# Capstone Project 2 Final Report

## Problem statement:

The goal is to understand what products are famous among customers and recommend them more similar products and increase the sales percentage. Also, to reduce the customer churn, go through the reviews provided by customers for the products and reduce the concerns.

## Dataset:

- Data is obtained from

http://jmcauley.ucsd.edu/data/amazon/

https://cseweb.ucsd.edu/~jmcauley/datasets.html

## Data Cleaning:

As the data file provided is in json format, convert it to csv format for easy access of data.

**Duplicates** –

Though it seems there are many duplicates, there are no repeated reviews each review is unique for different products by a customer.

**Missing Data** –

Only reviewText and sumary has missing values which are 24 and 1 respectively. As summary can be considered as minimal version of reviewtext. we can ignore the reviewText as we will be dropping the column going forward. And, summary has only 1 missing value which can be ingnored.

After looking at the overview of the data, changed few column names for better understanding.

- Asin – ProductID
- reviewerID_y - total_reviewers
- overall_x - overall_rating
- summary_x - review_summary

count how many total reviews are given for each product and sort the dataframe accordingly to later make selections

create a new df with the count data merged in to actual df for further analysis.

**Data Selection -**

Selecting only the products with more than 100 reviews for better results. As any products with less than 100 reviews might not be very popular and there is no need of recommending those products.

It is observed that we have reduced the data to 13379 rows, which is easy to handle and enough to make recommendations.

Single product might have received many reviews by different customers. Hence, grouping all the summary Reviews by product ID.

**Text Cleaning –**

Clean the reviews_summary column data by removing blank spaces and converting all the data into lower case.

Drop the duplicate rows and reset the index.

Using the CountVectorizer method to tokenize the data and build a vocabulary of known words.

**Modelling** –

Create a data frame with transformed reviews. Split the data into train and test data.

Using KNN(K Nearest Neighbor) find the similar products

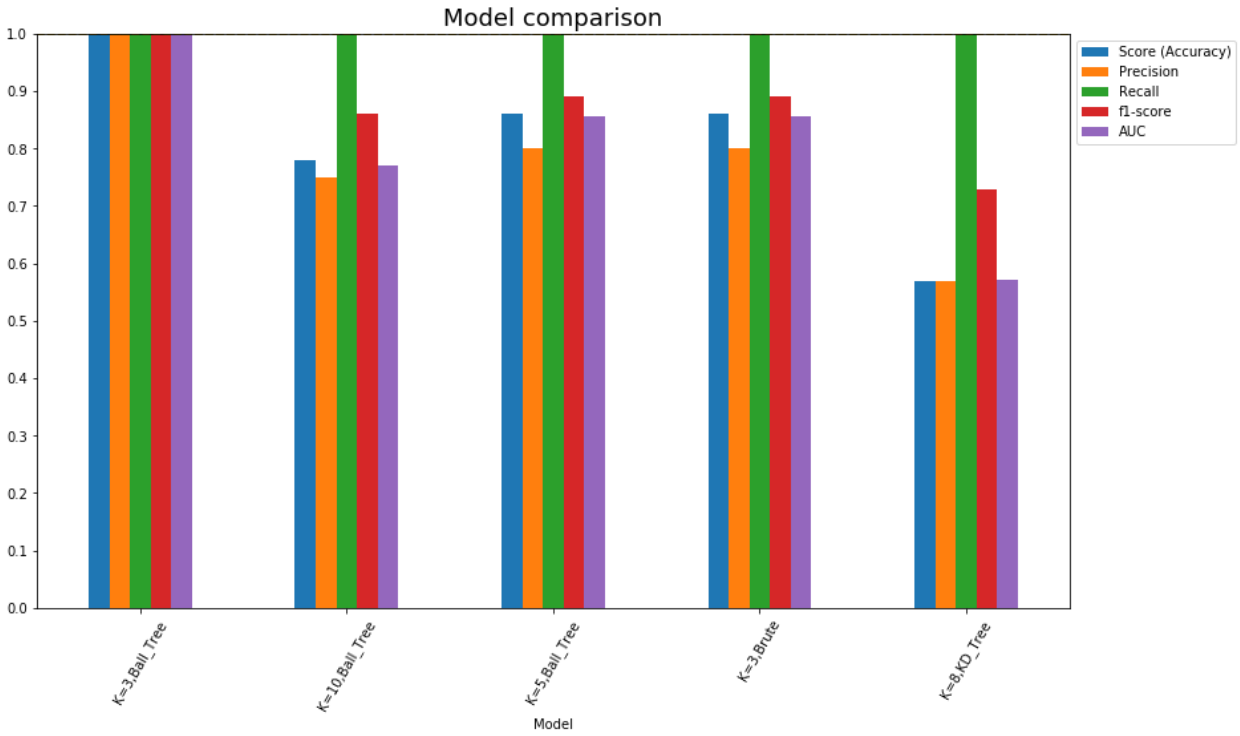
**kNN with k = 3, Algorithm = ball_tree**

- Based on product reviews, for B0051U15E4 average rating is 4.412280701754386
- The first similar product is B003YBHF82 average rating is 4.21
- The second similar product is B000FH4JJQ average rating is 4.536363636363636

**Predicting reviews with 85, 15 train, test split and k = 5**

- Based on product reviews, for B00DQYNS3I average rating is 4.526315789473684
- The first similar product is B003YBHF82 average rating is 4.21
- The second similar product is B000FH4JJQ average rating is 4.536363636363636

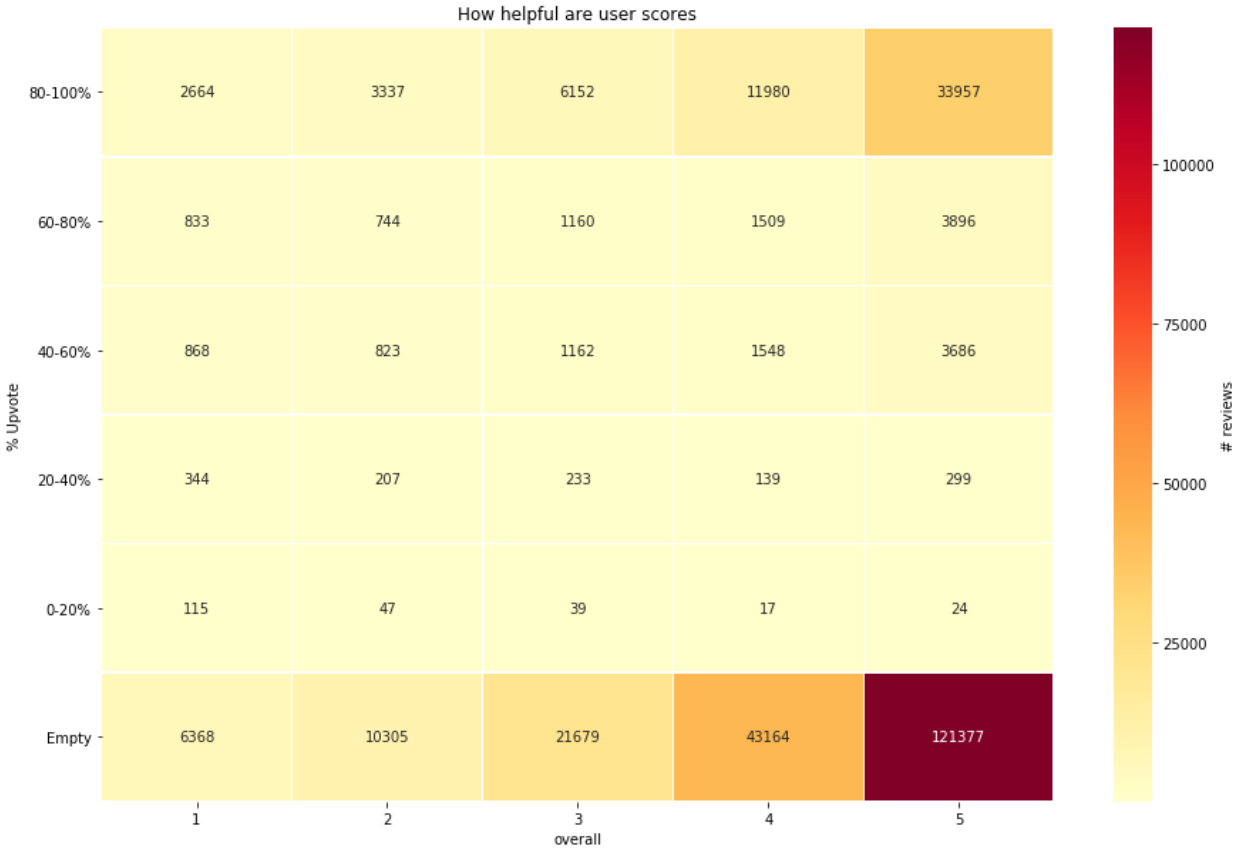
**Models comparison Results** –




**Sentimental Analysis** –

- To understand how customers are feeling about the products, sentimental analysis is performed using the reviews given by customers for the products.
- Adding in new columns to perform efficiency of helpfulness metrics ‘HelpfulnessNumerator’, ‘HelpfulnessDenominator’
- Remove the duplicates if any, based on the reviewerID, productID (asin) and unix timestamp. Adding the upvote metrics to analyze the dataset.
- Adding the helpfulness and upvote percentages for metrics.

## find how useful are the user reviews




We can remove the rating of 3 and convert the reviews into binary, 1 - positive, 0 - negative for better analysis as rating 3 is neither good nor bad and we don't get much out of it

**Modelling –**

**Performing logistic regression on word count**.

# features: 73968

# train records: 186189

# test records: 62063

Model Accuracy: 0.9370639511464157

It is observed that the few words with highest positive and negative coefficient doesn't make sense such as (complaint - 1.836577, worried - 1.806284, goodwill - -2.294153).

**Baseline accuracy of the model is as follows:**

# features: 73968

# train records: 186189

# test records: 62063

Model Accuracy: 0.8101445305576591

TF-IDF vectorizer is added to logistic regression to improve the model accuracy

# features: 73968

# train records: 186189

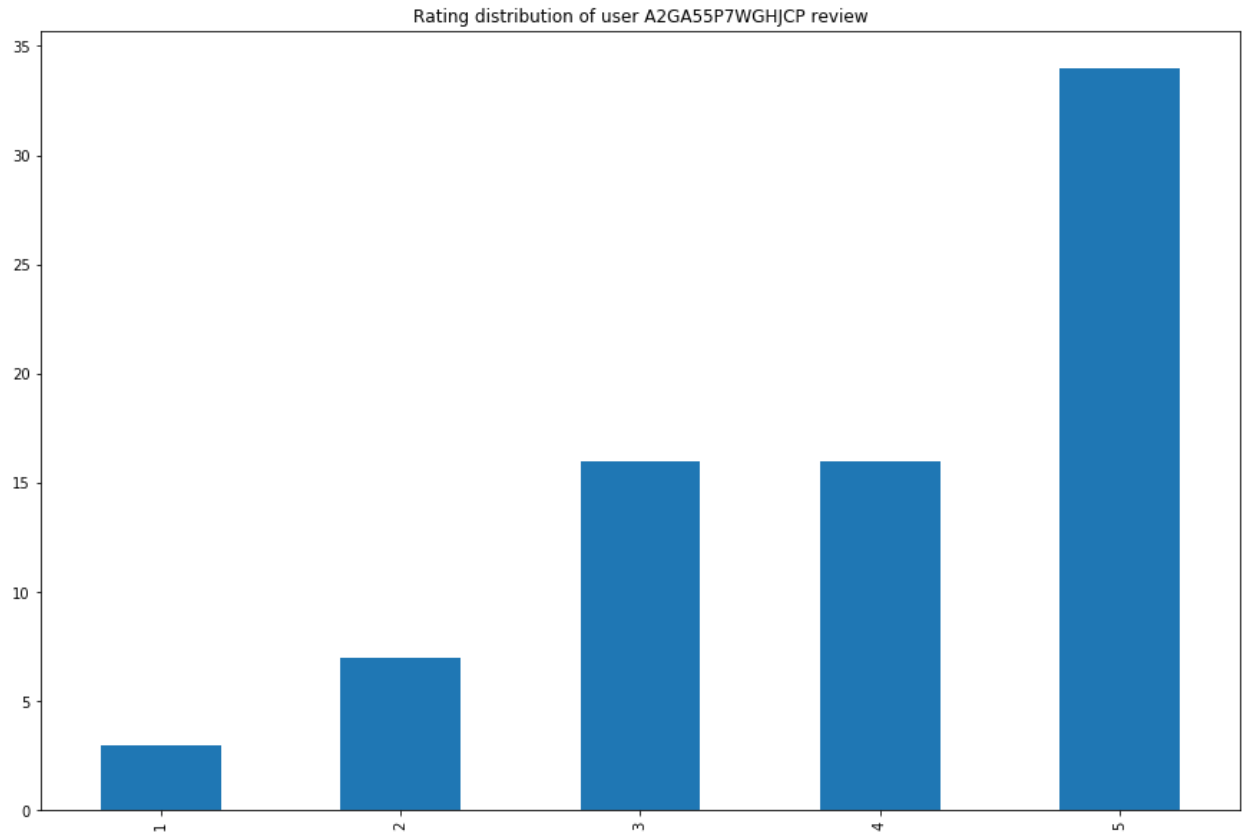
# test records: 62063

Model Accuracy: 0.9383046259446047

Accuracy with **tf-idf** has increased from 81% to 93.8%. It can also be observed that words that don't indicate polarity of the sentiment are removed.

**Study of the user behavior –**

The user behavior must be analyzed to improve the model performance and understand the underlying reasons for the bad or good reviews. This also gives importance to word choices of a user when giving the reviews.




It can be observed that the user is liked of most of the products.

Most popular words used by the user for different ratings are observed. (2-grams and 3-grams are chosen for analysis)

```
score 1 review most popular 2-gram / 3-gram:
     Count Occur %                      Phrase
0      1.0  33.33%         (list, prominently)
1      1.0  33.33%             (slot, abdomen)
2      1.0  33.33%                (month, old)
3      1.0  33.33%     (legs, purchase, large)
4      1.0  33.33%              (skin, slough)
..     ...     ...                         ...
136    1.0  33.33%             (itch, redness)
137    1.0  33.33%       (prominently, amazon)
138    1.0  33.33%  (amazon, incur, liability)
139    1.0  33.33%               (large, look)
140    1.0  33.33%     (death, please, amazon)
```

Now, only the adjectives are taken into consideration, as those express opinion and nouns don't.

score 1 reviews most popular adjectives word:

| | Count | Occur % | Phrase |
|---|---|---|---|
| 0 | 2.0 | 66.67% | large |
| 1 | 2.0 | 66.67% | wear |
| 2 | 1.0 | 33.33% | skin |
| 3 | 1.0 | 33.33% | avoid |
| 4 | 1.0 | 33.33% | chart |
| 5 | 1.0 | 33.33% | -could |
| 6 | 1.0 | 33.33% | big |
| 7 | 1.0 | 33.33% | expensive |
| 8 | 1.0 | 33.33% | latex |
| 9 | 1.0 | 33.33% | tiny |
| 10 | 1.0 | 33.33% | old |
| 11 | 1.0 | 33.33% | upper |
| 12 | 1.0 | 33.33% | much |
| 13 | 1.0 | 33.33% | anaphylactic |
| 14 | 1.0 | 33.33% | smoky |
| 15 | 1.0 | 33.33% | amazon |

score 2 reviews most popular adjectives word:

| | Count | Occur % | Phrase |
|---|---|---|---|
| 0 | 5.0 | 71.43% | good |
| 1 | 3.0 | 42.86% | small |
| 2 | 2.0 | 28.57% | big |
| 3 | 2.0 | 28.57% | large |
| 4 | 2.0 | 28.57% | great |
| 5 | 2.0 | 28.57% | much |
| 6 | 2.0 | 28.57% | thin |
| 7 | 2.0 | 28.57% | long |
| 8 | 2.0 | 28.57% | true |
| 9 | 2.0 | 28.57% | comfortable |

score 3 reviews most popular adjectives word:

| | Count | Occur % | Phrase |
|---|---|---|---|
| 0 | 8.0 | 50.0% | good |
| 1 | 7.0 | 43.75% | nice |
| 2 | 5.0 | 31.25% | fine |
| 3 | 5.0 | 31.25% | comfortable |
| 4 | 4.0 | 25.0% | little |

score 4 reviews most popular adjectives word:

| | Count | Occur % | Phrase |
|---|---|---|---|
| 0 | 11.0 | 68.75% | good |
| 1 | 7.0 | 43.75% | wear |
| 2 | 5.0 | 31.25% | true |
| 3 | 4.0 | 25.0% | top |
| 4 | 4.0 | 25.0% | nice |
| 5 | 4.0 | 25.0% | black |
| 6 | 4.0 | 25.0% | little |
| 7 | 4.0 | 25.0% | fine |
| 8 | 4.0 | 25.0% | comfortable |
| 9 | 4.0 | 25.0% | small |

score 5 reviews most popular adjectives word:

| | Count | Occur % | Phrase |
|---|---|---|---|
| 0 | 27.0 | 79.41% | good |

| | | | |
|---|---|---|---|
| 1 | 13.0 | 38.24% | nice |
| 2 | 12.0 | 35.29% | beautiful |
| 3 | 12.0 | 35.29% | great |
| 4 | 11.0 | 32.35% | black |
| 5 | 11.0 | 32.35% | comfortable |
| 6 | 10.0 | 29.41% | wear |
| 7 | 9.0 | 26.47% | little |