THE UNIVERSITY OF CHICAGO

Big Data Analysis of Amazon Review

Big Data Platform | Autumn 2023

Team Members
Nicole Young, Xiaobing Xu
Yixuan Chen, Boya Zeng

AGENDA

01 Project Overview

02 Data Process&EDA

03 Machine Learning Models

04 Conclusion

01

PROJECT OVERVIEW

Executive Summary

Business Problem:In the rapidly evolving e-commerce landscape, Amazon, a global leader in online retail, faces intense competition and ever-increasing customer expectations. The ability to understand customer preferences, predict their needs, and tailor the shopping experience is paramount. Amazon's vast repository of customer reviews and interaction data is underutilized, offering untapped potential to drive sales and improve customer satisfaction.

Goal: Leverage data analytics and machine learning to boost customer satisfaction, streamline marketing efforts, and drive sales growth.

Sentiment Analysis: Spark NLP pretrained model (`sentimentdl_use_twitter`) was employed to classify reviews into positive, negative, or neutral categories.

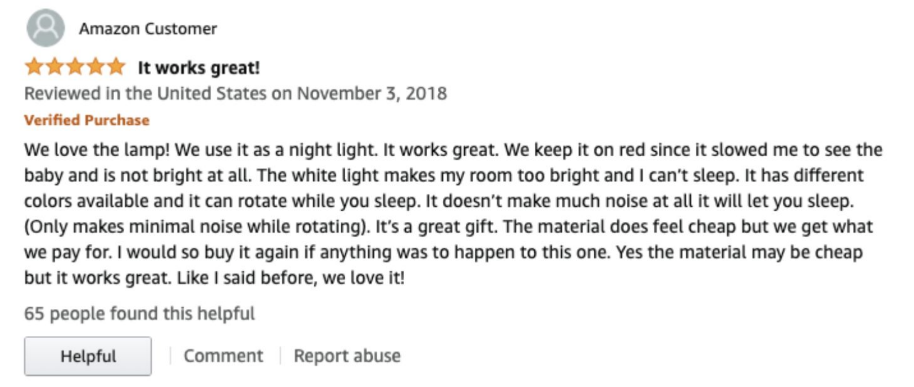
Product Recommendation System:Implemented using the Alternating Least Squares (ALS) algorithm in Apache Spark for collaborative filtering to develop a system that suggests products to users based on their preferences and behavior.

Customer Segmentation: Applied clustering technique - K-Means to segment customers into distinct groups based on purchasing behavior and preferences.

Business Objectives

- Sentiment Analysis:
  - Objective: Analyze customer reviews to gauge sentiment towards products.
  - Business Impact: Helps in identifying products that are well-received or those that require improvements. Guides product development and customer support strategies.
- Product Recommendation System:
  - Objective: Develop a model to recommend products to customers based on their past reviews and ratings.
  - Business Impact: Enhances customer experience through personalized recommendations, increasing sales and customer loyalty.
- Customer Segmentation:
  - Objective: Segment customers based on their purchasing patterns and review behaviors.
  - Business Impact: Enables targeted marketing campaigns, improves customer engagement, and identifies key customer segments.

Data Source



★★★★★ **It works great!**

Reviewed in the United States on November 3, 2018

Verified Purchase

We love the lamp! We use it as a night light. It works great. We keep it on red since it slowed me to see the baby and is not bright at all. The white light makes my room too bright and I can't sleep. It has different colors available and it can rotate while you sleep. It doesn't make much noise at all it will let you sleep. (Only makes minimal noise while rotating). It's a great gift. The material does feel cheap but we get what we pay for. I would so buy it again if anything was to happen to this one. Yes the material may be cheap but it works great. Like I said before, we love it!

65 people found this helpful

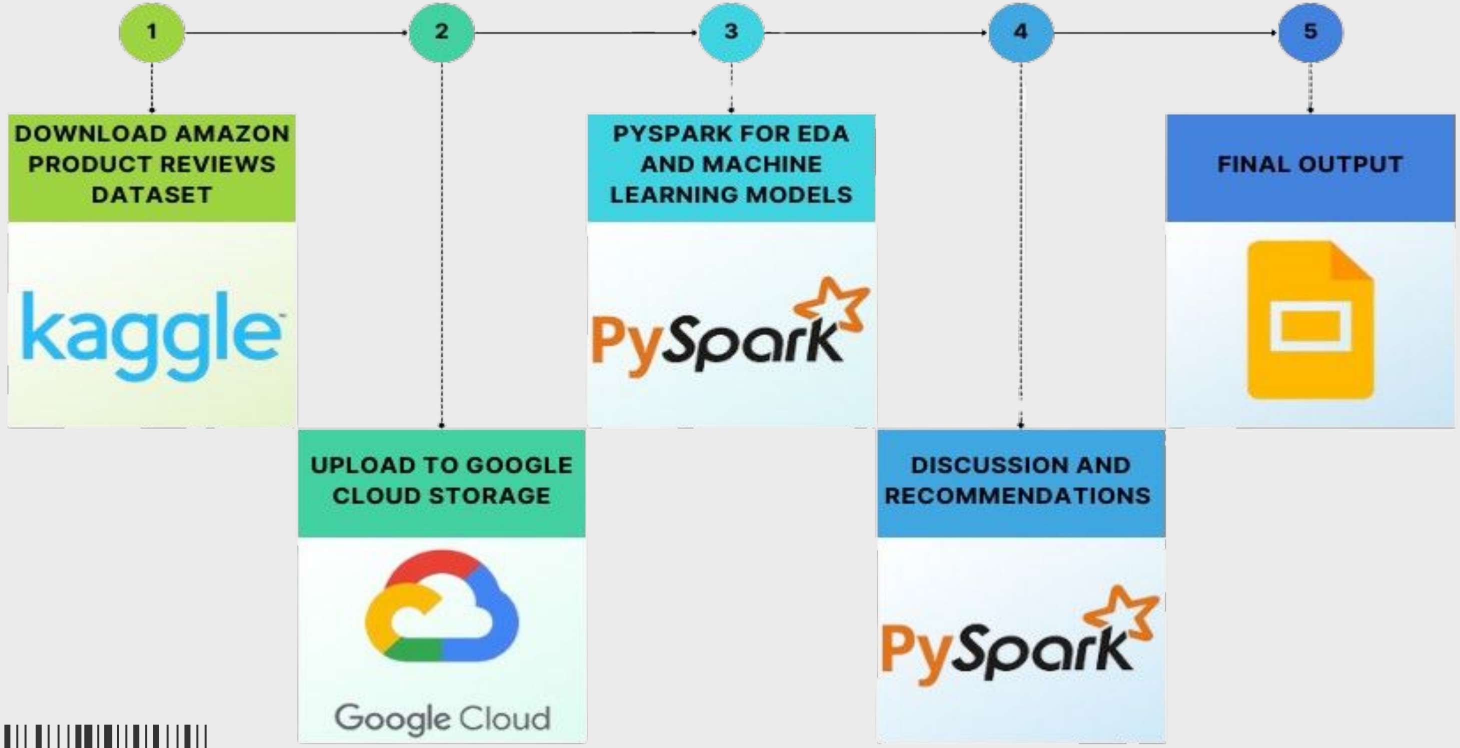
Amazon Customer Reviews Dataset.

Fields in Dataset:

- `marketplace`: Identifies the country of the Amazon marketplace.
- `customer_id`: A unique identifier for the author of the review.
- `review_id`: Unique ID of each review.
- `product_id` and `product_parent`: Identifiers for the reviewed products.
- `product_title` and `product_category`: Provide context about the product.
- `star_rating`, `helpful_votes`, `total_votes`: Metrics for review helpfulness and rating.
- `vine` and `verified_purchase`: Indicators of review's source and authenticity.
- `review_headline`, `review_body`, `review_date`: Text and metadata of the review.

https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-d~~ataset/data~~?select=amazon_reviews_us_Apparel_v1_00.tsv

Project Steps



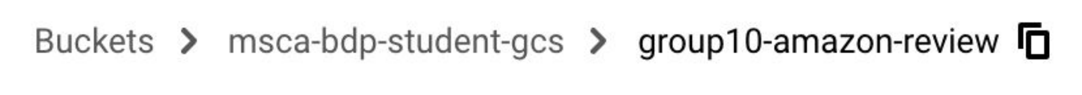
1. DOWNLOAD AMAZON PRODUCT REVIEWS DATASET
2. UPLOAD TO GOOGLE CLOUD STORAGE
3. PYSPARK FOR EDA AND MACHINE LEARNING MODELS
4. DISCUSSION AND RECOMMENDATIONS
5. FINAL OUTPUT

Big Data Platform | Fall 2023

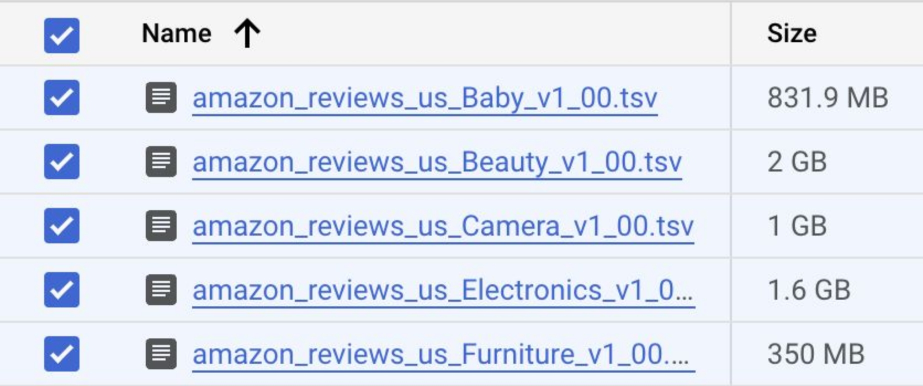
02

Data Process & EDA

Raw Data



Buckets > msca-bdp-student-gcs > group10-amazon-review



☑	Name ↑	Size
☑	amazon_reviews_us_Baby_v1_00.tsv	831.9 MB
☑	amazon_reviews_us_Beauty_v1_00.tsv	2 GB
☑	amazon_reviews_us_Camera_v1_00.tsv	1 GB
☑	amazon_reviews_us_Electronics_v1_0...	1.6 GB
☑	amazon_reviews_us_Furniture_v1_00....	350 MB

Step1 — **Use Kaggle API download zip file to GCP**

Step2 — **Unzip the file in Cloud shell**

Step3 — **Upload unzipped file back to GCS**

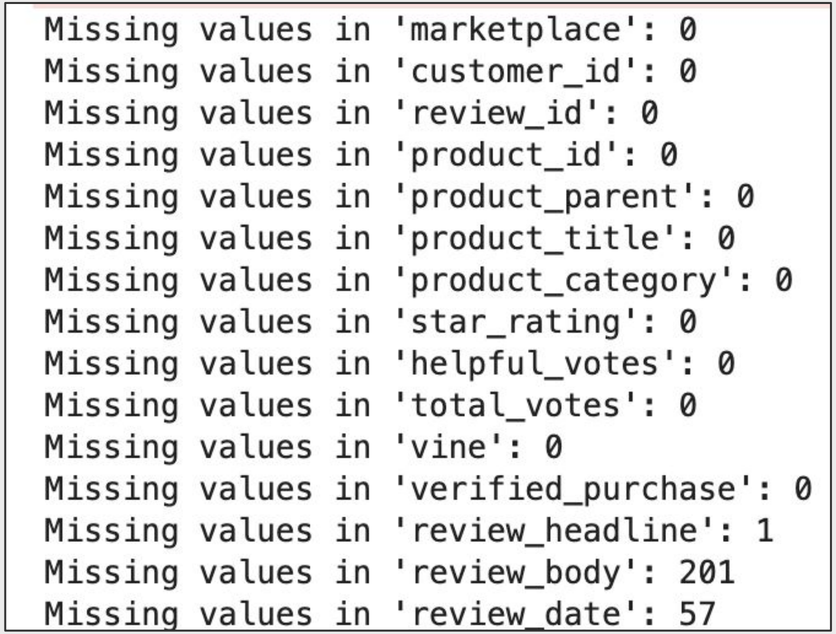
Data Cleaning

- Remove any unwanted characters
  - #, %, ^, //, etc.

- Convert to all lowercase text for consistency

- Check for duplicates based on a column

- Remove missing values

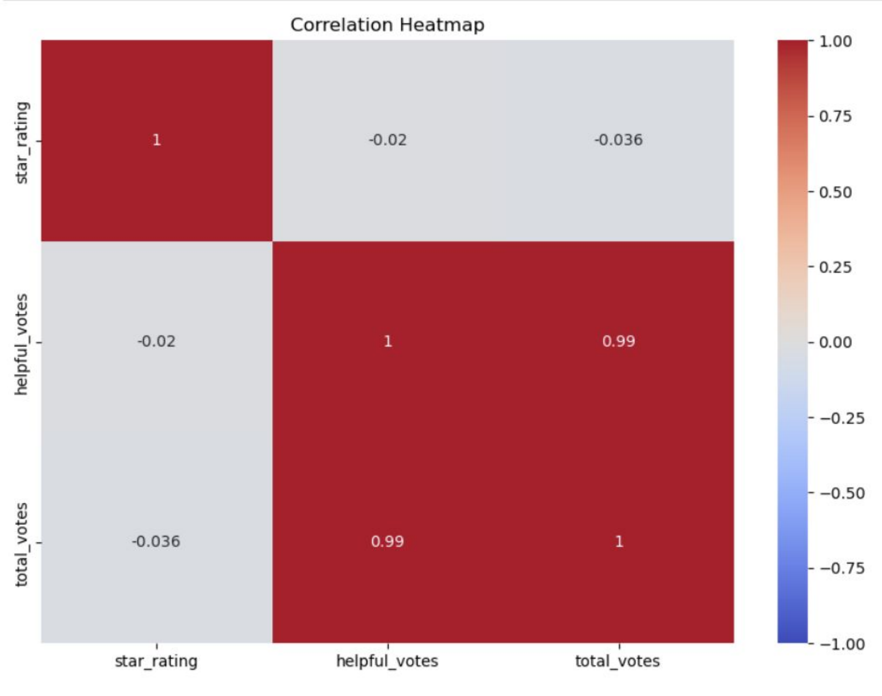


```
Missing values in 'marketplace': 0
Missing values in 'customer_id': 0
Missing values in 'review_id': 0
Missing values in 'product_id': 0
Missing values in 'product_parent': 0
Missing values in 'product_title': 0
Missing values in 'product_category': 0
Missing values in 'star_rating': 0
Missing values in 'helpful_votes': 0
Missing values in 'total_votes': 0
Missing values in 'vine': 0
Missing values in 'verified_purchase': 0
Missing values in 'review_headline': 1
Missing values in 'review_body': 201
Missing values in 'review_date': 57
```

Summary Statistics

	star_rating	helpful_votes	total_votes
count	619630	619630	619630
mean	4.13	1.97	2.47
std	1.33	21.39	22.61
min	1	0	0
max	5	8937	9072

Correlation heatmap

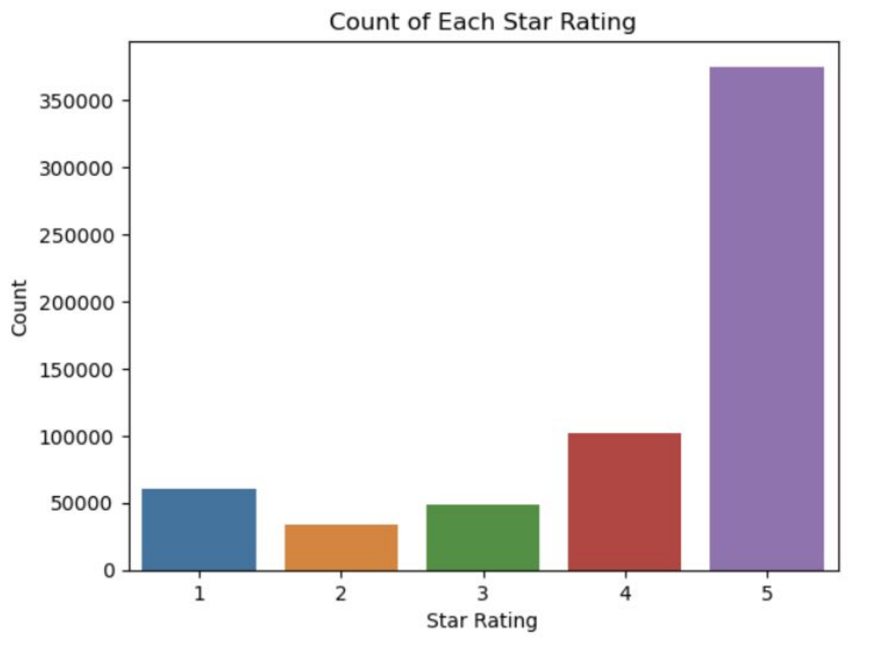


1. No significant correlation between 'star_rating' and 'helpful_votes' (-0.02).

2. No significant correlation between 'star_rating' and 'total_votes' (-0.036).

3. The correlation coefficient between 'helpful_votes' and 'total_votes' is very high at **0.99**, indicating a very strong positive linear relationship.

Distribution of Star Ratings

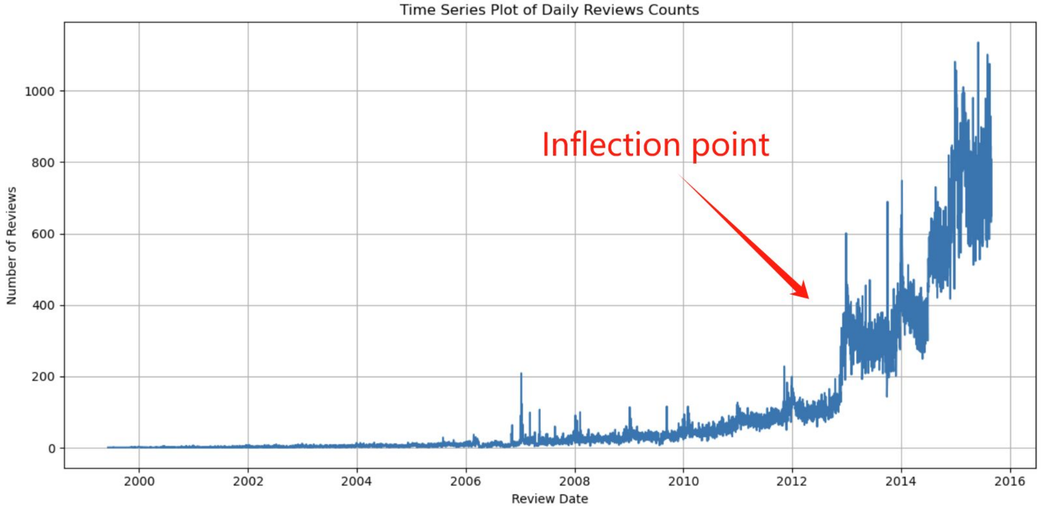


1. The majority of our reviews are 5-star, indicating **high customer satisfaction**.

2. Approximately 15.15% of the reviews are in the 1 or 2-star rating range, indicating potential areas for improvement.

3. Reviews with a 3-star rating, which account for roughly 8%, suggest a moderate viewpoint and could offer valuable insights for enhancement.

Time Series Plot of Review Counts

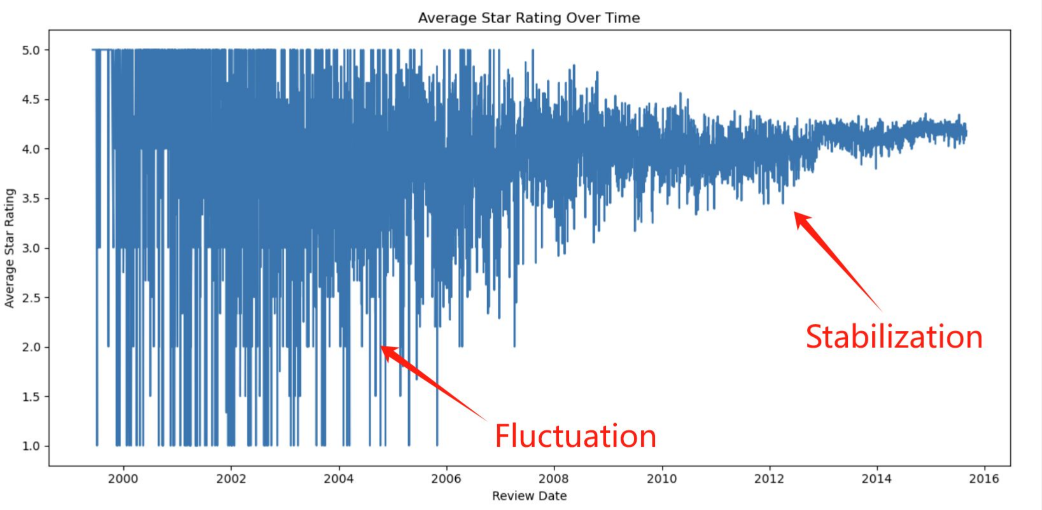


1. There's a clear trend of increasing review counts over time.

2. The graph depicts an sudden rise in review counts beginning in 2013, marking a distinct inflection point in the trend.

3. There may be patterns or seasonality in the data, such as certain times of year (holidays), but it is not immediately clear from the graph without further context.

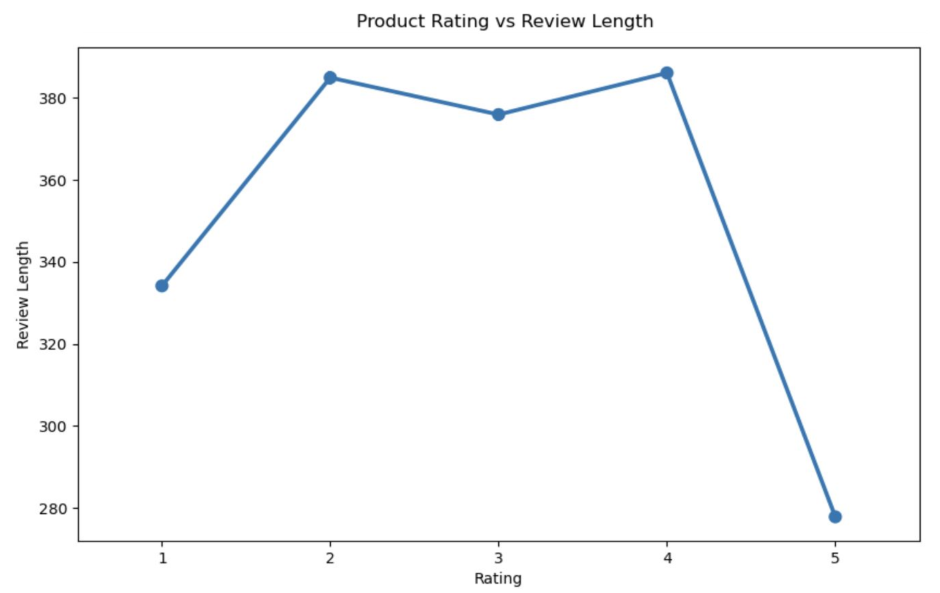
Average Star Rating Over Time



1. Initial Fluctuations (2000-2002): There's a high variability in the ratings, indicating inconsistent feedback during this period.

2. Stabilization (2003-now): The average rating seems to stabilize around the 3.5 to 4.5 range with less drastic fluctuations. This could indicate a period of consistent product or service quality, or perhaps more consistent expectations from reviewers.

Star Ratings & Review Length

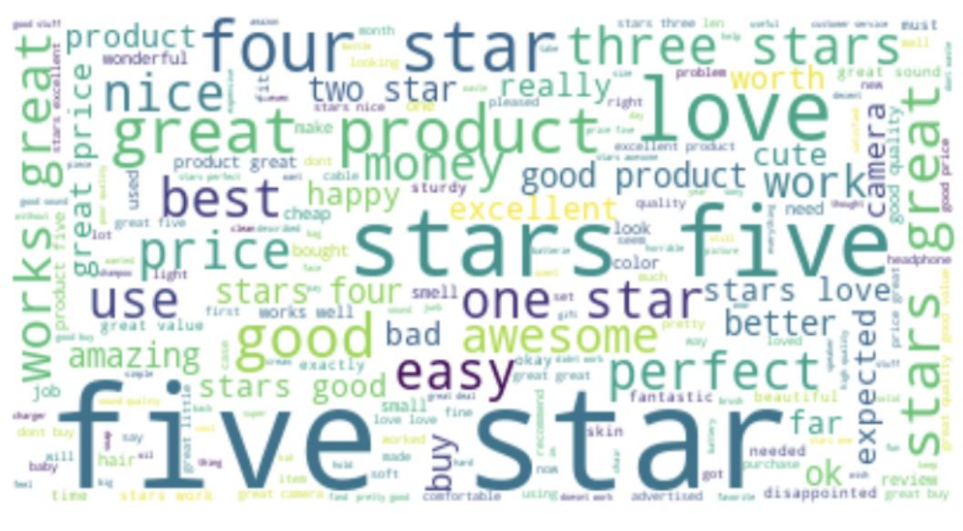


1. The product review word length are longer when the rating is below 5, indicating that people might write longer posts for unsatisfying products.

2. People also tend to write less and just give 5 star ratings when they satisfy the product.

Word Clouds

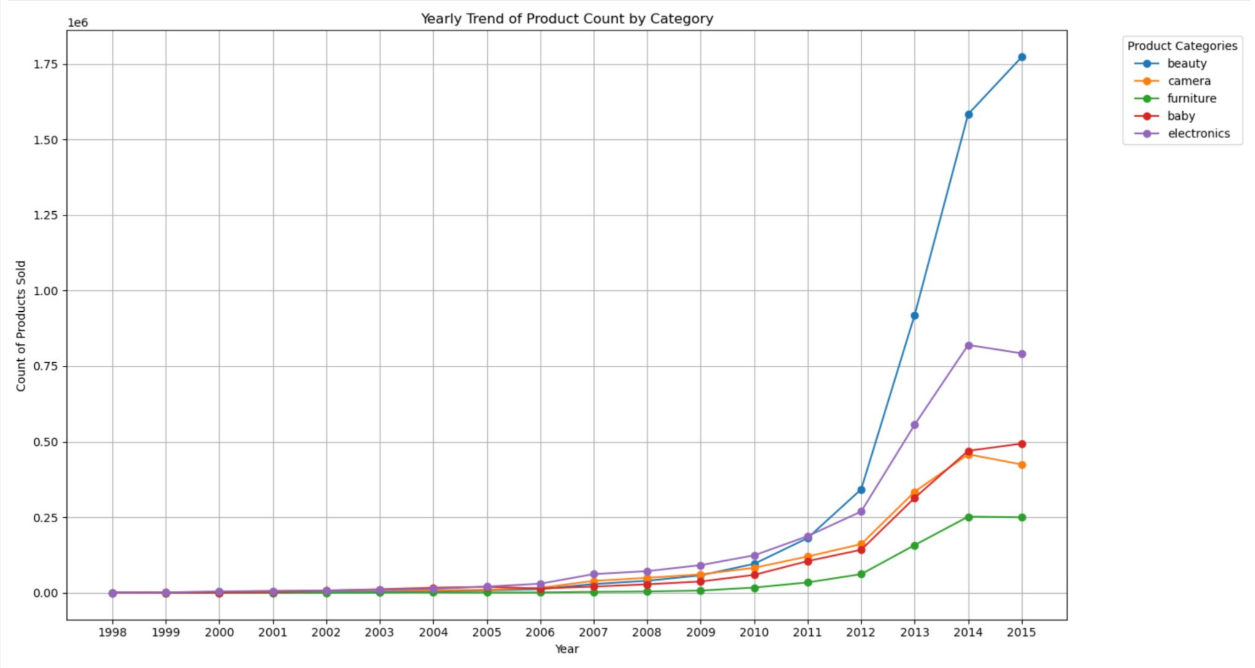
Word clouds for review headline



1. The most prominent words, such as "five", "star", "love", "great", "good", and "best", suggest that many of the reviews are positive. Words like "amazing", "perfect", "excellent", "easy", and "nice" contribute to the understanding that the sentiment of the reviews is generally positive.

2.Smaller words, such as "disappointed", "okay", "bad", and "cheap", negative reviews appear less frequently than positive comments.

Sold Products by Category



- The beauty category has the largest increase on trend of product count by category.

Models

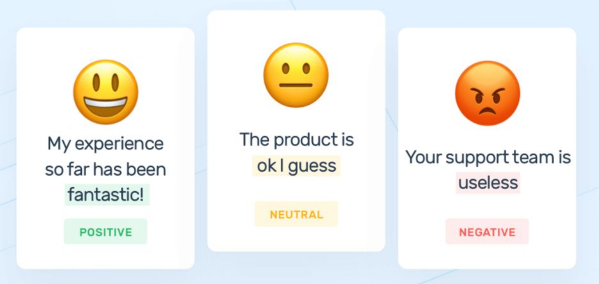
1. Sentiment Analysis
2. Customer Segmentation
3. Recommendation System

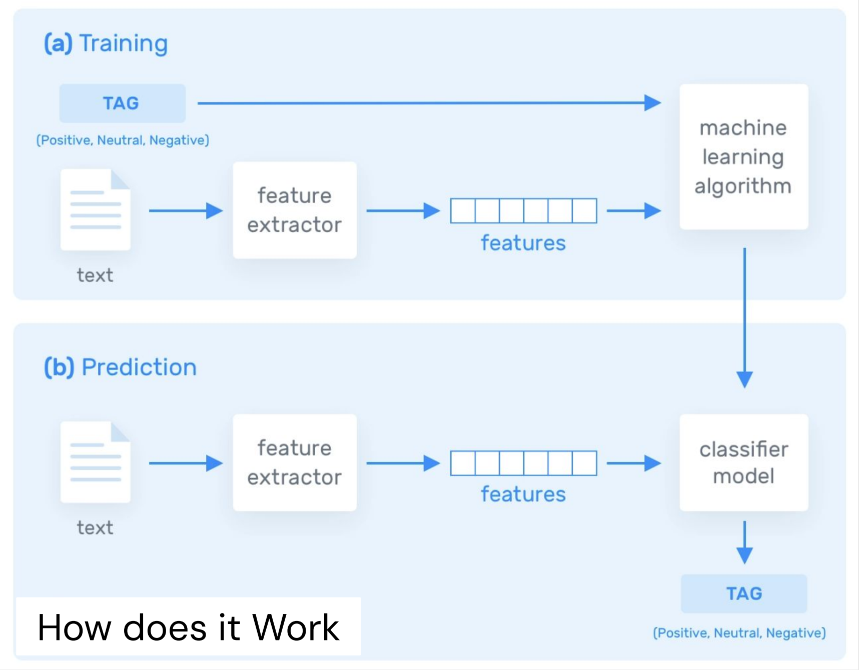
Sentiment Analysis

- The process of using NLP and machine learning to classify sentiments (positive, negative, neutral) in text data

Our Problem

- Classify the sentiment of an Amazon product review as Positive, Neutral, or Negative.

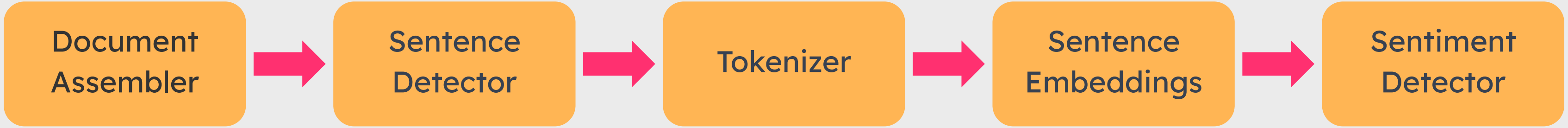

My experience so far has been fantastic!
POSITIVE
The product is ok I guess
NEUTRAL
Your support team is useless
NEGATIVE



How does it Work

Sentiment Analysis

Model Pipeline



Converts raw text into a structured format

Splits the text into individual sentences

Breaks sentences into tokens (words)

Uses **Universal Sentence Encoder** to convert sentences into numerical representations

Analyzes the embeddings to determine sentiment, utilizing a **Spark NLP pre-trained model (sentimentdl_use_twitter)**

Sentiment Analysis

Model Evaluation

	Logistic Regression Model
Star Rating	Accuracy = 0.6711058784664937
Sentiment	Accuracy = 0.7800860932363265

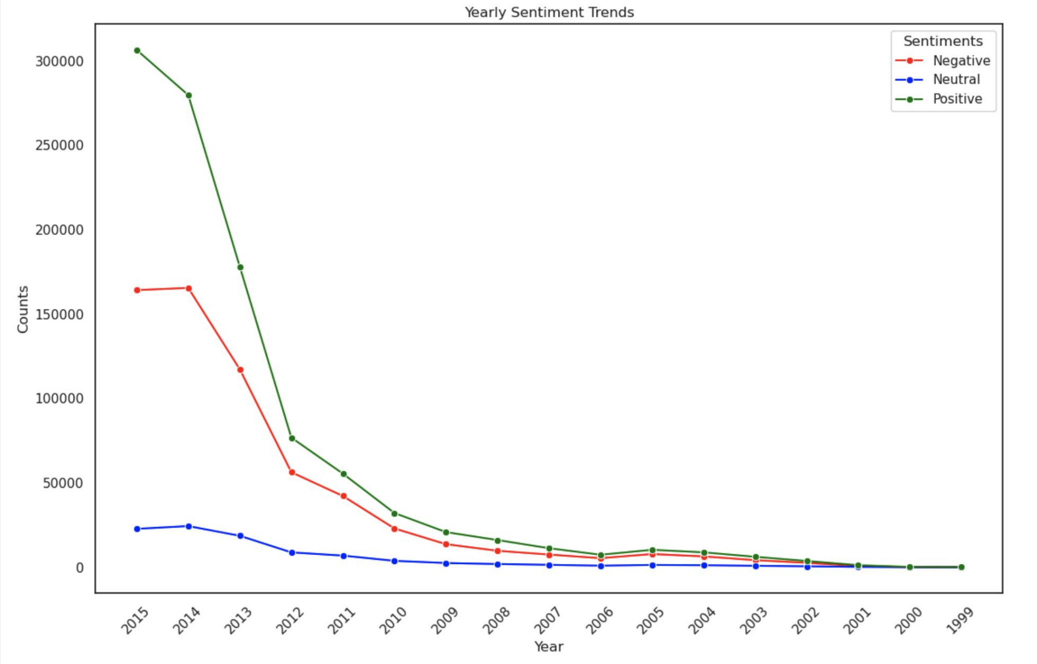
Class positive (Index 0)	Precision = 0.8134483314231166, Recall = 0.8790540895710681, F1 Score = 0.8449796863182262
Class negative (Index 1)	Precision = 0.763232640513266, Recall = 0.7294681876008027, F1 Score = 0.7459685430688879
Class neutral (Index 2)	Precision = 0.14800381740098617, Recall = 0.06459114257948077, F1 Score = 0.08993379403663074

Modeling Improvement

- Imbalance issues /Missing data /ngram for processing data/parameters adjusted for predicting
- Drop Neutral, Improved Accuracy: 0.8776766997207375

Sentiment Analysis

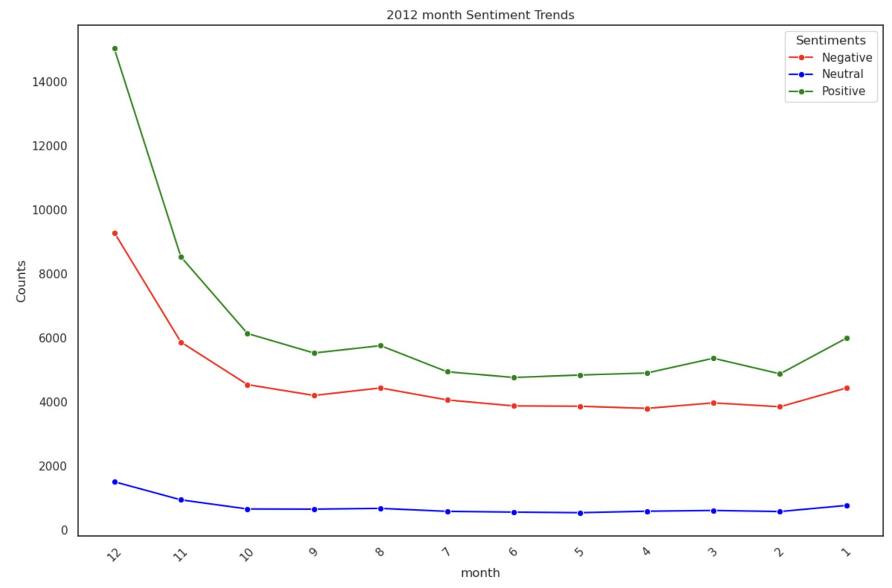
Trend Analysis



- The number of positive and negative sentiment significantly increased over years especially after 2012.
- Higher increment for positive reviews compared to negative reviews after 2012
- The negative reviews become stable after 2014

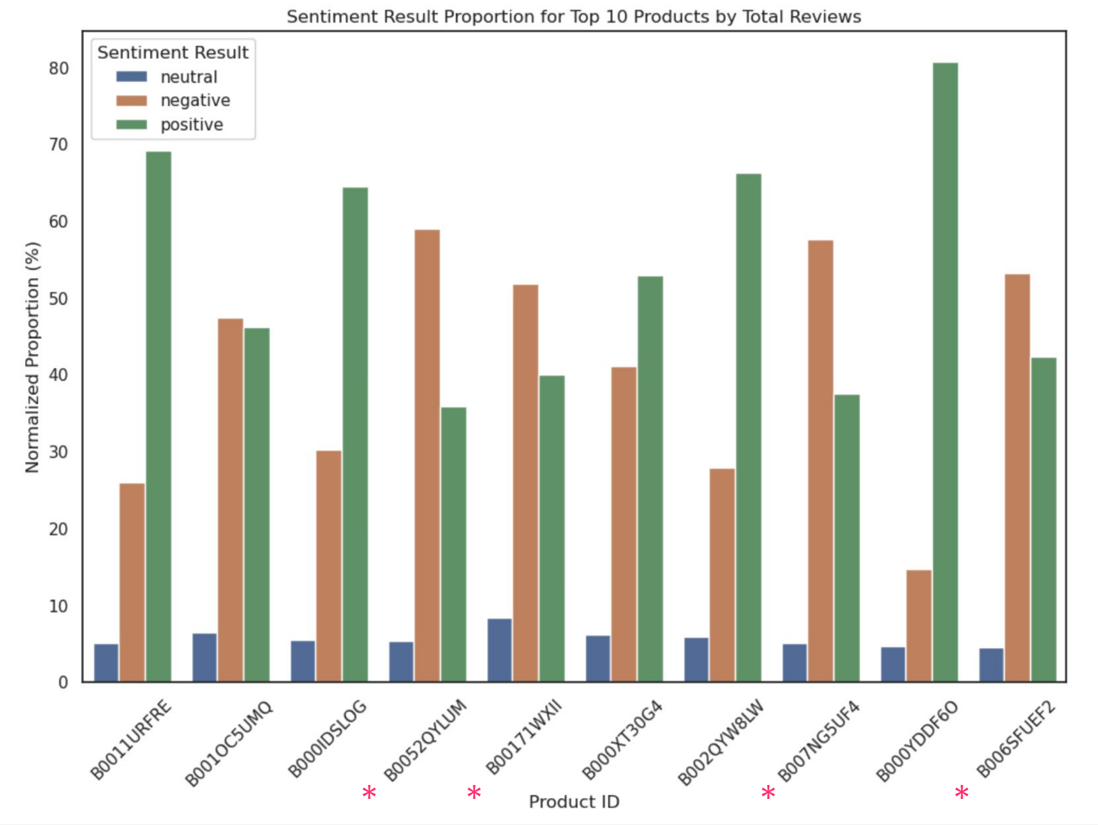
Sentiment Analysis

Trend Analysis



- **Seasonal or Monthly Trends**: There appears to be a significant decline in the counts of sentiments from the beginning of the year towards the following months. However, after the November, both positive and negative reviews increased.

Sentiment Analysis



- 4 products have high proportion of negative sentiment result but with fair star_rating scores

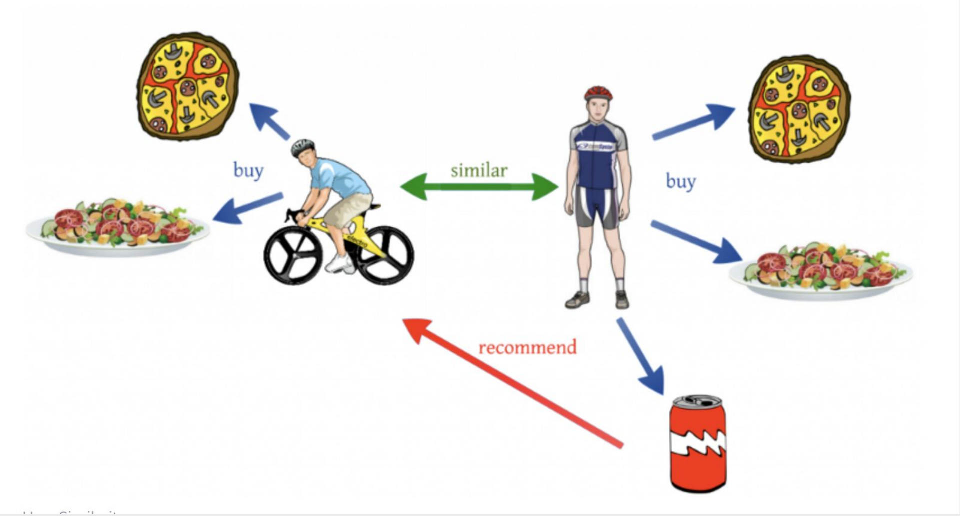
	product_id	avg_star_rating
1	B0052QYLUM	3.874745228830832
2	B00171WXII	4.656218402426694
3	B006SFUEF2	3.8251057827926656
4	B007NG5UF4	3.9677570093457946

Recommendation System

- The process of recommending products to users based on their **past reviews and ratings.**

Our Problem

- Recommend products to users using ALS (Alternating Least Squares



Recommendation System

Method:Alternating Least Squares(ALS)

ALS is a popular algorithm used in collaborative filtering for making recommendations, particularly in the context of large-scale matrix factorization problems.

Why ALS?

- **Scalability**: ALS can be parallelized and distributed across multiple nodes, making it suitable for large datasets.
- **Avoiding Overfitting**: The ALS algorithm in Spark includes regularization parameters that help prevent overfitting.
- **Cold Start Problem Mitigation**: While ALS faces challenges with new users or items (cold start problem), it can still provide reasonable recommendations based on the overall structure of the user-item interaction matrix.

Recommendation System

Process Description

Data Preparation:

- StringIndexer: Converts string identifiers to numeric indices.(ALS in Spark requires numeric IDs.)

ALS Model Initialization:

- The model is configured to drop any cold start occurrences and only produce non-negative predictions.

Hyperparameter Tuning:

- A parameter grid is built for cross-validation. It tests combinations of the ALS hyperparameters: rank, max iterations, and regularization parameter.
- The process evaluates the model's performance using the Root Mean Square Error (RMSE) metric.

Cross-validation:

- 4-fold cross-validation used to find the best model parameters.

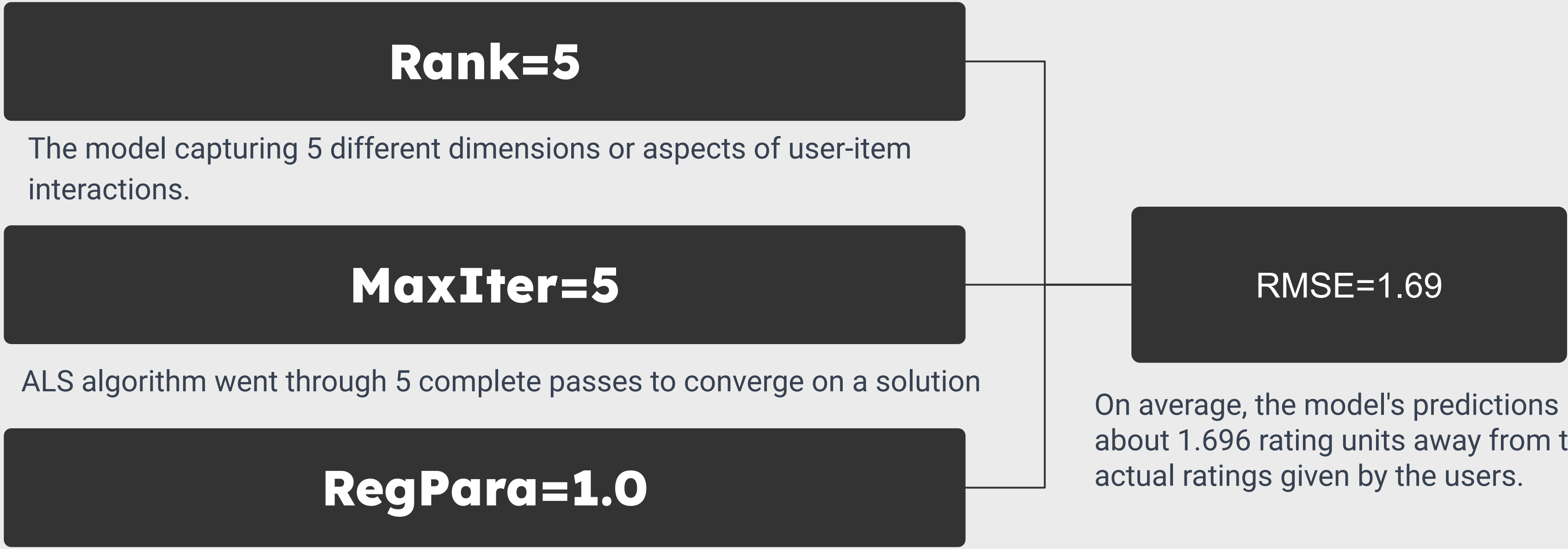
Best Model Selection:

- The best model from the cross-validation is selected based on the performance (lowest RMSE).

Making Predictions:

- The best ALS model is used to make predictions on the test dataset.

Recommendation System



Rank=5

The model capturing 5 different dimensions or aspects of user-item interactions.

MaxIter=5

ALS algorithm went through 5 complete passes to converge on a solution

RegPara=1.0

Regularization parameter =1.0 suggests that the model has applied a significant level of regularization.

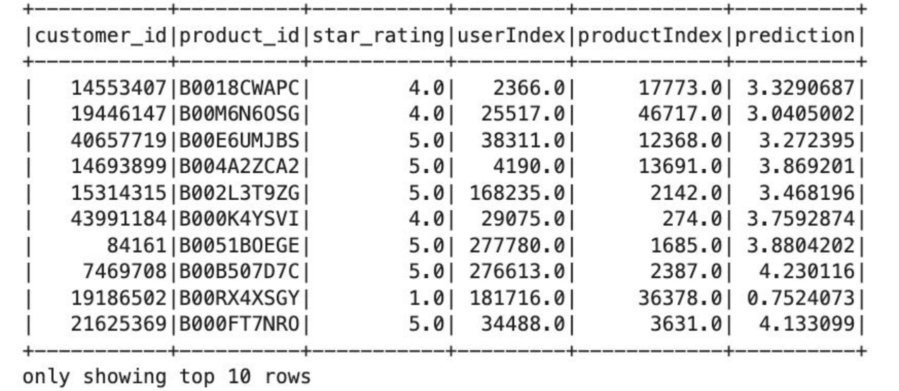
RMSE=1.69

On average, the model's predictions are about 1.696 rating units away from the actual ratings given by the users.

The RMSE value should be considered in the context of your dataset's rating scale. For example, if ratings are on a 1-5 scale, an average error of 1.696 could be seen as quite significant.

Recommendation System

Result



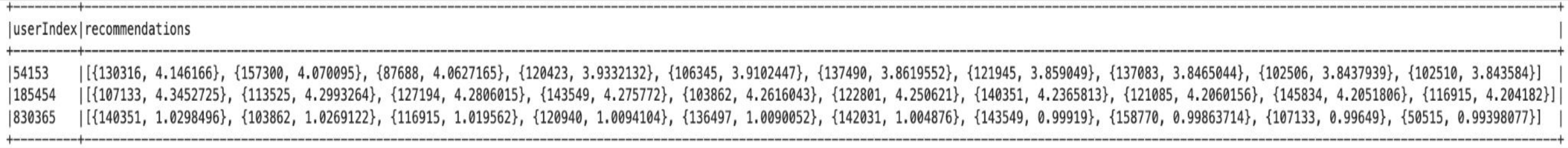
```
+-----------+----------+-----------+---------+------------+----------+
|customer_id|product_id|star_rating|userIndex|productIndex|prediction|
+-----------+----------+-----------+---------+------------+----------+
|   14553407|B0018CWAPC|        4.0|   2366.0|     17773.0| 3.3290687|
|   19446147|B00M6N6OSG|        4.0|  25517.0|     46717.0| 3.0405002|
|   40657719|B00E6UMJBS|        5.0|  38311.0|     12368.0|  3.272395|
|   14693899|B004A2ZCA2|        5.0|   4190.0|     13691.0|  3.869201|
|   15314315|B002L3T9ZG|        5.0| 168235.0|      2142.0|  3.468196|
|   43991184|B000K4YSVI|        4.0|  29075.0|       274.0| 3.7592874|
|      84161|B0051BOEGE|        5.0| 277780.0|      1685.0| 3.8804202|
|    7469708|B00B507D7C|        5.0| 276613.0|      2387.0|  4.230116|
|   19186502|B00RX4XSGY|        1.0| 181716.0|     36378.0| 0.7524073|
|   21625369|B000FT7NRO|        5.0|  34488.0|      3631.0|  4.133099|
+-----------+----------+-----------+---------+------------+----------+
only showing top 10 rows
```

Personalized Product Recommendations:

- E-commerce and Retail: Use the system to suggest products to customers based on their past purchases or browsing history. This can increase sales by showing customers items they are likely to buy.

Targeted Marketing:

- Personalized Promotions: Use the recommendation engine to identify products that specific customer segments are likely to purchase and target them with specialized marketing campaigns or promotions.



```
+---------+----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|userIndex|recommendations                                                                                                                                                                                                             |
+---------+----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|54153    |[{130316, 4.146166}, {157300, 4.070095}, {87688, 4.0627165}, {120423, 3.9332132}, {106345, 3.9102447}, {137490, 3.8619552}, {121945, 3.859049}, {137083, 3.8465044}, {102506, 3.8437939}, {102510, 3.843584}]  |
|185454   |[{107133, 4.3452725}, {113525, 4.2993264}, {127194, 4.2806015}, {143549, 4.275772}, {103862, 4.2616043}, {122801, 4.250621}, {140351, 4.2365813}, {121085, 4.2060156}, {145834, 4.2051806}, {116915, 4.204182}]|
|830365   |[{140351, 1.0298496}, {103862, 1.0269122}, {116915, 1.019562}, {120940, 1.0094104}, {136497, 1.0090052}, {142031, 1.004876}, {143549, 0.99919}, {158770, 0.99863714}, {107133, 0.99649}, {50515, 0.99398077}]  |
+---------+----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

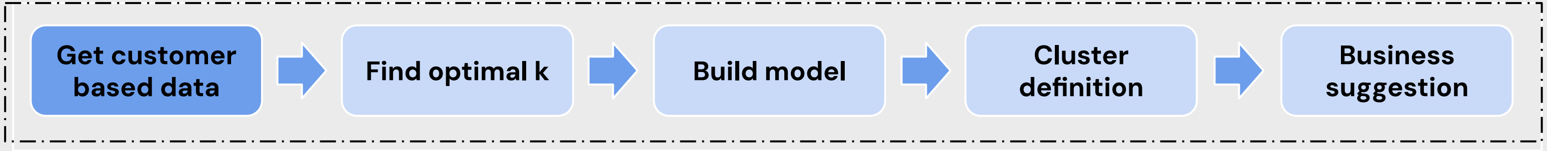
Customer Segmentation

- The process of dividing a customer base into groups with similar characteristics or behaviors

Our Problem

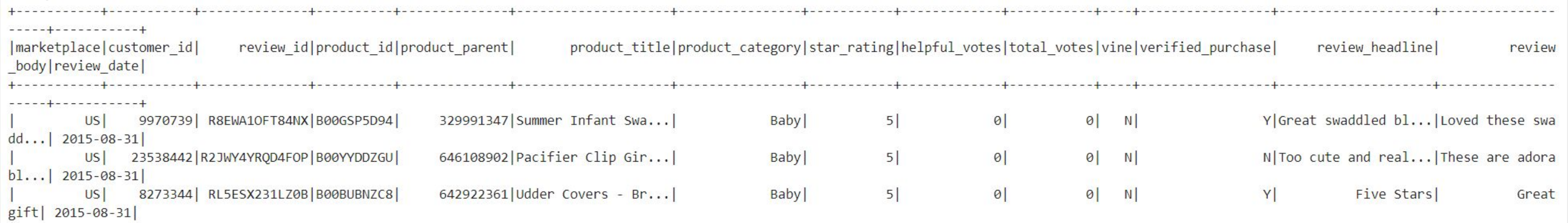
- Cluster Amazon customers based on their reviewing behavior using **K-means Clustering**

1. Create customer based dataset



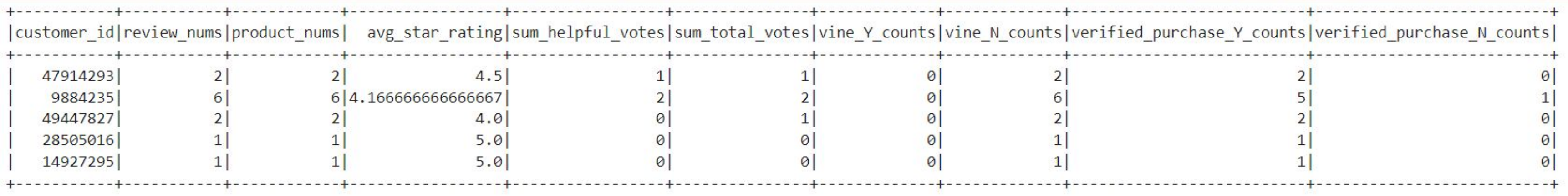
Get customer based data → Find optimal k → Build model → Cluster definition → Business suggestion

Original dataset:



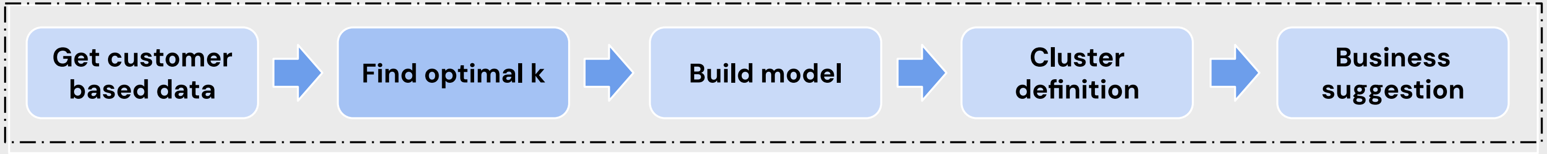
```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+---------------
-----+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review
_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+---------------
-----+-----------+
|         US|    9970739| R8EWA1OFT84NX|B00GSP5D94|     329991347|Summer Infant Swa...|            Baby|          5|            0|          0|   N|                Y|Great swaddled bl...|Loved these swa
dd...| 2015-08-31|
|         US|   23538442|R2JWY4YRQD4FOP|B00YYDDZGU|     646108902|Pacifier Clip Gir...|            Baby|          5|            0|          0|   N|                N|Too cute and real...|These are adora
bl...| 2015-08-31|
|         US|    8273344| RL5ESX231LZ0B|B00BUBNZC8|     642922361|Udder Covers - Br...|            Baby|          5|            0|          0|   N|                Y|          Five Stars|          Great
gift| 2015-08-31|
```

Customer based dataset:

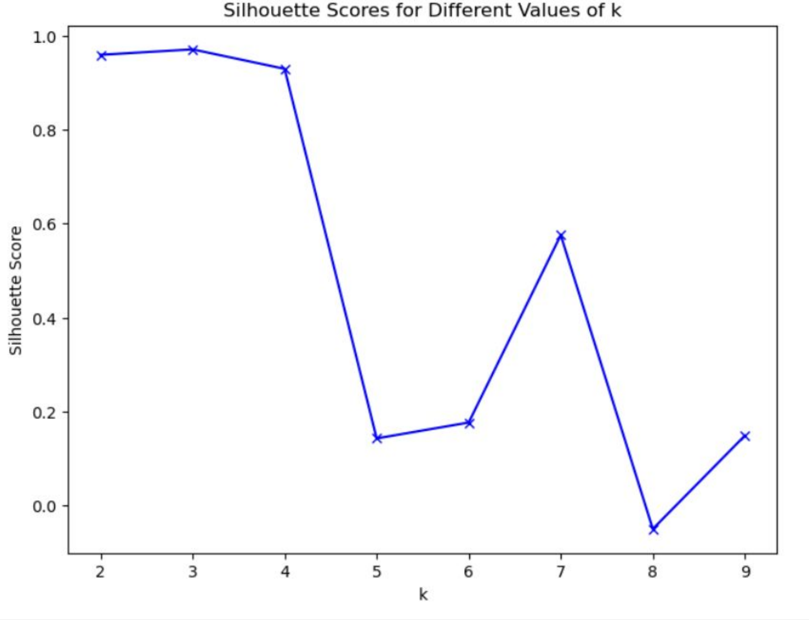


```
+-----------+-----------+------------+----------------+-----------------+---------------+-------------+-------------+--------------------------+--------------------------+
|customer_id|review_nums|product_nums| avg_star_rating|sum_helpful_votes|sum_total_votes|vine_Y_counts|vine_N_counts|verified_purchase_Y_counts|verified_purchase_N_counts|
+-----------+-----------+------------+----------------+-----------------+---------------+-------------+-------------+--------------------------+--------------------------+
|   47914293|          2|           2|             4.5|                1|              1|            0|            2|                         2|                         0|
|    9884235|          6|           6|4.16666666666667|                2|              2|            0|            6|                         5|                         1|
|   49447827|          2|           2|             4.0|                0|              1|            0|            2|                         2|                         0|
|   28505016|          1|           1|             5.0|                0|              0|            0|            1|                         1|                         0|
|   14927295|          1|           1|             5.0|                0|              0|            0|            1|                         1|                         0|
+-----------+-----------+------------+----------------+-----------------+---------------+-------------+-------------+--------------------------+--------------------------+
```

2. Find optimal k value



Get customer based data → Find optimal k → Build model → Cluster definition → Business suggestion

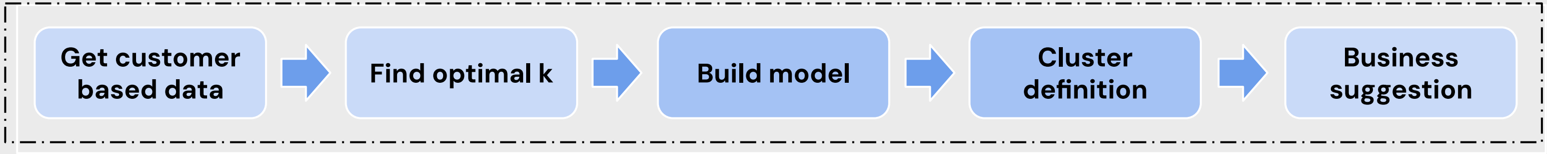


Silhouette Scores for Different Values of k

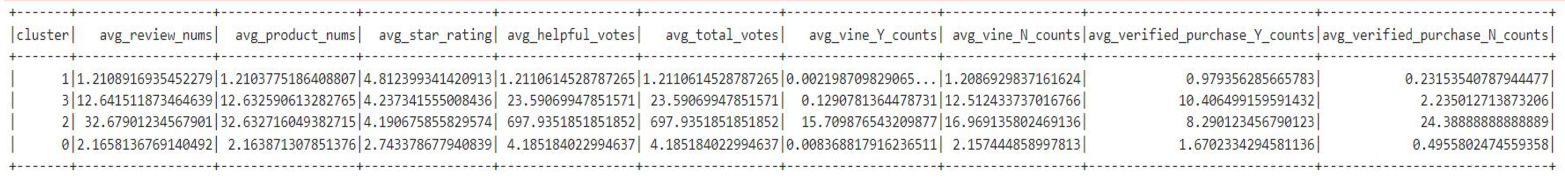
1. The silhouette score is a metric used to evaluate the quality of clusters in a clustering analysis. It measures how similar each data point in one cluster is to the data points in the same cluster compared to the nearest neighboring cluster. The silhouette score ranges from -1 to 1, with higher values indicating better cluster separation and cohesion.

2. Inspection point at **k=4**, with 0.93 silhouette score.

3&4. Build model & Cluster definition


Get customer based data
Find optimal k
Build model
Cluster definition
Business suggestion

Check statistics of each cluster:

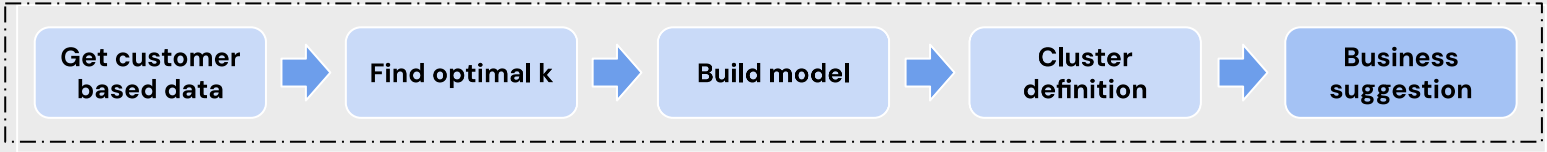


```
+-------+------------------+------------------+-----------------+------------------+------------------+--------------------+------------------+------------------------------+------------------------------+
|cluster|   avg_review_nums|  avg_product_nums|  avg_star_rating| avg_helpful_votes|   avg_total_votes|   avg_vine_Y_counts| avg_vine_N_counts|avg_verified_purchase_Y_counts|avg_verified_purchase_N_counts|
+-------+------------------+------------------+-----------------+------------------+------------------+--------------------+------------------+------------------------------+------------------------------+
|      1|1.2108916935452279|1.2103775186408807|4.812399341420913|1.2110614528787265|1.2110614528787265|0.002198709829065...|1.2086929837161624|             0.979356285665783|           0.23153540787944477|
|      3|12.641511873464639|12.632590613282765|4.237341555008436| 23.59069947851571| 23.59069947851571|  0.1290781364478731|12.512433737016766|            10.406499159591432|             2.235012713873206|
|      2| 32.67901234567901|32.632716049382715|4.190675855829574| 697.9351851851852| 697.9351851851852|  15.709876543209877|16.969135802469136|             8.290123456790123|             24.38888888888889|
|      0|2.1658136769140492| 2.163871307851376|2.743378677940839| 4.185184022994637| 4.185184022994637|0.008368817916236511| 2.157444858997813|            1.6702334294581136|            0.4955802474559358|
+-------+------------------+------------------+-----------------+------------------+------------------+--------------------+------------------+------------------------------+------------------------------+
```

Define a new variable: product review engagement

0/1: low engagement
3: middle engagement
2: high engagement

5. Business suggestion



1. **Low engagement**

Target marketing strategies: Implement targeted marketing strategies to increase customer engagement, such as email campaigns that encourage reviews in exchange for a discount on future purchases.
Feedback Loop: Send out surveys or feedback requests to understand why their engagement is low and what could be improved in their experience.
Apply New Recommendation System: To increase their interaction with our products or services.

2. **Middle & High engagement**

Community Building: Invite these customers to a VIP club or community where they can share their experiences, get early access to new products, and feel more connected to the brand.
Extra Access: Provide early access to new products, special editions, or exclusive previews as a reward for their high engagement. Put them into experiment group while performing A/B testing.

Future Work

Advanced Machine Learning and NLP Techniques

- **Aspect-Based Sentiment Analysis**: Move beyond overall sentiment to analyze specific aspects of products, like quality, price, or usability.
- **Deep Learning**: Incorporate deep learning models for more nuanced sentiment analysis and enhanced recommendation systems.

Larger and More Diverse Data Integration

- **Cross-Platform Analysis**: Incorporate reviews and ratings from other e-commerce platforms for a more comprehensive analysis.

Real-Time Data Processing

- **Streaming Data**: Implement real-time analysis of customer reviews using streaming technologies like Apache Kafka and Spark Streaming.

Conclusion

The Amazon Review Analysis project successfully integrated Sentiment Analysis, Product Recommendation System, and Customer Segmentation, showcasing the immense value of data analytics in enhancing customer experiences and guiding business strategies.

- **Sentiment Analysis** provided crucial insights into customer opinions, aiding in product improvement and responsive decision-making.
- **The Product Recommendation System** personalized shopping experiences, significantly boosting sales through targeted product suggestions.
- **Customer Segmentation** enabled precise marketing strategies and improved service offerings by categorizing customers based on behavior and preferences.

Thank You!