# In Consumer Reviews We Trust:
# Decision Making with Sentiment Analysis and Machine Learning

## 1 - Introduction:

As consumers, we live in a day and age where everyone seems to have a voice or an opinion, where no product or business is off limits. Whether it be that scathing review you see from a questionable restaurant on Yelp, or that praiseful review on an Airbnb experience, there are now endless amounts of web services to jot down your anecdotes for products and services. A company like Amazon has built its platform on encouraging its customers to rate and review their products in order to inform the consumer of what they are getting themselves into. With very little restriction, the consumer can express his/her own experience with text through a star rating system, corresponding to the positive or negative experience.

## 2 - The Problem:

Given the text of the review and which features of text are indicative of its context being positive or negative, natural language processing and sentiment analysis can be applied to Amazon's seemingly endless data and develop models to take advantage of these features. This project plans to investigate the various popular classification models to help analyze Amazon's product reviews and empower decision making for both the consumer and the manufacturer. I decided to investigate Amazon's Sports and Outdoors department and explore all descriptive reviews within this field. Ultimately, all sentiment analysis and machine learning models to help empower decision making for future products would be presented to my client, the largest sporting goods chain, Dicks Sporting Goods.

## 3 - The Approach:

These are the following topics which I covered:

1) Deciding which variables have the strongest association to the review classification model.
2) Breaking reviews into sentences while mapping categories and sentiment.
3) Implementation of Count Vectorization & Term Frequency Inverse Document Frequency (TF-IDF) on the dataset.
4) Analyze the various classification models for accuracy and precision.
5) Processing textual data through assigning a polarity score for both positive and negative reviews for each product, based on the type of vocabulary associated with each review.

**4 - The Dataset:**

The data was obtained through a collection of product reviews and metadata from the UCSD Database of Dr. Julian McAuley, consisting of ~296,000 product reviews. These Sports and Outdoors reviews are in a compressed JSON gz file. To convert this compressed file into a data frame I used the pandas package. Before proceeding with any further EDA, this data frame required a bit of cleaning. See below for details.

**4.1 - Data Wrangling:**

The important variables that were collected from these product reviews were the following:

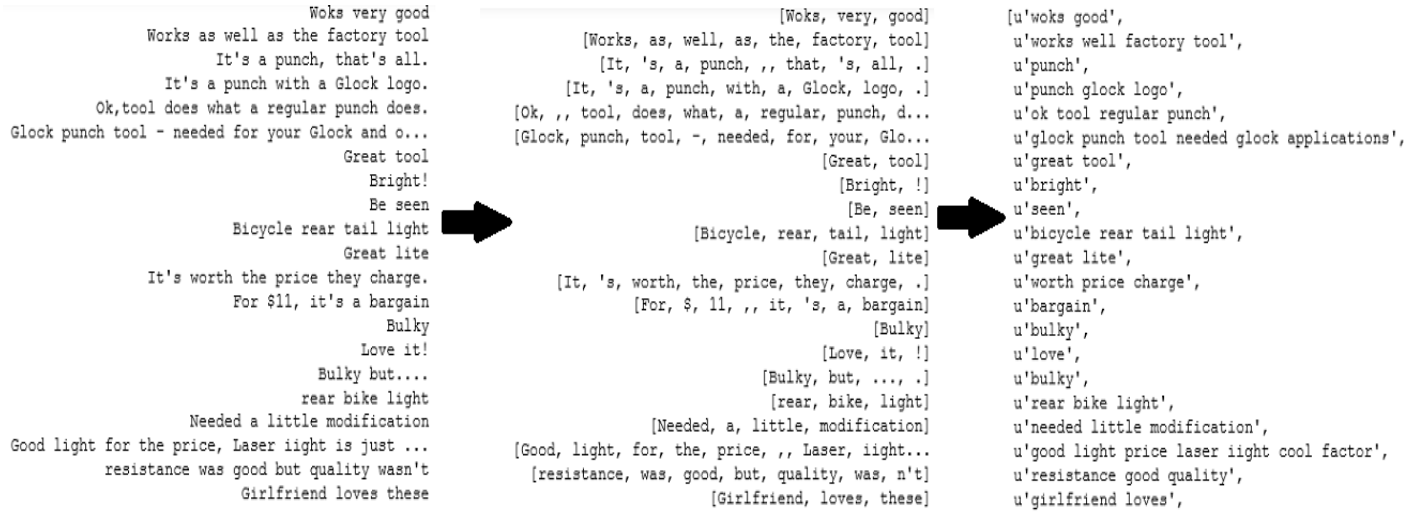
| | asin | helpful | overall | reviewText | reviewTime | reviewerID | reviewerName | summary | unixReviewTime |
|---|---|---|---|---|---|---|---|---|---|

A majority of data points came in raw form ie)*reviewTime* = 01,26,2014 instead of 01/26/2014, *reviewText* consisted of upper and lower case words, punctuation and stop words, ie) "the", "a", "and", "but", "what" "or", etc... Stop words don't really serve much of a purpose when delving into linguistic analysis. Removing all stop words helps with analyzing sentiment in a further concise manner.

Next, I took the revised product review data frame and used a natural language toolkit (Nltk) package, called *Tokenize* to create a new column "tokenizing" each word in its isolated form, rather than in complete sentence form. I then created a binary classifier in a new column solely to assess whether the review was deemed "Positive" or "Negative" based on star rating. What constitutes a positive negative column is where positive review (4-5) are 1 and negative reviews (1-3) are 0.

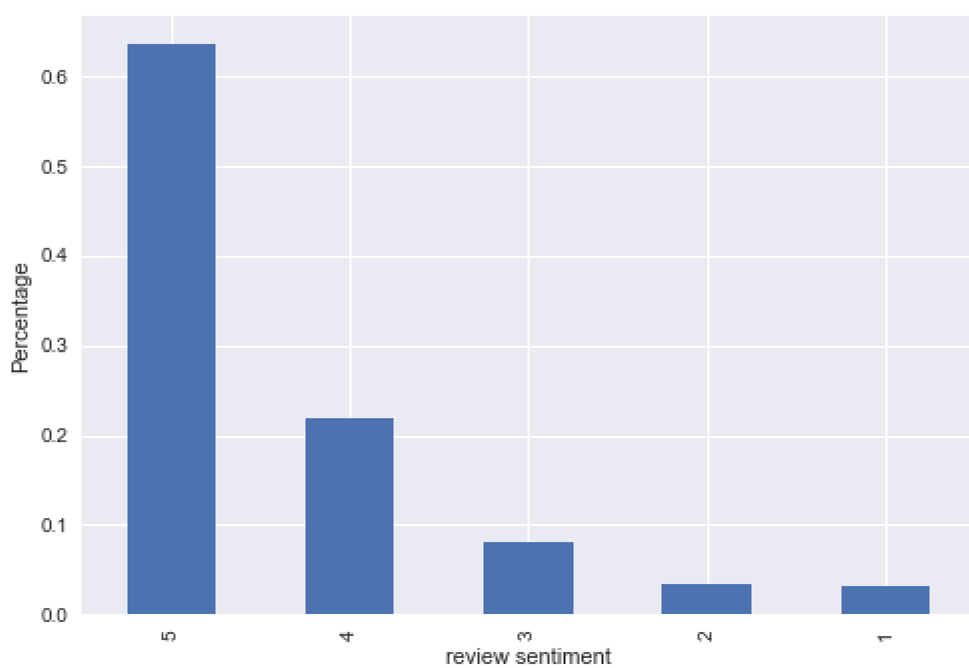
**4.2 - Data Preprocessing:**

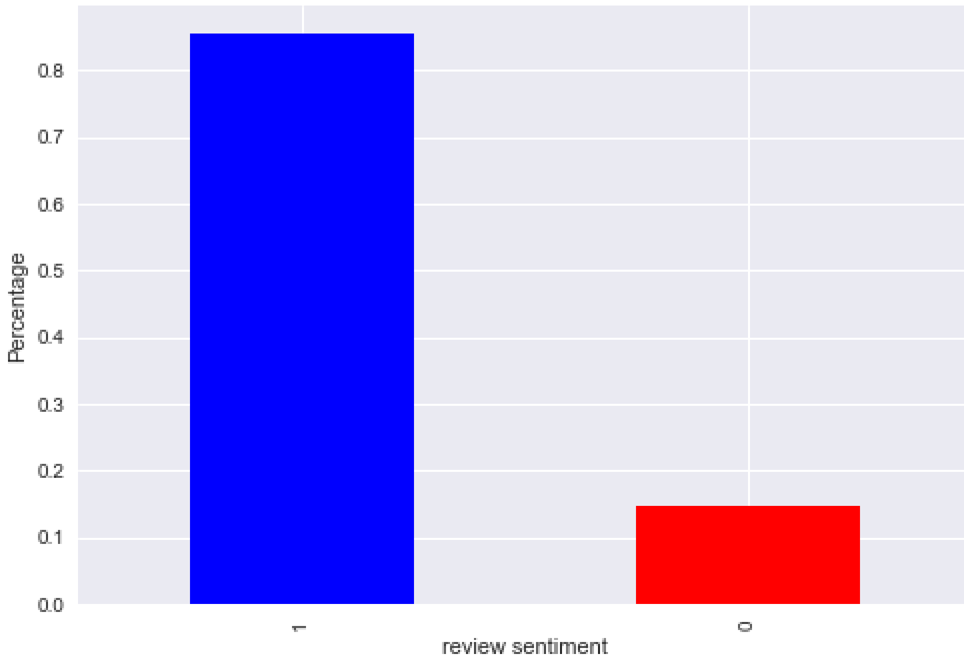
Preprocessing natural language processing (NLP) data can be time consuming and certain decisions need to be made such as: Should I remove stop words? Do you remove capitalization? How to handle contractions? I ultimately decided to remove stop words, capitalization, and punctuation, implementing lemmatization and stemming since it made the data easier to work with. This made my data a bag of words which means that the order of the words were no longer preserved.

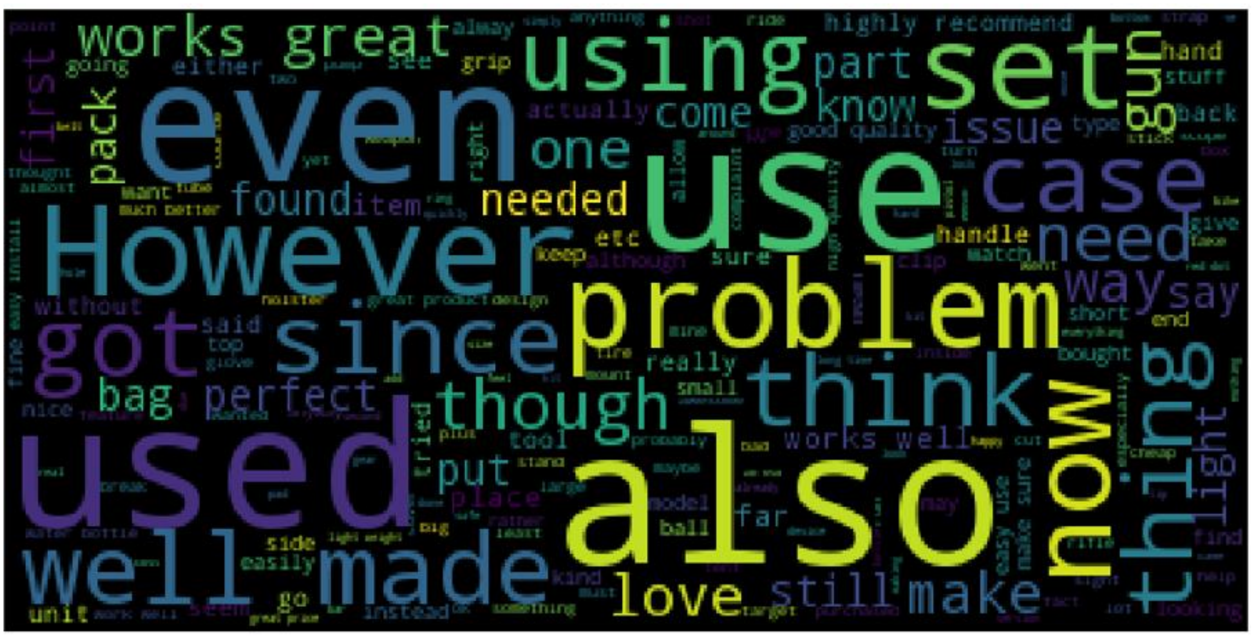



Once the data preprocessing is complete, I then had a clean corpus and could begin with my exploratory data analysis.

## **5 - Exploratory Data Analysis:**




Percentage
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
0
review sentiment


works great
using
set
even
use
However
case
problem
got
since
think
used
also
now
thing
well made
love
still
make
need
needed
though
one
know
issue
part
found
perfect
bag
pack
gun
way
say
highly recommend
light
come
put
really
easily
instead
unit

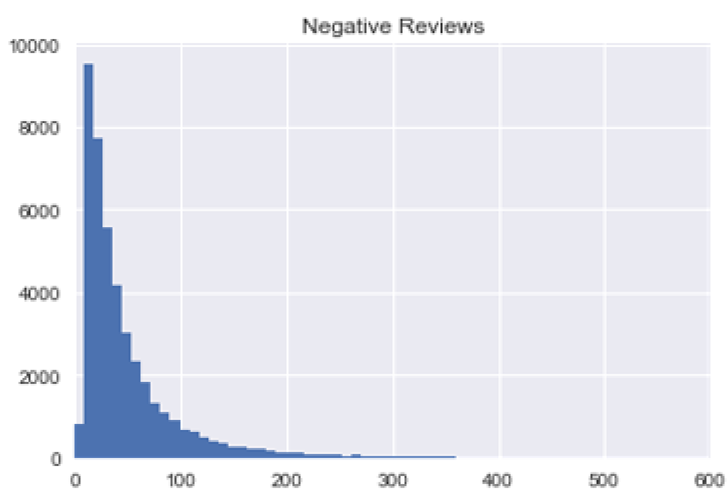
```
Top 20 Reviewed Products:

asin
B001HBHNHE    1042
B001T7QJ9O     763
B000S5ODN2     647
B0010O748Q     513
B0000C50K3     427
B002ZYRV2E     401
B002OKWHVO     398
B000GCRWCG     393
B001HBHNHY     372
B0035L35A8     359
B004U8CP88     357
B001WJ577O     355
B004TNWD40     349
B006X9DLQM     344
B00178CS4K     343
B006QF3TW4     323
B003NFI092     309
B00200E0HM     307
B001949TKS     298
B000JZ7JM8     293
Name: overall, dtype: int64

Most Reviewed Product, B001HBHNHE has 1042 reviews
```

```
Bottom 20 Reviewed Products:

asin
B003VN1U7K    5
B003VZIWZQ    5
B00BFXAIEY    5
B000T1VZCW    5
B000TH4MMG    5
B003XDWY32    5
B000T29STI    5
B003WXR992    5
B003WXJOEA    5
B000TAADTO    5
B00BGI7R38    5
B003WHHXAS    5
B000TTR0JG    5
B003WHA1N4    5
B000TRKTUK    5
B000TTHVYA    5
B003WGUEEG    5
B000TTM3OI    5
B003WDTI78    5
B003Z6HUZE    5
Name: overall, dtype: int64

Most Reviewed Product, B003Z6HUZE has 5 reviews
```

| ID | Product | ID | Product |
|---|---|---|---|
| B001HBHNHE | 1006409 Maglula UpLULA Universal Pistol Magazine Loader | B003VN1U7K | UTG Laser/Flashlight Ring w/Slide-on Picatinny Mounting Base |
| B001T7QJ9O | Howard Leight Impact Sport OD Electric Earmuff, Green | B003VZIWZQ | Last Punch Heavy Duty Punching Bag with Chains |
| B000S5ODN2 | Rothco 550lb. Type III Nylon Paracord | B00BFXAIEY | Unknown |
| B0010O748Q | SE FS374 All-Weather Emergency 2-IN-1 Fire Starter & Magnesium Fuel Bar (Everything you need to start a fire!) | B000T1VZCW | Unknown |
| B0000C50K3 | Big Rock Sports 24011 Bore Snake Rifle Cleaner, M16 - .22-Caliber | B000TH4MMG | Unknown |
| B002ZYRV2E | SET Tactical Weapon Mount | B003XDWY32 | Unknown |
| B002OKWHVO | Unknown | B000T29STI | Unknown |
| B000GCRWCG | Unknown | B003WXR992 | Unknown |
| B001HBHNHY | Unknown | B003WXJOEA | Valeo Chin Up Bar |
| B0035L35A8 | UTG Tactical OP Bipod, Rubber Feet, Center Height 8.3"-12.7" | B000TAADTO | Wigwam Men's Super 60 Tube 3-Pack Over-the-Calf Length Socks |
| B004U8CP88 | Icetek Sports Ultralight Portable Outdoor Backpacking Camping Stoves with Piezo Ignition, Butane / | B00BGI7R38 | Wheeler Anti Cant Indicator Pic Rail |
| B001WJ577O | UTG Shooter's SWAT Bipod, Rubber Feet, Height 6.2"-6.7" | B003WHHXAS | Grim Reaper Razorcut SS Whitetail Spec 2" Cut 3 Bld 100gr |
| B004TNWD40 | Morakniv Companion Fixed Blade Outdoor Knife with Carbon Steel Blade, Military Green, 4.1-Inch | B000TTR0JG | Harris Engineering Hinged Base 9 - 13-Inch BiPod |
| B006X9DLQM | Ade Advanced Optics Front/Rear 45-Degree Rapid Transition BUIS Backup Iron Sight | B003WHA1N4 | Hurricane Bag Archery Target - Taking the Archery World by Storm - Available in 3 Sizes |
| B00178CS4K | SE KHK6320 Outdoor Tanto Knife with Fire Starter | B000TRKTUK | Unknown |
| B006QF3TW4 | LifeStraw Personal Water Filter for Hiking, Camping, Travel, and Emergency Preparedness | B003WGUEEG | Trampoline Replacement Jumping Mat, fits for Round Frames with V-Rings - MAT ONLY |
| B003NFI092 | Ultimate Arms Gear Tactical 4 Reticle Red Dot Open Reflex Sight with Weaver-Picatinny Rail Mount | B000TTM3OI | Allen Tactical Rifle Case, 5 Pocket |
| B00200E0HM | Bushnell Trophy TRS-25 Red Dot Sight Riflescope, 1x 25mm (tilted front lens) | B003WDTI78 | Avid BB5 Mountain w/ G2 Rotor |
| B001949TKS | Potable Aqua Water Purification Germicidal Tablets - For Hiking, Camping, and Emergency Drinking Water | B003Z6HUZE | RUKO 2-Position Web Nylon Knife Sheath |

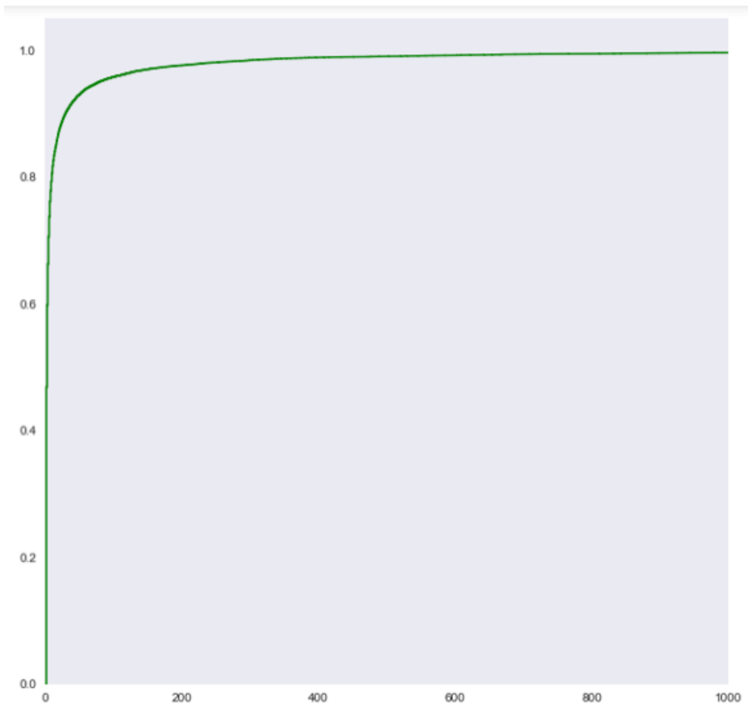

Length of Reviews
100000
80000
60000
40000
20000
0
0
100
200
300
400
500
600


Positive Reviews
80000
60000
40000
20000
0
0
100
200
300
400
500
600


Negative Reviews
10000
8000
6000
4000
2000
0
0
100
200
300
400
500
600

**5.1 - Variable Selection:**

Prior to constructing the various models and algorithms on my dataset, I decided to select a subset of features to optimize performance, “Overall” and “Summary”. By using the Term Frequency Inverse Document Frequency (TF-IDF) technique I could then demonstrate numerical value to how important a word is to a collection of a corpus. To do so, I constructed a graph demonstrating the frequency percentage to document count and examined where the minimum document frequency was plotted. See below for details




**6 - Training and Machine Learning:**

Once I had the proper variables selected I then decided to delve into supervised learning by splitting the data into Train and Test samples. I split the data into 5-fold and computed scikitlearn package, Stratified K Fold Cross Validation on the data set to maintain the same proportions in each fold for accuracy estimation and model selection.

I used 7 machine learning models to train my classifiers. Before doing so, I decided to tune my models and optimized my hyperparameters for each model. Afterwards I plan to evaluate each model’s strength with various performance metrics:

- Bernoulli NB
- Multinomial NB
- Ada Boost
- Linear Regression
- Decision Tree
- Random Forest
- Perceptron

**6.1 - Results:**

For this part of the project I decided to implement another sklearn package, called "accuracy score", which does a great job to catalogue the actual metrics and their true value for each model. The harmonic mean of precision and recall (F-1 Score) is a diagnostic test included in "accuracy score" to evaluate the Precision and Recall, simultaneously with a final score.

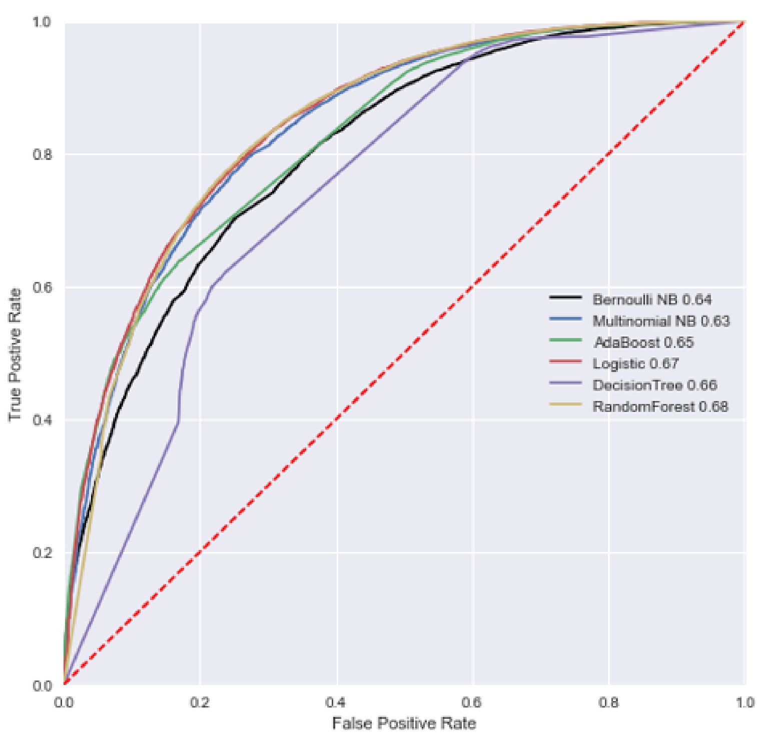
| Classifier | Recall | Precision | F-1 Score |
|---|---|---|---|
| Bernoulli Naïve Bayes | 0.92 | 0.87 | 0.43 |
| Multinomial Naïve Bayes | 0.95 | 0.88 | 0.91 |
| Ada Boost | 0.94 | 0.88 | 0.90 |
| Logistic Regression | 0.93 | 0.89 | 0.90 |
| Decision Tree | 0.92 | 0.88 | 0.89 |
| Random Forest | 0.92 | 0.89 | 0.90 |
| Perceptron | 0.85 | 0.84 | 0.84 |

| Classifier | Split | Accuracy |
|---|---|---|
| Bernoulli NB | Train | 88% |
| | Test | 87% |
| Multinomial NB | Train | 89% |
| | Test | 88% |
| Ada Boost | Train | 88% |
| | Test | 88% |
| Logistic Regression | Train | 89% |
| | Test | 89% |
| Decision Tree | Train | 92% |
| | Test | 88% |
| Random Forest | Train | 96% |
| | Test | 89% |

| Perceptron | Train | 84% |
|---|---|---|
| | Test | 84% |

| Classifier | Area Under the Curve (AUC) |
|---|---|
| Bernoulli NB | 0.64 |
| Multinomial NB | 0.63 |
| Ada Boost | 0.65 |
| Logistic Regression | 0.67 |
| Decision Tree | 0.66 |
| Random Forest | 0.68 |
| Perceptron | 0.65 |

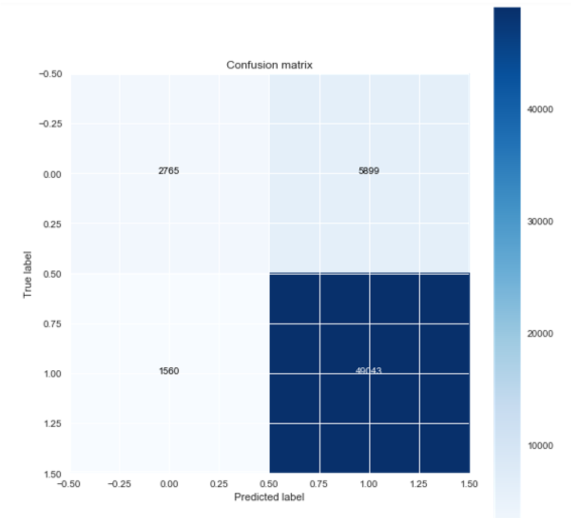
**ROC Curve:**



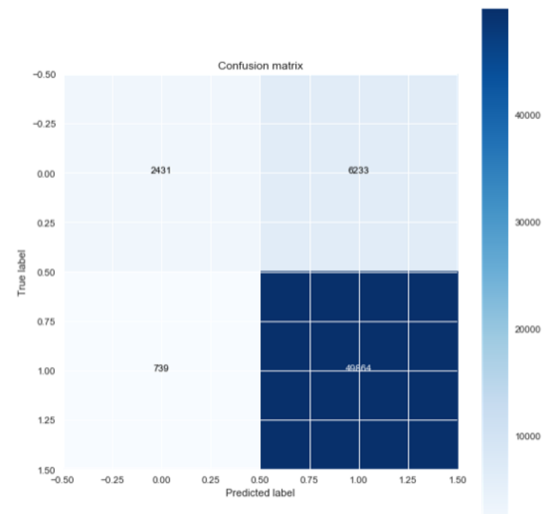
## 6.2 - Confusion Matrix:

To get a better visual understanding of the performance on the algorithms, I ran a confusion matrix on each model.

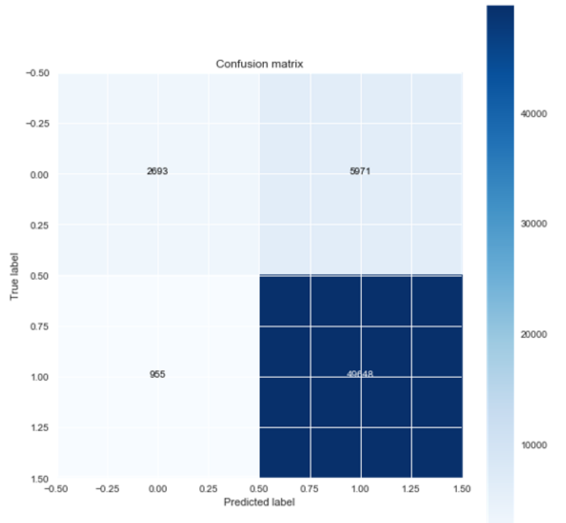
Bernoulli NB



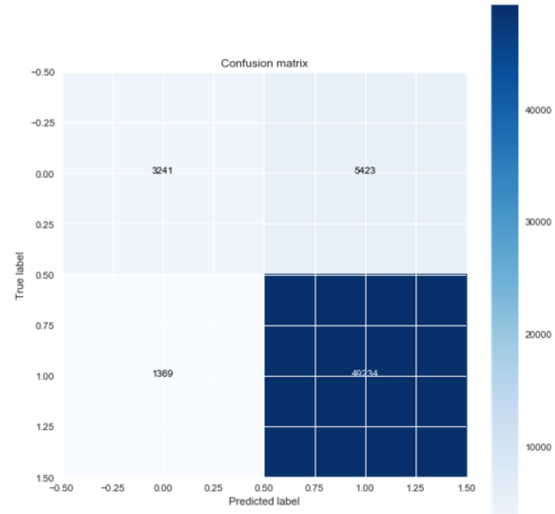

Multinomial NB



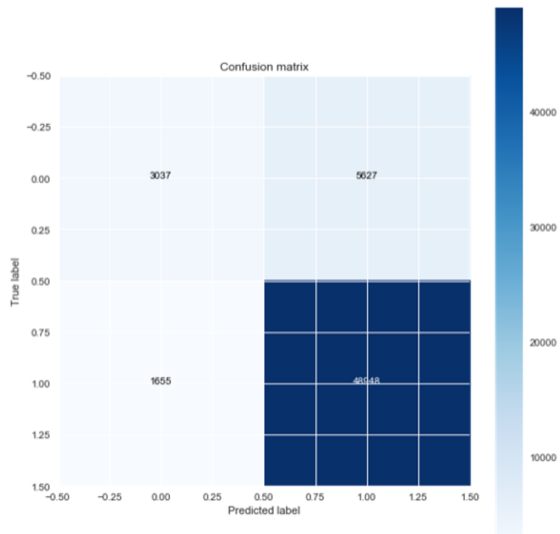

Ada Boost



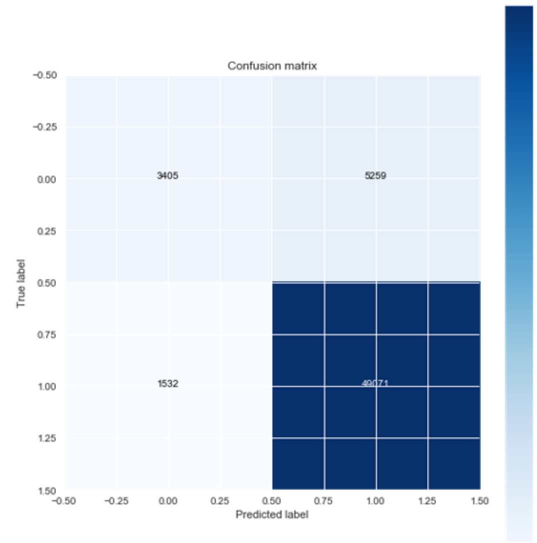

Logistic Regression



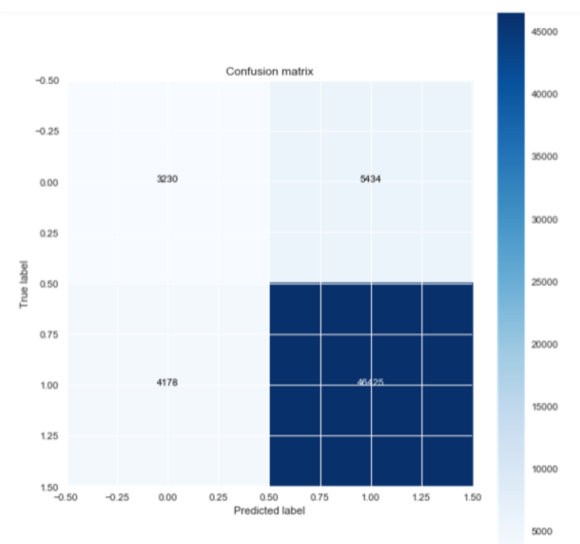
Decision Tree




Random Forest



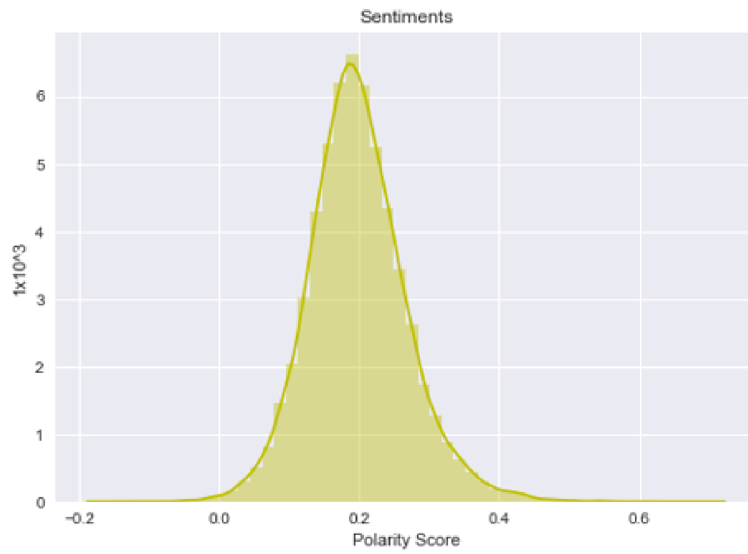

Perceptron

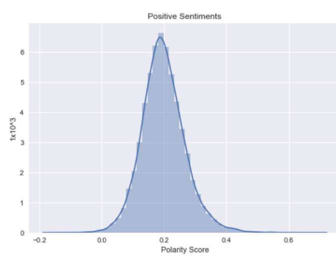


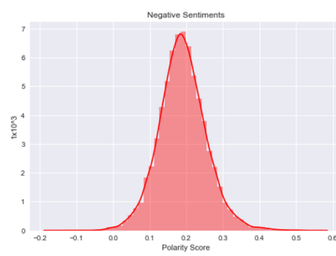

## 7 - Text Blob:

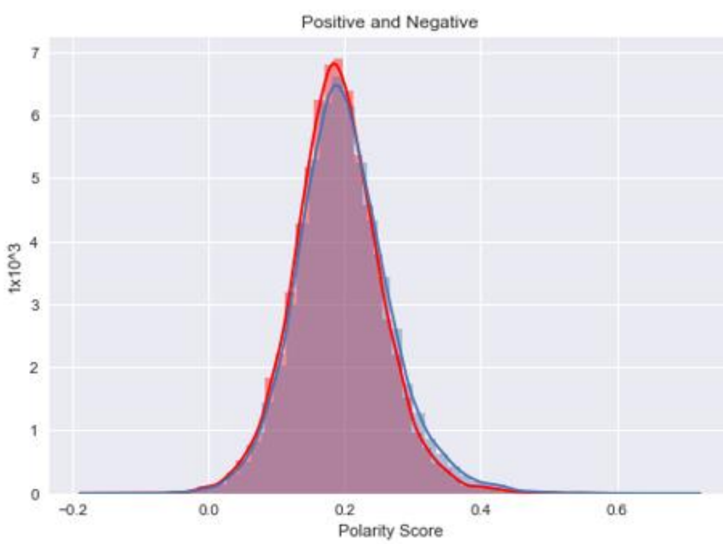
The next part of my project was to look at sentiment analysis through TextBlob which is a library in Python used for processing textual data. It is a simple API used for many NLP tasks including sentiment analysis.

For this part of the analysis, I grouped all reviews by product (~18,000) and looked at the vocabulary from all of the reviews for each product. Through the use of TextBlob, I assigned a polarity score to each item with negative polarity correlating to negative reviews/vocabulary and a positive polarity corresponding to positive reviews/vocabulary.




Based on the figures above, I separated the positive and negative reviews into two groups and plotted the distributions of each group's' product polarity scores. These two distributions followed a normal curve and found that the two distributions were statistically different after running a t test. Concluding that both groups were more positive than negative since the mean polarity score for both was approximately 0.19. This could be explained by the fact that Amazon typically removes products that are low performing, in order to keep up their reputation.

## 8 - Results:

After computing the various algorithms corresponding as the best recommendation to be used by Dicks Sporting Goods, we can conclude that Multinomial Naïve Bayes and Random Forest performed with the best results. Therefore, my recommendation would be to apply these methods to their sporting and outdoors equipment reviews to optimize the best results when analyzing how consumers are feeling about their products and how they can proactively improve as a company.

## 9 - Conclusion:

Generally, the results of this experiment were altogether successful, excluding the Bernoulli Naïve Bayes model.

## Potential Next Steps:

Since reviews are constantly increasing by the day, we have a necessary desire for rapid increase in Natural Language Processing technologies and methodology. NLP's development has significantly improved in bridging the gap between human insight and digital data. Once we better understand the intricacies of the various algorithms, we can move forward beyond consumer data exclusive to reviews and apply it to various applications beyond this market. It can and is being studied for improvements to Artificial Intelligence, extracting information in our financial markets and help making improvements in our healthcare framework.