Qualetics
DATA MACHINES

Data Intelligence in Retail Use Case

Sentiment Analysis of Amazon's Product Reviews

Use Case

Introduction

Online product reviews are important for both buyers and sellers in eCommerce. A detailed analysis of the reviews can help retailers extract actionable insights for product improvements, marketing campaigns, and consumer satisfaction, etc. The reviews help a buyer take informed purchase decisions based on the product experience of other buyers.

Deep data science technologies like AI, NLP and Machine Learning can be leveraged to mine the data, consumer reviews in this case, and provide valuable insights. Text mining can help businesses in exploring and analyzing vast amounts of unstructured text data and identify concepts, patterns, topics, etc.

This use case leverages Data Mining, Natural Language Processing, Machine Learning, and Data Visualization, to build algorithms that analyze online reviews and help us understand the consumer sentiments on electronic products available on Amazon. The model can be helpful for any ecommerce business to ascertain the consumer sentiment towards its products and brands.

Problem Statement/Challenges:

Amazon's newest set of electronic devices was a trending topic in recent times. With the total product review count showing an upward trend on the graph, it leaves an enormous trail of data behind. Analyzing the review data will help the company in enhancing the quality of its products and build a closer relationship with its consumers.
The objective of this use case is to analyze Amazon's most successful consumer electronics product launches, over the years, and discover insights into consumer reviews by performing consumer sentiment analysis.

About the Dataset:

The dataset contains over 34,000 consumer reviews for Amazon products like Kindle, Fire TV Stick, and other products provided by Datafiniti's Product Database. The dataset includes information about the products, ratings, review texts, and more for each product. For more details on the dataset and sources, please click here.

Data Cleansing:

After studying the data from the sample dataset carefully, we identified the most important columns for our analysis, for e.g. (reviews.rating, reviews.title).

We then performed data cleansing on the dataset, with the help of Natural Language Toolkit, to find the compound score for most occurring words. Identifying the accurate and inaccurate data from the dataset and correcting or deleting inaccurate data delivered us a cleaner data set required for our analysis. We also removed some columns like id, asins, etc as they do not provide us with any information needed for our analysis. During the process, we made assumptions like:

Assumptions

Sentimental Analysis can be performed on the present dataset while taking into account the following assumption-

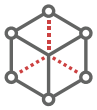
- Dropping the columns that do not play a role in our analysis for e.g. (id, asins etc).
- Imputing the values which were not known.

Technologies Used:

We leveraged the following technologies to realize the objective behind this use case-

Data Preprocessing

Exploratory Data Analysis (EDA)



Machine Learning algorithms

Natural Language Toolkit (NLTK)

Our Process:

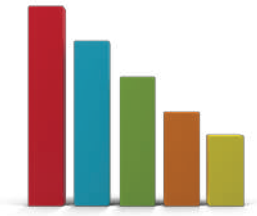
Machine Learning models cannot differentiate between noise and valuable data. Data must be cleaned before using it for analysis. We started our analysis by understanding and cleaning the data, where we checked for any null values and columns irrelevant for our study and deleted them. This delivered us with more filtered and refined data for our analysis.

We used NLTK (Tokenization, Stemming, NPS and TFIDF vectorizer), to ascertain the genuineness of ratings and reviews available in the dataset.

We used Machine Learning to build the model.

Models used :

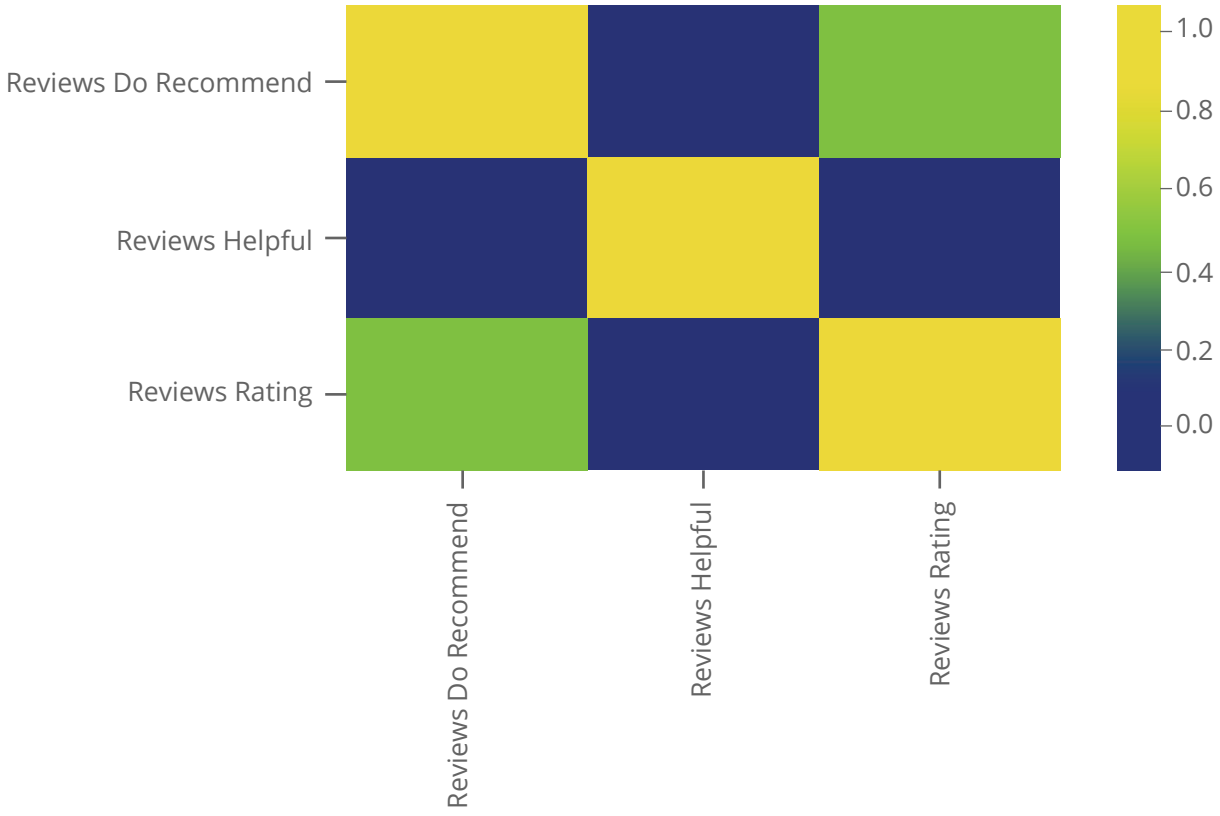
- Logistic Regression
- Multinomial Naive Bayes
- Bernoulli Naive Bayes



Correlation matrix for Amazon consumer reviews:

There are instances where a consumer has provided multiple ratings/reviews for a single product and recommend the product to other consumers. An NPS (Net Promoter Score) measures the willingness of the consumer to recommend a company's products or services.

To understand the correlation between ratings and recommending the products, we tried to analyze the data and prepare a correlation matrix.



Inference - From the above correlation matrix it can be observed that recommending a product is highly correlated with ratings.

Calculating Net Promoter Score with a 5-point NPS Breakdown

To arrive at the NPS score, we subtracted the percentage total of Detractors from the percentage total of Promoters. These percentages are calculated by considering the group total and dividing it by the total number of survey responses. With a higher NPS, it can be derived that the consumer has a high probability of recommending the product.

The classification of consumers into Detractors, Passive and Promoters is done based on the ratings given by them.

Ratings between 0-3: Detractors

Ratings above 4: Passive 80.324

Ratings above 5: Promoters

Promoters have an NPS of 80.324.

The average rating received by amazon is 4.58/5.

80.324
Average NPS

4.58/5
Amazon Ratings

Understanding the ratings

Below are our findings on ratings, the number of users, and the bulk ratings.

Total ratings: 34658

Total users: 26789

Users giving bulk ratings (more than 10): 146

Bulk ratings: 3160

Populations of bulk ratings: 9.117664031392463

Populations of bulk users: 0.5449998133562283

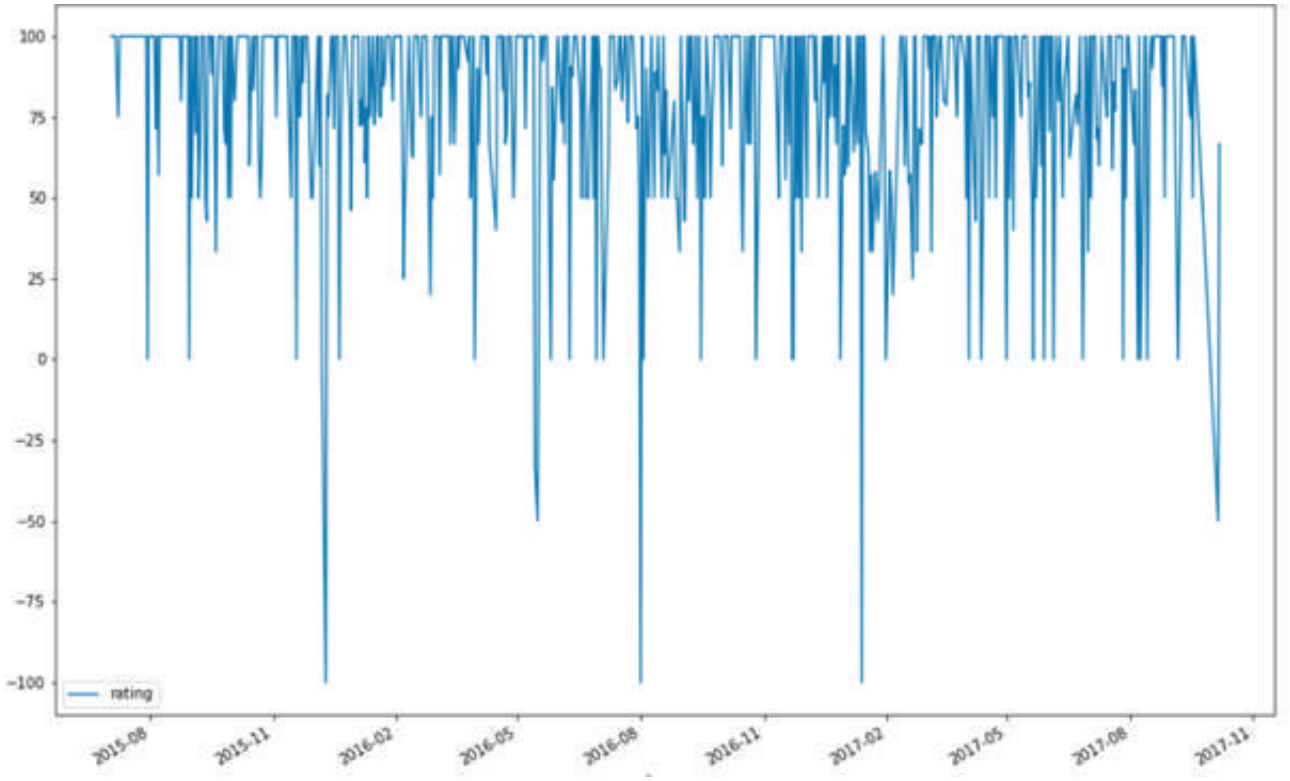
users

Ratings

Inference - It can be seen that around 0.55% are bulk users and 9% ratings are submitted by these bulk users.

The above analysis generated insights into the peak sale period for Amazon. With December and January being the holiday seasons, more sales were generated during the Christmas and New Year. There was a high degree of variance in reviews added over time.

Let's visualize how the rating of Kindle PaperWhite has changed over time



Based on the product IDs available in our sample dataset, we observed that there are a total of 42 products. It is obvious that each product will have a mix of reviews, i.e., positive, neutral, and negative. We now try to identify the most common words used in these reviews by using a word cloud.

Identifying Positive & Negative Words

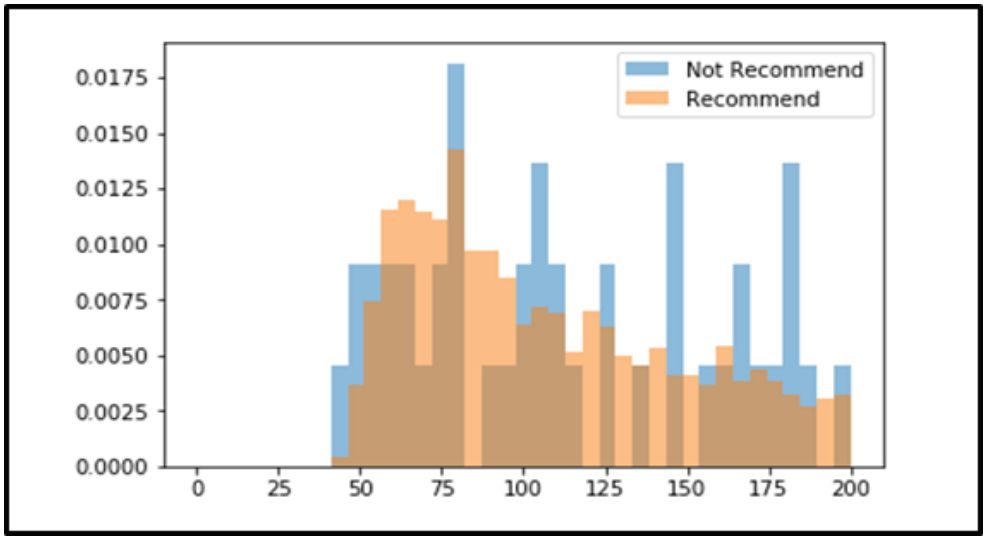
From the words collected from the word cloud, we performed a word count of certain words that described the sentiments of the consumers in a better and positive way.

Positive Words

Negative Words

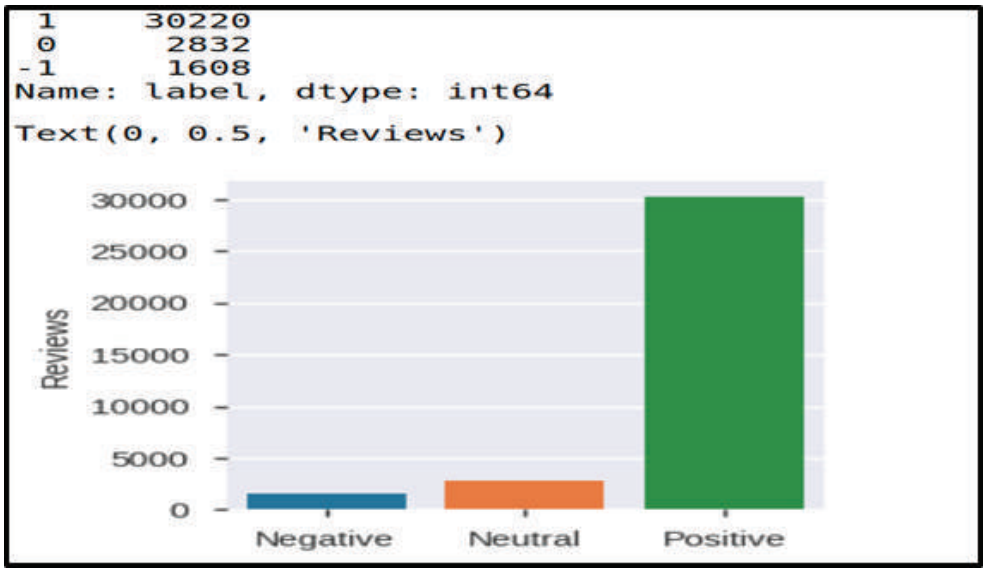
People who do not recommend the product have more to talk about its features

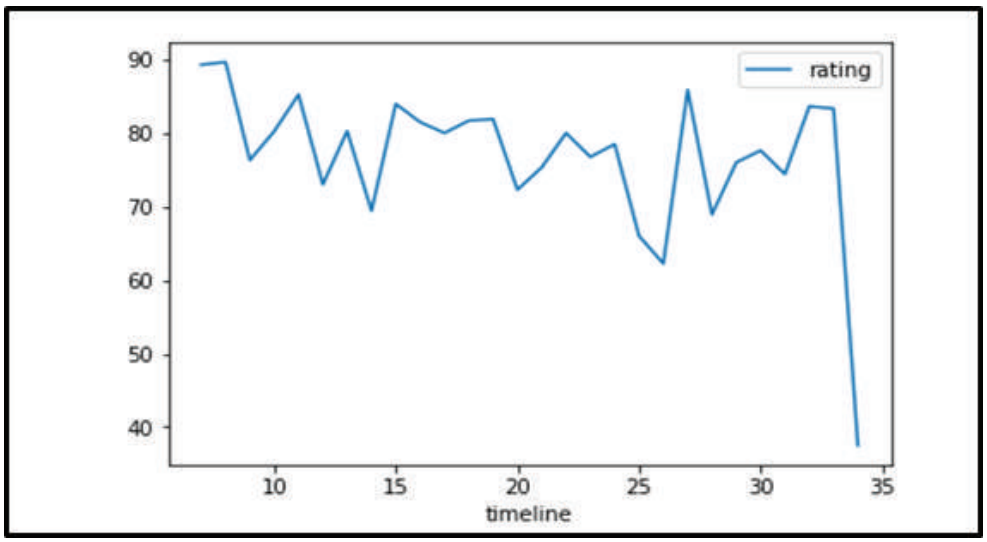
Consumers who are not satisfied with a certain product will provide a negative review of the product and company. To justify their ratings, they talk about the product features they are not satisfied with.



The plot above shows that the products that have positive reviews and high ratings are recommended by the consumers who provided those ratings.

Amazon sells millions of products and many products have positive responses. There are more positive responses than negative ones given for these products.





Rating of products over the years

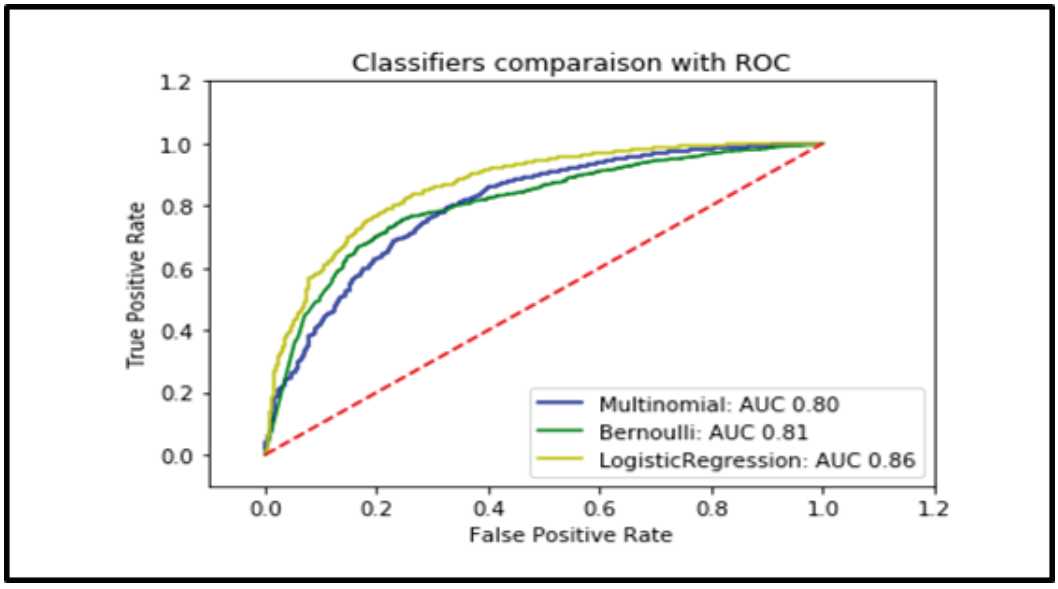
Model Metrics

Machine Learning is a technology that is part art and part science. There is no one solution or one approach that will fit to all the problems. Hence, choosing a model will depend on many factors to narrow down the search for the most feasible and accurate models. The models should reflect their ability to meet the business goals, accuracy in achieving the required results, their ability to make swift predictions, etc.

Model	Accuracy
Logistic Regression	93.7%
Multinomial Naive Bayes	93.2%
Bernoulli Naive Bayes	92.04%

Three models were selected for our analysis based upon various factors and the type of sample dataset available with us. Logistic Regression achieved the best accuracy levels of the models selected.

The Logistic Regression Model gave us the best performance with an accuracy of 93.7%.



The above plot tells us the area under the curve covered by each classifier.

To test the performance of each model We created a new column considering it as our target variable. We split the data, calculating count vector and TF-IDF vector on the training and test data.

Logistic Regression :	Precision	Recall	F1-score	Support
Positive	0.56	0.33	0.41	464
Negative	0.95	0.98	0.97	6461

1. Precision means the percentage of your results that are relevant.
2. Recall refers to the percentage of total relevant results correctly classified by your algorithm.
3. F1-score is the weighted average of the above two.

Qualetics As A Solution:

This analysis helps us understand the consumer reviews for each product of Amazon. Data obtained from consumer reviews provide valuable insights into buying decisions and understanding of the associated sentiment provides businesses with market awareness and the ability to proactively address issues.

As online reviews are often large in numbers and are unstructured, it becomes critical for a business to build a system capable of processing large data and extract actionable data intelligence. Qualetics provides you with the exact solution by giving you deep insights into consumer feedback and accurately understand their sentiments.

There will be many barriers in the process of capturing the true sentiments from the reviews, such as language ambiguity through cryptic dialogue, sarcasm, and irony, as well as the emotion icons (emojis/emoticons) which may not be analyzed in a pure text capture. All these barriers can make understanding the sentiment more difficult. Qualetics enables an easy and seamless integration of our solution into your applications or analytics solutions. It also provides the flexibility for immediate adaptations to be applied in order to meet the dynamically changing consumer needs.

- **Data Extraction and Processing Pipelines**
- **Data Analysis**
- **Visualization and Integration Platform**

Data Extraction and Processing Pipelines

Our proprietary architecture built using some of the industry-standard protocols and tools like MQTT, Kafka, NoSQL data stores allows us to set up a data pipeline from the source to the analytical data store with relative ease. This allows us to apply pre-designed learning models to data as it is being generated.

Data Analysis

In addition to pre-designed and pre-trained models, having a continuous data stream that can pick up new data points or variations allows us to refine and tweak the models over time or develop new models altogether. Our expert Data Science team can periodically monitor the efficiency and performance of the models and update them over time

Visualization and Integration Platform

The results of the analysis are delivered through our proprietary Visualization and Integration platform. The platform can be integrated into your Products, Systems, and Processes by Single Sign-On (SSO) or APIs.

Along with features such as continuous monitoring and automated alerts, we can automate an otherwise manual and tedious process such as data analysis and help you build intelligent frameworks from your most valuable resource, your data.

About Qualetics Data Machines

Qualetics Data Machines is a venture-funded startup established in July 2018, in the state of New Jersey in the USA. The company was founded and is led by the CEO, Sumanth Vakada, a software industry veteran with 20-years of experience in building modern software applications across various business domains. Qualetics currently has its core team based in the USA and an extended Application Development and Data Science team based in Hyderabad, India.

Our mission is to enable organizations to make an easier transition into the fields of Data Analytics and Artificial Intelligence. We aim to achieve this more efficiently than an organization might expect, were it to invest in the manpower and technology to build such complex platforms. For a universal need such as Data Science and AI, a dedicated focus is absolutely critical and Qualetics empowers our clients with the knowledge that allows them to apply in their core business offerings.

Key Features

Data Ingestion Platform

Visual Insights Delivery

License Pre-Existing Models

API based Intelligence Delivery

Integration with Apps

Develop Custom Models & Insights

For Inquiries, please contact:

Mike Fowler

Chief Commercial Officer

mike@qualetics.com

630.715.4540

www.qualetics.com