Exploring Customer Feedback for Strategic Insights on Amazon Review Dataset

Mousumy Kundu, Aditi Das, Hayley Baek

DSE230: Final Project Presentation

May 30th,2025

Background & Problem Definition

- Amazon sells millions of products and generates even more millions of dollars in revenue every day
- Within a competitive marketplace, what are the types of products that are most likely to be successfully selling?
- How does a certain product stand in the marketplace compared to other similar products?
- Are we able to predict a product’s success by their review metrics?

Dataset Source & Description

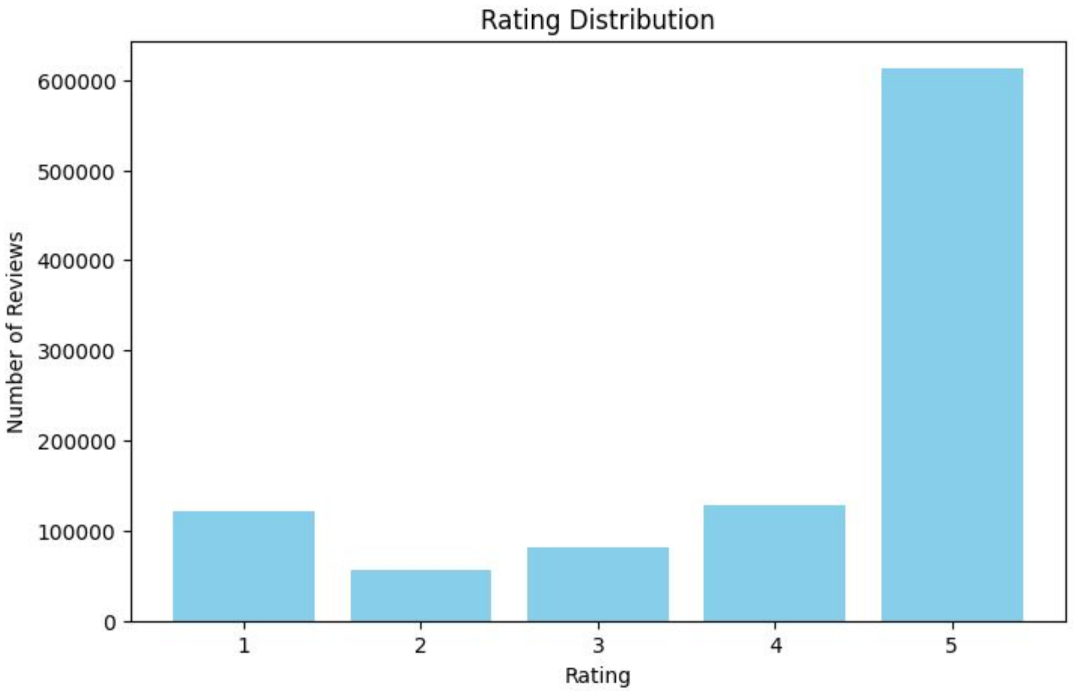
- **Data Source:** Amazon Product Data (All Beauty + Beauty and Personal Care category) from (https://amazon-reviews-2023.github.io/)
- **Description**:
  This dataset contains approximately 2 million customer reviews for beauty and personal care products. It includes the following columns

Items	Item Metadata		
rating	features	rating	images
title	description	title	videos
timestamp	price	main_category	bought_together
verified_purchase	subtitle	average_rating	details
helpful_vote	store	rating_number	
asin	categories	parents_asin	

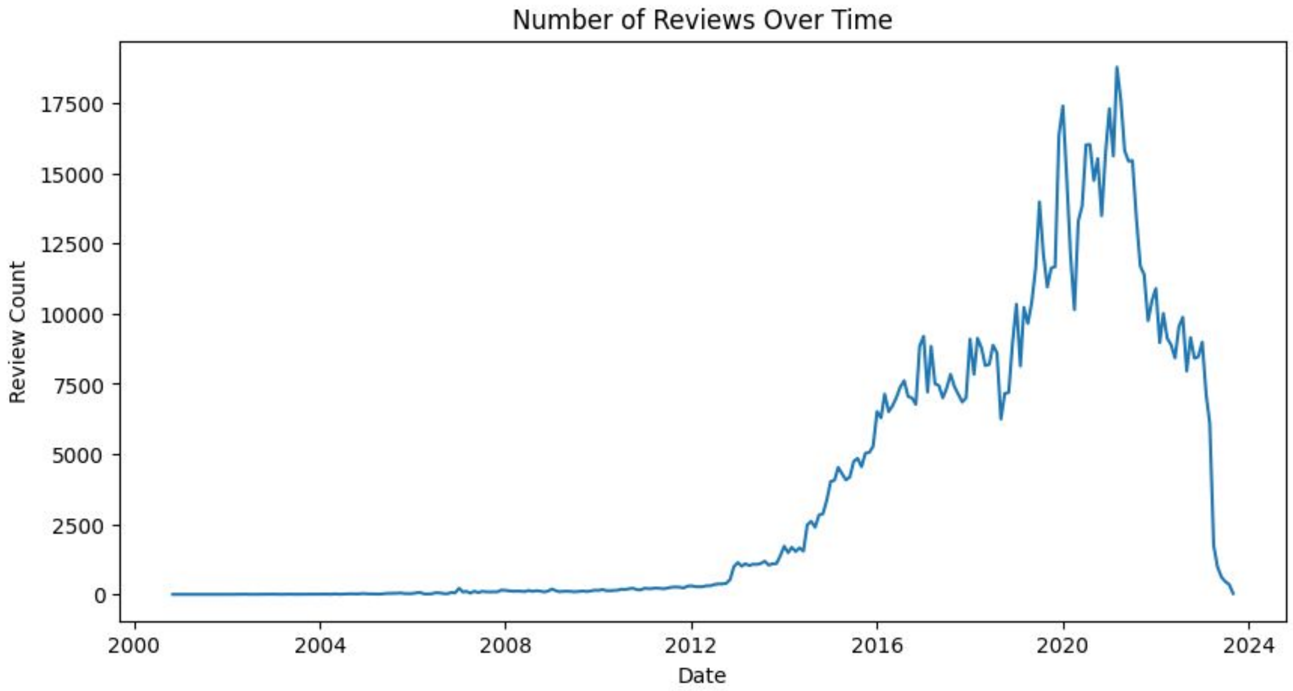
Data Cleaning and Preprocessing

- ➔ Type handling
  - ◆ Convert timestamp from long to datetime
  - ◆ Many columns had lists and dictionaries with more granular data
- ➔ Null value handling
  - ◆ Approximately 300k products had null values for price
- ➔ Some categorical features were evaluated numerically
  - ◆ Count of images and videos
  - ◆ Product score
  - ◆ Product rank
- ➔ Some columns disregarded altogether
  - ◆ User ID, bought_together

Exploratory Data Analysis

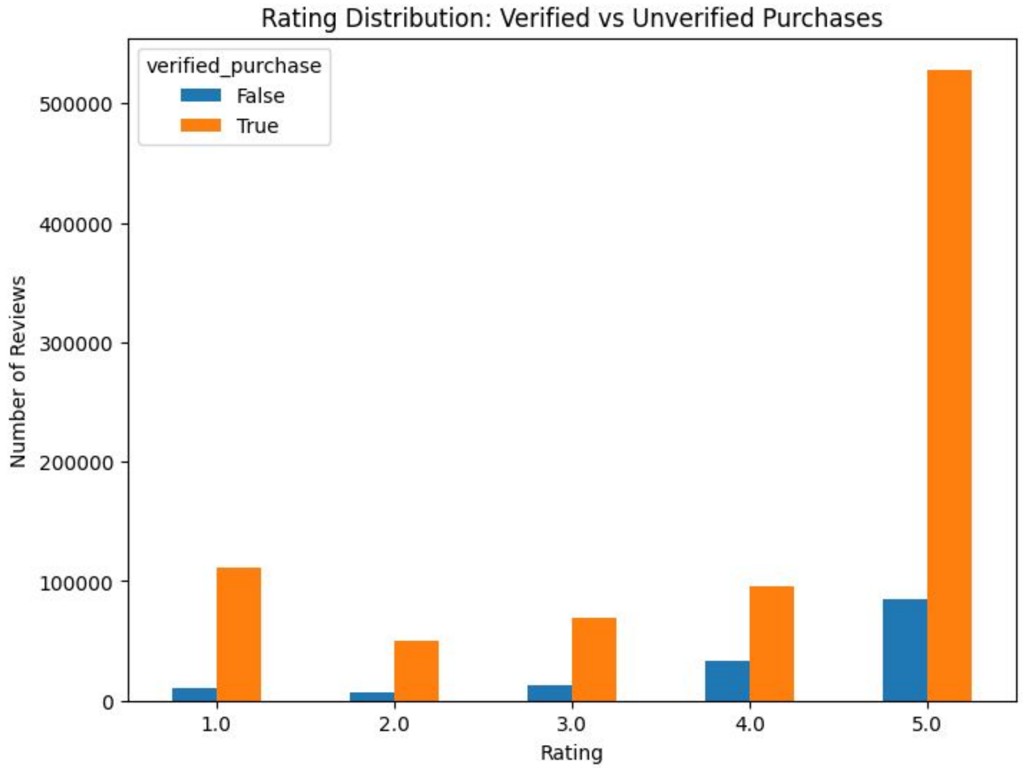


Most products are highly rated



Most reviews around COVID pandemic

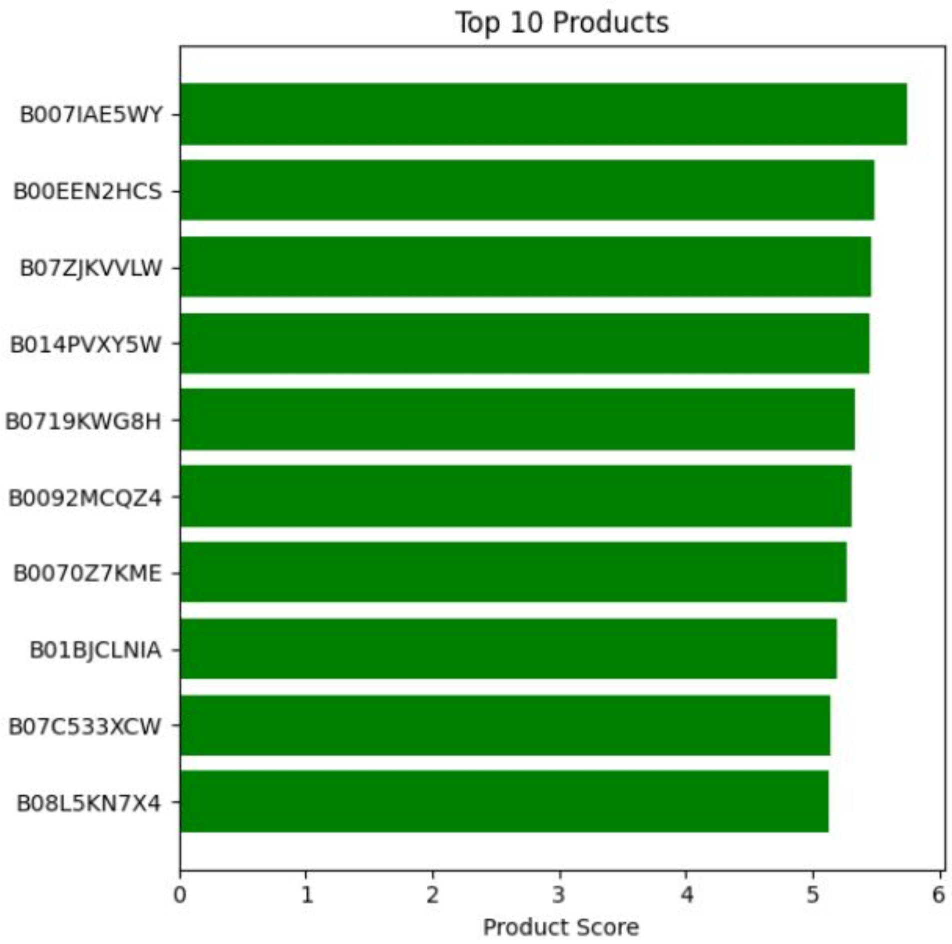
Exploratory Data Analysis

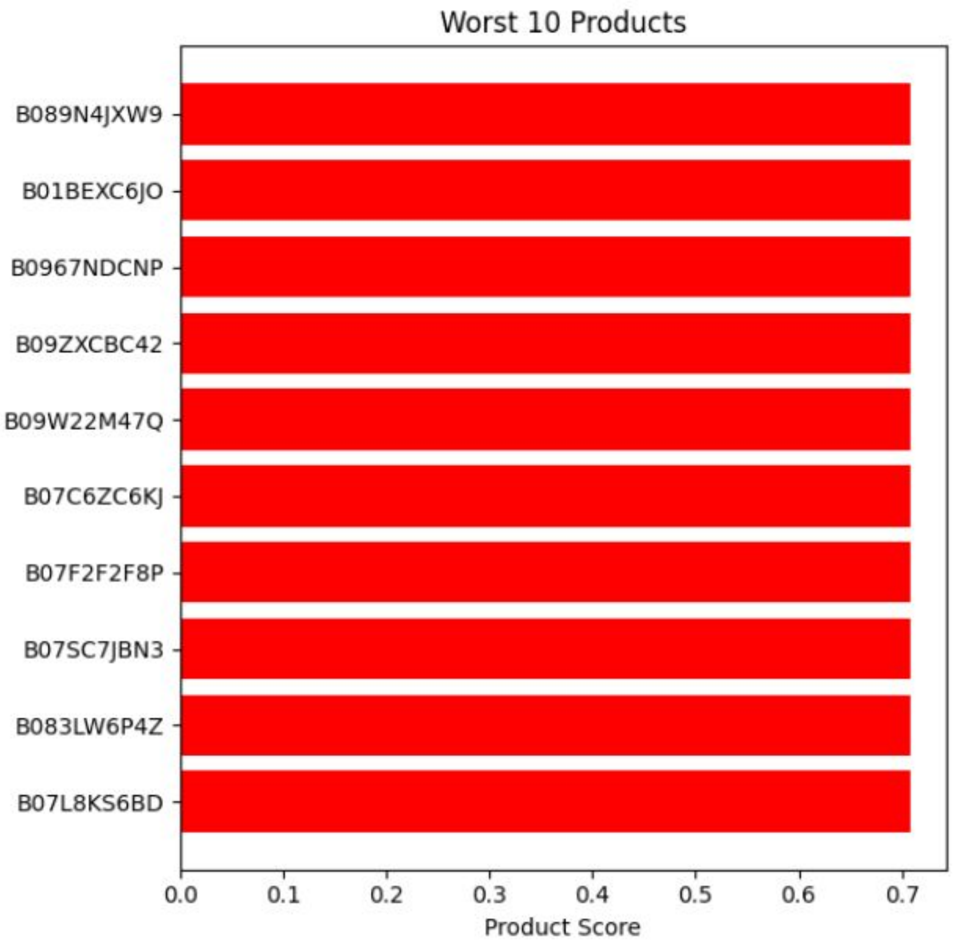


Longest review length from 4* ratings

Most ratings are verified

Best & Worst Selling Products



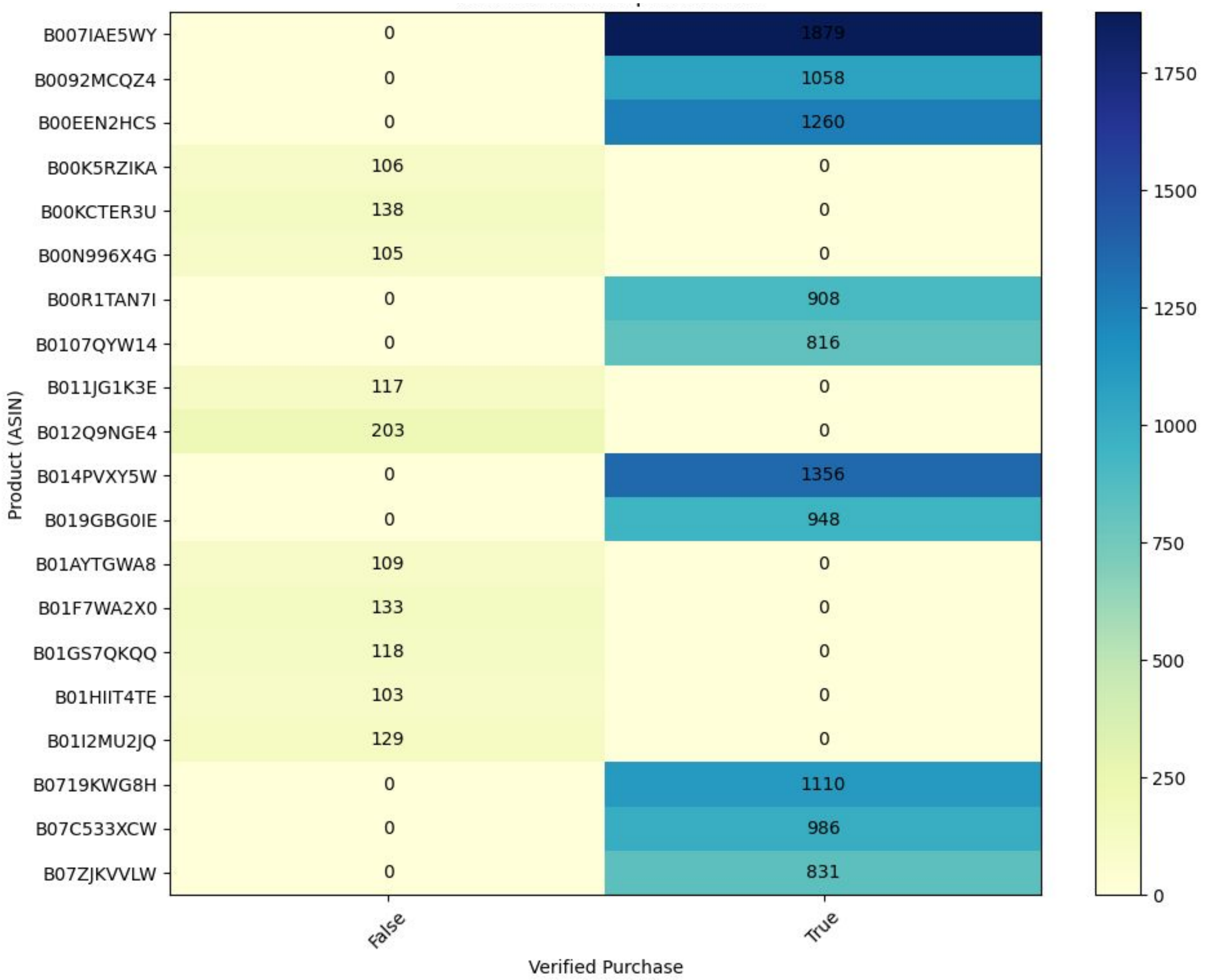


<u>Product Score:</u>

Calculated based on average rating, total number of reviews, total helpful votes and verified purchase.

Each were given a weightage of 50%, 30%, 15% & 5%

Verified Vs Unverified Purchase



- Top 10 products by number of verified and unverified reviews
- Product B007IAE5WY has the most verified reviews(1879) suggesting strong product engagement
- Product B00K5RZIKA has the most number(203) of unverified reviews suggesting manipulation
- The contrast between verified and unverified counts may help flag suspicious review patterns or assess product authenticity.

Tasks

Task	Main Insight	Business Value
Predicting Product Ratings	Drivers of ratings, early quality signals	Proactive quality control, smarter recommendations
Price vs Rating Analysis	How price affects perception	Strategic pricing
Product Popularity Analysis	More sought out products	Guide promotion and support

Product Rating Classification

Why Predict Ratings?

Why Predict Ratings

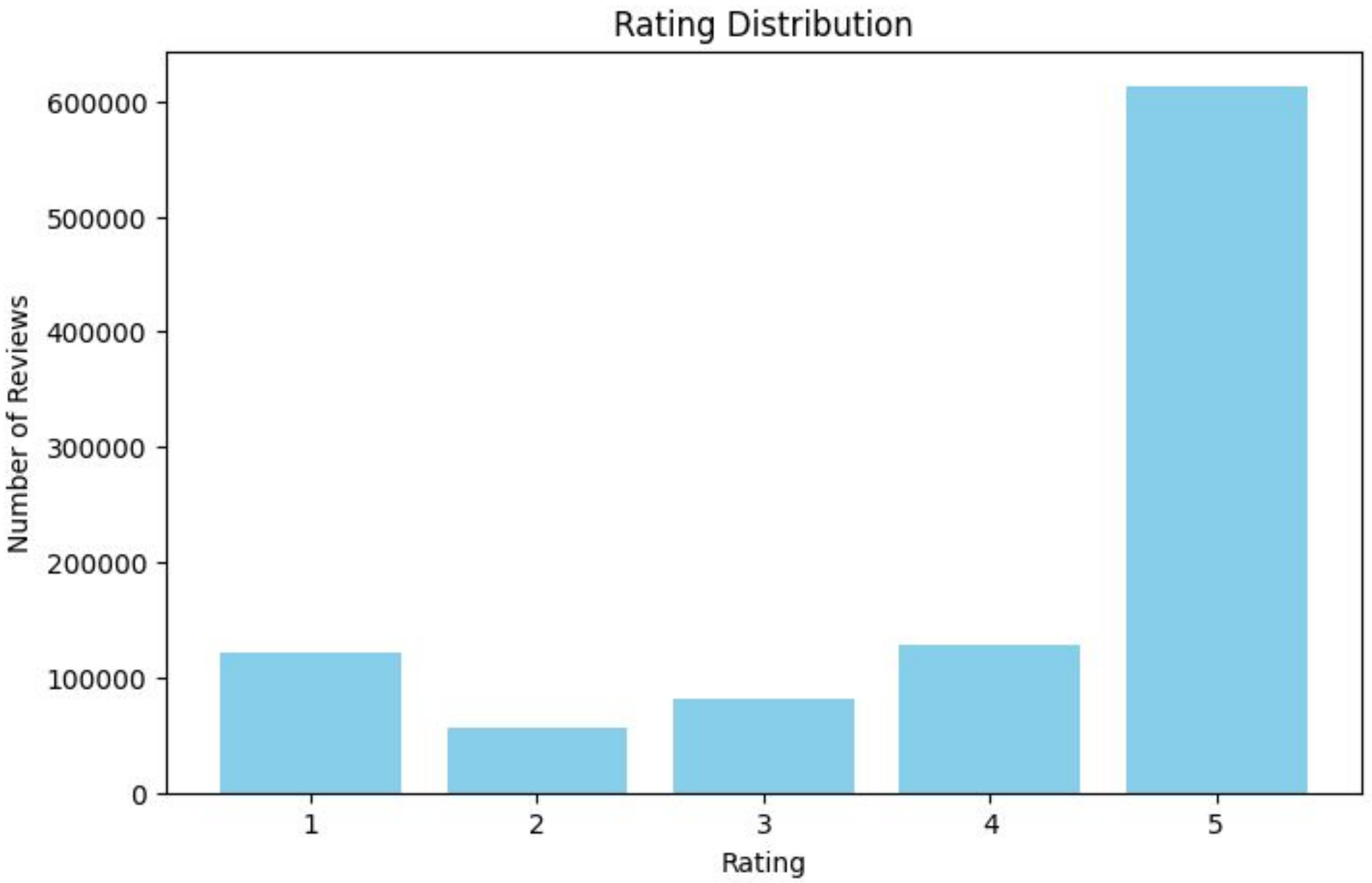
Understanding what kinds of reviews lead to lower or higher ratings helps companies:

- Identify problems with products
- Detect fake or low-quality review

Insights Gained

- Identify **key drivers of dissatisfaction or satisfaction** in customer feedback
- Discover patterns across verified purchases, helpful votes, and product engagement
- Spot **fake or misleading reviews** through abnormal rating behavior

Rating Distribution



Workflow for both Imbalance and Balanced Dataset

Data Cleaning and Preprocessing

- Handling missing data
- Dropping duplicates

Data preparation

- Convert rating to 3-class target: low, medium, high
- Feature engineering:
  - has_image, verified_purchase, helpful_vote, review_year

Text Preprocessing

1. **Tokenization**: Split title and review text into individual words.
2. **Stopword Removal**: Eliminate common, non-informative words.
3. **TF-IDF Vectorization**: Convert text into numerical features using term frequency and inverse document frequency.

Feature Assembly

- Combine numerical + text features into a single vector

Model Building and Evaluation

- Split the dataset into train and test
- Train models
- Predict on test data
- Evaluate the model performance

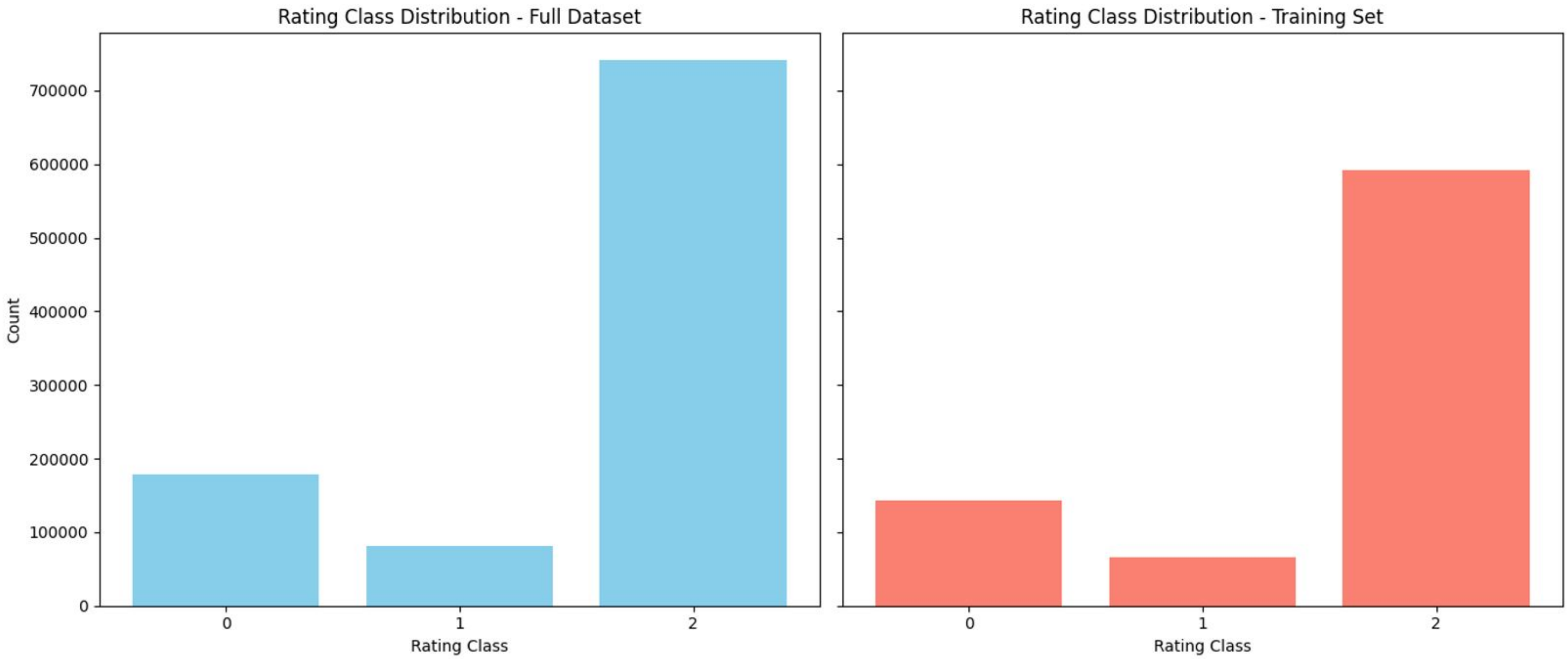
Interpretation and Insights

- Analyze the results to draw conclusions or actionable insights.
  Present findings with visualizations

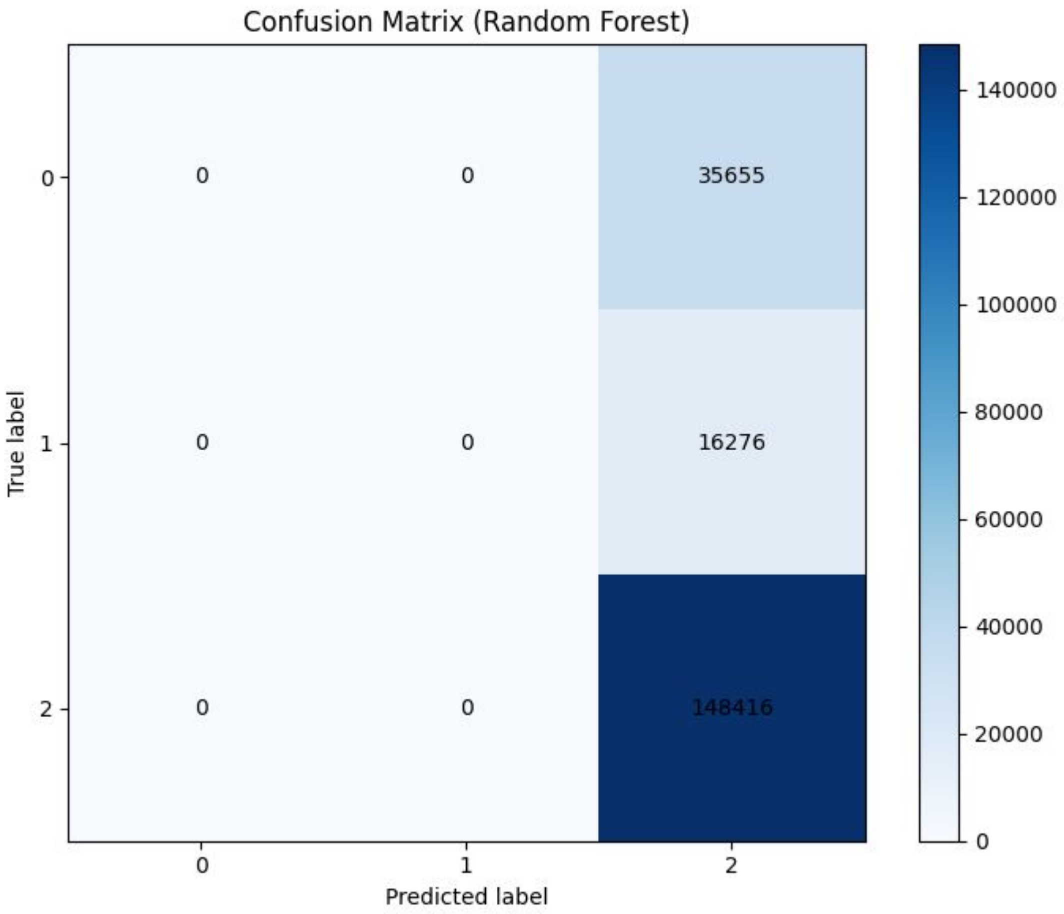
Class Mapping

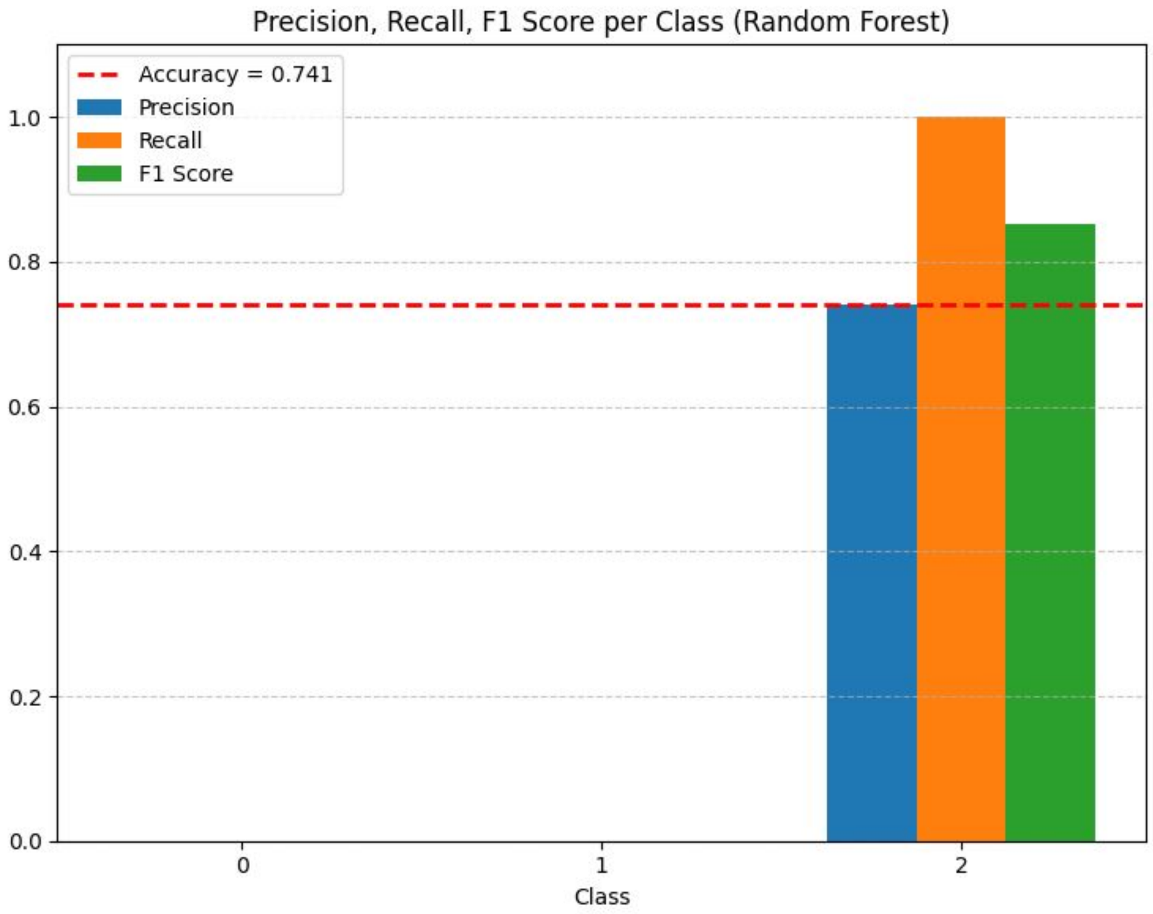
Rating Range	Category Label (rating_class)	Interpretation
4.0 and above	2	high
3.0 to <4.0	1	medium
Less than 3.0	0	low

Data Distribution of the Imbalanced Dataset

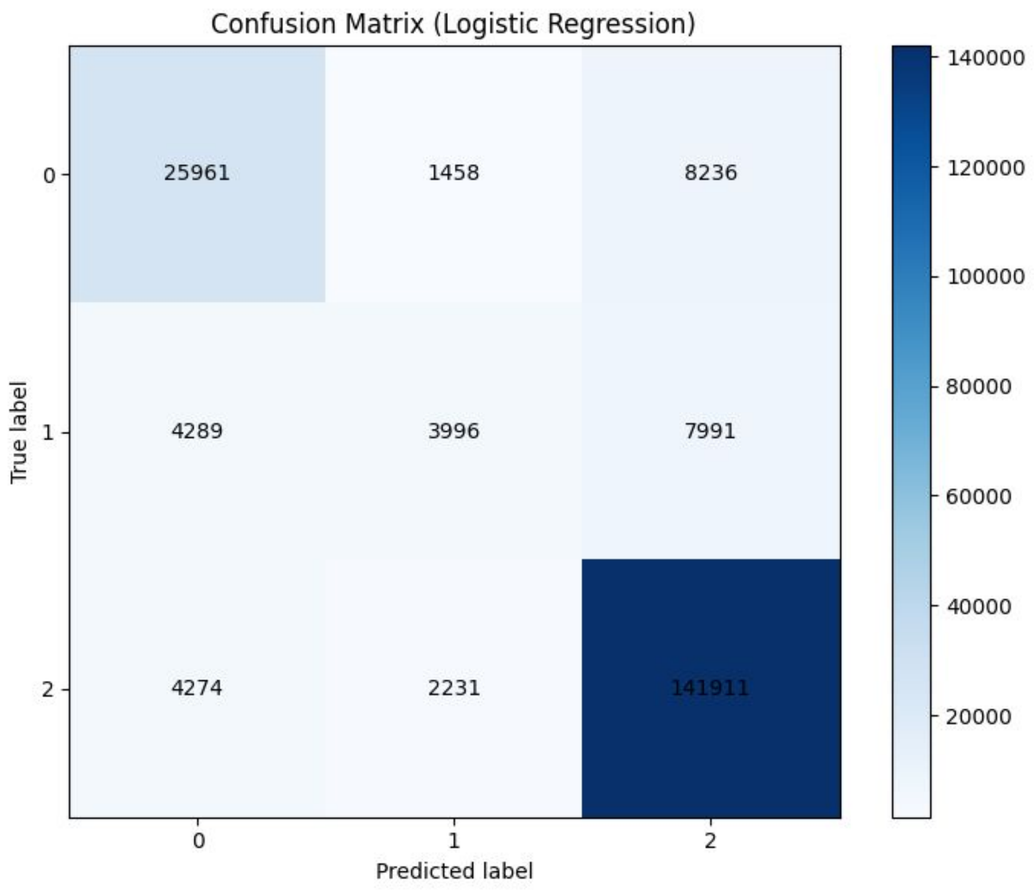


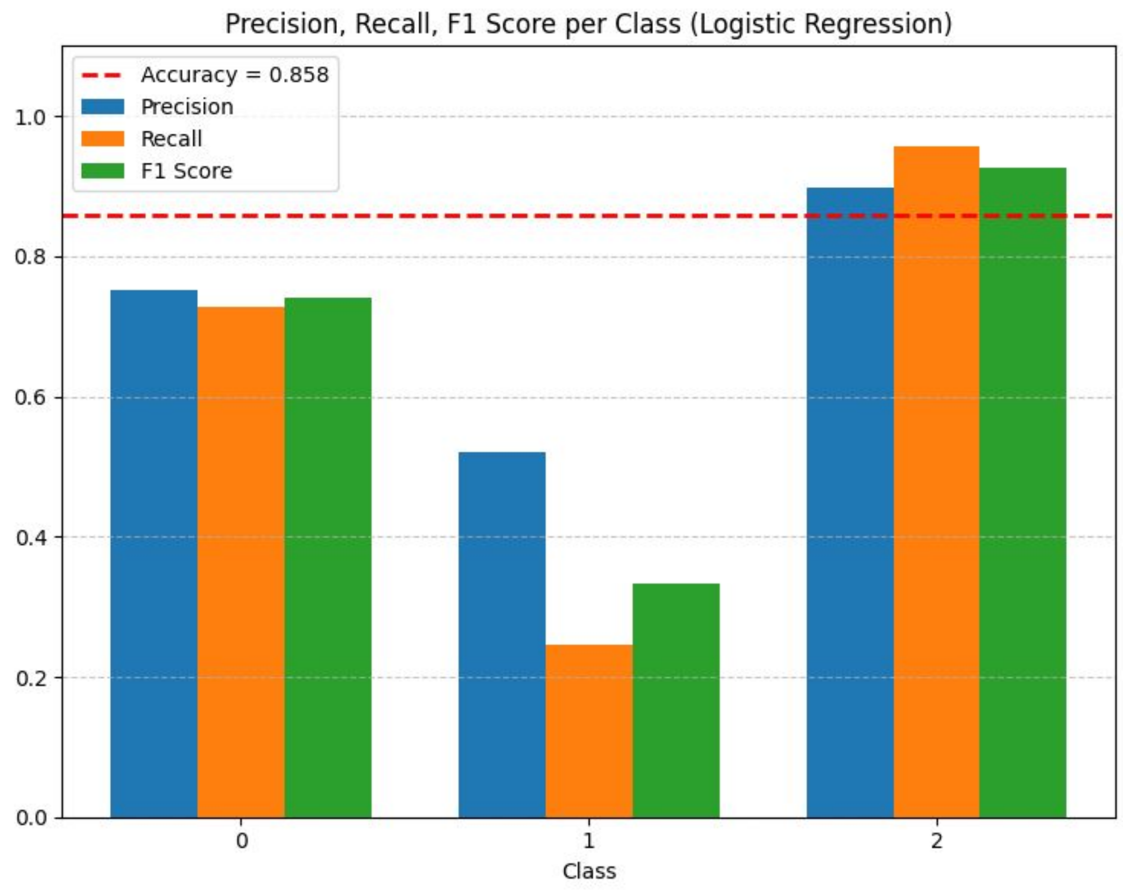
Applying Random Forest



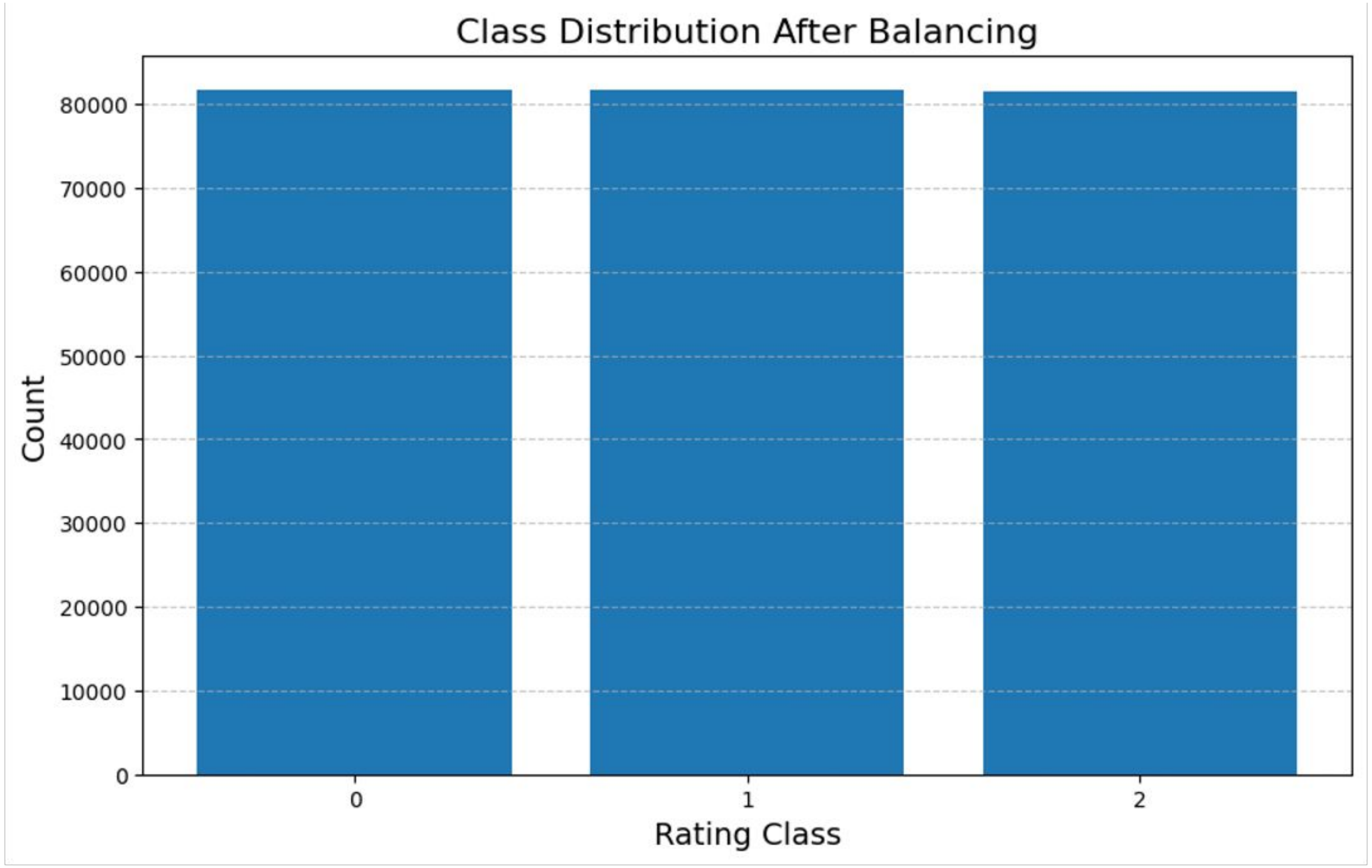


Applying Linear Regression

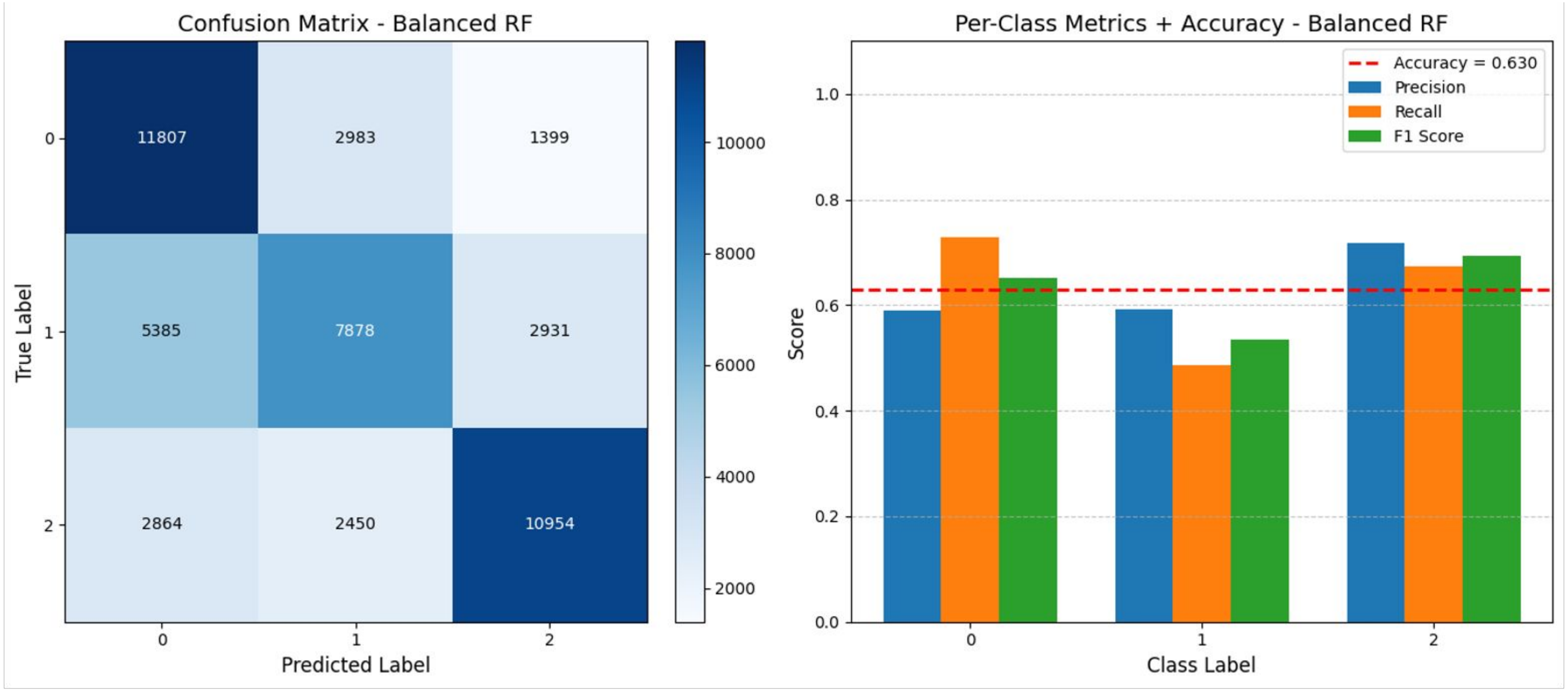




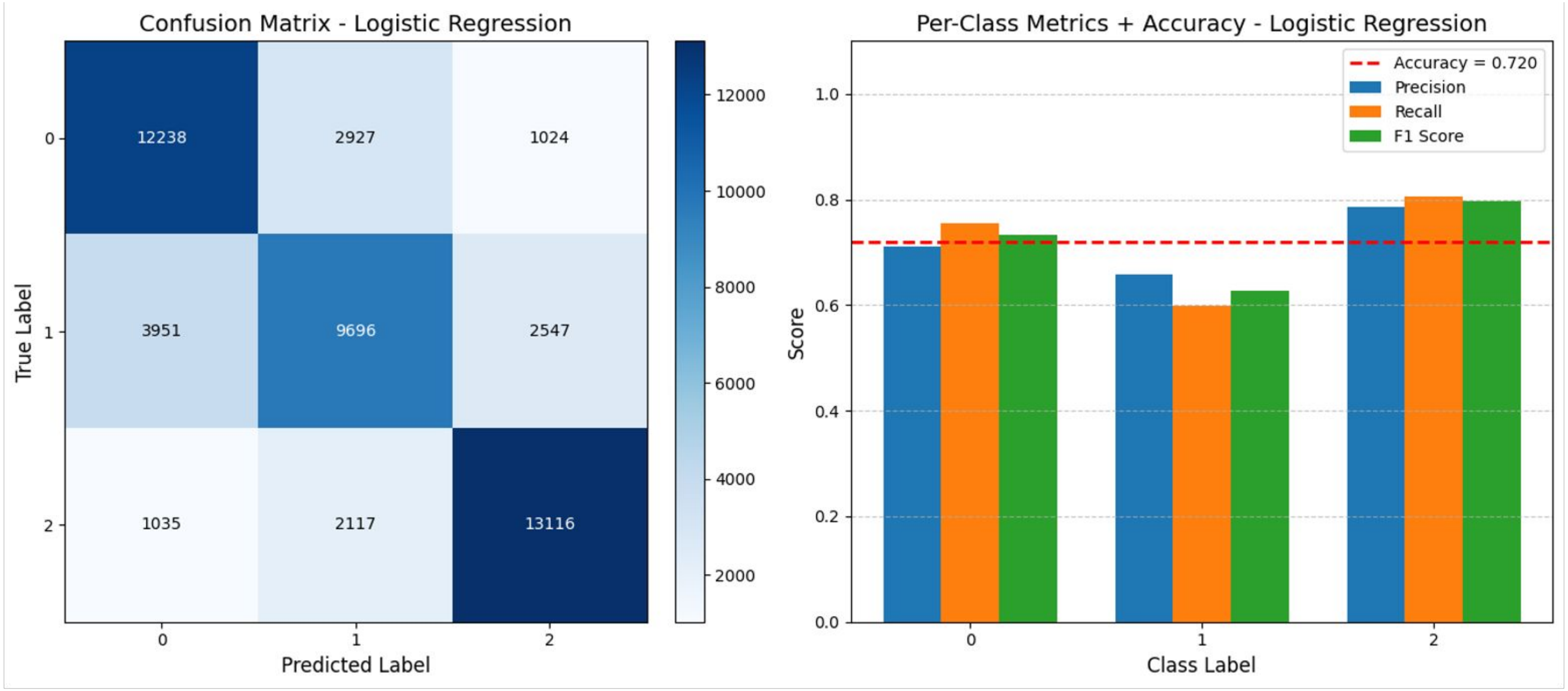
Data Distribution After Balancing



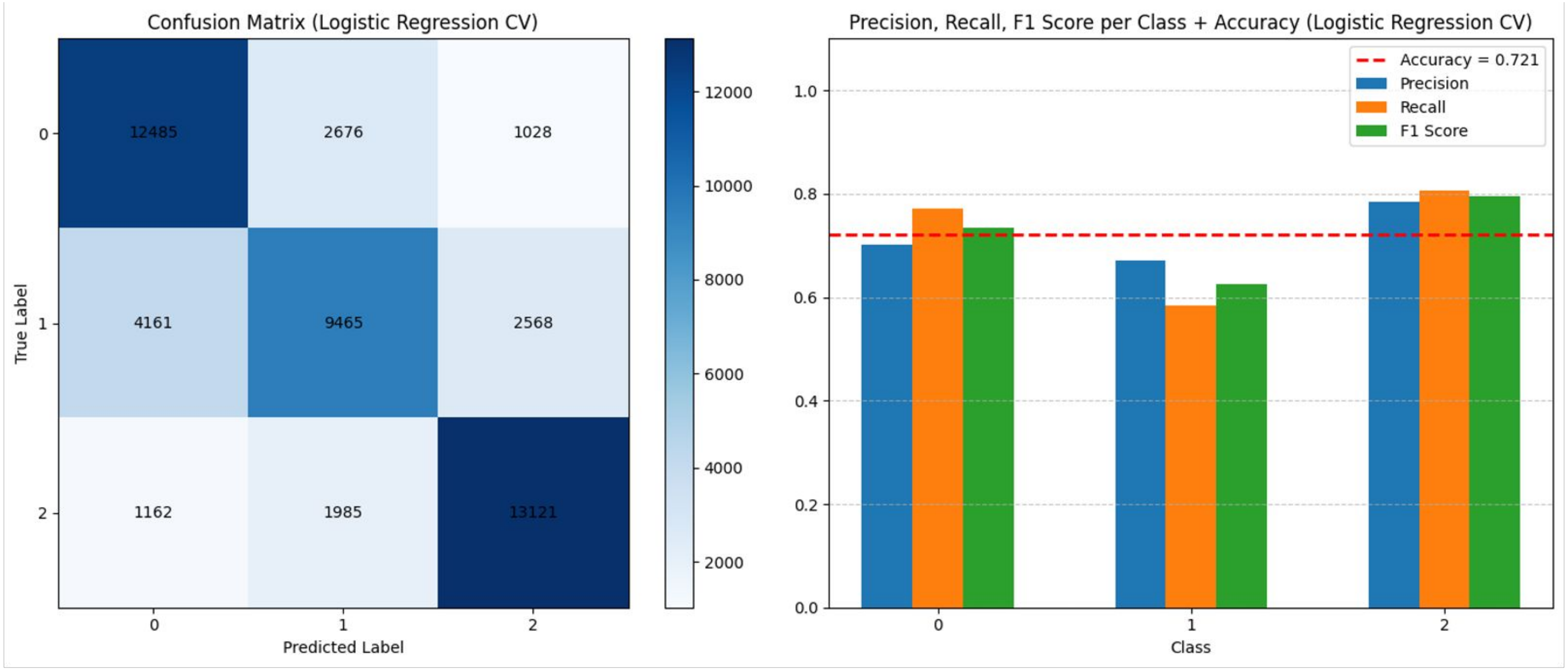
Applying Random Forest



Applying Linear Regression



Linear Regression with Cross Validation



Price vs Rating Analysis

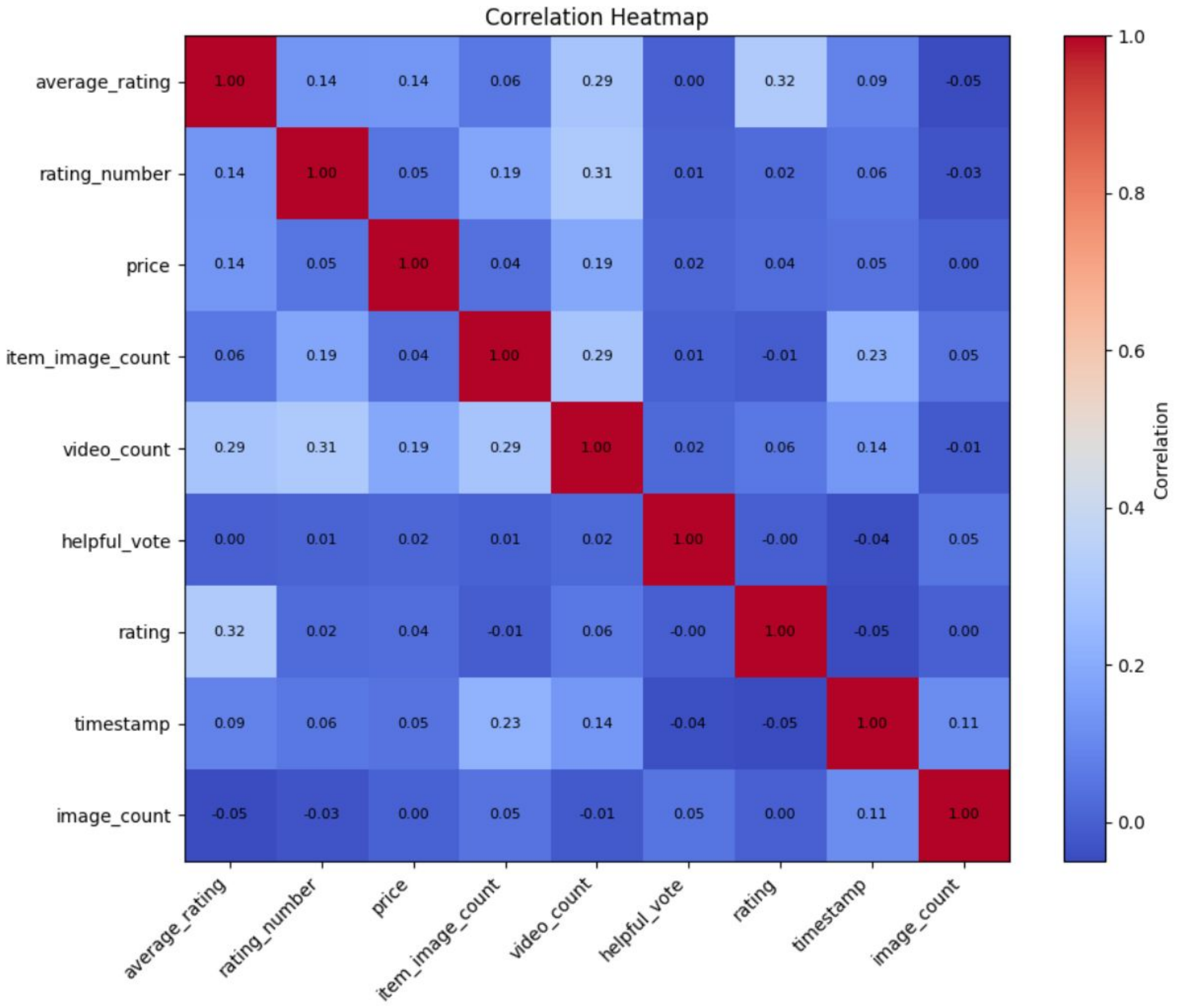
Price vs Rating Analysis

- How do price and ratings affect perception?
- Finding correlations
  - Price and average rating
  - Average rating and number of ratings
- Return on ratings (high ratings for low prices)
- Predictive model using linear regression

Correlations

Rating vs average rating
Image and time of day
Video count

Key takeaway:
Optimize page holistically
for user engagement



Return on ratings

Prices grouped into ranges

Highest average rating for \$20-\$50 products

Key takeaway:

Products too cheap or too expensive yield worse average ratings than mid-price range products

price_group	avg_rating	count
< \$5	4.3095	33789
\$5 - \$10	4.3566	230291
\$10 - \$20	4.3807	302294
\$20 - \$50	4.403	246989
> \$50	4.363	54762

Return on ratings (cont)

Very cheap or very expensive products have highly variant ratings

$20-$50 products least variant ratings

Key takeaway:

Premium products or cheaply produced products cannot expect consistent average ratings

price_group	stdev_rating
< $5	0.3601
$5 - $10	0.3447
$10 - $20	0.3321
$20 - $50	0.3171
> $50	0.3706

Regression Analysis

PySpark ML library for linear regression

Categorical values: Store name, category

Numeric values: Rating, price, helpful votes, assets counts (image, video)

Ultimately: Can we build a regression model that can predict ratings?

Regression Analysis

Prediction model works well

Mean squared error:
Average of squares of differences between estimated and true values

Mean absolute error:
Average magnitude of error between estimated and true values

R2:
Proportion of variance in dependent variable explained by independent variable

```
+--------------+------------+
|average_rating| prediction |
+--------------+------------+
|           4.1| 4.0999     |
|           4.1| 4.0999     |
|           4.2| 4.1999     |
|           4.2| 4.1999     |
|           4.2| 4.1999     |
|           3.9| 3.8999     |
|           4.2| 4.2000     |
|           4.2| 4.2000     |
|           4.3| 4.2999     |
|           4.4| 4.3999     |
+--------------+------------+
```

RMSE	MSE	MAE	R2
5.03e-7	2.54e-13	2.75e-7	0.999

Price Popularity Analysis

Goal

Identifying Popular Products Earlier :

- Using review data early we can predict how successful a product is likely to be based on the custom score value for each product

This is useful for:

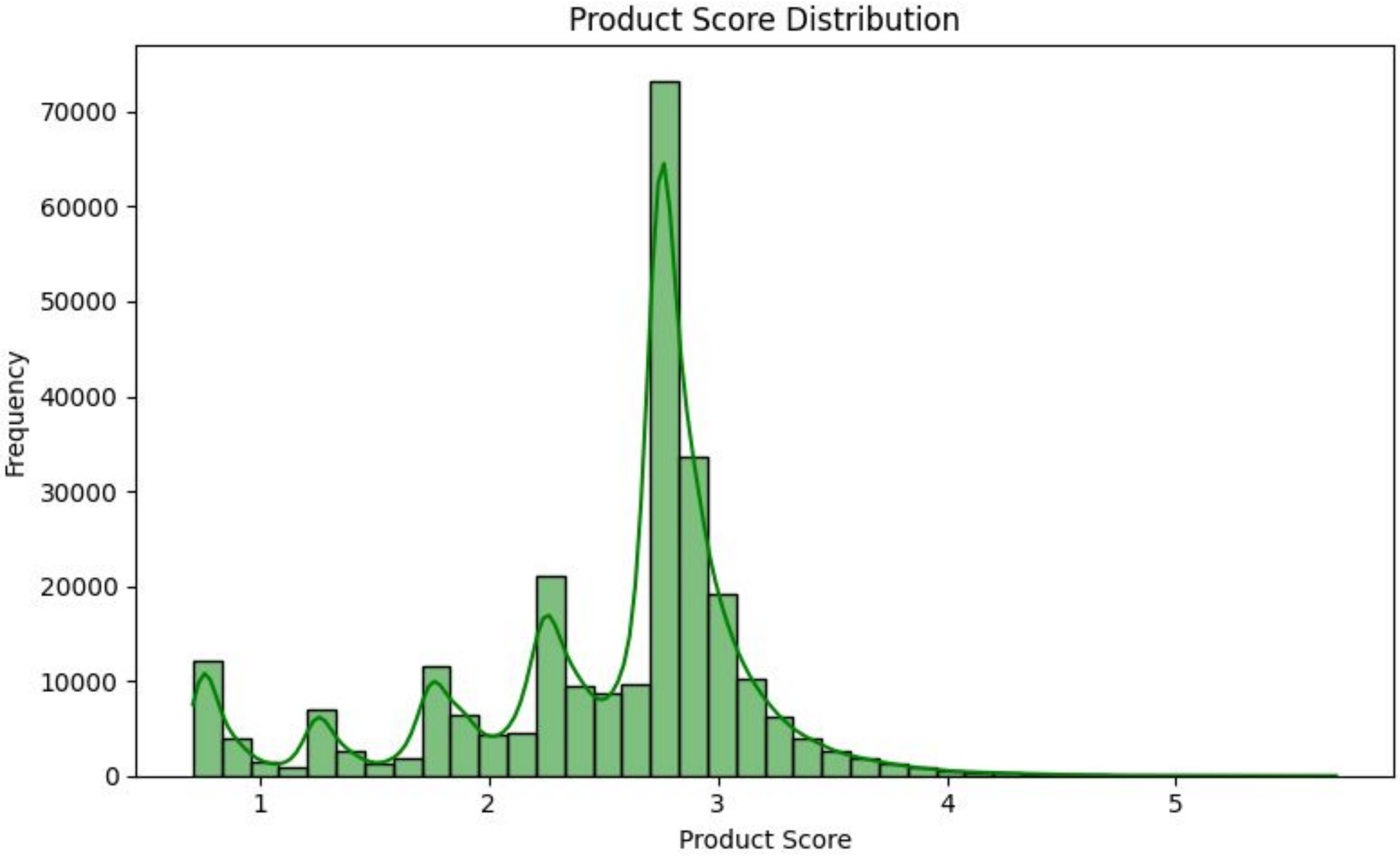
- Helps platforms like Amazon flag or promote products early
- Assists in inventory planning and recommendation systems
- Enables early alerts for poor-performing products

Product Score Analysis

Column Name	Weight
Average Rating	50%
Total Rating Count	30%
Total votes	15%
Verified Product	5%

Custom metric score calculated for product performance based on weightage of other numeric columns

Distribution of Product Score



Product Score is divided into 3 classes: 0(low), 1(medium) & 2(high).

Majority of the product score lies between 2.5 to 3 (medium class) which suggested high data imbalance.

Data Processing & Pipeline

Data collected
(>1M rows)

Cleaned data

- dropped unnecessary columns
- casted datatype into appropriate string & int type
- dropped rows with 0 pricing

Balanced dataset using product class & down-sampled data

Created encoding for the string column

Created the features list and target label

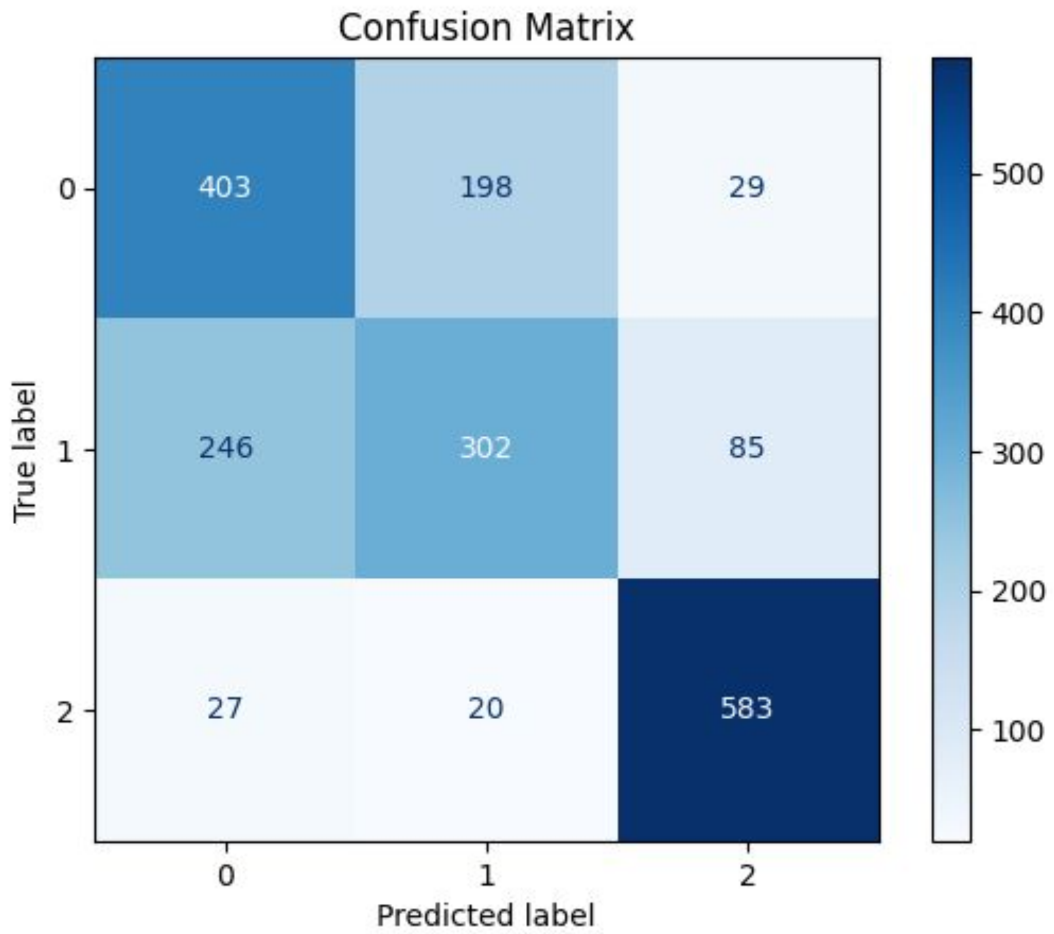
Splitted the data into training and test data by 80/20

Applied RandomForestClassifier Model and predicted the targets

Feature Engineering

- verified_purchases - field to determine whether purchase was verified
- price - price of the product
- title_length - length of the review title
- text_length - length of the review text
- review_year - year of the review obtained from timestamp column
- review_month - month of the review obtained from timestamp column
- days_since_review - number of days passed since review, obtained from timestamp column
- feature_length - length of the features column
- description_length - length of the description column
- main_category_encoded - encoded values of main_category field

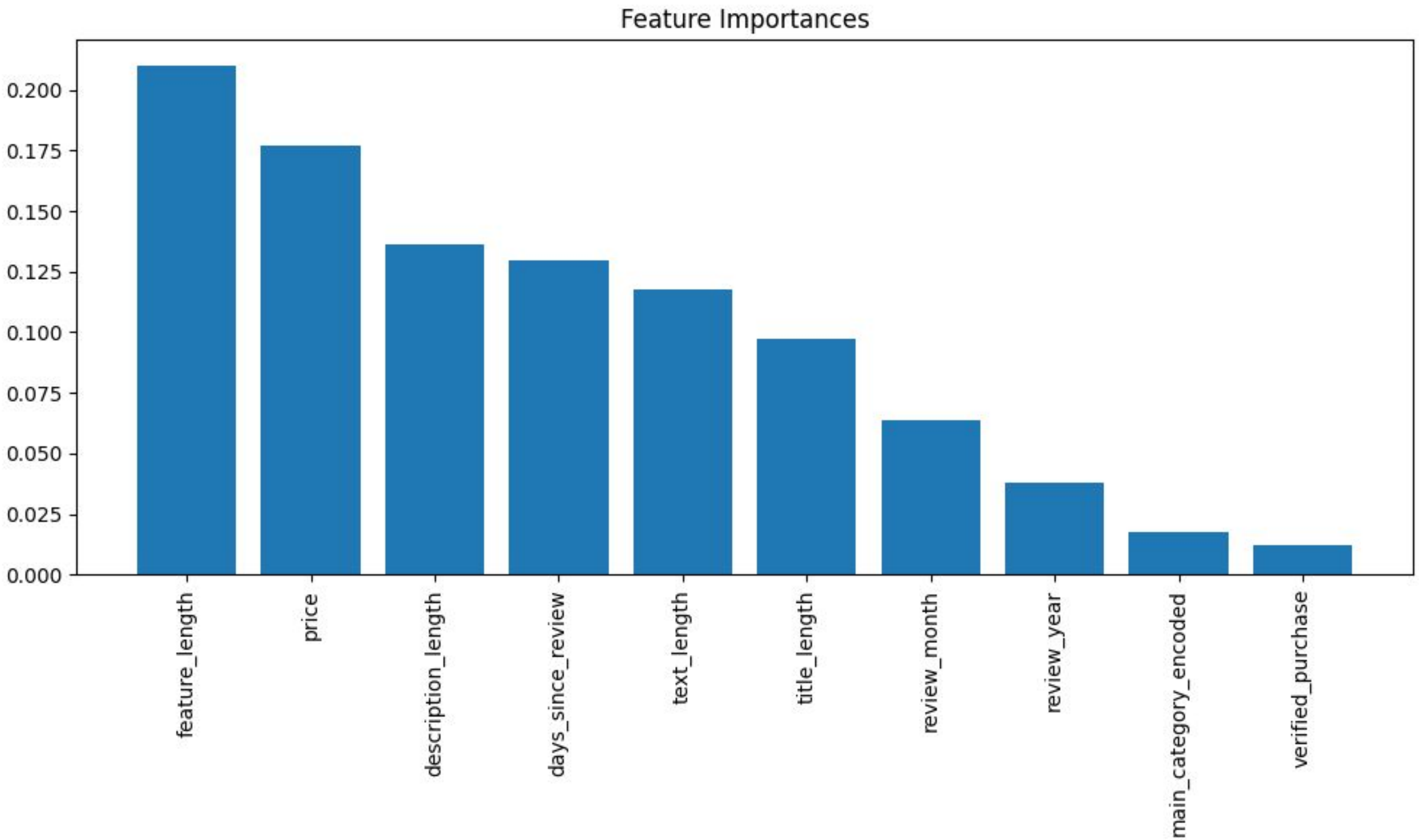
Model Performance



Model Used: Random Forest Classifier
3-class classification: *Low*, *Medium*, *High* popularity class

- The model performs well on Class 2 (High) with F1-score of 0.88 and Recall of 93%.
- Class 1 (Medium) shows the weakest performance, with only 48% recall, indicating many are misclassified.
- Class 0 (Low) has moderate performance (F1-score: 0.62), suggesting room for improvement.
- Overall accuracy is 68%, with balanced but variable class-wise performance. Further tuning and feature enhancement could help.

Feature Importance



Product presentation matters most: feature_length, price, and description_length have strong influence on outcomes

Recent reviews and review length also play significant roles, showing that both older and longer feedback affect predictions.

Categorical fields like main_category have minimal influence, suggesting they contribute little compared to descriptive or behavioral features.

Insights

- Predicting Product Ratings
  - Factors like review length, helpful votes, and verified purchase affect product ratings.
  - Balancing the dataset improved the model's ability to learn from all rating classes more effectively. While overall accuracy did not significantly increase, the model became better at predicting underrepresented classes, leading to a more balanced and fair classification.
- Price Analysis and Rating
  - Too cheap or too expensive products generally yield lower average ratings
  - High number of ratings appear more trustworthy
- Product Popularity Analysis
  - Product with higher ratings and older reviews or longer review texts are more influential for a product popularity
  - Longer feature descriptions and higher prices are strong indicators of product performance.

Challenges and Future Work

Challenges:

- Large dataset size led to slow processing and frequent crashes on the local machine.
- Local environment struggled with memory and performance due to data scale.
- High imbalance created a bias in the model.
- Some fields were missing or inconsistent.
- Unstructured Text: Reviews contain typos, slang, and inconsistent phrasing.
- Basic Feature Representation: TF-IDF doesn't capture context or sentiment depth.

Future Work:

- **Advanced NLP Models:** Explore BERT/RoBERTa to better understand review context.
- **Sentiment Integration:** Add sentiment scores to improve classification quality

Thank you!