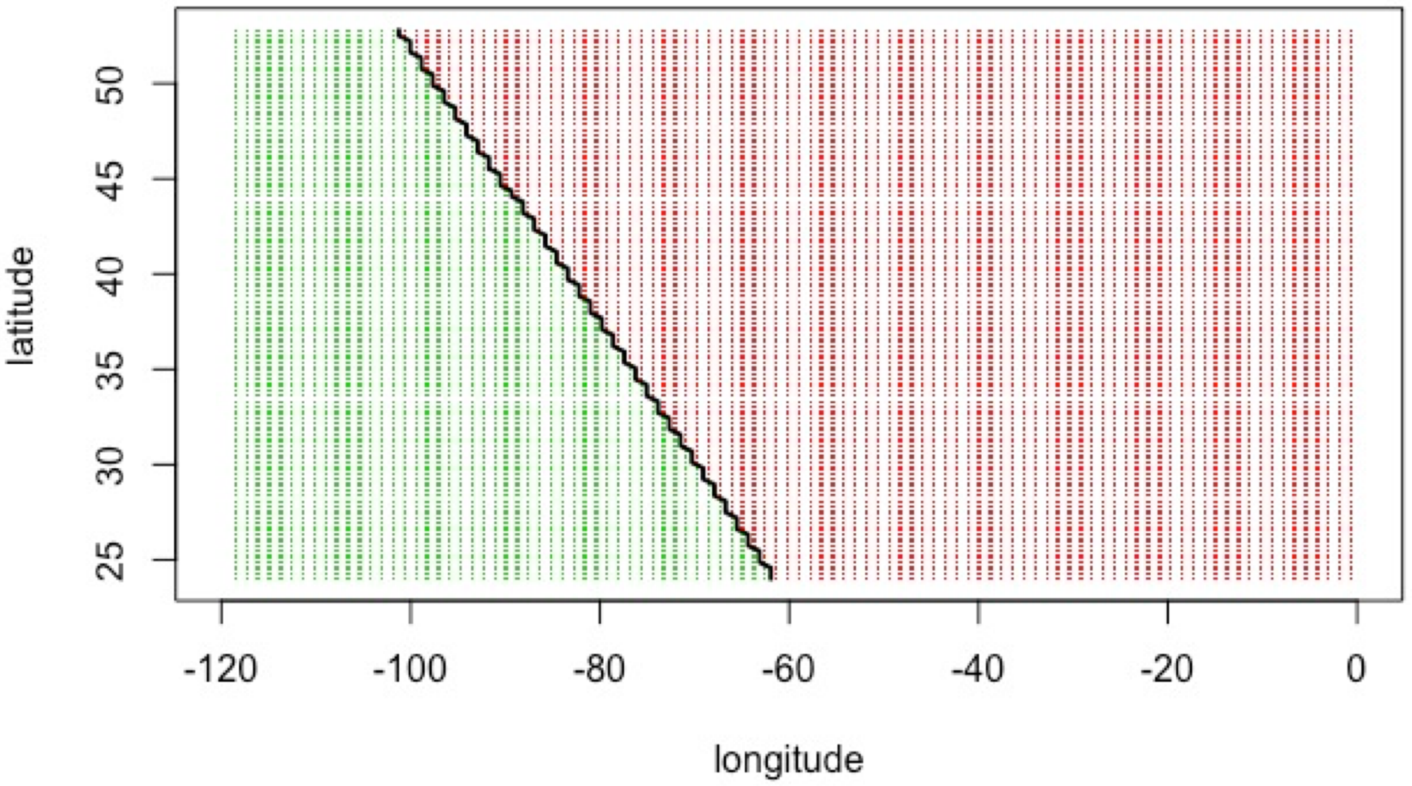
- **Data Background:**
  - About 1.5 million reviews with features on Yelp
  - Include stars, id, name, text, data, categories, city, etc.



FEATURES PROCESSING


Monthly Average Rating
4.4
4.2
4.0
3.8
3.6
2005
2007
2009
2011
2013
2015
2017
date

-- DATE --

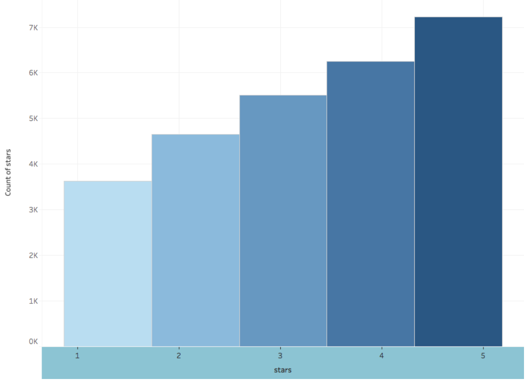


-- CITY, LONGITUDE & LATITUDE --

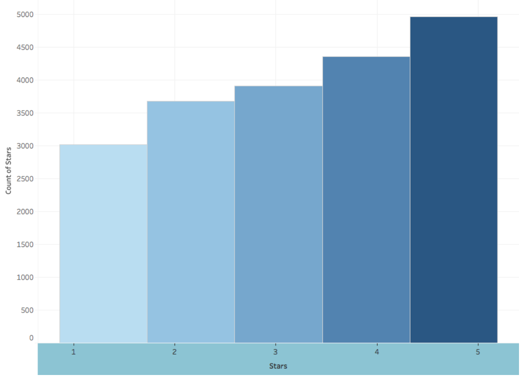
Categories Processing

Coffee & teas
↓
coffeeteas

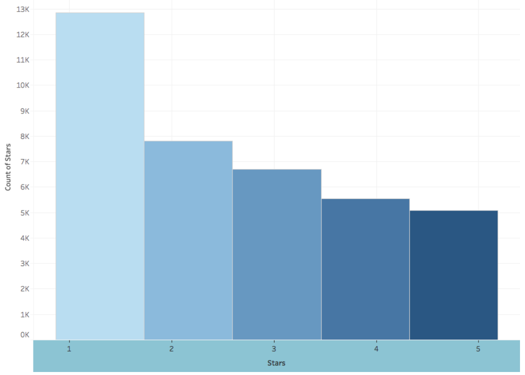
Fast food
↓
fastfood



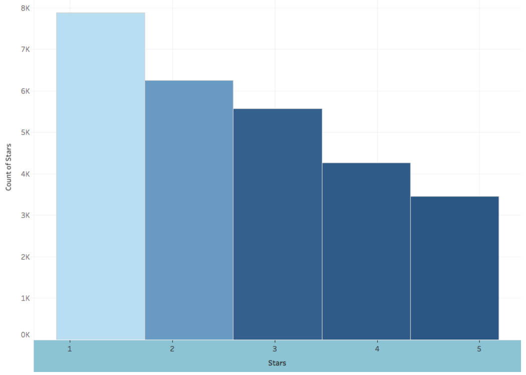
Coffee & teas



Wine bars

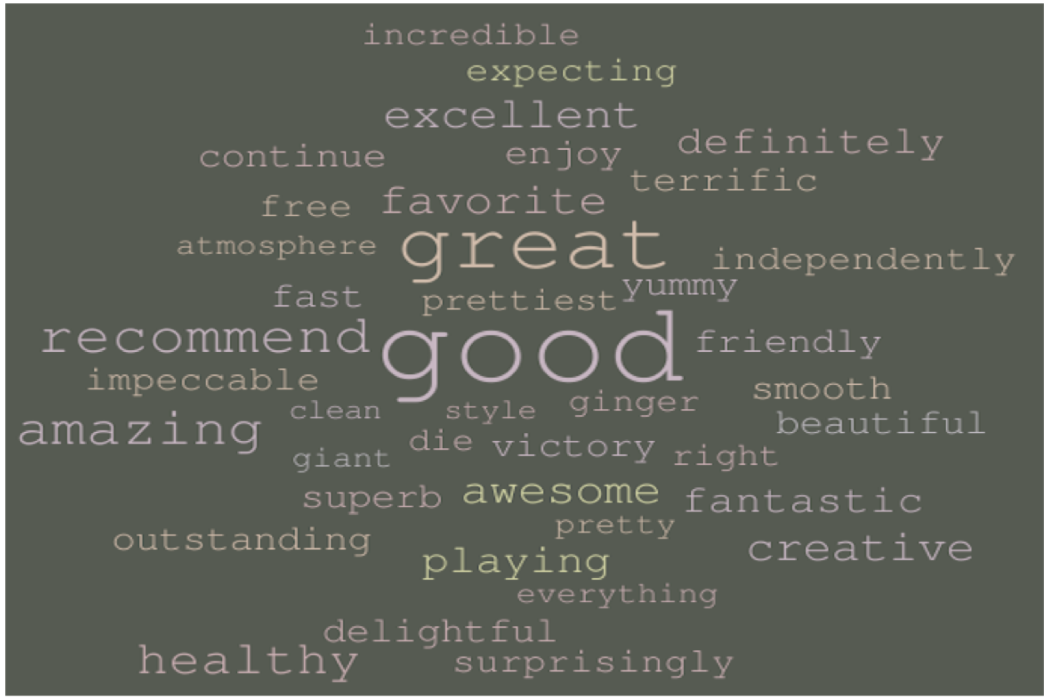


Fast food

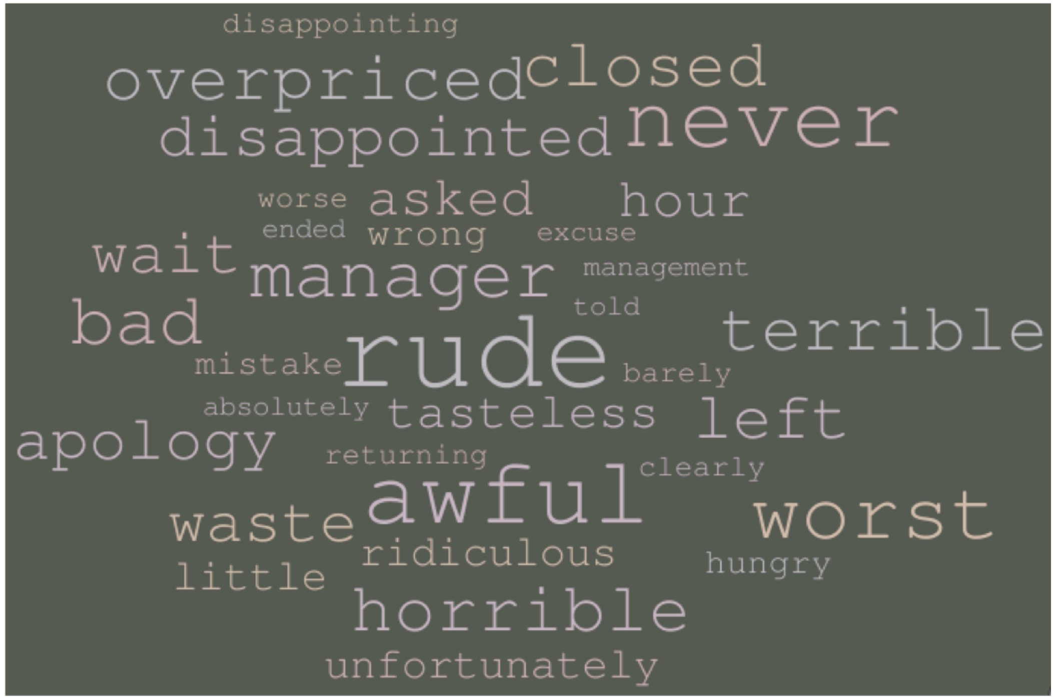


Sports bars

-- CATEGORIES --



Positive Word



Negative Word

-- Word Cloud --

A Bag of Word Matrix:

- Lowercase
- Extract the word stem: ‘sitting’ >> ‘sit’, etc.
- Remove stopwords: ’is’, ‘was’, ‘were’, etc.
- Remove punctuations

Dimension Reduction:

- Chi-Square Test
- K most informative columns

-- TEXT --

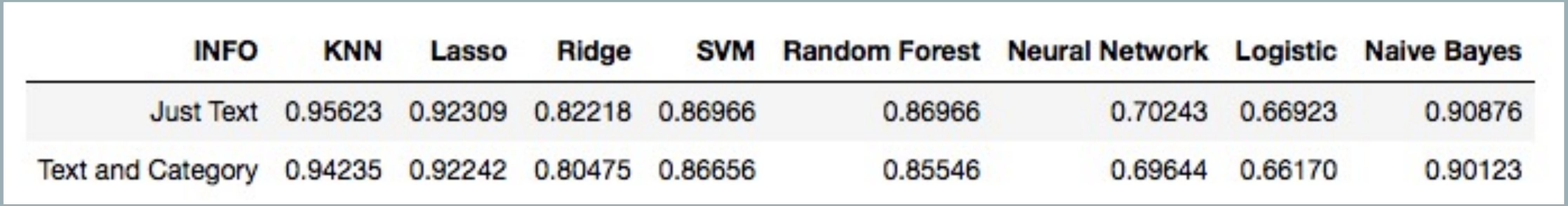
- Number of specific punctuations and expressions: ‘?’, ’!’, ’:)’, ’:D’, etc.
- Number of all caps: ‘GOOD’, ’NOT’, etc.
- Text length

-- ADD MORE FEATURES --

MODEL FITTING

PREDICTION MODEL

RMSE COMPARISON FOR DIFFERENT METHODS



INFO	KNN	Lasso	Ridge	SVM	Random Forest	Neural Network	Logistic	Naive Bayes
Just Text	0.95623	0.92309	0.82218	0.86966	0.86966	0.70243	0.66923	0.90876
Text and Category	0.94235	0.92242	0.80475	0.86656	0.85546	0.69644	0.66170	0.90123

MODEL FITTING

INTERPRETABLE MODEL

$$Stars = 3.741 + 1.108723 \times TextScore + 0.042362 \times CategoryScore$$
$$-0.001122 \times longitude \times Month + 0.000629 \times Day$$

Final Model:

$$Stars = 3.741 + 1.108723 \times TextScore + 0.043080 \times CategoryScore$$

STRENGTHS AND WEAKNESS

1. Strengths

- Prediction Model: Small RMSE
- Interpretable Model: Easy to understand

2. Weaknesses

- Prediction Model: Hard to interpret
- Interpretable Model: Low accuracy