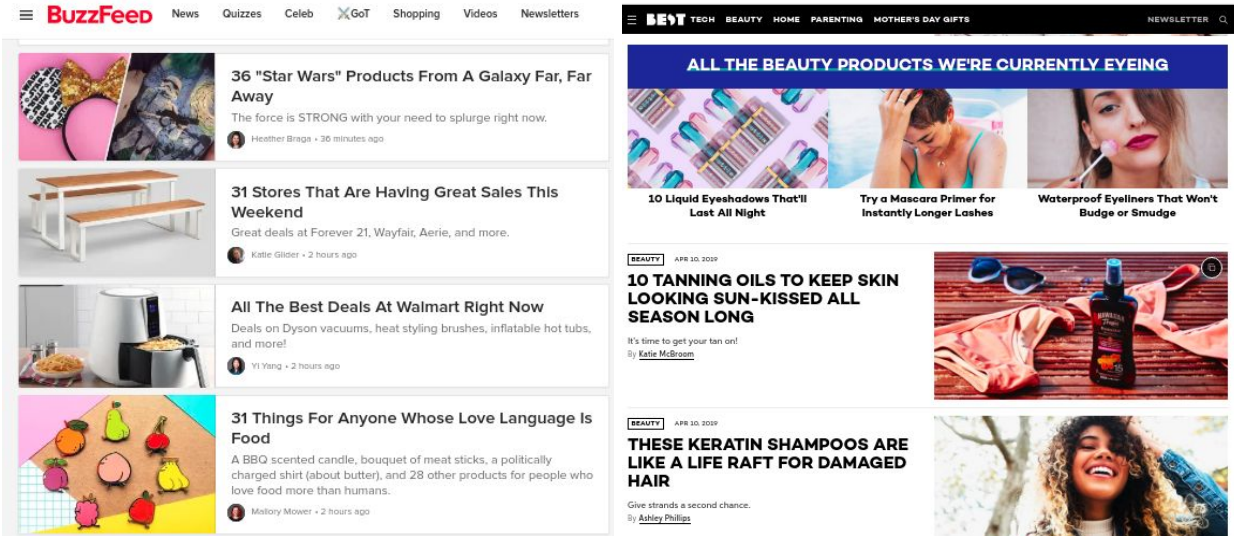
INTERNET MEDIA INFLUENCE

Aroun Amitabh Dalawat(301366807), Aisuluu Alymbekova(301378870), Shreejata Bhattacharjee(301376324)

Motivation and Background:

Internet media platforms have evolved from low-quality entertainment content to global media and tech companies, whose articles go viral and have great influence on people's opinions all around the world. Every company needs efficient marketing to thrive, grow and effectively communicate to their potential customers. With a rapid growth of e-commerce segment, the influence of internet media platforms can be leveraged as a strong marketing tool to promote goods. Hence, platforms such as Buzzfeed, BestProducts.com, etc. can be used for digital advertising in e-commerce. These are the websites you look to when you're trying to get information, opinions, even suggestions on the kind of products that we want to buy or should buy.

The project is focused on two things in particular. First, identification and evaluation of the impact of internet media platforms on e-commerce. Second, development of a tool that will automate the creation of articles for internet media platforms. So, for example, from the point of view of a Buzzfeed employee, the time and labor spent in manually searching for potential products to be featured in articles and writing descriptions individually for each of them will be reduced. Hence, we might say that the practical application of this project will be in the digital marketing sphere.



36 "Star Wars" Products From A Galaxy Far, Far Away

The force is STRONG with your need to splurge right now.

Heather Braga • 36 minutes ago

31 Stores That Are Having Great Sales This Weekend

Great deals at Forever 21, Wayfair, Aerie, and more.

Katie Glider • 2 hours ago

All The Best Deals At Walmart Right Now

Deals on Dyson vacuums, heat styling brushes, inflatable hot tubs, and more!

Yi Yang • 2 hours ago

31 Things For Anyone Whose Love Language Is Food

A BBQ scented candle, bouquet of meat sticks, a politically charged shirt (about butter), and 28 other products for people who love food more than humans.

Mallory Mower • 2 hours ago

ALL THE BEAUTY PRODUCTS WE'RE CURRENTLY EYEING

10 Liquid Eyeshadows That'll Last All Night

Try a Mascara Primer for Instantly Longer Lashes

Waterproof Eyeliners That Won't Budge or Smudge

BEAUTY APR 10, 2019

10 TANNING OILS TO KEEP SKIN LOOKING SUN-KISSED ALL SEASON LONG

It's time to get your tan on!

By Katie McBroom

BEAUTY APR 10, 2019

THESE KERATIN SHAMPOOS ARE LIKE A LIFE RAFT FOR DAMAGED HAIR

Give strands a second chance.

By Ashley Phillips

Examples of scraped websites

Problem Statement:

The questions that we want to answer through the completion of this project are:

- What influence do marketing strategies of internet media platforms have on e-commerce?
- Is there a correlation between the release of articles on platforms such as Buzzfeed, BestProducts, etc. and the increase of interest in that product and its sales?
- Sales of which category of products are affected most by internet media platforms?
- What products can potentially draw attention of internet users?
- Is it possible to automatize the process of article creation for Internet Media Platforms?
- Can indirect advertisement be impactful and be the future of marketing?

The challenge to answer stated questions is for the reason that it is difficult to find the direct correlation between sales of products and release of articles. The sales of the products can be affected by many factors and it may not be justified to claim that only the release of articles and listings are affecting the change in sales of the product. Moreover, Amazon does not release their sales data, we have to do some reverse engineering and consider the ratings or number of reviews, assuming that they signify the variation in sales of products.

Data Science Pipeline:

1. Data Collection:

First, we started with scraping the data. We scraped websites like Buzzfeed, BestProducts.com, which contained articles and listings of recommended products. We scraped the links of the articles at first and then for each link we scraped the names of the products featured in the article, the date of the article, the description of the product and the link of the product on Amazon.com. We scraped more than 400 articles and got the product details of more than 7000 products listed in those articles.

After that, we retrieved product ID's (ASIN), scraped corresponding categories and all ratings, reviews and reviews' dates from Amazon. We decided to use number of reviews on each day as an indicator of rise in interest of the product as Amazon does not provide their actual sales data. We also had to understand the structure of different websites for different scrapers to avoid discrepancies and errors, while scraping data for a long duration.

We also collected data from the amazon reviews dataset. For that, we created a table in AWS Athena with required partitions. Then we executed queries to load parquet files with amazon reviews from S3 to created table.

Challenges:

- Firstly, for web scraping, we had to understand the structure of web pages and check if there are any special cases that need to handled (infinite scrolling pages, sliders, differently structured DOM objects).

- Most websites block requests if they find that there are multiple concurrent requests from the same IP address.
- Moreover, the Amazon reviews dataset provides reviews only until 2015, so we had to scrape the data from Amazon website. However, Amazon has services that detect web crawlers and blocks all such requests.

To resolve these issues and scrape all article links from infinite scrolling pages, we used Selenium to automate the web browser interactions during the whole process of web scraping. While Selenium can mimic user oriented actions to trigger desired events, BeautifulSoup helps us to scrape specific structured information easily.

2. Data Cleaning:

To prepare the data we cleaned the scraped data to filter out products that are linked to different retail websites. Then we transformed reviews dates to relative days before and after the release of the article where the day of release of the article is the 0th day. Some articles had product links to websites other than Amazon. We removed these entries as we did not have the reviews data from shopping websites other than Amazon. We had to handle various exceptions and errors, where the fields were empty or the data format didn't match. We also obtained general categories of each product from Amazon Reviews Dataset.

Challenges:

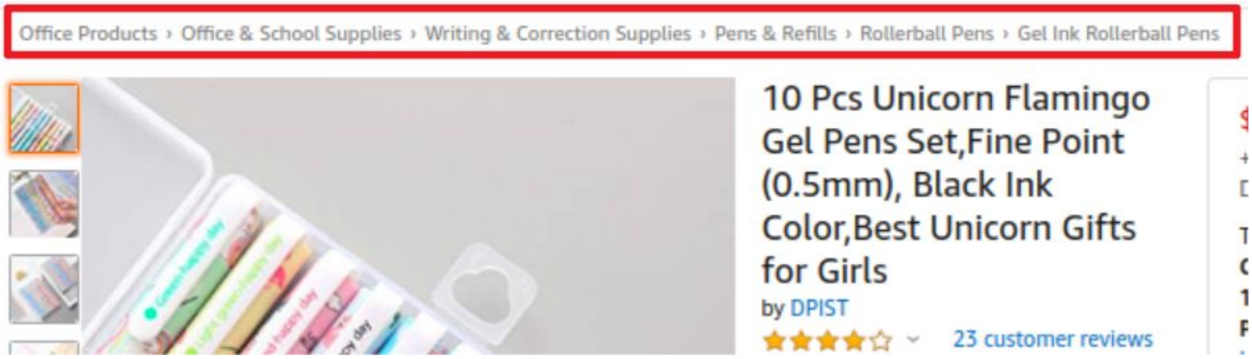
- We had to handle missing values in the dataset while cleaning the data. The missing values were MCAR(Missing Completely At Random), which means that the missingness is unrelated to any study variable and we can simply ignore the values in the further analysis as it will not affect the analysis in any way.

3. Data Integration:

We started with combining the amazon links and the descriptions of products from scraped articles with reviews on ASIN's which are IDs that uniquely identify each product. After that we combined the categories in scraped products with generalized categories from amazon reviews dataset.

Challenges:

- The main challenge here was that the data from the amazon reviews had general categories and the scraped data had specific categories of each product. To resolve this problem, we had to perform entity resolution on the scraped product information to generalize the categories for each product to pass into the machine learning model. For that, we calculated the Jaccard similarity coefficient and took the category with the highest similarity. In case of a tie, we took both categories, since 1 product can belong to several categories.



Office Products › Office & School Supplies › Writing & Correction Supplies › Pens & Refills › Rollerball Pens › Gel Ink Rollerball Pens

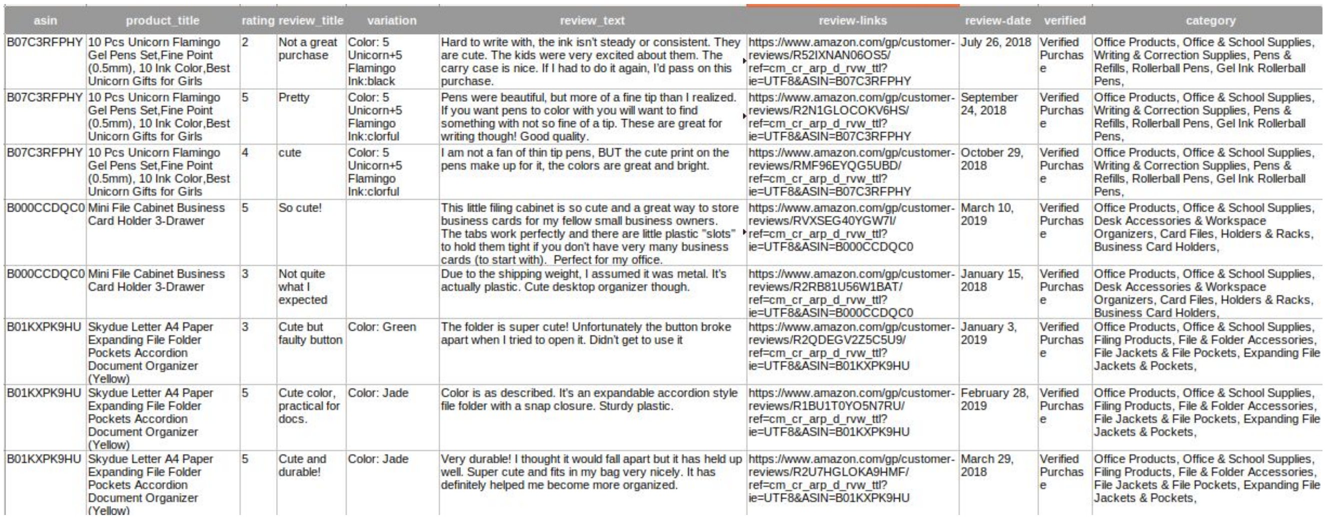
10 Pcs Unicorn Flamingo Gel Pens Set,Fine Point (0.5mm), Black Ink Color,Best Unicorn Gifts for Girls

by DPIST

23 customer reviews

Example of categories on Amazon.com

Finally, we got a clean dataset which contained Product IDs, Product Title, Rating, Review Title, Review Text, Review Date, Article Date and Relative Days.

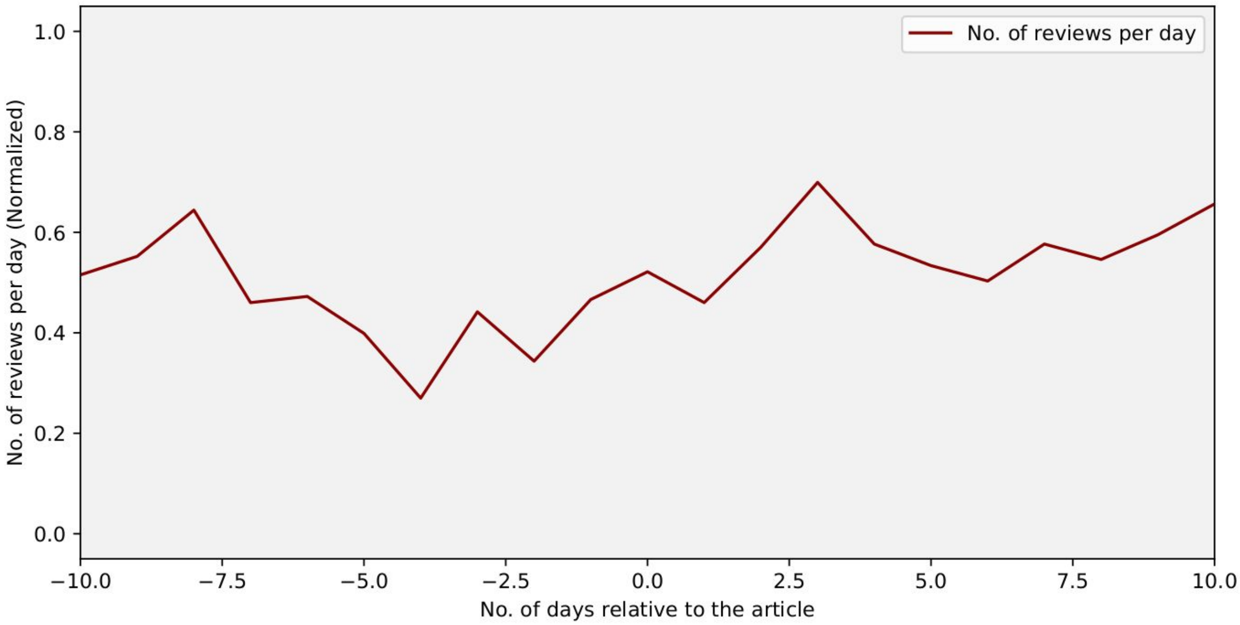


asin	product_title	rating	review_title	variation	review_text	review-links	review-date	verified	category
B07C3RFPHY	10 Pcs Unicorn Flamingo Gel Pens Set,Fine Point (0.5mm), 10 Ink Color,Best Unicorn Gifts for Girls	2	Not a great purchase	Color: 5 Unicorn+5 Flamingo Ink:black	Hard to write with, the ink isn't steady or consistent. They are cute. The kids were very excited about them. The carry case is nice. If I had to do it again, I'd pass on this purchase.	https://www.amazon.com/gp/customer-reviews/R52IXNAN06OS5/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B07C3RFPHY	July 26, 2018	Verified Purchase	Office Products, Office & School Supplies, Writing & Correction Supplies, Pens & Refills, Rollerball Pens, Gel Ink Rollerball Pens,
B07C3RFPHY	10 Pcs Unicorn Flamingo Gel Pens Set,Fine Point (0.5mm), 10 Ink Color,Best Unicorn Gifts for Girls	5	Pretty	Color: 5 Unicorn+5 Flamingo Ink:clorful	Pens were beautiful, but more of a fine tip than I realized. If you want pens to color with you will want to find something with not so fine of a tip. These are great for writing though! Good quality.	https://www.amazon.com/gp/customer-reviews/R2N1GLOCOKV6HS/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B07C3RFPHY	September 24, 2018	Verified Purchase	Office Products, Office & School Supplies, Writing & Correction Supplies, Pens & Refills, Rollerball Pens, Gel Ink Rollerball Pens,
B07C3RFPHY	10 Pcs Unicorn Flamingo Gel Pens Set,Fine Point (0.5mm), 10 Ink Color,Best Unicorn Gifts for Girls	4	cute	Color: 5 Unicorn+5 Flamingo Ink:clorful	I am not a fan of thin tip pens, BUT the cute print on the pens make up for it, the colors are great and bright.	https://www.amazon.com/gp/customer-reviews/RMF96EYQG5UBD/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B07C3RFPHY	October 29, 2018	Verified Purchase	Office Products, Office & School Supplies, Writing & Correction Supplies, Pens & Refills, Rollerball Pens, Gel Ink Rollerball Pens,
B000CCDQC0	Mini File Cabinet Business Card Holder 3-Drawer	5	So cute!		This little filing cabinet is so cute and a great way to store business cards for my fellow small business owners. The tabs work perfectly and there are little plastic "slots" to hold them tight if you don't have very many business cards (to start with). Perfect for my office.	https://www.amazon.com/gp/customer-reviews/RVXSEG40YGW7I/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B000CCDQC0	March 10, 2019	Verified Purchase	Office Products, Office & School Supplies, Desk Accessories & Workspace Organizers, Card Files, Holders & Racks, Business Card Holders,
B000CCDQC0	Mini File Cabinet Business Card Holder 3-Drawer	3	Not quite what I expected		Due to the shipping weight, I assumed it was metal. It's actually plastic. Cute desktop organizer though.	https://www.amazon.com/gp/customer-reviews/R2RB81U56W1BAT/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B000CCDQC0	January 15, 2018	Verified Purchase	Office Products, Office & School Supplies, Desk Accessories & Workspace Organizers, Card Files, Holders & Racks, Business Card Holders,
B01KXPK9HU	Skydue Letter A4 Paper Expanding File Folder Pockets Accordion Document Organizer (Yellow)	3	Cute but faulty button	Color: Green	The folder is super cute! Unfortunately the button broke apart when I tried to open it. Didn't get to use it	https://www.amazon.com/gp/customer-reviews/R2QDEGV2Z5C5U9/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B01KXPK9HU	January 3, 2019	Verified Purchase	Office Products, Office & School Supplies, Filing Products, File & Folder Accessories, File Jackets & File Pockets, Expanding File Jackets & Pockets,
B01KXPK9HU	Skydue Letter A4 Paper Expanding File Folder Pockets Accordion Document Organizer (Yellow)	5	Cute color, practical for docs.	Color: Jade	Color is as described. It's an expandable accordion style file folder with a snap closure. Sturdy plastic.	https://www.amazon.com/gp/customer-reviews/R1BU1T0YO5N7RU/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B01KXPK9HU	February 28, 2019	Verified Purchase	Office Products, Office & School Supplies, Filing Products, File & Folder Accessories, File Jackets & File Pockets, Expanding File Jackets & Pockets,
B01KXPK9HU	Skydue Letter A4 Paper Expanding File Folder Pockets Accordion Document Organizer (Yellow)	5	Cute and durable!	Color: Jade	Very durable! I thought it would fall apart but it has held up well. Super cute and fits in my bag very nicely. It has definitely helped me become more organized.	https://www.amazon.com/gp/customer-reviews/R2U7HGLOKA9HMF/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B01KXPK9HU	March 29, 2018	Verified Purchase	Office Products, Office & School Supplies, Filing Products, File & Folder Accessories, File Jackets & File Pockets, Expanding File Jackets & Pockets,

Example of scraped data

4. Data Analysis:

In order to analyse the influence of the articles on the increase in popularity of the product, we performed EDA to see different trends across categories, days of the reviews, etc. Since amazon does not release the actual sales data, we take into consideration the number of reviews on the product on each day before and after the release of the article as it is may be proportional to the sales of the product. After that we analyse the trend in reviews of product before and after the release of the article using line graph where we plot the scaled count of reviews per day against the days.

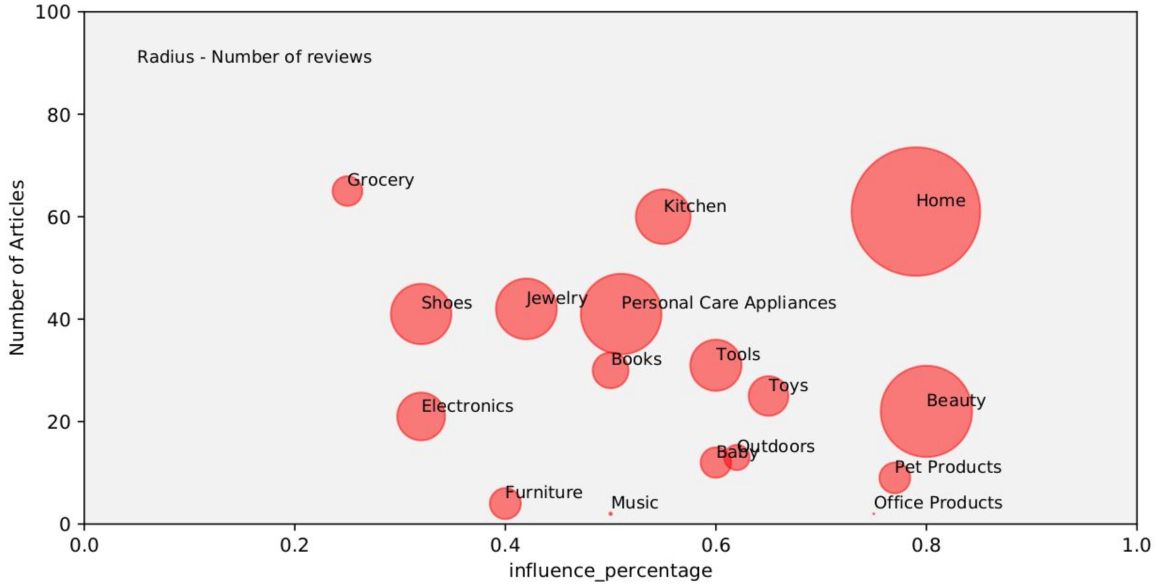


The change in number of reviews before and after the release of the article

Next we wanted to see how influential each product category was after the release of an article. So, for the task, we randomly sampled a number of articles and grouped the data by product and categories. We proceeded to find the percentage change in the following manner:

1. Removing data that appears 10 days before and after the release of the articles
2. Broke the data into before and after the release of the article
    a. From the first section, for each product, found the cumulative review count on first day and the last day (day -10 to 0)
    b. Similarly for the second section, for each product got the corresponding cumulative review count on the first day and last day (day 0 to 10)
3. For each product, found the percentage increase in the number of comments for both the sets of data. The Percentage increase was computed in the following manner:
    a. Increase = comment count on last day - comment count on first day
    b. % increase = Increase ÷ Original Number × 100
4. For each product, subtracted the percentage increase of the first set from the percentage increase from the second set.

Based on the above calculations, the following bubble chart was constructed. The radius of the circles indicates the number of reviews received on amazon.



Number of articles released under each category and their influence

The influence percentage and number of articles were plotted to get some insights about which category of articles have more influence on the sales. So, for example, the category ‘Home’ has the most influence and the number of articles in this category is also high. However, the number of articles in “Beauty” category is less, even though the influence is high. Also, number of articles in categories like “grocery” and “kitchen” are high but they do not have much influence on the sales of the products. So, this analysis is really helpful to determine which articles on which categories should be published more that can have a greater impact on the sales of the product in that category.

We also wanted to analyse if there was an actual correlation or statistical relationship between the release of the article and the popularity of the product. Hence, we performed correlation analysis and hypothesis testing. While correlation analysis determines if there is any statistical relationship between values of two attributes, hypothesis testing calculates the probability that a given hypothesis is true.

For hypothesis testing, the contingency table was constructed in the following manner:

1. A random number of articles were selected.
2. For each product, only the data for ten days before and after the release of the article was kept.
3. For each product, we found the number of reviews received on each day
4. For each product, the mean of the number of reviews was found
5. For each product, the mean of the number of reviews per day was compared to the number of reviews each day
    a. If the number of reviews on that day was less than the mean, it was given a 0
    b. If the number of reviews on that day was greater than the mean, it was given a 1

6. For each product, each day was given a 0 or 1 depending on whether it is before or after the release of the article
7. Now we have two columns. First column for days which tell us by 0 and 1 whether the article has been released or not on that day. Second column indicating whether the reviews received on that day were more or less than the average number of reviews received for that product.

Based on the two columns we can construct a contingency table with the null hypothesis that "there is no statistically significant difference between the number of reviews before and after the release of the article".

	Instances where numbers of reviews is less than average	**Instances where number of reviews is greater than average**
Before the release of the article	137	99
After release of article	129	139

Contingency table

The contingency table was created and the computed p-value was 0.03, which is greater than the picked threshold of 0.01.

This p-value is the probability that the variables are independent. The higher the value, the higher the chances of the variables being independent. We received a p-value of 0.03, which is slightly greater than 0.01, meaning that we cannot reject the null hypothesis that there is no statistical difference between the number of reviews before and after the release of the article, since we don't have a p-value which is less than the chosen threshold.

The threshold is picked to be 0.01, so that we can be totally sure that the chosen variable are dependent on each other. However, it is common practice for people to choose the threshold as 0.05, which would have been ideal for us, but since we were only analyzing the reviews per day and not the actual sales for each day before and after the release of the article, we had to be strict with our threshold and didn't want to p-hack.

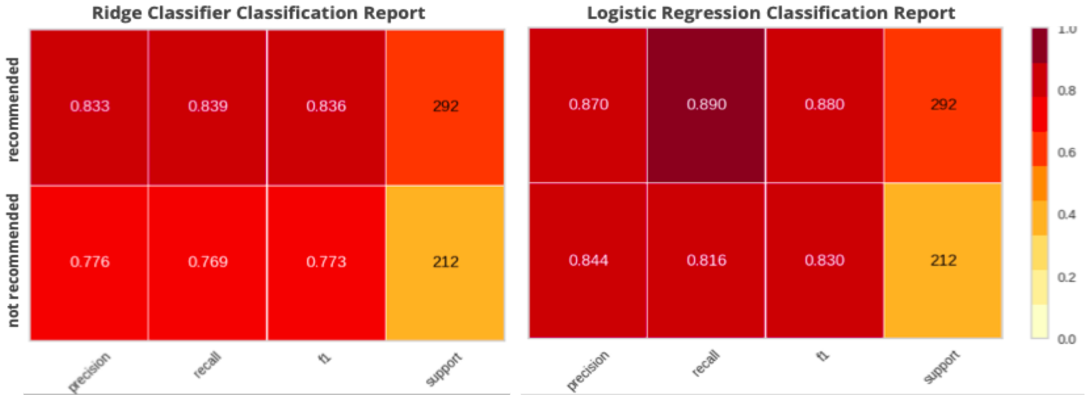
Challenges:

- The first challenge we faced while trying to analyze the data was of the difference in the number of reviews for each product. For example, if a product was popular among buyers, it would be getting about about 50 reviews per day compared to other products which are not that popular which might just be getting 5-10 reviews per day. But the problem arises when we trying to analyze the comment increase after the release of an article. If a product was getting about 2 reviews per day and then that changed to 10 reviews per day, it is a 400% increase and if a product was getting 50 reviews per day and then started getting 100 reviews per day, it's just a 100% increase. So while plotting the review increase, we had to find a way to put the review increase for each product in a way that the increase is represented in a fair manner and that the less popular products aren't under-represented. To solve this issue, we normalized the review count for each product so that all the counts lay between 0 and 1.

- The second challenge arose when we were trying to find the number of reviews received for each product 10 days before and after the release of the article on which it was mentioned. The problem was that some products had comments on day -9 or day -8 and so on. To solve the problem, for each product in each category, we had to find the earliest day in the range of -10 to 0 and the last day in the same range. Then for those days we found the cumulative review counts. The same procedure was performed for the ten days after the release of the article. A lot of computation was involved so that we obtained the correct influence of the articles in each category.

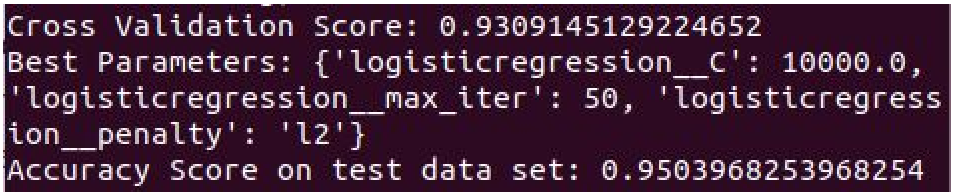
Machine Learning - Identifying potential products that can be recommended in future articles:

After that, to prepare the data, we labeled the scraped products from articles as '1' ("good recommended products" that is, the product can be featured in an article) and randomly selected products from amazon dataset and labeled them as "0" ("bad products" that is the product cannot be featured in an article). We created a pipeline with numerical and transformed categorical features. We split this data into training and testing datasets. First we got a baseline estimate of metrics for further optimization. The calculated baseline accuracy was 0.606, which served as a reference in further steps of model validation. Our aim was to choose a model that can efficiently classify the "good" and "bad" products with highest accuracy. Among the most influential features that make a product to be classified as "good" were how many reviews did the product received, what is the rating of these reviews, under what category is the product listed. After training and testing multiple classification models (Logistic Regression, SGD Classifier, Ridge Classifier, etc.) on the transformed data, Logistic Regression demonstrated the best performance in terms of accuracy, precision, etc. The color-coded heatmap classification reports gave us opportunity to easily compare the classification models, where models that are "redder" have stronger classification metrics. Reports for Ridge Classifier and Logistic Regression provided below.



Visualized classification reports

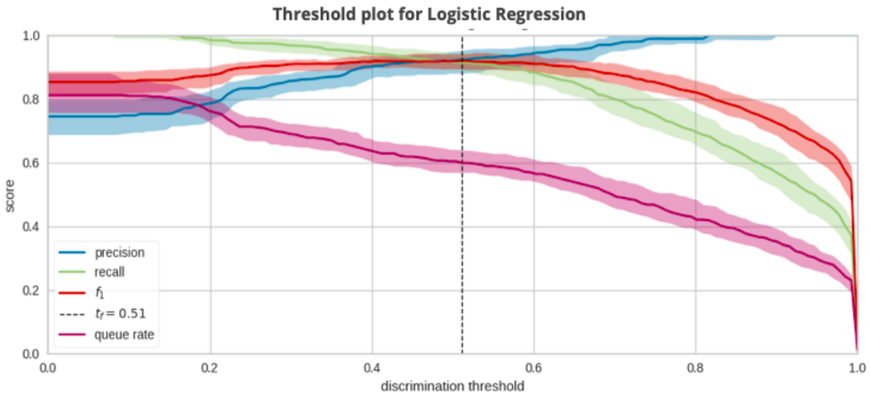
After choosing the model, we used GridSearchCV for hyper-parameter tuning, that performs search over specified parameter values for the chosen model. The scoring was based on the accuracy, since we wanted to observe how accurate the model is. The best parameters are as follows:



```
Cross Validation Score: 0.9309145129224652
Best Parameters: {'logisticregression__C': 10000.0,
'logisticregression__max_iter': 50, 'logisticregress
ion__penalty': 'l2'}
Accuracy Score on test data set: 0.9503968253968254
```

Hyper-parameters chosen with the use of GridSearchCV

Since we were performing a binary classification, we were able to plot a discrimination threshold to visualize the probability at which the positive class is chosen over the negative class to properly handle the sensitivity to false positives. The results of that graph showed us that the threshold of around 50% is the optimal one.



After obtaining the model that demonstrated the best performance, we were able to make predictions on newly obtained data. (i.e., classify the products into “good recommended” and “bad” products)

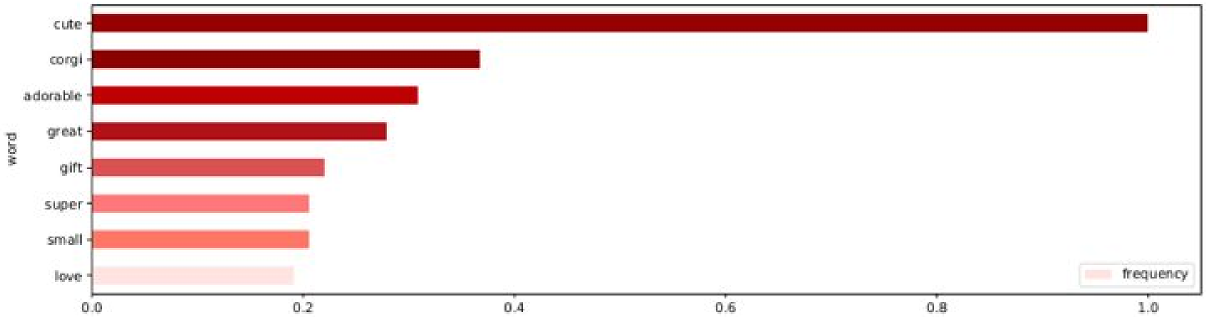
ASIN	Product Category	Count Reviews	Rating	Prediction	Not Featured Probability	Featured Probability
B0021XTP2S	Furniture	1	5	0	0.704	0.295
B007XY2JF0	MobileApps	40	3.375	0	0.787	0.212
B000JCGYD6	**Tools**	**236**	**4.38**	**1**	**0.412**	**0.587**
B007WHWMX	Home	28	2.427	0	0.855	0.144
B00GNVPJKG	Digital Music	1	5	0	0.704	0.295
B00C0NFMKI	**Home**	**260**	**4.346**	**1**	**0.381**	**0.618**
B000GF219M	Sports	2	4	0	0.782	0.217
B007FTHUM0	Toys	2	4	0	0.782	0.217

B00004R9RX	Tools	34	3.85	0	0.758	0.241
B00168ZIBQ	Music	13	3.30	0	0.818	0.181

Predictions made by the model

NLP - auto-generated articles

Then we moved on to creating descriptions of products that could potentially be listed in articles, that is the “good” products by restructuring the comments in reviews data. From the combined dataset with all reviews, for each product that was picked by the model, we analyze the comments to generate a summary which would then go on to the article as the descriptions of potentially featured products. For each product all the reviews were obtained, which were then broken down into sentences and separately the frequencies of each of the word were calculated. Then these frequencies were put back to the corresponding words and sentences with the highest frequency were used to make a description of the product.



Top 10 frequent words for a product

5. Data Product:

All the three modules of our project were combined to get the final data product. That is, the graphs that were plotted demonstrate whether there was any impact on the sales of products, validation curves, precision vs. recall curves, etc. along with the created set of products that can be potentially featured in articles of recommended goods and their descriptions generated. Finally, an HTML webpage that contains the list of selected products, generated product titles and descriptions for each product, scraped product image was constructed.

For creating descriptions of products that could potentially be listed in articles, that is the “good” products, comments were restructured in reviews data. From the combined dataset with all reviews, for each product that was picked by the model, we analyze the comments to generate a summary which would then go on to the article as the descriptions of potentially featured products.

The description of each product was generated in the following manner:

1. For each product all the reviews were obtained,
2. The reviews were broken down into sentences
3. Separately the frequency of each of the word was calculated.

4. The maximum frequency word was found
5. The frequency of each word was divided by the maximum frequency to get the weighted frequencies
6. Then these frequencies were put back to the corresponding words
7. The top four sentences with the highest frequency score were used to make a description of the product.

An example of how the above process works is as follows. Suppose we have the following text:
"keep working. keep striving. never give. fall seven time get eight. ease greater threat progress hardship. ease greater threat progress hardship. keep moving keep growing keep learning. see work."

We can break the above text into sentences and then find the frequencies of each word.
Keep appear 5 times, so the frequency of every other word would be divided by Keep's frequency. So after putting back the frequencies on the sentences we get:
"So, keep moving, keep growing, keep learning" equates to 1 + 0.20 + 1 + 0.20 + 1 + 0.20 = 3.60

Using this, we can take the sentences with the highest score to form a description of the product.

<u>Methodology</u>:

- The Amazon comments database was loaded onto the AWS S3. We used S3 as it is fully scalable, fast, and reliable. Since, the Amazon dataset provides reviews only until 2015, had to scrape data from Amazon website.
- Initially, we started web scraping using BeautifulSoup. However, we realised that we need an automation tool for scraping data from Amazon website as Amazon blocks all scraper requests and to scrape infinite scrolling pages on internet media websites. Thus, we used Selenium with BeautifulSoup to efficiently web scrape the data.
- We used AWS Athena query service to retrieve data from the Amazon reviews dataset which was stored as an object on AWS S3, as it enables us to run ad-hoc queries using ANSI SQL, without the need to aggregate or load the data to Athena. It can process unstructured, semi-structured and structured data sets. To query data through python script, we used PyAthenaJDBC, which serves as a wrapper for Amazon Athena JDBC driver.
- To perform entity resolution, we used Jaccard similarity measure as we were comparing binary vectors obtained by vectorizing the categories data.
- For analysis and predicting if a product can be included in a list of recommended items in an article, we have used Scikit-learn library which provides easy to use machine learning toolkit (Pipelines, Column Transformers, Simple Imputers, One Hot Encoders and various classification models), pandas and numpy. Yellowbrick suite, which extends the Scikit-Learn API, was used for visualizations to assist with model selection and optimization.

- In order to generate descriptions of articles from reviews, we have used the Natural Language Toolkit which provides a suite of programs and libraries (sent_tokenize, word_tokenize, etc.) which were used to break down the reviews and form a description of them.

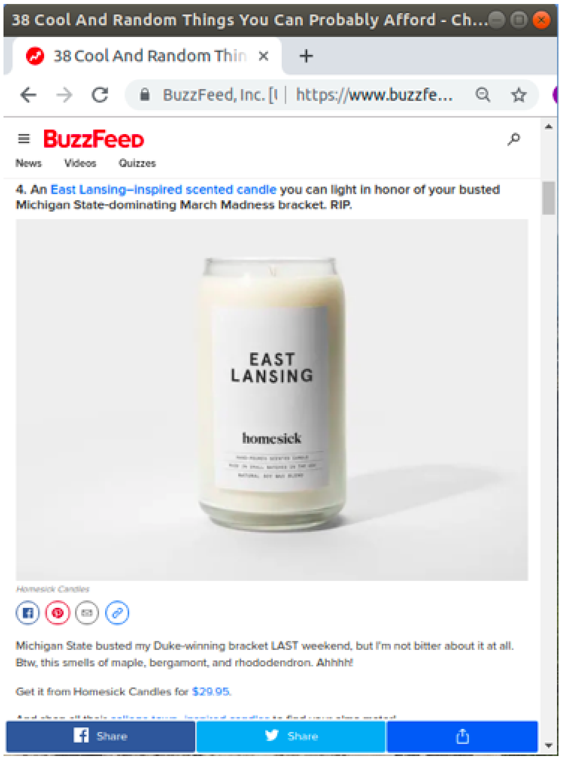
Evaluation:

The developed solution can be used as a beneficial tool in digital advertising to reduce the labor cost and make marketing strategies more efficient by automatizing the redundant work that internet media platform employees are doing every day.

Data Product:

The final data product is a tool that automatically creates marketing articles with product listings and descriptions for them. So it simplifies the task of choosing the potential products that can be listed in an article as well as constructing a description for them.

Example of original article

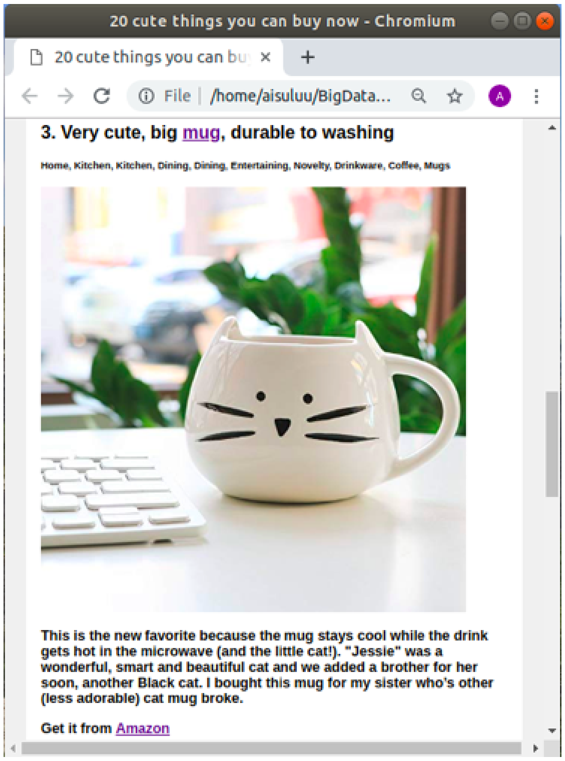

38 Cool And Random Things You Can Probably Afford - Ch...
38 Cool And Random Thin
BuzzFeed, Inc. [I | https://www.buzzfe...
BuzzFeed
News
Videos
Quizzes
EAST LANSING
homesick
Homesick Candles
Share
Share

4. An East Lansing–inspired scented candle you can light in honor of your busted Michigan State-dominating March Madness bracket. RIP.

Michigan State busted my Duke-winning bracket LAST weekend, but I'm not bitter about it at all. Btw, this smells of maple, bergamont, and rhododendron. Ahhhh!

Get it from Homesick Candles for $29.95.

Auto-generated article with recommended products and their descriptions


20 cute things you can buy now - Chromium
20 cute things you can bu
File | /home/aisuluu/BigData...
A

3. Very cute, big mug, durable to washing

Home, Kitchen, Kitchen, Dining, Dining, Entertaining, Novelty, Drinkware, Coffee, Mugs

This is the new favorite because the mug stays cool while the drink gets hot in the microwave (and the little cat!). "Jessie" was a wonderful, smart and beautiful cat and we added a brother for her soon, another Black cat. I bought this mug for my sister who's other (less adorable) cat mug broke.

Get it from Amazon

Lessons Learnt:

The key lessons we have learned are that any Data Science project requires effective planning and requirement analysis, through the completion of the project, it is important to keep in mind the initial questions we want to answer and the problem statement.

We learned that real-world data is never as clean and structured, as the datasets we receive in class, For data collection, if we are using web scraping, we must understand the structure of the web pages first. We must also make sure that the websites do not block our requests. Large websites deploy services that can detect crawling on the site and if we send concurrent requests from same host, they will classify it as Denial of Service attack on their website and block future requests. To deal with this problem, we can either chain the requests to make them more human-like or use Selenium to automate the process of scraping.

Once we have the data, it is important to deal with missing values, anomalies and ambiguous, sparse data. To disambiguate real world entities in various records, we can use entity resolution.

Along with that, we learned that it is better to optimize the query before running it on a big dataset, as it not only may take a long time to return the results, but also can lead to charges for big queries (on AWS Athena).

Summary:

The main aim of the project was to determine how we can effectively use recommended list of products on articles on internet media platforms for digital marketing.
The results of completed project can find multiple applications:

— Businesses and sellers will find the results from data analytics stage useful to plan out their marketing strategies based on the observations that were made throughout the data analysis and make data-driven decisions to identify which marketing approach will give them the highest Return of Investment (ROI);

— Internet media platforms can exploit the developed tool model to automate the process of picking products that can potentially be featured in future articles. Along with that, the NLP tool will also construct the article itself, which will reduce the labor cost that is spent on the creation of such articles.