McDonald's Review Predictions

MS-AAI-501 Team 1 Final Project

Paul Parks, Lishi Wang, Ivan Steklov

Introduction

Introduction

Customer feedback and online **reviews** significantly influence consumer behavior.

Understanding and analyzing this feedback is essential for businesses aiming to thrive in competitive markets.

Can we determine customer **sentiment** purely based on text **reviews**?

Objective

- Conduct **sentiment analysis** on McDonald's customer reviews

- Explore **geographic trends** in reviews and ratings

- **Predict star ratings** based on review text.

Target Audience

Our target audience is **business leaders in the Food industry** who wish to improve the reviews of their storefronts or create an easier way for customers to create reviews.

The ability to calculate sentiment based on text reviews can have positive benefits including:

- Customer Insights
- Product and Service Improvement
- Customer Segmentation

Dataset

Dataset

Dataset collection of over 33,000 anonymized Google reviews for various McDonald's locations within the United States

- 10 Features
- Kaggle dataset from public Google reviews
    - https://www.kaggle.com/datasets/nelgiriyewithana/mcdonalds-store-reviews

Dataset

- **reviewer_id**: A unique identifier for each reviewer, ensuring anonymity.
- **store_name**: The name of the specific McDonald's store under review.
- **category**: The category or type of store, providing context to the location.
- **store_address**: The physical address of the store.
- **latitude**: The latitude coordinate of the store's geographical location.
- **longitude**: The longitude coordinate of the store's geographical location.

Dataset

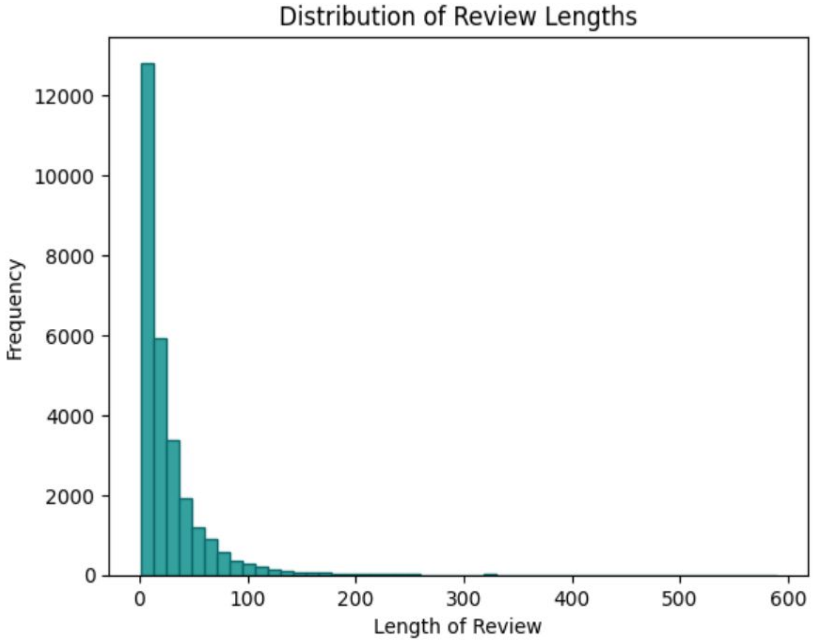
- **rating_count**: The total number of ratings or reviews for the particular store.
- **review_time**: The timestamp when the review was posted.
- **review**: The textual content of the customer's review.
- **rating**: The rating provided by the reviewer on a predetermined scale.

Dataset

	store_name	category	store_address	latitude	longitude	rating_count	review_time	review	rating
reviewer_id									
1	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	3 months ago	Why does it look like someone spit on my food?...	1 star
2	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	5 days ago	It'd McDonalds. It is what it is as far as the...	4 stars
3	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	5 days ago	Made a mobile order got to the speaker and che...	1 star
4	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	a month ago	My mc. Crispy chicken sandwich was ï¿½ï¿½ï¿½ï¿...	5 stars
5	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	2 months ago	I repeat my order 3 times in the drive thru, a...	1 star

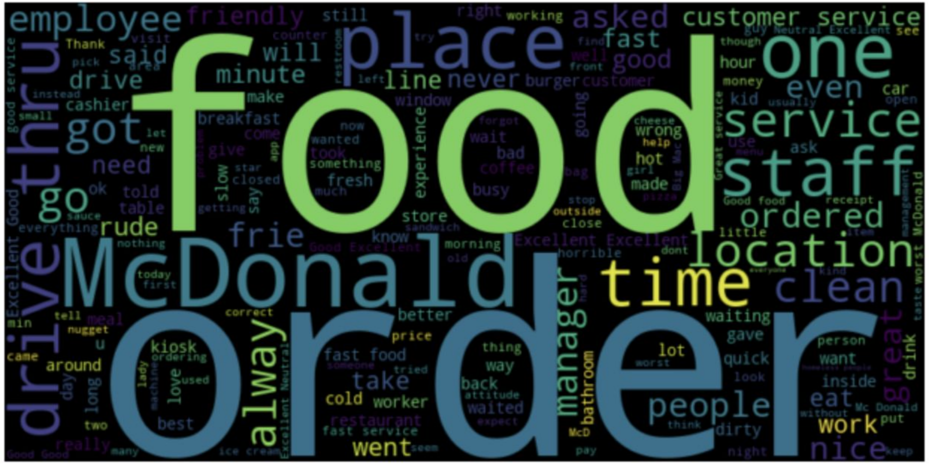
Dataset - Review

The review lengths range from a minimum of **1** word to a maximum of **589 words**. The majority of reviews are only a few words long.



Dataset - Review

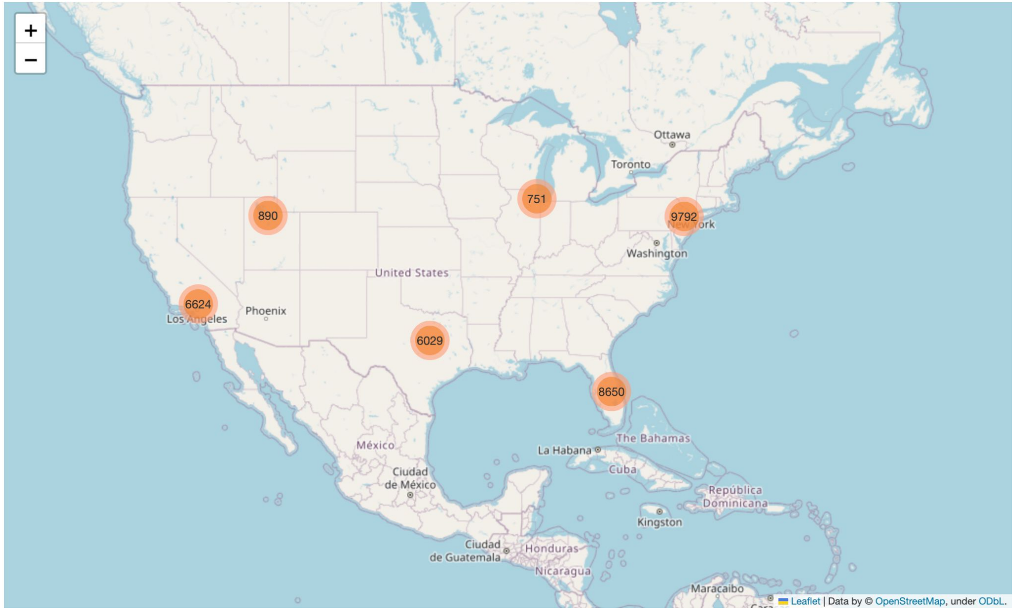
The most commonly used words are “I”, “food”, “order”, “service”, “The”, “McDonald”, “get”, “place”, “good”, and “time”.



Location Analysis

Dataset - Locations

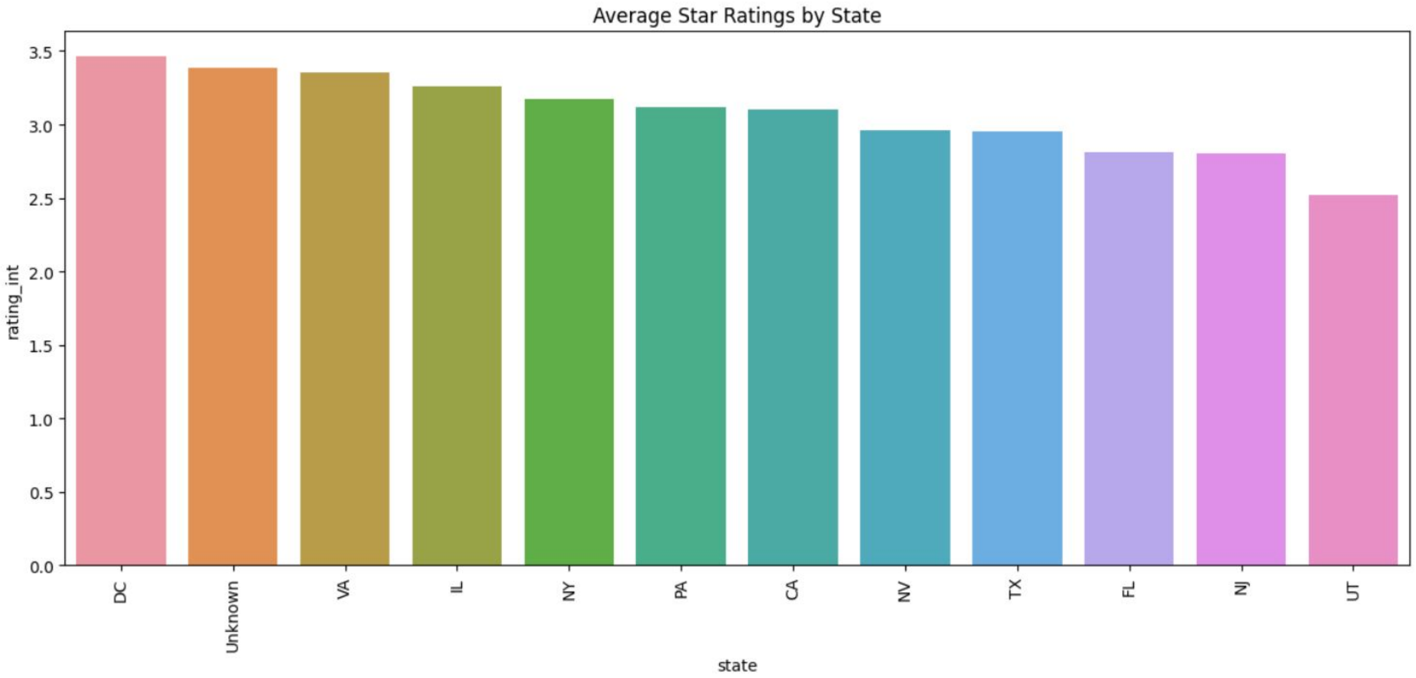
The dataset contains 33,000 anonymized Google reviews from 11 States.



Location - State

A geolocation analysis was conducted to understand the geographical patterns of customer satisfaction within the dataset, focusing on the average star ratings across different states in the United States.

Location - State



Cleaning

Dataset Cleaning and Preprocessing

Renamed the 'latitude ' column, which had an extraneous space, to 'latitude' to ensure consistency within the dataset.

Several unwanted characters (e.g., '½ï', 'ï', '½', '¿') were identified in the 'review' and 'store_address' columns. These are errors in encoding and rendering, and are removed.

A new column 'review_string_list' was added, slicing the review into a list of it's individual words, removing punctuation.

Dataset Cleaning and Preprocessing

Many reviews did not have written content. In those cases, the star rating was converted to a default sentiment. The scale is 1=Bad/Terrible, 2=Poor, 3=Neutral, 4=Good, 5=Excellent. These default reviews were removed we want to train the model based on actual text content.

A new column, 'rating_int,' was created, converting the rating data from string to integer format.

A new 'state' column was created, extracting the two-letter state abbreviation from the 'store_address' column

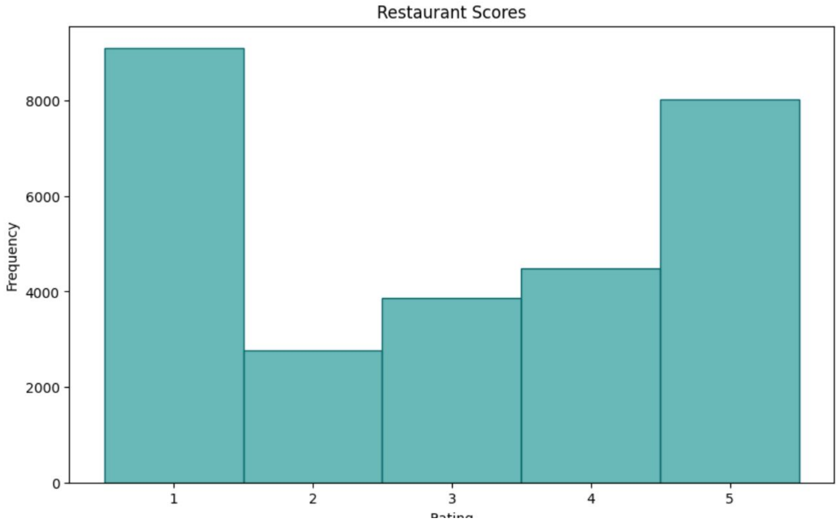
Dataset Cleaning and Preprocessing

After the initial cleaning and preprocessing, we see imbalance in the data. The majority of reviews are either 1 or 5 stars. We will combine 2, 3, 4 into one category. The sentiment scale we will use is:

1 = Terrible

2 = Neutral

3 = Excellent



Sentiment Analysis

Training - Split

The dataset was divided into training and testing subsets, with **80%** allocated for training and **20%** for testing. The training set contained **22582** samples, and the testing set included **5646** samples.

Training - Embedding

The Google Universal Sentence Encoder was used to embed the reviews which would allow the regression model to predict review scores.

Embedding is the process of converting a sentence into a numerical vector by capturing the semantic meaning of the words and the structure of the sentence, allowing it to be mathematically manipulated and analyzed.

Training - Ridge Regression

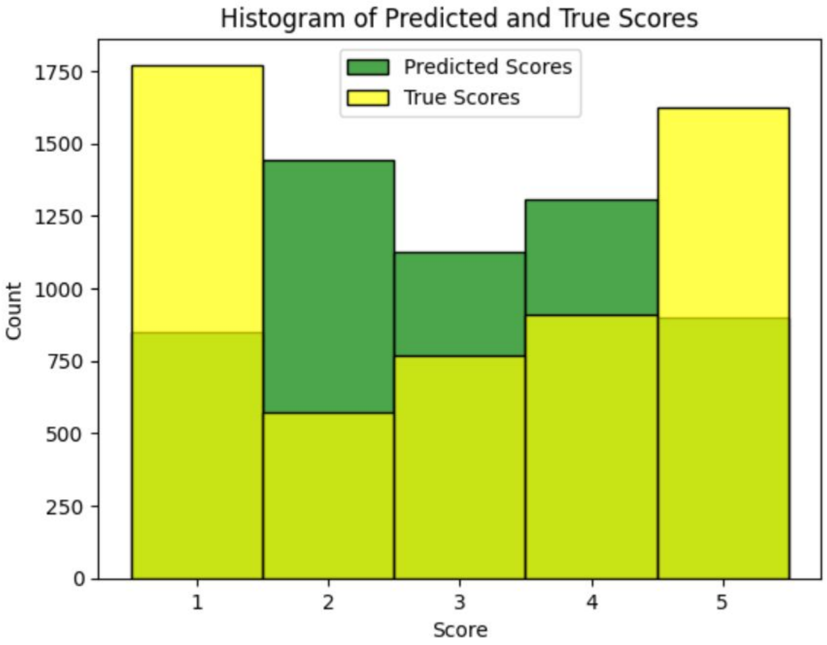
The SciKit Ridge Regression model was used for the Regression model.
We will use this model to try to predict the review score based on text.

The coefficient of determination (R^2 score) for the Ridge Regression model trained on embeddings was 0.6339

Training - Ridge Regression

Predicted scores vs True scores in a histogram.

The histogram shows some overestimation in the neutral (2-4) range, and some underestimation of bad (1) and excellent (3).



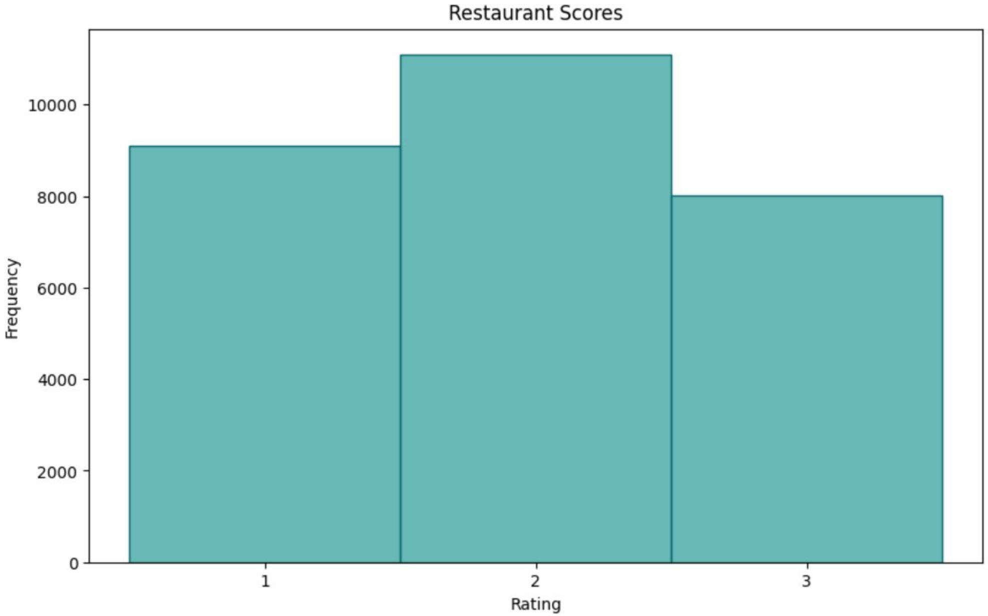
Training - Bayesian

The SciKit Bayesian Ridge Regression model was used for the Bayesian model, to compare to simple Ridge Regression. This produced a model that achieved a r^2 score of 0.641, so it was not a notable improvement over simple Ridge Regression.

Training - SVM

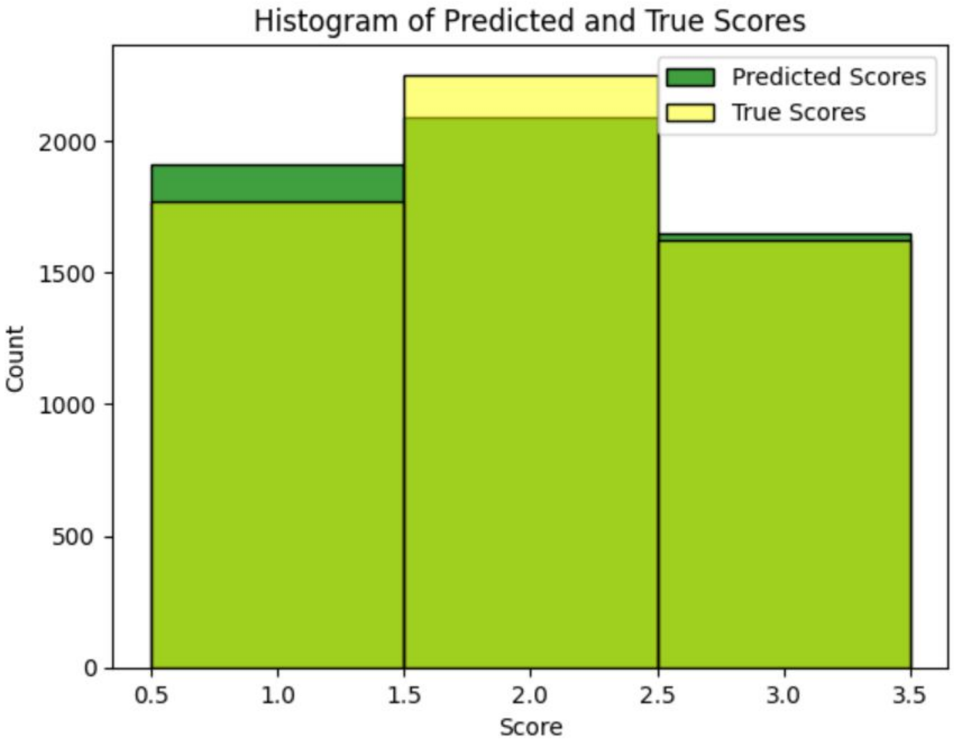
Ridge Regression calculates a continuous score, as a result of calculating regression distance, so it does not do explicit classification. In addition, the imbalance in the data may have cause a loss in accuracy with trying to predict the exact review score.
We trained a SVM model to see how it performs in sentiment classification, and combining the scarcer categories together. The result was that it performed better than Ridge, with an accuracy score of 71.7%. SVM did perform worse in terms of speed, which should be noted.

Dataset Cleaning and Preprocessing



Data visualization after correcting imbalance.

Training - SVM



Conclusion

Conclusions

The implementation of the Ridge regression model aimed at predicting review scores using the Google Universal Sentence Encoder embeddings. This approach resulted in a coefficient of determination (R^2 score) of 0.6334.

The Bayesian Ridge model was also implemented, resulting in an R^2 score of 0.6341. The proximity of these scores suggests a similar level of performance between the Ridge and Bayesian Ridge models in predicting review ratings.

However, it was the SVM model that performed best, as it is a classifier instead of calculating a score that is subject to some subjective analysis (what is the difference between 1 and 1.3, for example. Correcting the data imbalance and performing a sentiment analysis versus a score prediction.

Outcome - Demo

The model produced in this project can be used to improve business by calculating sentiment from customer reviews without requiring customers to input a star rating.