MACHINE LEARNING APPLIED: RATING PREDICTION BASED ON USER REVIEWS.

Submitted by:

S. Dilip Kumar

ACKNOWLEDGMENT

With brands vying to win customers over with aggressive marketing, online reviews have become an important consideration for consumers. Over 88% of Ecommerce shoppers buy only after reading reviews of products. Product ratings and reviews secure the customer journey by giving them the confidence to complete the checkout process.

User-generated content provides valuable consumer insights. It helps Ecommerce retailers understand the needs of consumers. Brands can use this information to ideate, improve and innovate products or services, foster customer loyalty, and improve their business.

User-generated content (UGC) is any form of content created voluntarily by users. This includes reviews, images, videos, comments, and questions. Creating platforms for users to discuss the products and services that improve the relationship between a brand and its consumers.

A Nielsen study undertaken to determine the Consumer Trust Index found that 92% of shoppers trust organic user-generated content more than traditional advertising. By outsourcing content creation to its consumers, brands can save time and cost and increase brand credibility. UGC, when leveraged smartly can be an asset to content marketers. At the outset, it might help to understand the motivation for people to create and share content for free.

- Consumers like to share their opinions on products or services that they've bought. The internet provides a platform for them to express themselves and feel heard.
- Even though consumers don't get paid for user-generated content, they enjoy other incentives like social recognition and appreciation.
- Some brands invite users to try their products and share their reviews with a specific hashtag and photo. In contests like these, a winner is chosen and gets gifts or coupons.
- In general, people consume a lot of content before buying. But user reviews are not just for consumers. Let's see the benefits brands can reap from User-Generated Content.

Shoppers read at least 10 online customer reviews before they fee

(https://targetbay.com/blog/customer-reviews/)

INTRODUCTION

Business Problem Framing

When you go to make an online purchase, what's the first thing you do? In an ecommerce-driven world where customers can't physically experience products before purchasing, many consumers turn to online product reviews.

As online review sites such as Yelp! and Facebook have expanded, finding an opinion on just about anything is only a few clicks away. The proliferation of reviews has even gone so far as to shape how businesses are perceived online.

As Chris Anderson, businessman and current head of TED, puts it, "Your brand isn't what you say it is — it's what Google says it is."

For any company that exists in the digital space, online reviews are critically important when it comes to winning business and maintaining a positive reputation.

Conceptual Background of the Domain Problem

Who is Reading Online Reviews?

In today's web-based world, virtually everyone is reading online reviews. In fact, 91% of people read them and 84% trust them as much as they would a personal recommendation. The effects of reviews are measurable, too.

The average customer is willing to spend 31% more on a retailer that has excellent reviews.

Negative reviews can carry as much weight as positive ones. One study found that 82% of those who read online reviews specifically seek out negative reviews.

That may sound alarming — this stat only emphasizes that negative reviews aren't going unnoticed — but there are some benefits: Research indicates that users spend five times as long on sites when interacting with negative reviews, with an 85% increase in conversion rate.

Customers like to see lots of reviews. A single review with a few positive words makes up an opinion, but a few dozen that say the same thing make a consensus.

The more reviews, the better, and one study found that consumers want to see at least 40 reviews to justify trusting an average star rating. However, a few reviews are still better than no reviews.

One study found that, on average, products are 270% more likely to sell with as few as five reviews.

With the vast array of review sites and the level of trust most consumers have in reviews, it's a safe assumption that virtually everyone considering your products, no matter your target demographic, industry, or market, is reading online reviews before making a purchase.

Review of Literature

Significance Of Reviews

Reviews are able to garner trust because they represent personal and unbiased users' experiences. Written by customers, they constitute unique content that search engines love. Naturally, a good number of Customer reviews will drive higher organic traffic to Ecommerce websites. There is a great demand for original content that equally impresses the customers and search engines. User-generated content is authentic and distinctive. Since it is created by customers, brands need not use much of their resources to build it. User-Generated Content is the solution to the ongoing search for original content. It is a steady stream of quality, searchable content for a brand's website. User-Generated Content also helps brands learn about the latest trends and the preferences of their customers.

We've analyzed and studied 500+ eCommerce stores, the majority of them were Shopify stores. However, given the huge amount of data we gathered, it should come as no surprise that the Shopify product reviews app improves trust in a brand, boosts conversion rates, and also gets more traffic from search engines. Since most customer journeys start on search engines, brands must leverage User Generated Content (their strongest tool for marketing) to power SEO and create social proof. Since most customer journeys start on search engines, brands must leverage User Generated Content (their strongest tool for marketing) to power SEO and create social proof.

In addition to how it influences SEO, customer reviews have an effect on driving real-world shoppers to trust-based purchase decisions. The Independent SaaS

and SMB industry research studied factors that improve trust in website visitors, and here's what they found:

How Does the Review Collection Process Work?

How User-Generated Content Can Boost SEO for Your Ecommerce Store?

Search Engines love fresh content. At the same time, the content ideation and creation process take time and resources. It is not easy to continuously create new content for SEO. There is a great demand for original content that equally impresses the customers and search engines. User-Generated Content is authentic and distinctive. Since it is created by customers, brands need not use much of their resources to build it. User-Generated Content is the solution to the ongoing search for original content. It is a steady stream of quality, searchable content for a brand's website. User-Generated Content also helps brands learn about the latest trends and the preferences of their customers.

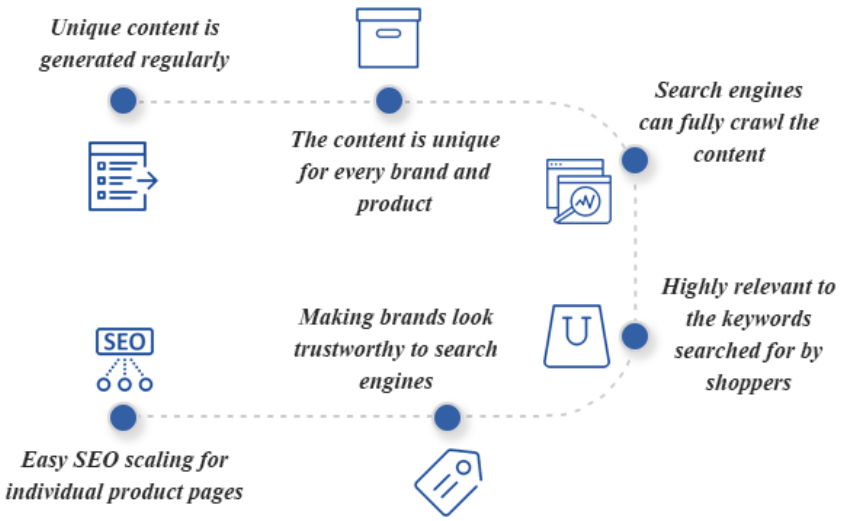
User-Generated Content plays a valuable role in SEO and marketing strategies. Here's how it works:

Organically Strengthens SEO

The search engine optimization for Ecommerce websites involves several attributes like keywords, internal links, and backlinks.

When customers write reviews, they often use phrases that are closely associated with a product or service. Appropriate keywords and links are included naturally in customer reviews. This is the best way to shape and strengthen SEO.

From the organic search standpoint, customer reviews provide several key benefits for Ecommerce retailers:



- Unique content is generated regularly
- The content is unique for every brand and product
- Search engines can fully crawl the content
- Highly relevant to the keywords searched for by shoppers
- Making brands look trustworthy to search engines
- Easy SEO scaling for individual product pages

Thus, the organic traffic to the website improves over time resulting in increased conversions and order value. Brands can observe a significant improvement to their bottom line while also satiating the demand for genuine product reviews.

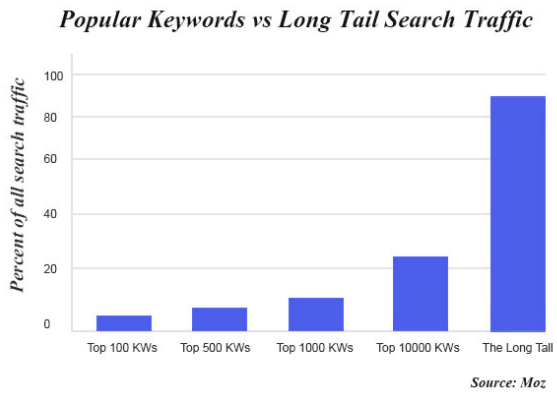
Motivation for the Problem Undertaken

Helps Ecommerce Websites Rank For Long-Tail Keywords

Let us imagine that you've returned from a spectacular vacation in Spain and have a sudden craving for tapas. What would you search for, 'restaurants', 'Spanish restaurants near me' or 'best Spanish restaurant near me'? If you pick the last option, you belong to the majority of people who use more than four words in their search query. These are called long-tail keywords.

Long-tail keywords are phrases that have low search competition and a high potential for conversion. Typically, a search query that is longer than four words produces traffic through long-tail keywords. Such phrasal search queries are frequently used by shoppers but brands are largely unaware of them and end up omitting long-tail keywords in consumer marketing.

As keywords become more specific, the search volume becomes less competitive and the customer intent is much higher. Long-tail keywords are very specific and can help Ecommerce websites rank higher in search results. They can be used to target specific demographics and produce excellent short and long-term results by attracting qualified shoppers who are highly likely to convert.



Analytical Problem Framing

Data Sources and their formats.

We have a client who has a website where people write different reviews for technical products. Now they are adding a new feature to their website i.e. The reviewer will have to add stars(rating) as well with the review. The rating is out 5 stars and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Now they want to predict ratings for the reviews which were written in the past and they don't have a rating. So, we have to build an application which can predict the rating by seeing the review.

On seeing the above problem definition, I have tried in building a successful machine learning model that will predict the rating based on the reviews given by the customer. But to train the model we require some data to play around. So, I have scraped some data from multiple ecommerce websites along with the ratting for the reviews. Web scraping or data collection is done with the help of Beautiful Soup and Selenium, and I have collected more than 50000 records with ratings and reviews and have stored in CSV format. There are some records from the data which has regional language, on using NLP technique I have tried in cleaning all the collected data, the process of data cleansing and pre processing are as follows.

Data Pre-processing Done.

Using NLTK library have followed multiple technique for data pre-processing.

Data Cleansing.

In [6]:
```
# Convert all messages to lower case
df['Review'] = df['Review'].str.lower()
```

In [7]:
```
# Replace email addresses with 'email'
df['Review'] = df['Review'].str.replace(r'^.+@[^\.].*\.[a-z]{2,}$','emailaddress')
```

In [8]:
```
# Replace URLs with 'webaddress'
df['Review'] = df['Review'].str.replace(r'^http\://[a-zA-Z0-9\-\.]+\.[a-zA-Z]{2,3}(/\S*)?$','webaddress')
```

In [9]:
```
# Replace money symbols with 'moneysymb' (£ can by typed with ALT key + 156)
df['Review'] = df['Review'].str.replace(r'£|\$', 'dollers')
```

In [10]:
```
# Replace 10 digit phone numbers (formats include paranthesis, spaces, no spaces, dashes) with 'phonenumber'
df['Review'] = df['Review'].str.replace(r'^\(?[\d]{3}\)?[\s-]?[\d]{3}[\s-]?[\d]{4}$','phonenumber')
```

In [11]:
```
# Replace numbers with 'numbr'
df['Review'] = df['Review'].str.replace(r'\d+(\.\d+)?', 'numbr')
```

As live we see above, I have cleaned the data in following these steps.

- Converting all the letters into small case.
- Have converted all the email address present in the reviews as “email address”.
- Have converted all the web address present in the reviews as “Web address”.
- Have converted all the currency present in the reviews as “dollars”.
- Have converted all phone numbers present in the reviews as “phonenumber”.
- Have converted all the number present in the reviews as “number”.

In [12]:
```
df['Review']= df['Review'].astype(str)
```

In [13]:
```
df['Review'] = df['Review'].apply(lambda x: ' '.join(term for term in x.split() if term not in string.punctuation))
```

In [14]:
```
stop_words = set(stopwords.words('english') + ['u', 'ü', 'ur', '4', '2', 'im', 'dont', 'doin', 'ure'])
df['Review'] = df['Review'].apply(lambda x: ' '.join(
    term for term in x.split() if term not in stop_words))
```

In [15]:
```
lem=WordNetLemmatizer()
df['Review'] = df['Review'].apply(lambda x: ' '.join(lem.lemmatize(t) for t in x.split()))
```

Post following the above method I have also removed all punctuations and stop words and lemawords from the reviews.

In [16]:
```
df['clean_length'] = df.Review.str.len() # checking the length of the words post cleaning.
df.head()
```

Out[16]:

	Rattings	Review	length	clean_length
0	4	received yesterday (numbr/numbr/numbr). prompt...	1661.0	1309
1	1	different charger sent box amazon ready provid...	349.0	235
2	5	amazing laptop.. ordered laptop released date ...	387.0	305
3	4	soon found numbrgen gen inumbr ~numbrk got stu...	2470.0	1864
4	5	bought day ago, glad decided make purchase. la...	2465.0	1810

In [17]:
```
print ('Origian Length', df.length.sum())
print ('Clean Length', df.clean_length.sum())
```

```
Origian Length 18423780.0
Clean Length 13286613
```

From above we can understand post cleansing the data how the length of the words is reduced.

Preprocessing

In [19]:
```
from sklearn.feature_extraction.text import CountVectorizer

cv = CountVectorizer(binary = True)
cv.fit(Text)
x = cv.transform(Text)
```

In [20]:
```
x
```

Out[20]:
```
<50593x35810 sparse matrix of type '<class 'numpy.int64'>'
        with 1545906 stored elements in Compressed Sparse Row format>
```

In [21]:
```
y = labels
```

Finally, before feeding the data into the model, I have stored the data as count vectorizer to make the training process easy, and have stored Ratings data in y variable.

Hardware and Software Requirements and Tools Used

Tools Used. (Hardware and Software Requirements):

1. Python 3.8.
2. NumPy.
3. Pandas.
4. Matplotlib.
5. Seaborn.
6. Data science.
7. SciPy
8. Sklearn.
9. NLTK library.
10. Machine learning.
11. CPU with RAM of 8GB.
12. Anaconda Environment.
13. Jupyter Notebook.
14. re (Regular expression).

Model/s Development and Evaluation

Selecting best random state parameters for training.

With logistic regression I have created a loop between 0-500 to sort the best random state parameters with highest accuracy.

Selecting parameters for training

[22]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

accu = 0
for i in range(0,500):
    x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = .25, random_state = i)
    mod = LogisticRegression()
    mod.fit(x_train,y_train)
    y_pred = mod.predict(x_test)
    acc = accuracy_score(y_test,y_pred)
    if acc> accu:
        accu= acc
        best_rstate=i

print(f"Best Accuracy {accu*100} found on randomstate {best_rstate}")
```

```
Best Accuracy 74.6699343821646 found on randomstate 429
```

[23]:
```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = .25, random_state = best_rstate)
```

As like we see above the highest accuracy occurs in random state of 429 using the same am splitting the data as train and test.

Testing of Identified Approaches (Algorithms)

I have used eight different algorithms and tried in sorting the top performed algorithm they are “Logistic Regression", "Naive Bayes Gaussian", "Random Forest", "Decision Tree", "Extra Tree", "Ada Boost", "Gradient Boosting", "XGBoost”.

	Model	accuracy	precision	recall	f1
2	Random Forest	0.750448	0.782204	0.750448	0.723917
0	LogisticRegression	0.721247	0.704191	0.721247	0.705472
3	Decision Tree	0.687566	0.679076	0.687566	0.681945
4	Extra Tree	0.657205	0.649585	0.657205	0.652436
7	XGBoost	0.692521	0.695257	0.692521	0.650717
6	Gradient Boosting	0.621943	0.641040	0.621943	0.549488
5	Ada Boost	0.579380	0.504720	0.579380	0.492523
1	Naive Bayes Gaussian	NaN	NaN	NaN	NaN

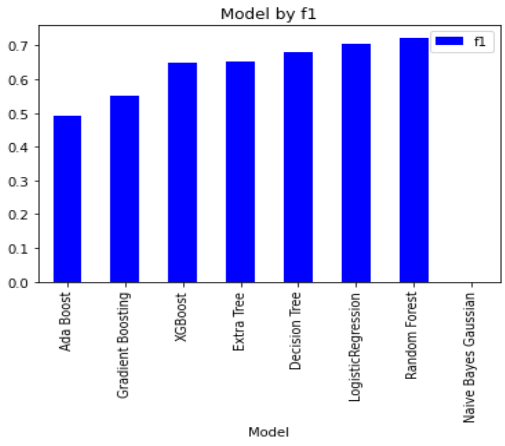
From above table its evident that Random Forest Algorithm tops the chart with the metrics are as follows

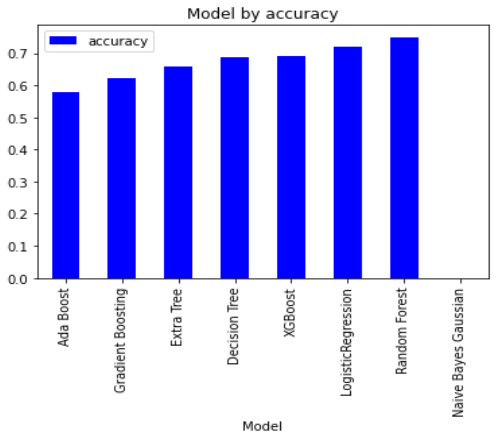
Random Forest

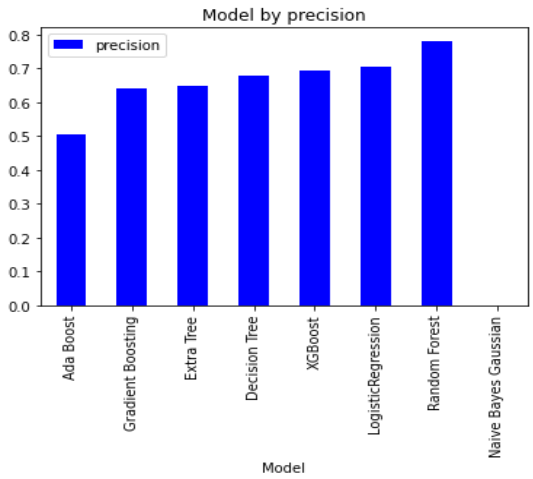
1. accuracy = 0.750448
2. precision = 0.782204
3. recall   = 0.750448
4. f1   = 0.723917

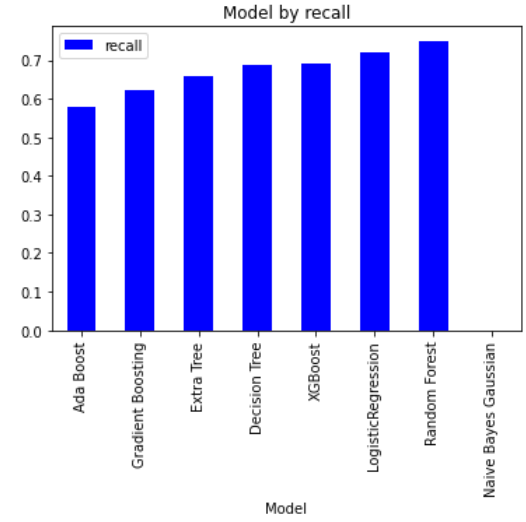
Visualizations.

As like said earlier I have trained using eight different models and finalized Random Forest as top performed model we will visualize the metrics of these models as below.









Seeing the performance, I have finalized Random Forest model and hyper parameter tuned and saved the same.

CONCLUSION

Key Findings and Conclusions of the Study

Online reviews are an effective word of mouth marketing strategy in the digital age, providing outside perspectives on products and services. While positive reviews can drive revenue and build a trustworthy reputation, negative reviews or the absence of reviews can do the opposite. Understanding the importance of reviews as well as how to leverage them to boost your business can be a critical way to get ahead in the competitive ecommerce marketplace, positioning yourself miles ahead of the competition.

Learning Outcomes of the Study in respect of Data Science

In this study we understand how the reviews plays major role in the business of E-Commerce. On a lay perspective how, ratings play a major role in promoting the business and most of the people are more interested in only knowing what is the rating of the products rather than the reviews.

Closing thoughts

Generating unique, high-quality content on every product page is a big challenge for Ecommerce businesses. Scaling content for thousands of products in the inventory is a Herculean task that online retailers cannot possibly handle on their own as it demands a lot of time and resources. To add to the complexity of search engine ranking, Google algorithms are becoming more intuitive with every update. They filter out websites that do not have original content and those with duplicate content within the site. But on the bright side, when the content on a website is unique and valuable, Google recognizes it for being trustworthy. So, if a brand wants to stand out in search results, it is not enough to just generate unique content for the website, but also to ensure that the unique content is compelling and valuable.

Collect more customer reviews and to derive maximum SEO benefits from user-generated content by ensuring it is crawled and checking for duplication.