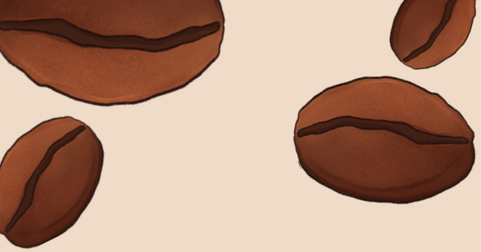


Project Setup

Using Reddit and Yelp Signals for Starbucks Popularity Prediction

Project Team Overview

The Moonbucks Team

Yuyang Chen
Applied Data Scientist
GitHub: yuyangchen11

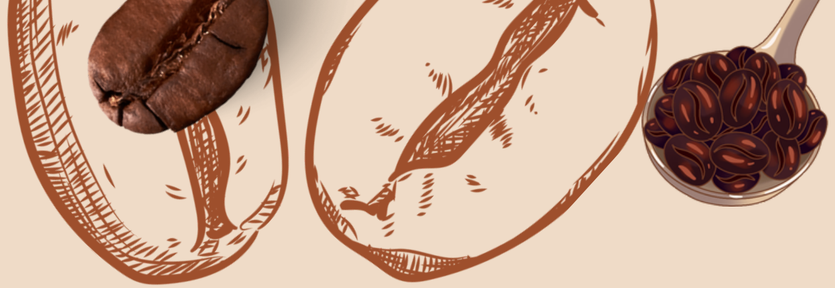
Ruihe Zhang
Product Manager
GitHub: Mudkipython

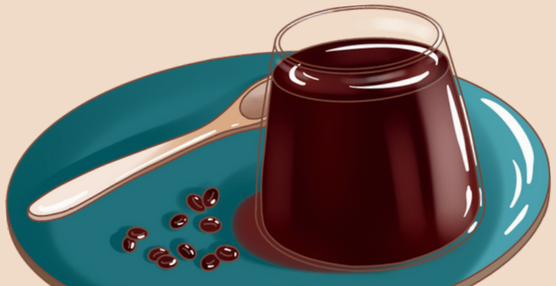
Zhaihan Gong
ML Engineer
GitHub: yuehuatong

Wendy Xu
Data Analyst
GitHub: ZihanXu12

Serena Sun
Analytics Data Scientist
GitHub: YS - 02

BUSINESS CONTEXT & PROBLEM





Business problem definition

- Starbucks operates in a highly competitive market where understanding consumer preferences is critical to product success.

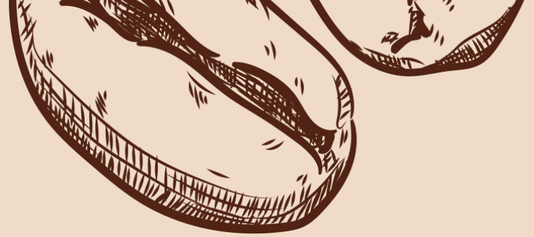
- Large volumes of consumer opinions are shared on platforms such as Reddit and Yelp, but this data is mostly unstructured and difficult to leverage.

- Existing product decision-making processes often fail to fully utilize online discussion and review signals.

- This project addresses the challenge of transforming unstructured text data into measurable indicators that support the prediction of product popularity.

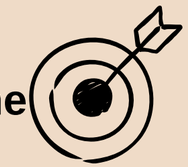
RESEARCH HYPOTHESIS & EVALUATION METRICS

Hypothesis



Reddit discussion signals provide actionable early insights that help prioritize Starbucks products for marketing attention, as reflected in subsequent Yelp ratings and review volume.

How We Validate the Hypothesis



Popularity Outcome

Product popularity is defined using Yelp-based signals only, such as:

- Average rating
- Review volume (e.g., Top X% threshold)

Reddit Discussion Signals

Early-stage consumer interest captured from Reddit, including:

- Mention frequency
- Discussion volume
- Temporal spikes in activity

Reddit Sentiment Signals

Aggregated sentiment polarity of Reddit discussions, representing early consumer perception of each product.

Evaluation & Validation

We evaluate whether Reddit signals support popularity prediction by:

- Comparing model performance with and without Reddit features
- Using classification metrics (e.g., Recall, ROC-AUC) to assess decision usefulness

Dataset & Approach

1. Starbucks Beverage & Nutrition Data

- Kaggle – https://www.kaggle.com/datasets/henryshan/starbucks
- Product-level information including drink names, categories, and nutritional attributes
- Structured tabular dataset
- Used to define and align Starbucks products across data sources

2. Reddit Discussion Data (Starbucks-related)

- Collected via Reddit API
- Posts and comments mentioning Starbucks products
- Includes timestamps and engagement signals (e.g., comments, upvotes)
- Unstructured text data reflecting consumer discussions

3. Yelp Review Data (Starbucks locations)

- Yelp Open Dataset
- Consumer reviews, star ratings, and review counts
- Textual reviews linked to Starbucks stores
- Used as a proxy for consumer sentiment and popularity

Target Variable

Product Popularity (Proxy-based)

- Defined using Yelp-based signals only, such as:
  - Review volume
  - Average rating (threshold-based)
- Represents relative popularity outcomes, not exact sales figures

Analytical Approach & Models

Sentiment Analysis

Aggregated sentiment scores from Reddit text

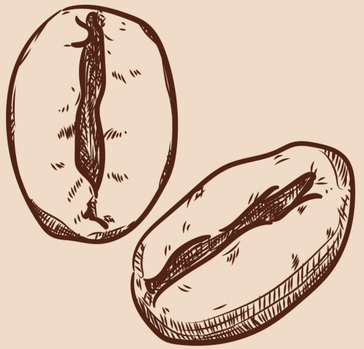
Logistic Regression

- Interpretable benchmark
- Establishes whether Reddit signals provide predictive value

Learning-to-Rank / Tree-based Models

- Capture non-linear relationships between discussion signals and popularity outcomes
- Used to improve ranking and classification performance

EXPECTED RESULTS & RISKS



01 Expected Outputs

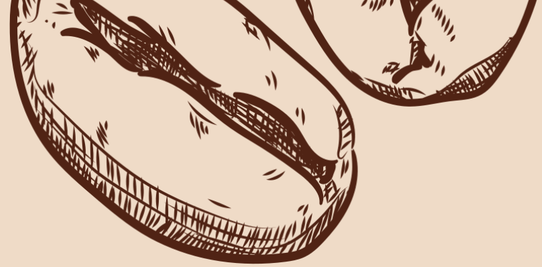
- Relative popularity scores or rankings of Starbucks products to support marketing prioritization decisions.
- Interpretable insights on how Reddit discussion volume and sentiment contribute to popularity predictions.

02 Threats to Validity

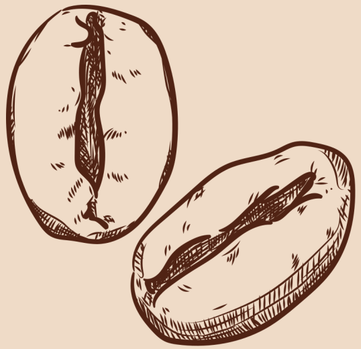
- Proxy limitation: Yelp ratings and review volume may not fully represent actual sales performance.
- Sampling bias: Reddit and Yelp users may not be representative of the broader Starbucks customer base.
- Text noise: Informal language and sarcasm in online discussions may affect sentiment extraction accuracy.

03 Validation Strategy

- Use train-test splits or cross-validation to evaluate predictive performance.
- Compare models with and without Reddit features to assess the incremental value of discussion signals.
- Benchmark results against simple baselines (e.g., popularity based on Yelp signals only).



GiTHUB REPO & BOARD



GitHub Repo & Board

GitHub Repository

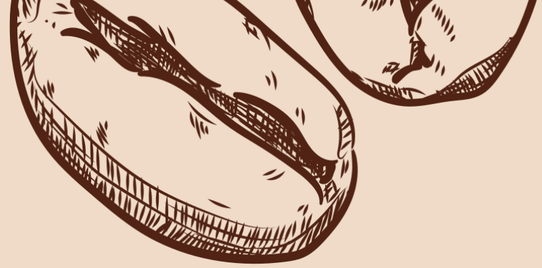
- Centralized repository
- Clear project structure
- README file describing project scope, data sources, and workflow

GitHub Project Board

- Task-based project board
  - track progress and collection
  - To Do, In Progress, and Done
- Issues linked to commits ensure transparency and accountability

Collaboration & Best Practices

- Repository designed to support reproducibility and future extensions
- Requests and code reviews applied for major changes



THANK YOU