Internet Media Influence

Aroun Amitabh Dalawat, Aisuluu Alymbekova, Shreejata Bhattacharjee

SFL

Motivation and application

Project motivation

With a rapid growth of e-commerce segment, the influence of internet media platforms can be used as a strong marketing tool to promote goods. 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 automatize the creation of articles for internet media platforms.

Real-world application scenario

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 automatize 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.

Work process and challenges

Data Collection

Scraping the data

Scrape links to articles with recommended products.

From each article link scrape the date of article, products, their description and amazon links.

From the retrieved product id's, scrape corresponding categories and all ratings, reviews and comment dates from Amazon.

Getting data from reviews dataset

Create a table in AWS Athena with required partitions.

Execute queries to load parquet files with amazon reviews from S3 to created table.

Data Cleaning

Preparing the data

Clean the scraped data to filter out products that are linked to different retail websites.

Transform reviews dates to relative days before and after the release of the article.

Form a new dataset with Product IDs, Product Title, Rating, Review Title, Review Text, Review Date, Article Date and Relative Days.

Obtain general categories from Amazon Reviews
Dataset.

Data Integration

Combining datasets

Combine amazon links and the descriptions of products from scraped articles with reviews on ASIN's.

Perform entity resolution on the scraped product information with general categories to get a more general category for each product to pass into the machine learning model.

NLP

From the combined dataset with all reviews, for each product that was picked by the model, analyze the comments to generate a summary which would then go on to the article as the descriptions of potentially featured products.

Data Analysis

Plotting graphs

Perform EDA to see different trends across categories, days of the reviews, etc. about the influence of the article on the actual sales of the products.

ML

As data preparation step, labe scraped products from articles as '1' ("good" products).
Randomly select products from amazon dataset and label them as "0". Split into training and testing data sets.

Transform categorical features using One-Hot Encoding approach.

Create Visual Pipeline with numerical and transformed categorical features.

Train and test multiple classification models (Logistic Regression, SGD Classifier, etc.) on the transformed data.

Data Product

Combining results

Plot graphs that demonstrate whether there was any impact on the sales of products.

Visualize learning, validation curves, precision vs. recall curves, etc. using Yellowbrick

Create a set of products that can be potentially featured in articles of recommended goods.

Automatically generate
HTML webpage that
contains the list of selected
products, generated
product titles and
descriptions for each
product, scraped product
image.

Challenges

Loading the comments database onto the AWS S3 and then finding out a way to query it

Amazon dataset provides reviews only until 2015, had to scrape data from amazon website.

Amazon website blocks all scrapers – need to access through the use of selenium.

Understanding structure of different websites for different scrapers to avoid discrepancies while scraping data for a long duration.

Generalizing the categories of the scraped data by performing ER with the categories in the amazon dataset.

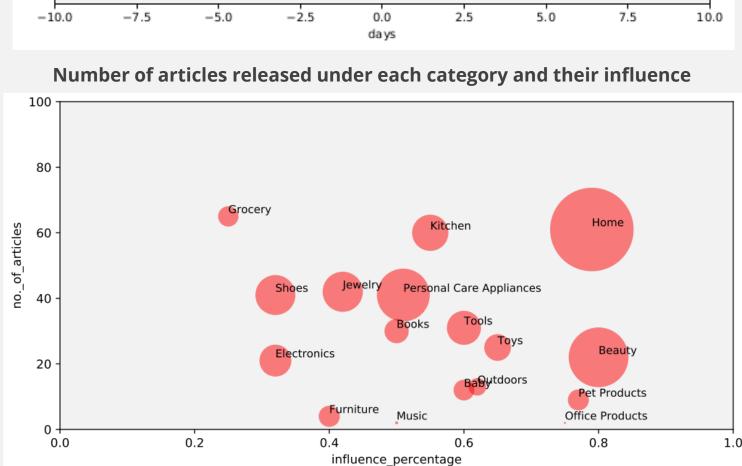
Generating keys for each product so that all team members are able to access the data properly

Choosing an efficient model that predicts products with the highest accuracy.

Restructuring the comments data and then using NLP to obtain a description of the product being put on the article

Experiments and results

The change in number of reviews before and after the release of the article 1.0 0.8 0.6 0.2 0.0 -10.0 -7.5 -5.0 -2.5 0.0 days



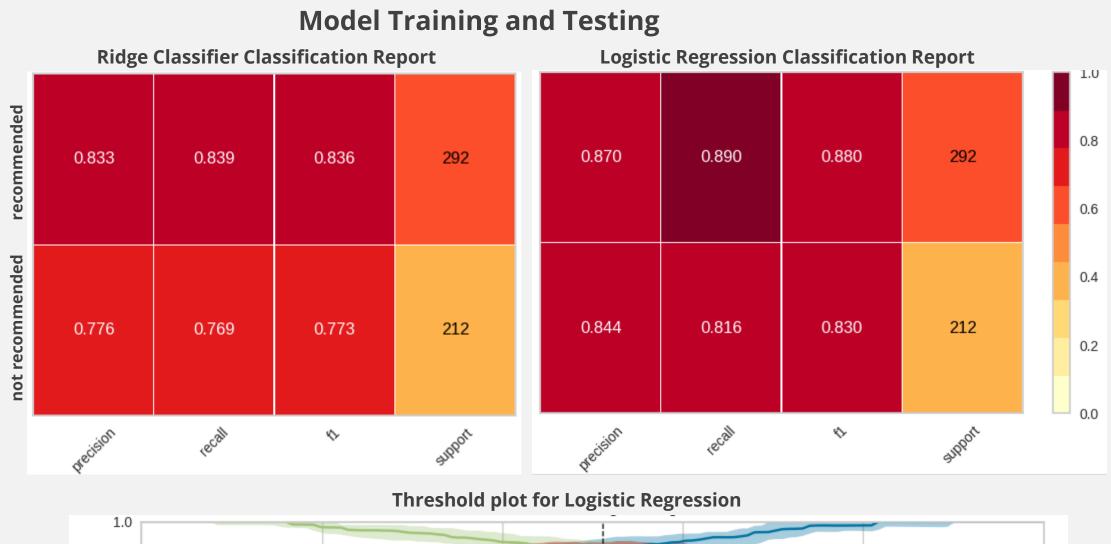
Correlation analysis and hypothesis testing

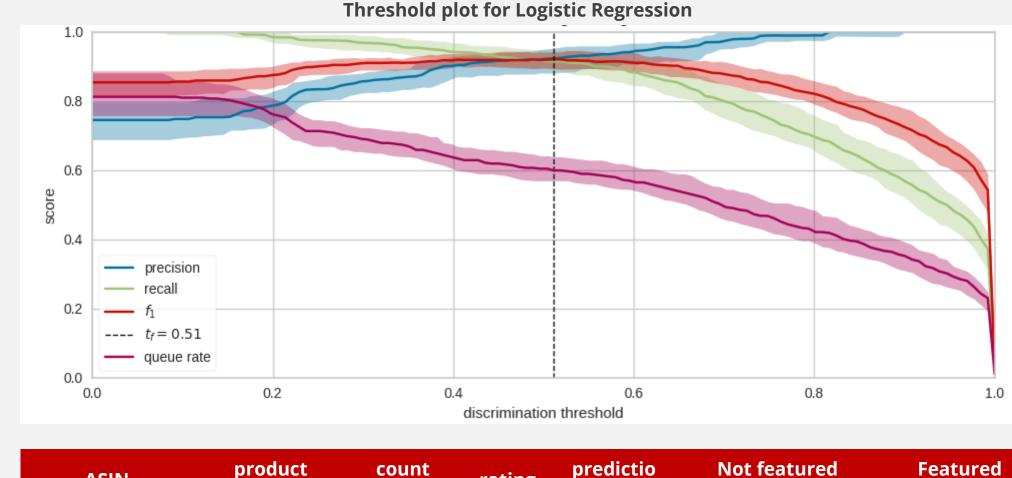
H0: no statistically significant difference between the number of reviews before and after the release of the article

H1: there is a 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 the release of the article	129	139					

The calculated p-value - 0.03269, which is higher than the picked threshold of



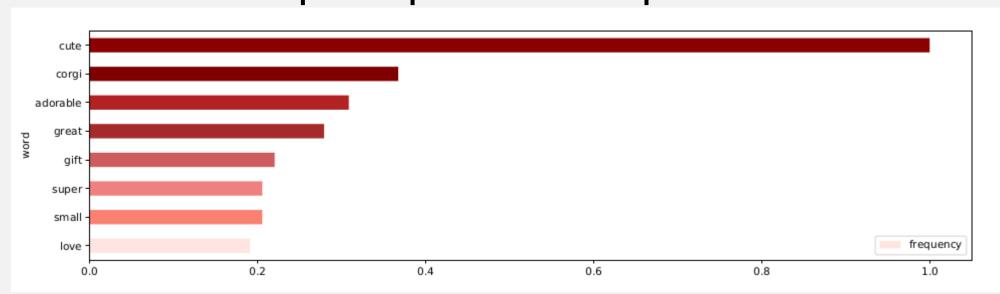


ASIN	product category	count reviews	rating	predictio n	Not featured probability	Featured probability
B0021XTP2S	Furniture	1	5	0	0.704	0.295
B007XY2JF0	MobileApps	40	3.375	0	0.787	0.212
B000JCGYD6	Tools	236	4.38	1	0.412	0.587
B007WHWMX	Home	28	2.427	0	0.855	0.144
B00GNVPJKG	Home	1	5	0	0.704	0.295
B00C0NFMKI	Digital Music	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

NLP approach

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, which is what you see in the bar chart. 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 1 product



Example of original article

Auto-generated article with recommended products and their descriptions

