FLIP ROBO

RATING PREDICTION PROJECT

BASED ON USER REVIEWS

Submitted by:

G. PREMSAGAR

ACKNOWLEDGMENT

With brands competing for customers' interest through aggressive marketing, internet reviews have become a crucial factor for consumers to consider. Over 88 percent of Ecommerce customers make purchases solely after reading product reviews. Customer journeys are secured by product ratings and reviews, which give them the confidence to continue the checkout procedure.

Consumer insights may be gained through user-generated material. It aids Ecommerce sellers in comprehending the wants of their customers. Brands may utilise this data to come up with new product ideas, enhance and reinvent existing items, increase consumer loyalty, and grow their business.

User-generated content (UGC) is any type of content contributed by people on their own initiative. Reviews, photos, videos, comments, and queries are all included. Creating forums for users to discuss products and services in order to strengthen a brand's engagement with its customers.

According to a Nielsen survey used to create the Consumer Prefer Index, 92 percent of customers trust organic user-generated content above traditional advertising. Brands may save time and money while also increasing brand trust by outsourcing content development to their customers. When used correctly, user-generated content (UGC) may be a valuable resource for content marketers. To begin, it may be helpful to understand why individuals generate and distribute free material.

- Consumers like sharing their thoughts on items and services they've purchased. They may use the internet as a platform to express themselves and be heard.

- Consumers benefit from additional incentives such as social recognition and admiration, despite the fact that user-generated material is not compensated.

- Some businesses encourage customers to sample their items and post reviews with a specific hashtag and photo. A winner is picked in these types of contests, and they are given prizes or discounts.

- People, on the whole, absorb a lot of stuff before making a purchase. However, user reviews aren't only for shoppers. Let's look at the advantages that User-Generated Content may provide for brands.

Before making a purchase, customers always check at least ten online customer reviews.

INTRODUCTION

- Business Problem Framing

What do you do initially if you go to make an online purchase? Many shoppers rely on online product reviews in an ecommerce-driven environment where they can't personally inspect things before buying.

With the growth of internet review sites like Yelp! and Facebook, getting an opinion on just about anything is now just a few clicks away. The rise of internet reviews has even influenced how businesses are regarded.

When it comes to earning business and keeping a great reputation, online reviews are vital for every company that operates in the digital arena.

- Conceptual Background of the Domain Problem

Who is it that reads online reviews?

Almost everyone reads internet reviews in today's web-based environment. In fact, 91% of people read them, and 84% trust them as much as a personal endorsement. The effects of reviews may also be measured.

The average client is prepared to pay 31% extra for a merchant with positive evaluations.

Negative reviews may be just as influential as favourable evaluations. According to one survey, 82 percent of those who read internet reviews intentionally look for unfavourable feedback.

This statistic just reinforces the fact that unfavourable reviews aren't going overlooked - but there are some advantages: According to research, when consumers interact with unfavourable reviews, they spend five times as much time on the site and convert at an 85 percent higher rate.

Customers like to see a large number of reviews. An opinion is formed by a single review with a few nice words, but a consensus is formed by a few dozen reviews that all say the same thing.

More reviews are better, according to one research, and customers need at least 40 reviews to believe an average star rating. A few reviews, on the other hand, are preferable than none.

According to one study, items with as little as five evaluations are 270 percent more likely to sell.

With the abundance of review sites and the high degree of faith that most customers place in reviews, it's reasonable to assume that almost everyone evaluating your items, regardless of their target demographic, industry, or market, is reading online reviews before making a purchase.

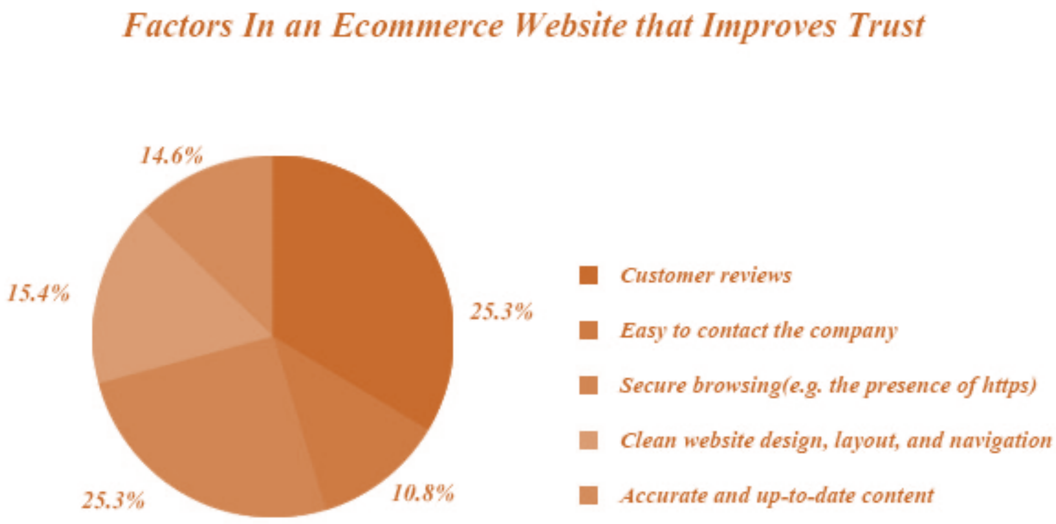
- ## Review of Literature

Significance Of Reviews

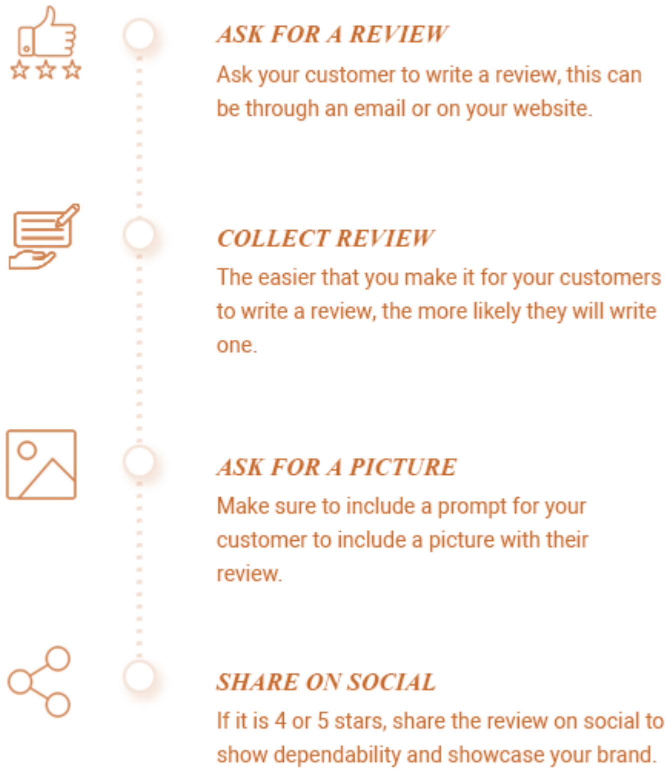
Because they represent personal and unbiased user experiences, reviews are able to gain credibility. They are original content that search engines adore since they are written by consumers. A large number of customer evaluations will, of course, result in increased organic traffic to Ecommerce websites. Original content that is appealing to both customers and search engines is in high demand. User-generated material is unique and authentic. Customers made it, therefore it's unique. It does not necessitate a significant investment of resources on the part of brands. User-Generated Material (UGC) is the answer to the never-ending hunt for unique content. It's a consistent flow of high-quality, searchable information for a company's website. User-Generated Content also aids companies in learning about current trends and client preferences.

We looked at and evaluated 500+ eCommerce sites, the bulk of which were powered by Shopify. Given the vast quantity of data we obtained, it should come as no surprise that the Shopify product reviews app enhances brand trust, increases conversion rates, and attracts more search engine traffic. Because most customer journeys begin with a search engine, organisations must use User Generated Content (their most powerful marketing weapon) to boost SEO and build social proof. Because most customer journeys begin with a search engine, organisations must use User Generated Content (their most powerful marketing weapon) to boost SEO and build social proof.

Customer reviews have an impact on influencing real-world buyers to make trust-based purchasing decisions, in addition to how they influence SEO. The following are the findings of an independent SaaS and SMB industry study on elements that boost website visitor trust:



How Does The Review Collection Process Work?



ASK FOR A REVIEW

Ask your customer to write a review, this can be through an email or on your website.

COLLECT REVIEW

The easier that you make it for your customers to write a review, the more likely they will write one.

ASK FOR A PICTURE

Make sure to include a prompt for your customer to include a picture with their review.

SHARE ON SOCIAL

If it is 4 or 5 stars, share the review on social to show dependability and showcase your brand.

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.

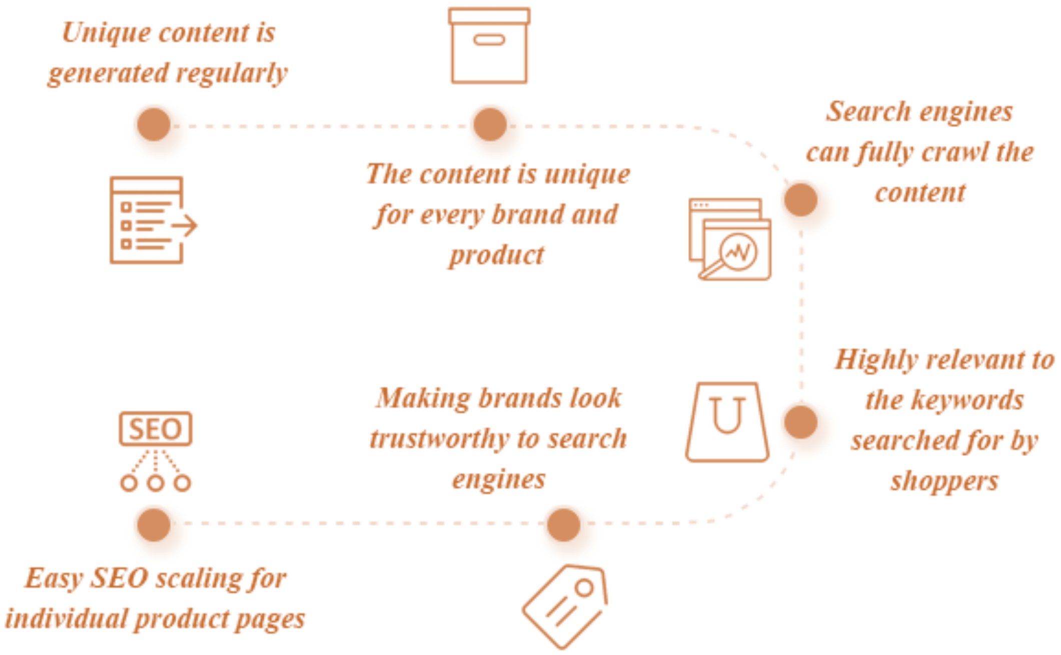
User-Generated Content plays a valuable role in SEO and marketing strategies. Here's how it works:

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
SEO
Easy SEO scaling for individual product pages

- 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.

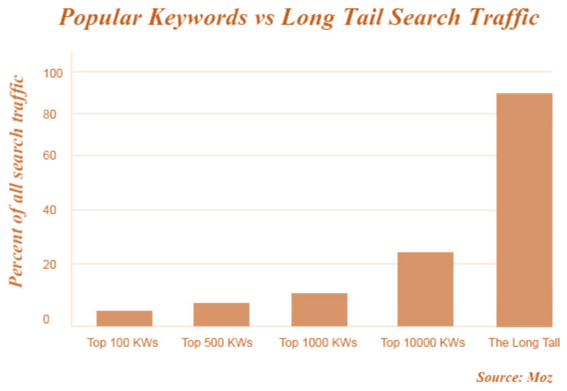
- Motivation for the Problem Undertaken

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

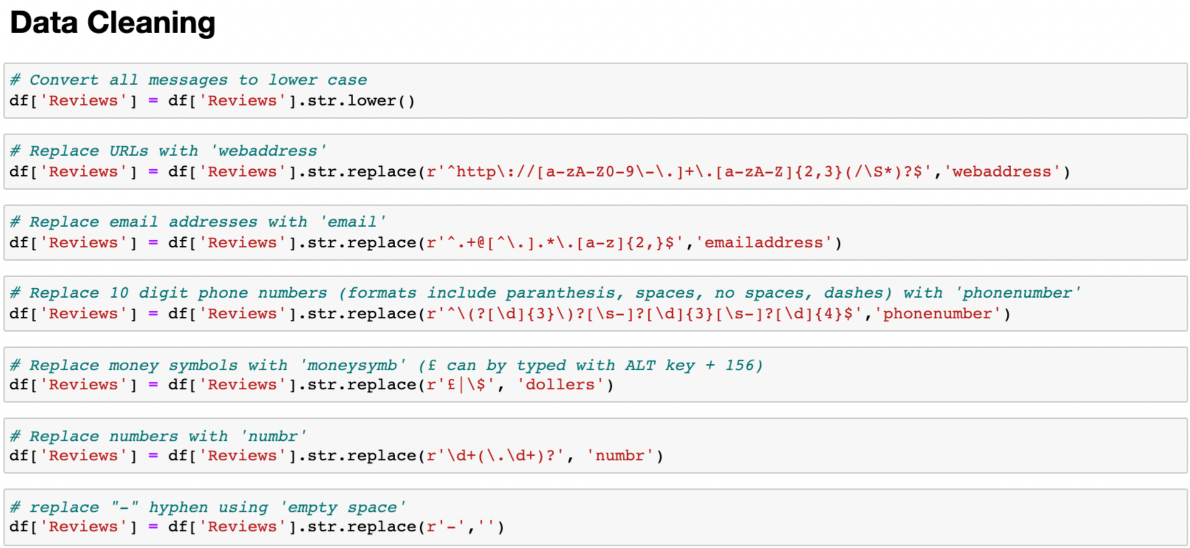
- Data Sources and their formats

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 Preprocessing Done

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



Data Cleaning

```
# Convert all messages to lower case
df['Reviews'] = df['Reviews'].str.lower()
```

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

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

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

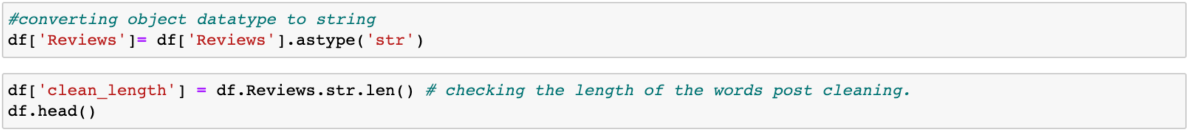
```
# Replace money symbols with 'moneysymb' (£ can by typed with ALT key + 156)
df['Reviews'] = df['Reviews'].str.replace(r'£|\$', 'dollers')
```

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

```
# replace "-" hyphen using 'empty space'
df['Reviews'] = df['Reviews'].str.replace(r'-','')
```

As you can see in the video above, I cleaned the data by following these procedures.

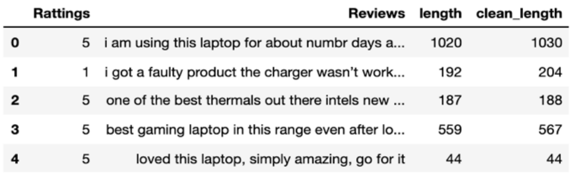
- Changing all of the letters to lower case.
- All email addresses in the reviews have been transformed to "email address."
- All web addresses in the reviews have been transformed to "Web address."
- All currency in the reviews has been converted to "dollars."
- All phone numbers in the reviews have been converted to "phonenumber."
- All numbers in the reviews have been translated to "number."



```
#converting object datatype to string
df['Reviews']= df['Reviews'].astype('str')
```

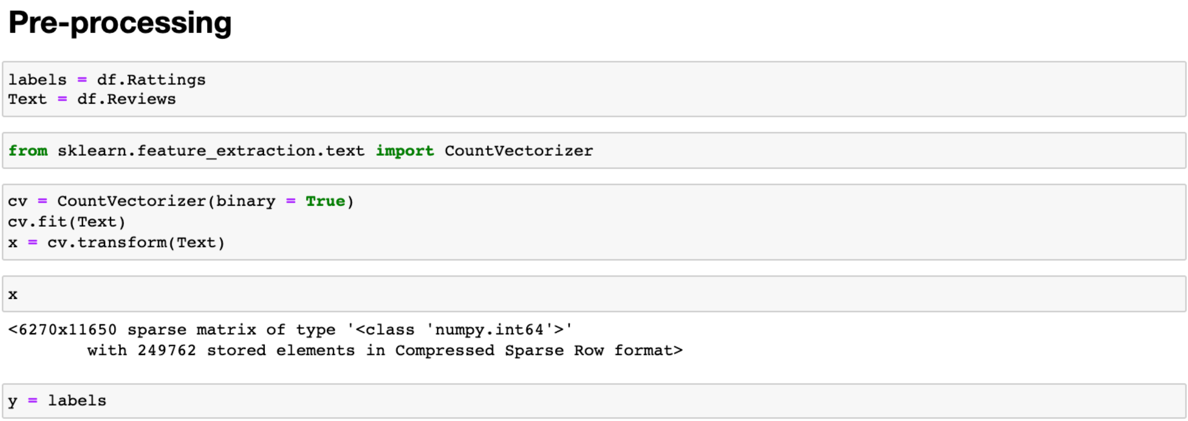
```
df['clean_length'] = df.Reviews.str.len() # checking the length of the words post cleaning.
df.head()
```

Converted Reviews column from object datatype to string datatype and then post cleaning.



	Rattings	Reviews	length	clean_length
0	5	i am using this laptop for about numbr days a...	1020	1030
1	1	i got a faulty product the charger wasn't work...	192	204
2	5	one of the best thermals out there intels new ...	187	188
3	5	best gaming laptop in this range even after lo...	559	567
4	5	loved this laptop, simply amazing, go for it	44	44

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



Pre-processing

```
labels = df.Rattings
Text = df.Reviews
```

```
from sklearn.feature_extraction.text import CountVectorizer
```

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

```
x
```

```
<6270x11650 sparse matrix of type '<class 'numpy.int64'>'
	with 249762 stored elements in Compressed Sparse Row format>
```

```
y = labels
```

Finally, before feeding the data into the model, I saved it as a count vectorizer to make the training process go faster, and I saved the Ratings data in the y variable.

- Hardware and Software Requirements and Tools Used

Hardware technology being used:-

CPU: MacBook Pro

Chip: Apple M1 - 8 (4 performance and 4 efficiency) GPU: 8GB

RAM: 8 GB

Software technology being used:-

Programming language: Python

Distribution: Anaconda Navigator

Browser based language shell: Jupyter Notebook

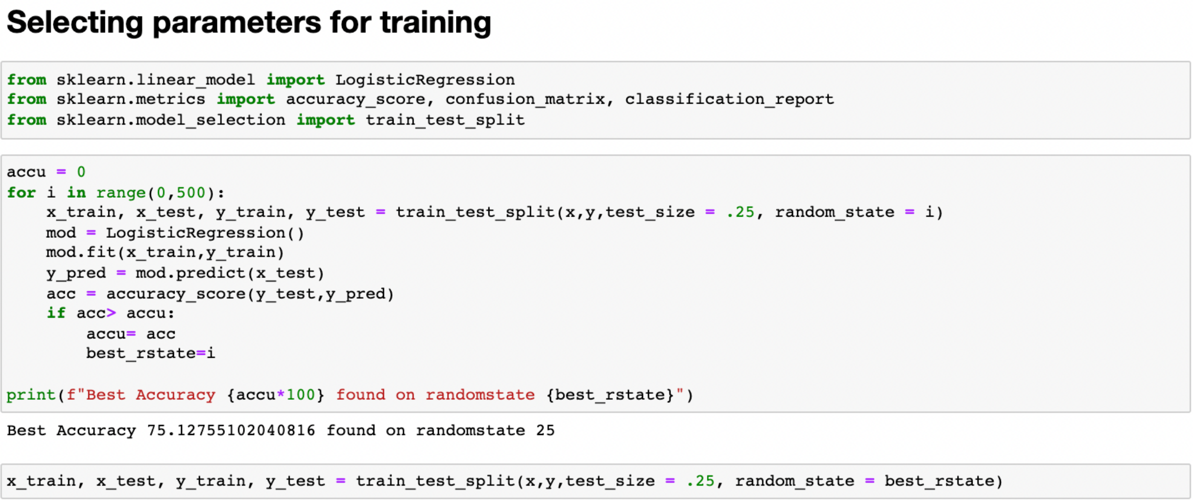
Libraries/Packages specifically being used:-

Pandas, NumPy, matplotlib, seaborn, Data science, scikit-learn, NLTK library, Machine Learning, Anaconda Environment, Jupyter Notebook.

Model/s Development and Evaluation

- **Selecting best random state parameters for training.**

  I used logistic regression to generate a loop ranging from 0-500 to sort the most accurate random state parameters.



Selecting parameters for training

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split
```

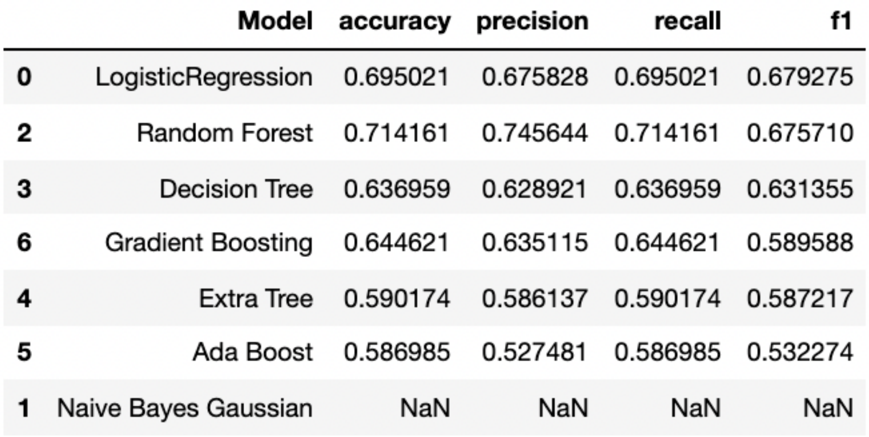
```
accu = 0
for i in range(0,500):
    x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = .25, random_state = i)
    mod = LogisticRegression()
    mod.fit(x_train,y_train)
    y_pred = mod.predict(x_test)
    acc = accuracy_score(y_test,y_pred)
    if acc> accu:
        accu= acc
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

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

```
Best Accuracy 75.12755102040816 found on randomstate 25
```

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

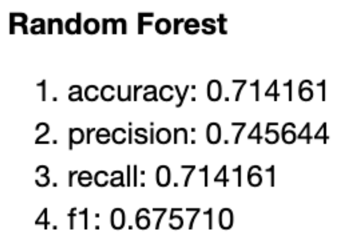
- **Testing of Identified Approaches (Algorithms)**

  I utilised seven different algorithms and tried to rank them in order of how well they worked. “Logistic Regression", "Naive Bayes Gaussian", "Random Forest", "Decision Tree", "Extra Tree", "Ada Boost", "Gradient Boosting".



	Model	accuracy	precision	recall	f1
0	LogisticRegression	0.695021	0.675828	0.695021	0.679275
2	Random Forest	0.714161	0.745644	0.714161	0.675710
3	Decision Tree	0.636959	0.628921	0.636959	0.631355
6	Gradient Boosting	0.644621	0.635115	0.644621	0.589588
4	Extra Tree	0.590174	0.586137	0.590174	0.587217
5	Ada Boost	0.586985	0.527481	0.586985	0.532274
1	Naive Bayes Gaussian	NaN	NaN	NaN	NaN

The Random Forest Algorithm, as seen in the table above, is at the top of the heap, with the following metrics:

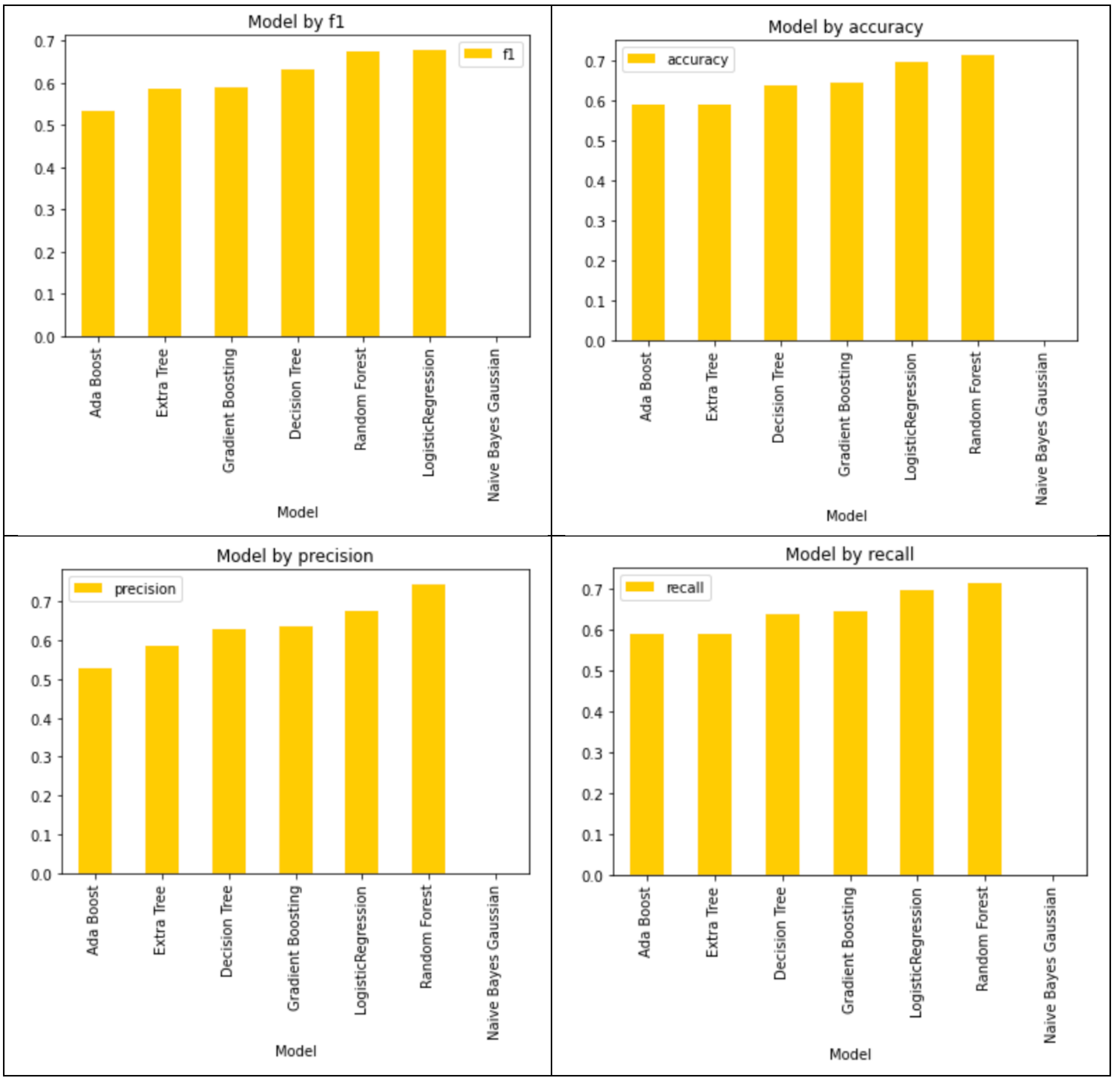


Random Forest

1. accuracy: 0.714161
2. precision: 0.745644
3. recall: 0.714161
4. f1: 0.675710

- **Visualizations:**

As previously said, I trained using seven different models, with Random Forest emerging as the best performer. We will illustrate the metrics of these models as follows.



After seeing the results, I finished the Random Forest model, tweaked the hyper parameters, and saved it.

CONCLUSION

- ## Key Findings and Conclusions of the Study

In the digital era, online evaluations are a powerful word-of-mouth marketing tactic for offering outside viewpoints on products and services. While favourable evaluations can boost income and establish a trustworthy reputation, negative reviews, or the lack thereof, can have the opposite effect. Understanding the value of reviews and how to use them to help your business may be a crucial part of getting ahead in the competitive ecommerce industry and putting yourself miles ahead of the competition.

- ## Learning Outcomes of the Study in respect of Data Science

In this research, we will learn how reviews play an important part in the E-Commerce industry. In a layman's perspective, ratings play a significant part in advertising a business, and most individuals are more interested in knowing what a product's rating is than in reading reviews.

- ## Closing thoughts

Ecommerce firms face a significant problem in creating unique, high-quality content for each product page. Scaling content for thousands of goods in inventory is a Herculean effort that online merchants cannot realistically do on their own due to the time and resources required. Google algorithms are getting more obvious with each update, adding to the intricacy of search engine ranking. They filter out websites with duplicate material and those that do not have original content. On the plus side, when a website's content is original and useful, Google regards it as trustworthy. So, if a company wants to be seen in search results, it's not enough to create original content for the website; it also has to be interesting and valuable.

Collect more customer reviews and ensure that user-generated material is indexed and checked for duplicate to get the most SEO benefits from it.