IST 707 PROJECT

Women's E-Commerce Clothing Reviews

Team: Andrea Bradshaw, Neal Bates, John Fields, James Alexander

Class: IST707 Date: 14 Sep 2019

Introduction

Women's clothing in e-commerce is a particularly interesting sector of online shopping as women face a myriad of challenges finding clothes that work well for their body type and fit as expected. Women constantly struggle with sizing variances between brands. A size 0 in one brand can be a size 4 or 6 in another brand. Because women cannot try on clothes purchased online before committing to the purchase, if they do purchase clothes that don't fit, they then must deal with return policies that can be extremely time-consuming and frustrating. As a result, women who want to purchase clothing online are heavily reliant on other customer reviews – particularly the written reviews from other customers indicating 'did it fit as expected', 'was it well made', 'did it bunch anywhere weird', etc.

Customer reviews in general are extremely impactful for purchases. Ninety percent of customers will decide whether to buy a product online or in the store based on online customer reviews. People are literally standing in the aisles of stores across the country looking up reviews on products they're about to purchase. So not only do e-commerce reviews impact e-commerce, but online reviews impact brick and mortar sales as well. In addition, 86% of customers will avoid purchasing a product based on negative reviews.[i]

Over 25% of fashion purchases occur online, therefore, retailers and marketers must adapt from the "brick and mortar" model that was prevalent in the 20th century.[ii] Forester Research predicts that by 2022, e-commerce spending will increase to $765 billion accounting for 36% of the total. With the increase in specialized online fashion retailers (e.g. Net-A-Porter, ASOS, OutdoorVoices) and general e-commerce retailers (e.g. Amazon, Target, Walmart), any company who survives in this fast-changing landscape will need to evolve.

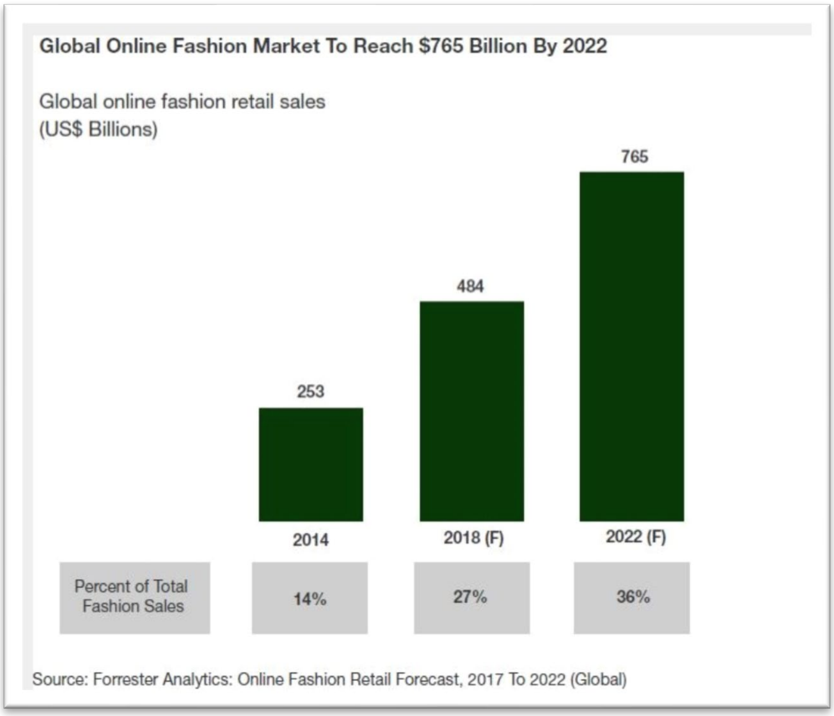


Figure 1

One strategy that can be leveraged by existing e-commerce companies is to mine the wealth of data entered on their websites through customer reviews. The crowd-sourced data science website Kaggle.com has compiled a dataset for these types of customer reviews and made it publicly available. These are actual reviews from an anonymous retailer, and they provide a detailed view into the buying habits and preferences of over 23,000 customers.

The goal for this analysis is to utilize various data analytics techniques to understand how an online retailer could increase sales and market share via the following methods:

1. Text analysis of positive and negative reviews
2. Age range correlations
3. Customer perception of other customer reviews
4. Rating correlations

Analysis and Models

About the Data

Initial Data Set

The Kaggle Women's E-Commerce Clothing Reviews dataset contains 23,486 customer reviews of clothing, shoes and accessories from an anonymous retailer.

Each record in the data set represents a review and associated data from a consumer on a specific item. Here is an example of one of the reviews:

> "Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc it's petite. i bought a petite and am 5'8". i love the length on me- hits just a little below the knee. would definitely be a true midi on someone who is truly petite."
>
> This reviewer is age 34, gave a 5-star rating and recommended the product to others. Her review was also rated as helpful by 4 other people for the dress that was purchased.

Figure 2

The initial data set was downloaded as shown in the chart below. The following missing data entities were identified and dealt with as stated.

- Title – 3810 missing values, substituted with “XTitle”
- ReviewText – 845 missing values – omitted from data set
- DivisionName, DepartmentName & ClassName – 14 missing values – omitted from data set

Attribute	Description	Data Type	Range of Values
ClothingID	Variable that refers to the specific piece being reviewed	Integer	0-1205
Age	The reviewers age	Integer	18-99
Title	The title of the review	Factor	13,993 unique values
ReviewText	The review body	Factor	22,634 unique values
Rating	Variable for the product score granted by the customer from 1 Worst, to 5 Best	Integer	1-5
Recommended	Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended	Integer	0 or 1
PositiveFeedbackCount	Number of other customers who found this review is helpful.	Integer	0 to 122

DivisionName	Product high level division	Factor	4 options
DepartmentName	Product department name	Factor	7 options
ClassName	Product class name	Factor	21 options

Table 1

Data Concerns and Cleaning/Prep

In addition to dealing with blank and missing data, the data type for several variables needed to be changed as shown below:

- ClothingID - character
- ReviewText - character
- DivisionName - factor
- DepartmentName - factor
- Rating - factor

There were also quotation marks and special characters that were discovered in the data and these needed to be removed prior to starting the analysis.

Once these steps were complete, here is a histogram summary of the cleansed dataset prior to creating the corpus for text mining:

Initial Data Analysis Visualizations

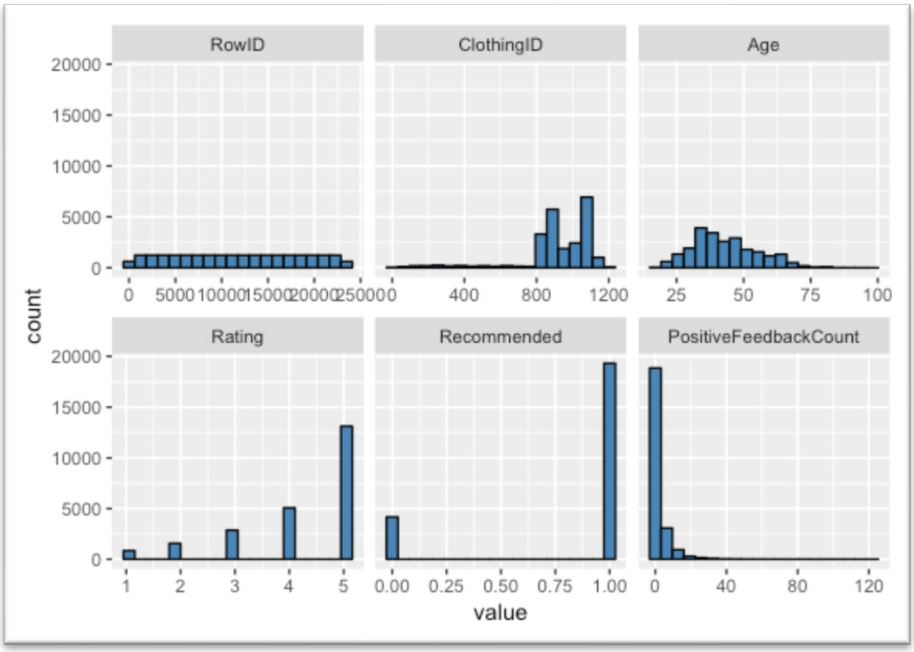
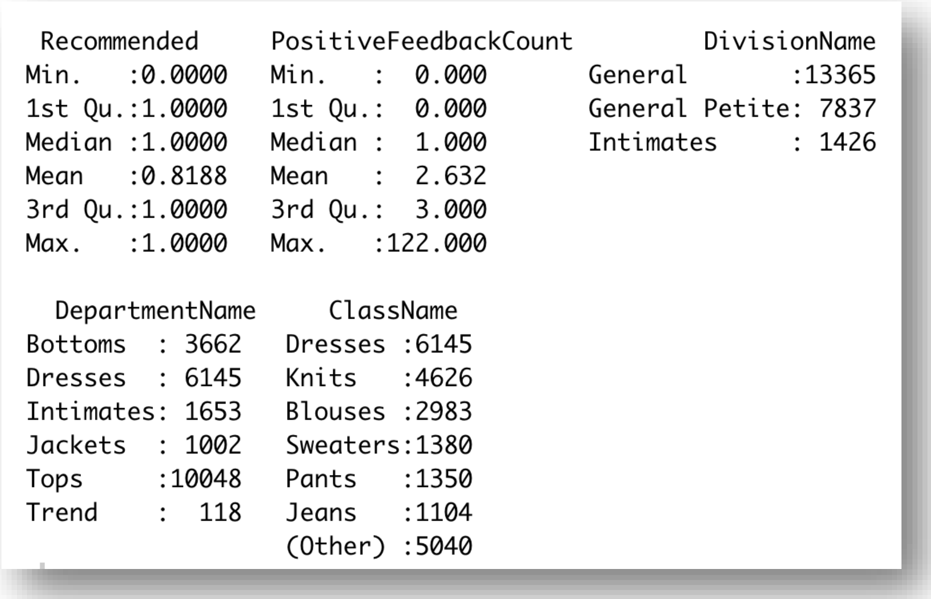


Figure 3

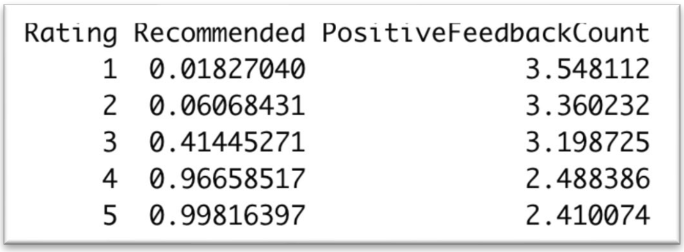
Additional data information is shown in Figure 4. Statistics of interest include a mean of 2.6 for Positive Feedback Count and a Max of 122. The Tops department has the most items mentioned in the reviews with 10,048 and ClassName is led by Dresses with 6145.



```
  Recommended      PositiveFeedbackCount         DivisionName
 Min.   :0.0000   Min.   :  0.000      General       :13365
 1st Qu.:1.0000   1st Qu.:  0.000      General Petite: 7837
 Median :1.0000   Median :  1.000      Intimates     : 1426
 Mean   :0.8188   Mean   :  2.632
 3rd Qu.:1.0000   3rd Qu.:  3.000
 Max.   :1.0000   Max.   :122.000

   DepartmentName     ClassName
 Bottoms  : 3662   Dresses :6145
 Dresses  : 6145   Knits   :4626
 Intimates: 1653   Blouses :2983
 Jackets  : 1002   Sweaters:1380
 Tops     :10048   Pants   :1350
 Trend    :  118   Jeans   :1104
                   (Other) :5040
```

Figure 4

Means of Recommended and Positive Feedback Count also show some correlations between Ratings and Recommended where a Rating of 4 or 5 has a 97%+ chance of being recommended. Also, the highest mean for Positive Feedback Count is for a Rating of 1. This is not surprising as people likely view these negative reviews as "helpful" because they provide feedback on shortcomings of the product related to fit, quality, etc.



```
Rating Recommended PositiveFeedbackCount
     1  0.01827040              3.548112
     2  0.06068431              3.360232
     3  0.41445271              3.198725
     4  0.96658517              2.488386
     5  0.99816397              2.410074
```

Table 2

Looking at Positive Feedback Count by Age Group shows a normal distribution with the 35-44 age group having the highest number of Positive Feedback Counts.

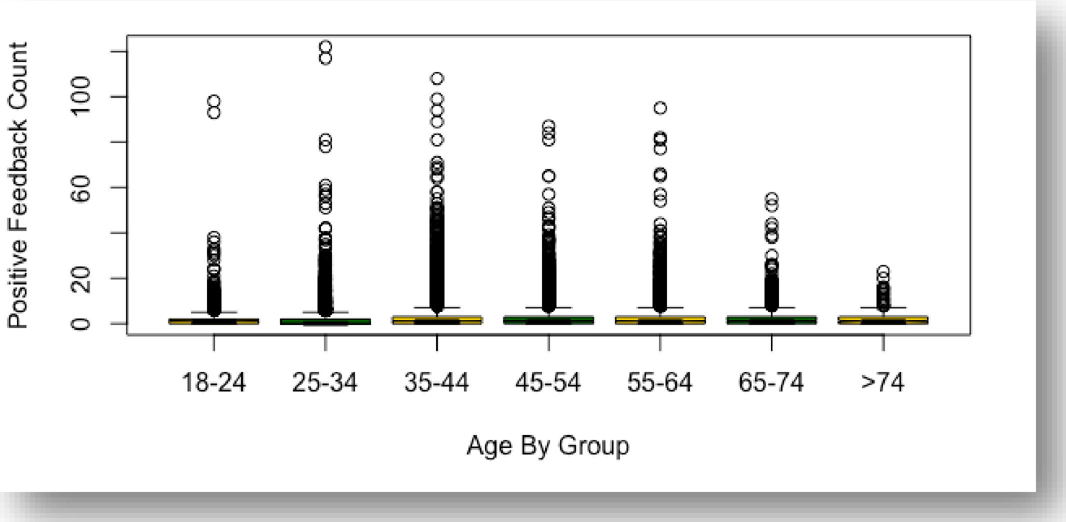


Figure 5

A review of the clothing ID frequency shows that most items are reviewed less than 100 times with some outliers in the 250-1000 reviews range.

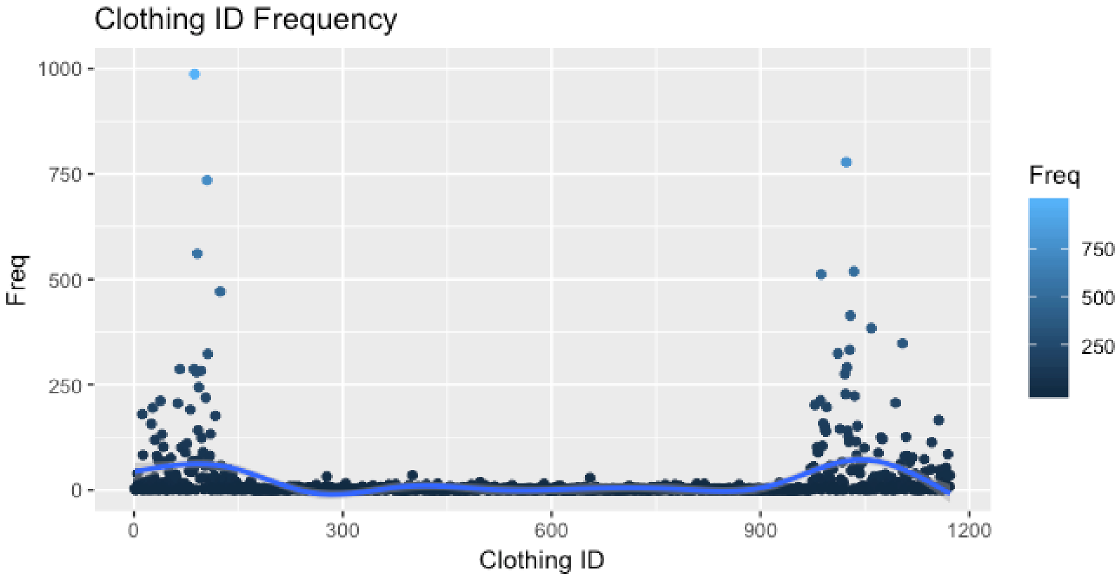


Figure 6

The next step in the data preparation was the creation of a word corpus for the text mining. The following steps were used to complete this task:

1. Merge the Title and Review Text into one variable
2. Remove numbers, punctuation, whitespace and convert to lower case
3. Remove standard "english" stop words
4. Remove additional stop words "like", "very", "can", "I", "also", "lot", "xtitle"
5. Lemmitization was considered but not implemented due to the structure of the data with run on sentences and many misspellings which couldn't be easily corrected in an automated manner
6. Once steps #1-4 were complete, a new clean data frame was created and exported to CSV for a final review.

In preparation for the Association Rule Mining, the following fields were added to bin the values of the originating attribute fields:

- AgeByGroup: "18-24", "25-34", "35-44", "45-54", "55-64", "65-74", ">74"
- RecommendedBins: “YES”, “NO”
- PositiveFeedbackCountBins: "0", "1-25", "26-50", "51-75", "76-100", "101-125", "125+"

On top of using texting mining techniques, there were also significant investments in the non-corpus features as well. These other data features are presented in a record dataset. Table 1 above describes what fields are part of this record set. The fields, ‘DivisionName’, ‘ClassName’, and ‘DepartmentName’ were also transformed to be numeric options through the ‘dummy’ package. Dummies is a data mining technique where a factor column is split into several columns depending on its levels (number of different options) into separate columns. And each individual column is represented with a 1 or 0 to signify if the record is a member of an individual option.

However, most of those data points ended up being removed as there was correlation between these points. This topic will be covered in greater detail in the model and results sections.

One last thing to note about the record dataset. When testing the algorithms, the data was randomly but also evenly split 80% for training and 20% for testing.

Model 1: Text Mining and Sentiment Analysis

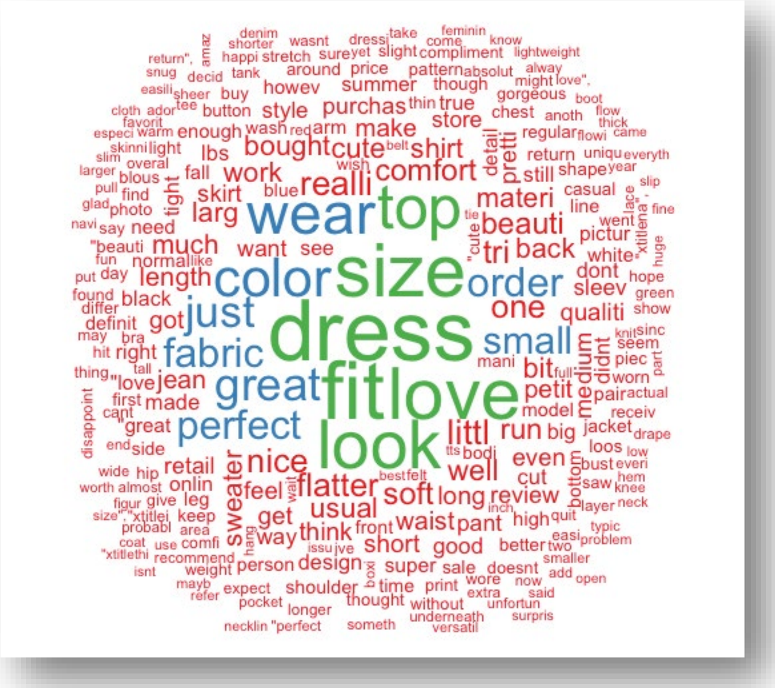
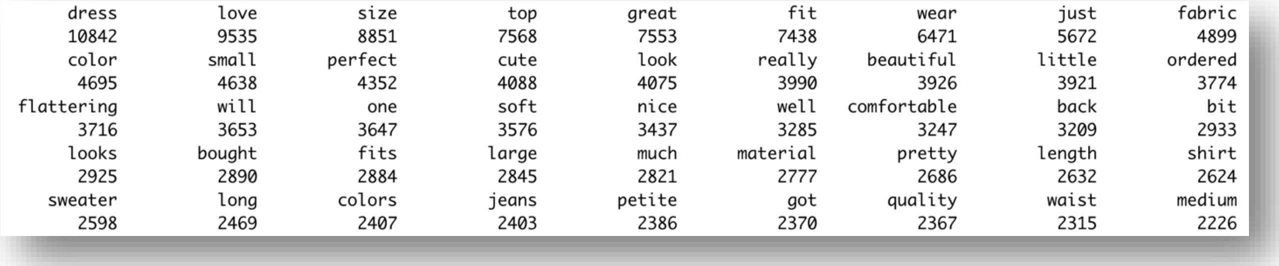


Figure 7

The text mining analysis utilized the combined Title and Review Text from 672,170 words. The word cloud above and Figure 8 below show that the terms dress, love, size, top, great and fit were most common.

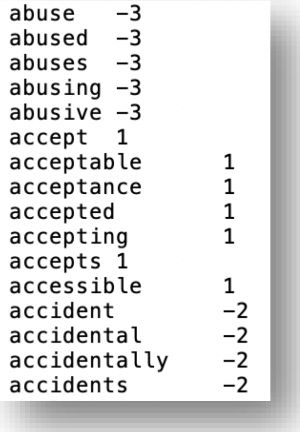


```
     dress        love        size         top       great         fit        wear        just      fabric
     10842        9535        8851        7568        7553        7438        6471        5672        4899
     color       small     perfect        cute        look      really   beautiful      little     ordered
      4695        4638        4352        4088        4075        3990        3926        3921        3774
flattering        will         one        soft        nice        well comfortable        back         bit
      3716        3653        3647        3576        3437        3285        3247        3209        2933
     looks      bought        fits       large        much    material      pretty      length       shirt
      2925        2890        2884        2845        2821        2777        2686        2632        2624
   sweater        long      colors       jeans      petite         got     quality       waist      medium
      2598        2469        2407        2403        2386        2370        2367        2315        2226
```

Figure 8

Sentiment analysis is a relatively new technique which has been applied to social media posts and other text to determine if the person is more positive or negative about a particular topic. For example, this technique has been utilized for posts on Twitter to determine how well political candidates are performing in an election. This technique was applied to the e-commerce reviews to understand the overall positive/negative sentiment and understand the correlations with ratings and recommendations.

There are several sentiment lists available such as Standford, NRC, Bing and AFINN. For this analysis, the AFINN list of 2477 English words was used for a comparison to the review texts. The AFINN list has a rating of +5 to -5 for each of the 2477 words and a sample is shown in Figure 9. This list is imported and then compared with the text to determine overall score of all reviews and by the rating groups of 1 to 5.



```
abuse   -3
abused  -3
abuses  -3
abusing -3
abusive -3
accept  1
acceptable      1
acceptance      1
accepted        1
accepting       1
accepts 1
accessible      1
accident        -2
accidental      -2
accidentally    -2
accidents       -2
```

Figure 9

After creating the text corpus (as described on Page 7) and performing the sentiment analysis on all reviews, the overall sentiment score for all reviews is 196,708. A grouping of the sentiment scores by rating is shown in Figures 10 and 11.

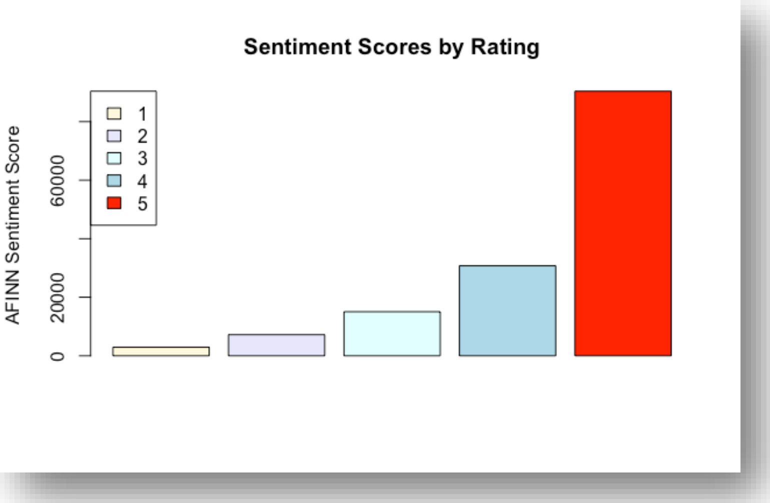


Figure 10

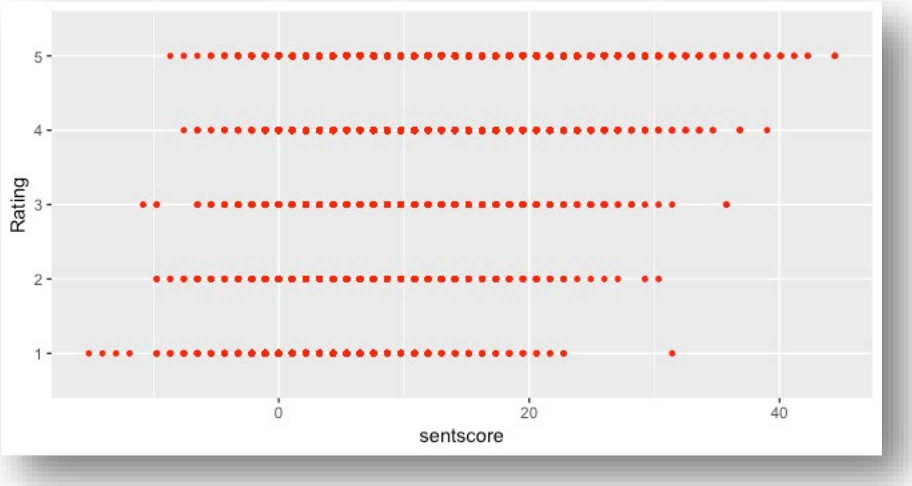
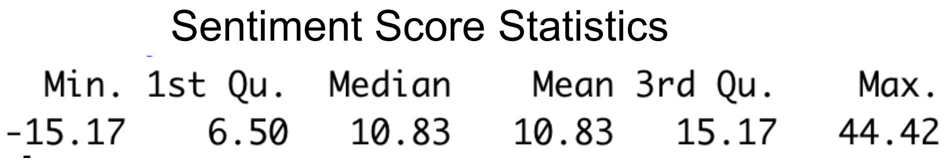


Figure 11

In addition to analyzing the sentiment scores by rating, a function was also written to calculate the sentiment scores by review. The result was a median sentiment score of 10.83 with a minimum of 15.17 and max of 44.42. The sentiment scores by review were also utilized in the decision tree analysis and will be discussed more in this section and the results section.



Sentiment Score Statistics

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-15.17    6.50   10.83   10.83   15.17   44.42 
```

Figure 12

Model 2: Association Rule Mining Analysis

Initial Dataset – First Run

Variables Assessed:

- Age
- Rating
- Recommended
- PositiveFeedbackCount
- Class

Division & Department were not included in the rule generation as the associations with their inclusion provided too many details about associations between departments, divisions and classes.

Association Rule Mining was used to identify 19 strong rules from the cleansed data. Based on the rules generated from the initial rule mining. The maximum frequency an attribute appeared in a record was about 55% of the time. However, the average frequency an attribute appeared in the records was 4.7% of the time. The maximum percentage of accuracy for a rule found to be true was 100% of the time.

The initial rules analyzed were created requiring a minimum frequency of an attribute's appearance in the records 0.1% of the time, and a rule accuracy percentage of 90%. Also, only 4 items could be present in any rule. To obtain the 19 strongest rules, a multitude of approaches were applied altering these factors until the strongest applicable rules were identified. Figure 13 shows the initial rules scatterplot, where **support** is the fraction of the rows of the database that contain all the items in the itemset for the rule. Support indicates the frequencies of the occurring patterns. **Confidence** is the number of times a rule is found to be true; it is often referred to as accuracy. A confidence level of 1 in the graph below indicates 100% accuracy. Finally, **lift** is a measure of the performance of the rule. The larger the number value for lift, the higher the performance. However, all three of these factors were evaluated in parallel to obtain meaningful rules.

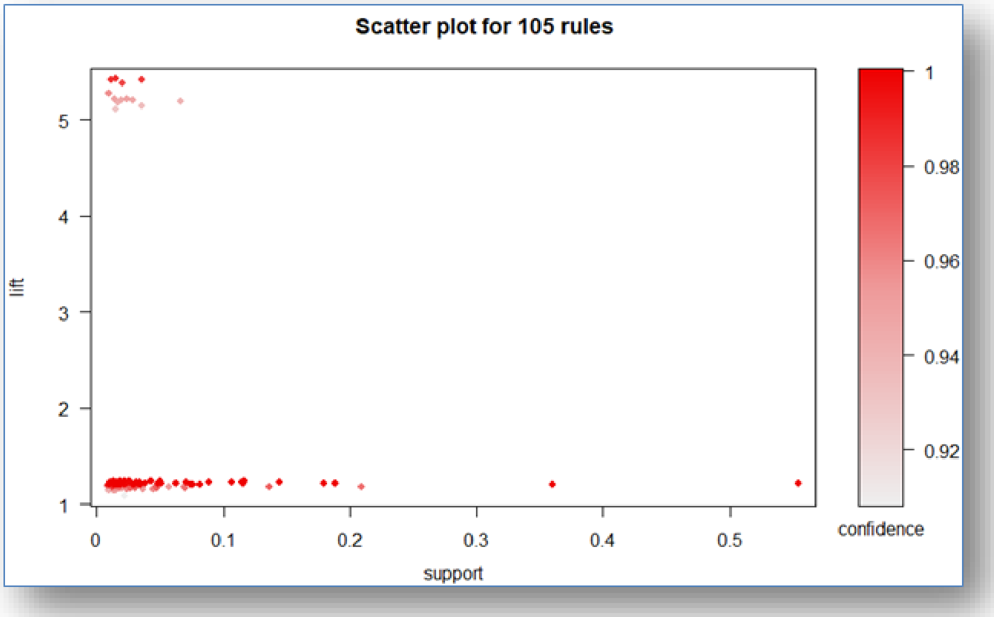
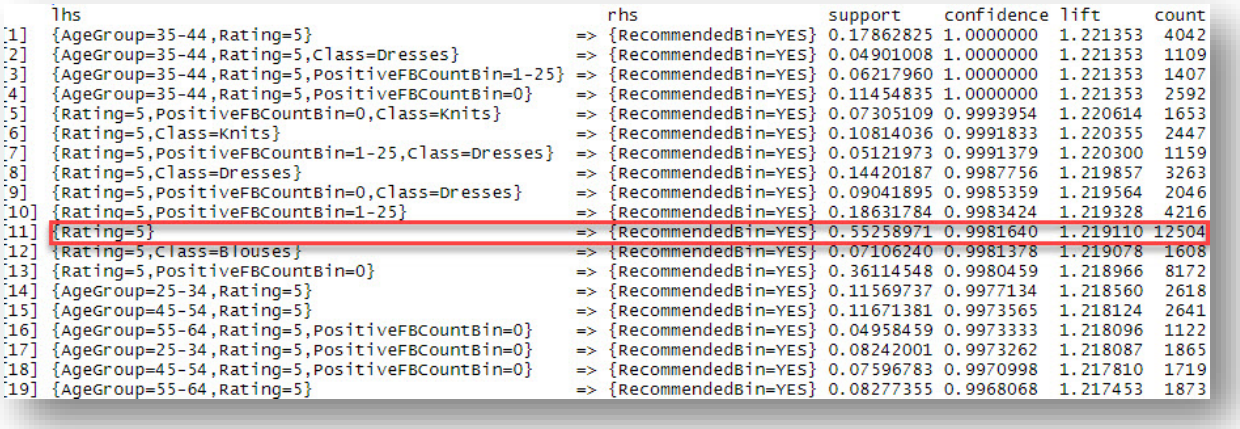


Figure 13

Tuned Dataset – Run 1:

First run generated all rules using highest lift by itself. This generated 105 rules. To better tune the initial model, apriori algorithm was re-run using highest lift again with a minimum support of 4.7% -

the mean support for all rules – and minimum confidence of 98.2% - the mean confidence for all rules. This significantly reduced the number of rules to 19, but limited the evaluation of attributes. Figure 14 shows the 19 generated rules and reveals additional insights to be explored further via a different approach. The lift of all rules was similar enough that the rules were resorted to display in order of descending confidence. This revealed a rating of 5 was strongly associated with Recommended = YES, regardless of other attributes. These rules are too limited to provide insights into questions such as "What attributes are associated strongly with Ratings 1-4?", "What attributes are associated strongly with higher Positive Feedback counts?", or "What attributes are associated strongly with Recommended = NO?". Therefore, additional runs were required to extract this data.



```
     lhs                                                 rhs                        support confidence     lift count
[1]  {AgeGroup=35-44,Rating=5}                        => {RecommendedBin=YES} 0.17862825 1.0000000 1.221353  4042
[2]  {AgeGroup=35-44,Rating=5,Class=Dresses}          => {RecommendedBin=YES} 0.04901008 1.0000000 1.221353  1109
[3]  {AgeGroup=35-44,Rating=5,PositiveFBCountBin=1-25} => {RecommendedBin=YES} 0.06217960 1.0000000 1.221353  1407
[4]  {AgeGroup=35-44,Rating=5,PositiveFBCountBin=0}   => {RecommendedBin=YES} 0.11454835 1.0000000 1.221353  2592
[5]  {Rating=5,PositiveFBCountBin=0,Class=Knits}      => {RecommendedBin=YES} 0.07305109 0.9993954 1.220614  1653
[6]  {Rating=5,Class=Knits}                           => {RecommendedBin=YES} 0.10814036 0.9991833 1.220355  2447
[7]  {Rating=5,PositiveFBCountBin=1-25,Class=Dresses} => {RecommendedBin=YES} 0.05121973 0.9991379 1.220300  1159
[8]  {Rating=5,Class=Dresses}                         => {RecommendedBin=YES} 0.14420187 0.9987756 1.219857  3263
[9]  {Rating=5,PositiveFBCountBin=0,Class=Dresses}    => {RecommendedBin=YES} 0.09041895 0.9985359 1.219564  2046
[10] {Rating=5,PositiveFBCountBin=1-25}               => {RecommendedBin=YES} 0.18631784 0.9983424 1.219328  4216
[11] {Rating=5}                                       => {RecommendedBin=YES} 0.55258971 0.9981640 1.219110 12504
[12] {Rating=5,Class=Blouses}                         => {RecommendedBin=YES} 0.07106240 0.9981378 1.219078  1608
[13] {Rating=5,PositiveFBCountBin=0}                  => {RecommendedBin=YES} 0.36114548 0.9980459 1.218966  8172
[14] {AgeGroup=25-34,Rating=5}                        => {RecommendedBin=YES} 0.11569737 0.9977134 1.218560  2618
[15] {AgeGroup=45-54,Rating=5}                        => {RecommendedBin=YES} 0.11671381 0.9973565 1.218124  2641
[16] {AgeGroup=55-64,Rating=5,PositiveFBCountBin=0}   => {RecommendedBin=YES} 0.04958459 0.9973333 1.218096  1122
[17] {AgeGroup=25-34,Rating=5,PositiveFBCountBin=0}   => {RecommendedBin=YES} 0.08242001 0.9973262 1.218087  1865
[18] {AgeGroup=45-54,Rating=5,PositiveFBCountBin=0}   => {RecommendedBin=YES} 0.07596783 0.9970998 1.217810  1719
[19] {AgeGroup=55-64,Rating=5}                        => {RecommendedBin=YES} 0.08277355 0.9968068 1.217453  1873
```

Figure 14

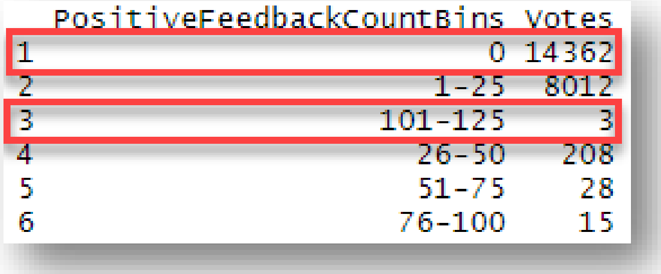
Secondary Runs – Rating Associations, Positive Feedback Counts, and Recommended = NO.

Rating Associations

First, apriori was re-run using Ratings = 1 – 5 were set for a lhs parameter in the algorithm to determine the attributes associated with these ratings. To ensure at least one rule was encountered for each rating, the support was initially lowered to .01% and the confidence was set to 5%. See **Results** Figure 36.

Positive Feedback Count Associations

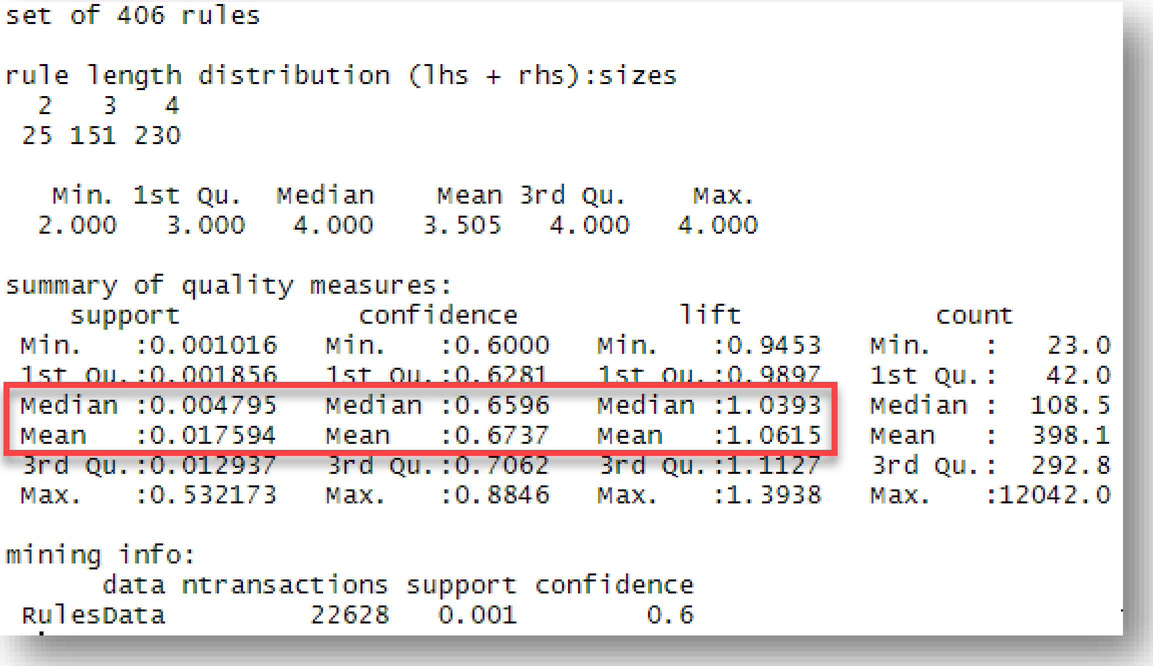
Next, the breakdown of the Positive Feedback Counts was evaluated to determine the breakdown of customer votes on review feedback. Figure 15 shows the breakdown of the Positive Feedback Count Bins.



```
  PositiveFeedbackCountBins Votes
1                         0 14362
2                      1-25  8012
3                   101-125     3
4                     26-50   208
5                     51-75    28
6                    76-100    15
```

Figure 15

As a result, it was determined to evaluate only the Positive Feedback Counts = 0, and 101-125. Next rules were generated once again using each of the PositiveFeedBackCountBins = 0 in the RHS using a low initial support of .1% and confidence of 60% ordered by lift descending. This generated 406 rules with a summary shown in Figure 16.

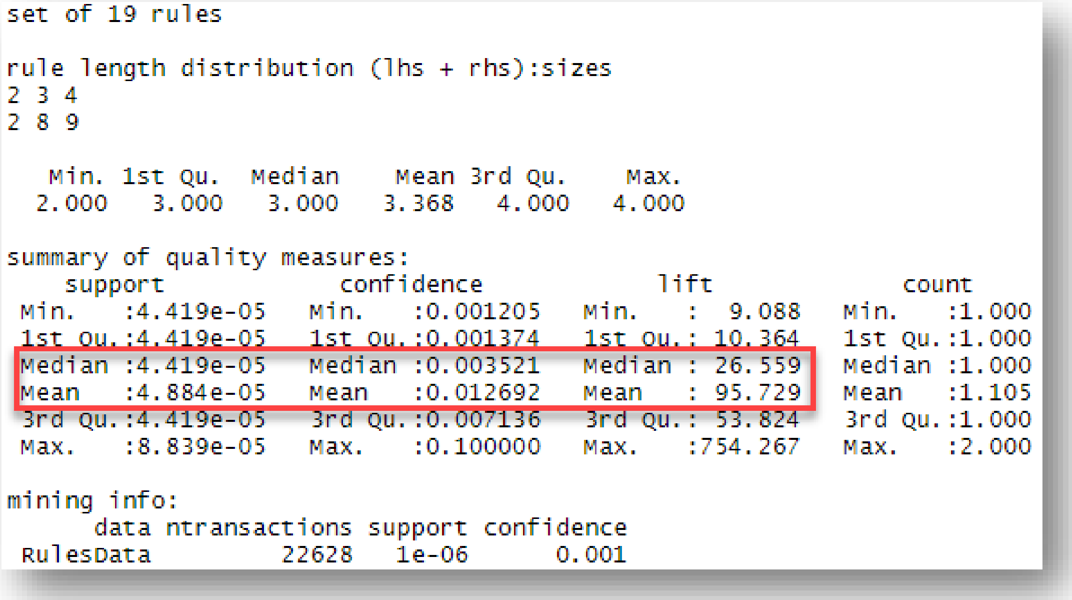


```
set of 406 rules

rule length distribution (lhs + rhs):sizes
  2   3   4 
 25 151 230 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   3.000   4.000   3.505   4.000   4.000 

summary of quality measures:
    support           confidence          lift            count        
 Min.   :0.001016   Min.   :0.6000   Min.   :0.9453   Min.   :   23.0  
 1st Qu.:0.001856   1st Qu.:0.6281   1st Qu.:0.9897   1st Qu.:   42.0  
 Median :0.004795   Median :0.6596   Median :1.0393   Median :  108.5  
 Mean   :0.017594   Mean   :0.6737   Mean   :1.0615   Mean   :  398.1  
 3rd Qu.:0.012937   3rd Qu.:0.7062   3rd Qu.:1.1127   3rd Qu.:  292.8  
 Max.   :0.532173   Max.   :0.8846   Max.   :1.3938   Max.   :12042.0  

mining info:
      data ntransactions support confidence
 RulesData         22628   0.001        0.6
```

Figure 16

As a result, the support was increased to the mean value of .1760 and the confidence was increased to .6737 to tune the data. The rules were generated again. See **Results** section Figure 37.

The rules were generated again with PositiveFBCountBins = 101-125 in the RHS with a very low support of .00001 and confidence of .001 as there were very few of these records in the dataset. The summary of this dataset is shown in Figure 17.



```
set of 19 rules

rule length distribution (lhs + rhs):sizes
2 3 4
2 8 9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.368   4.000   4.000

summary of quality measures:
    support             confidence             lift              count
 Min.   :4.419e-05   Min.   :0.001205   Min.   :  9.088   Min.   :1.000
 1st Qu.:4.419e-05   1st Qu.:0.001374   1st Qu.: 10.364   1st Qu.:1.000
 Median :4.419e-05   Median :0.003521   Median : 26.559   Median :1.000
 Mean   :4.884e-05   Mean   :0.012692   Mean   : 95.729   Mean   :1.105
 3rd Qu.:4.419e-05   3rd Qu.:0.007136   3rd Qu.: 53.824   3rd Qu.:1.000
 Max.   :8.839e-05   Max.   :0.100000   Max.   :754.267   Max.   :2.000

mining info:
      data ntransactions support confidence
 RulesData         22628   1e-06      0.001
```

Figure 17

To identify the strongest associations with Positive Feedback Counts from 101-125, apriori was run one last time with the confidence increased to the mean of .0127. See **Results** section Figure 38.

Recommended = NO Associations

To determine attributes associated with RecommendedBin=NO, apriori was run again with a support value of 1% and confidence of 90%. See **Results** Figure 39.

Model 3: K-Means Clustering Analysis

Variables Assessed:

- Rating
- Age
- Recommended
- PositiveFeedbackCount

K-means was used to cluster the review records into Ratings of 1-5. Centers = 5 was used as there were 5 ratings into which the data was to be clustered. The measurement type used for K-means was Euclidean, however measurement did not affect the clustering. Several different centroids were tested, and none provided improved clustering of reviews with a 3 rating. See **Results** Figure 40.

Model 4: Support Vector Machine Analysis

Support Vector Machines (SVM) uses the training data to identify the various groupings of data points that predict if the product is recommended. SVM looks for the attributes like age, or affinity, etc. data points that can be separated to determine the dependent variable and it draws a linear line (or plane) between the different features.

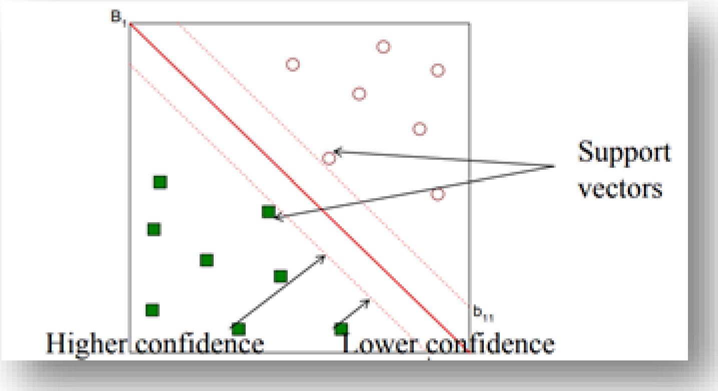


Figure 18[iii]

Figure 18 is demonstrating how an SVM draws a line between two classifications. The image also shows that points that are further from the line have a higher confidence in predicting the correct classification. SVMs do not rely on distance measures, which is common for other data mining algorithms, but on a method known as kernel.

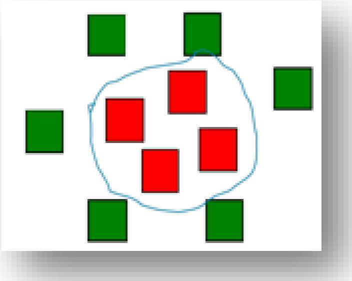


Figure 19[iv]

Figure 19 is representing datapoints that are not Linearly Separable. Hence; there is no linear process that can separate those lines. That would be an issue for SVM as it draws a linear line or plane, to separate the groupings. This dilemma is where a kernel comes into play, a kernel uses other mathematical algorithms to transform the data points from non-linear dataset to a linear dataset[v]. In this paper's SVM Model, Linear kernel preformed the best for the dataset. But Radial, Polynomial, and Sigmoid where also tested for kernels along with different parameters tuned.

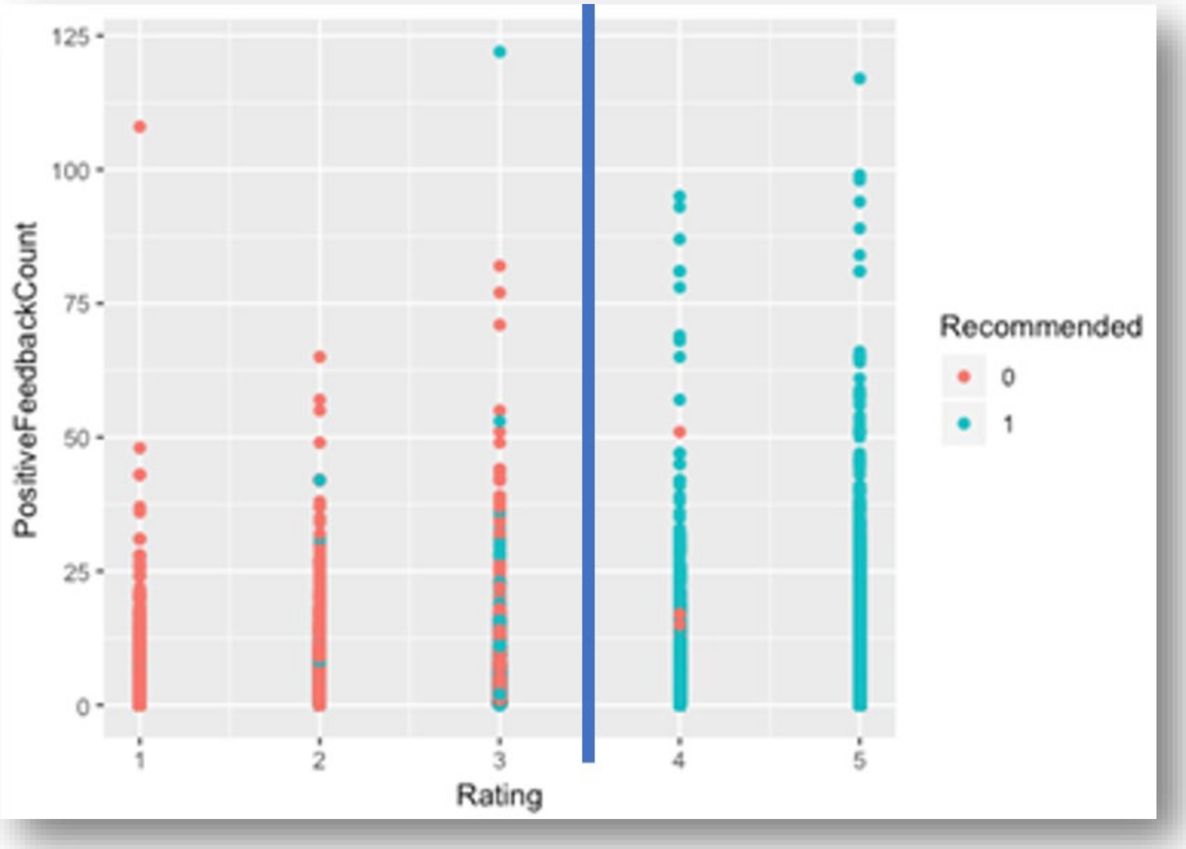


Figure 20

Figure 20 is an example of how an SVM might draw a line for a feature. The image is demonstrating the almost linear relationship between product 'Rating' and product 'Recommended'. In other words, if a rating for a product is greater than '3' it will most likely be Recommended and the inverse if the rating is below a '3' as demonstrated by the **blue** line.

A Linear Kernel preformed the best for SVM. Radial, Polynomial, and Sigmoid where also tested kernels along with different parameters tuned. However due to the length of the paper the other models and results will not be discussed besides the best model. The trimmed feature dataset, known as the 'numeric' dataset with a Linear Kernel, was the most accurate of all models, for if a product would be recommended or not.

Model 5: Naïve Bayes Analysis

Naïve Bayes (NB) uses the training data to identify the trends in probabilities for the independent variables in order to predict if the product is recommended.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Equation 1

Equation 1 is showing the Naïve Bayes (NB) probability calculation. P(c|x) is stating that given the x (independent variables) what is the probability of c (dependent variable) is expected to occur. P(x|c) is the probability of x existing when c is present. Or in other words what is the probability of x and c are present together. P(c) is the probability of c occurring in the dataset and P(x) is the probability of x occurring in the total dataset. What this means for this example is that NB calculate the probability of a product being recommendation by summing up its total probabilities by each individual field's confidence. This infers that there will be distinct patterns in the data points that are not present in other review types and that there is not a strong correlation between independent variables.

Naïve Bayes are models that predict the outcome utilizing the probabilities of the historical/structured data. The one drawback to Naïve Bayes is that it 'naively' assumes that the independent variables are independent of one another. Hence, removing fields that are highly correlated will assist a NB's accuracy.

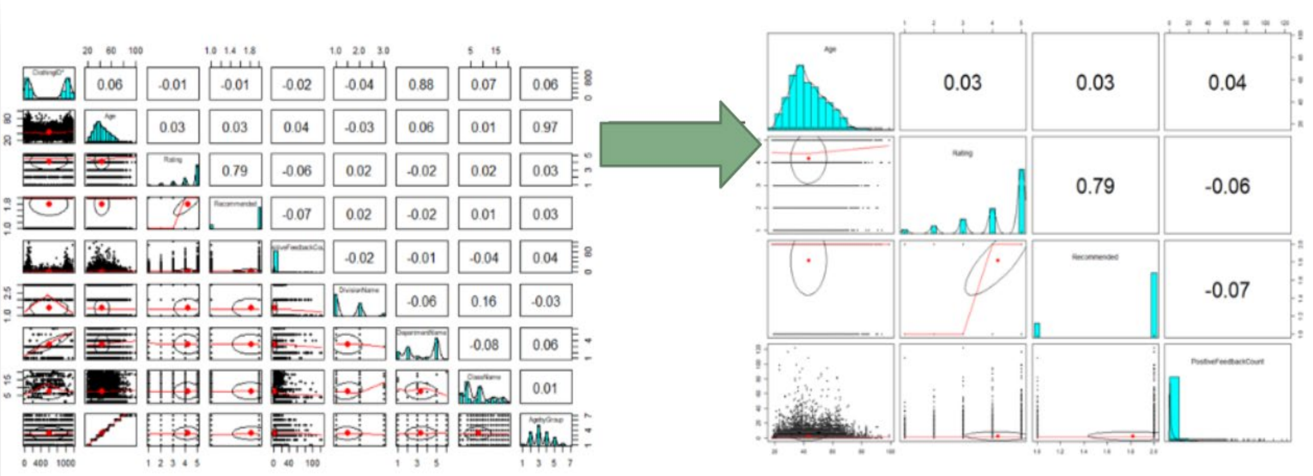


Figure 21

Figure 21 is displaying a feature selection effort. As stated in the last paragraph, Naïve Bayes preforms best when the independent variables are not correlated. Hence, fields 'DivisionName', 'ClassName', and 'DepartmentName' are all correlated to each other. On a similar note, fields that have no impact on dependent variable can also be removed. This feature selection effort ended with a subset of fields that had significant impact to determining if a product is recommended or not and ended with dataset of 'numeric'. In other words, these are the fields that are transformed into numeric values for SVM, are independent to each other, and has the greatest impact to the dependent variable.

The best preforming Naïve Bayes model is one that uses laplace of 1 and does not use kernels, and with the 'numeric' dataset. Naïve Bayes preforms over 2% better when the numeric dataset is utilized. That is probably since removing the correlated and non-impactful fields reduce the noise that NB must determine.

Model 6: Random Forest Analysis

Random Forest (RF) is a type of Ensemble Learning technique. RF looks at the dependent variable and makes a prediction, and tests its results looking at results match the dependent variable. It will then test its predictions and retrain the wrong tests in a process called, 'boosting'[vi]. Random Forest is known as an Ensemble Learning technique since it deploys multiple different techniques to predict and constructs its opinion by how many of those hypotheses are correct.

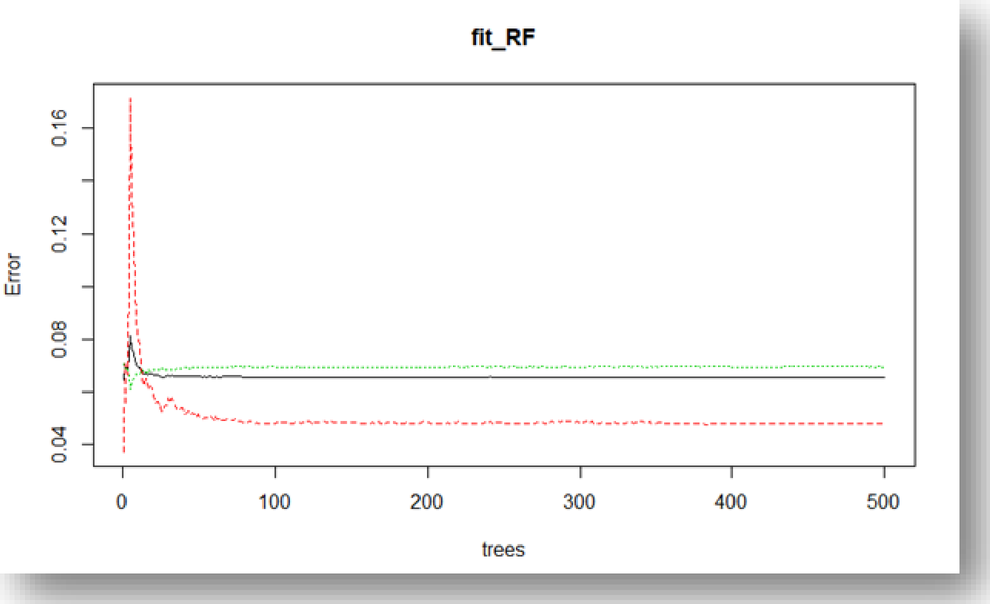


Figure 22

Random Forests operate differently from SVM and Naïve Bayes as there are not a distance measure or kernel to change. However; there are a number of variables that can be tuned to help a RF predict. Figure 22 is showing the error rate of each classification (recommended or not) and how those are being smoothed out throughout the number of trees used (decisions) and maturity. What is noticeable is that the green and red lines (those two numbers) are consistently accurate by 50 trees. Tuning will be covered in greater detail in the results but a maturity of 1 and 200 trees ended up helping RF predict the best. Technically the more trees used the better the overall prediction but after 200 trees less than 0.01 changes in the prediction. Meaning that after 200 trees, there is a huge point for diminishing returns.

Model 7: Decision Trees Analysis

A strength of decision tree analysis is the ability to predict an outcome, like ordering more medical tests vs. ruling out certain diagnosis. In this analysis decision trees reacted better when information gain is maximized to predict a yes/no, utilizing factors rather that raw numbers. Here is a diagram of how to read a decision tree.[vii]

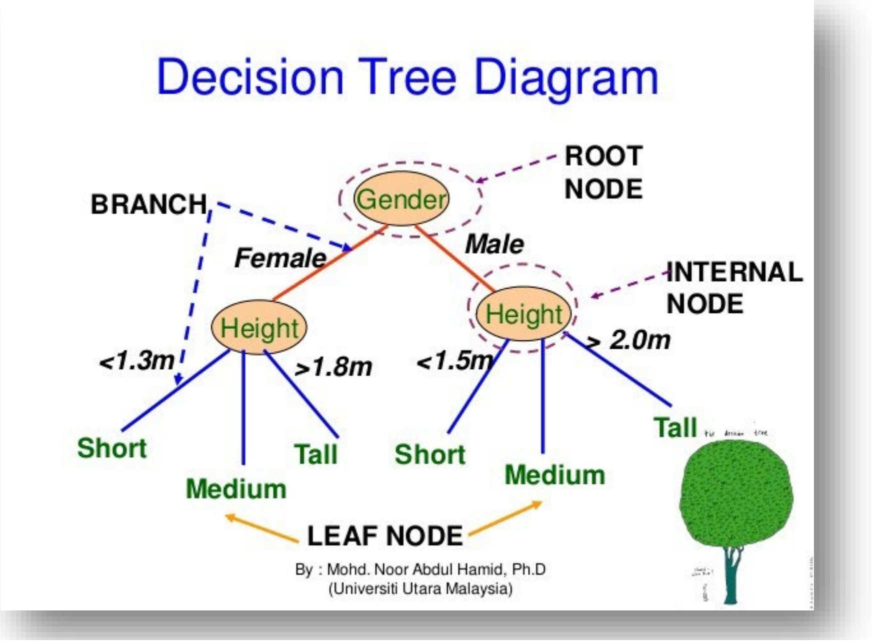


Figure 23

The approach for the Decision Tree (DT) algorithm was setup in such a way to test theories about the e-commerce data. A hypothesis test was used to see if a DT could determine if the purchaser recommendations and/or the product ratings are most valuable and will influence customers. What part of a customer review is most valuable? The data set includes a positive feedback button count and a decision tree is a possible approach to determining what is most likely to result in a helpful review. Just 37% of all reviews received a thumbs up. Yet some of those reviews received more positive feedback.

Decision trees models were created in order to meet the following objectives:

1. Determining what a consumer will do. Given we do not have sales, we will predict the closest variable that influences sales.

2. Identifying key features that influence purchases and recommendations.

3. Identifying the potential to introduce collaborative filtering based on the existing dataset.

The following are the labels (aka dependent variables) predicted by three decision trees:

1. The features that result in a positive rating.

2. The independent variables that determine positive feedback count.

3. The columns that are correlated to product preferences.

As mentioned in the data analysis section, the following are assumptions based on research of the category examination of the data:

1. PositiveFeedBackCount represents a thumbs up for that review in terms of it being helpful to someone reading the review. It's a count of the number of times the review received a thumbs up.
2. Written reviews and scores are from purchasers.
3. Rating is the topline review graphic on e-commerce sites like Amazon and thus it is most quickly influences a customer.
4. Positive reviews - especially RATING - equate to more sales.

Additional Cleaning/Prep/Pruning

As part of Decision Tree hypothesis testing, there were better results by binning sentiment scores. The following bins were generated for Sentiment Score by increasing the density for each bin. Three and then four larger breaks were created and the models run, to decrease the number of tree leaves and branches. The resulting bell-shaped distribution was more even around the mean of 10.8.

Sentiment Bins

Bins	Frequency
0-5	4200
6-10	7070
11-15	5260
16+	608

Table 3

Data frames were created for each of the three trees identifying the label and the independent variables for the hypothesis test. The data frames for each tree were updated to include only the most relevant variables as part of the pruning process. Narrowing the list was based on trial and error to prune the trees to maximize information gain accuracy and clarity. Here are the columns used in the data tree data frames:

A. "What Impacts Rating" ("Rating", "SentBins", "DepartmentName")

B. "What influences Positive Feedback?" ("PositiveFeedbackCountBins", "Rating", "RecommendedBins", "sentscore")

C. "What Age Purchases each Class of Clothing” ("AgebyGroup","ClassName")

Indices were created for train and testing data and then a separate data frame for each tree and its train and test sets. For example:

```
> nobs <- nrow(RtgTree)
> nobs
[1] 22628
> set.seed(6701)
> train.indices <- sample(nobs, 0.7*nobs)
> rtg.train.indices <- setdiff(1:nobs, train.indices)
> Rtgtrain<- RtgTree[rtg.train.indices,]
> Rtgtest<- RtgTree[-rtg.train.indices,]
```

Figure 24

Decision trees were created using the rpart model and the trees were rendered using Rattle. For example:

```
> rpart_modelC <- rpart(ClassName~ ., data = Agetrain, method = "class", model = TRUE, control = 
rpart.control(cp = 0))
> fancyRpartPlot(rpart_modelC, uniform=T, cex=.9, space=0, tweak=0.5, gap=0)
```

Figure 25

Error and accuracy were calculated by first using the model to create a confusion matrix (Figure 26) and then using the following formulas shown in Figure 27. Note this matrix produced a high error rate in as his tree predicted lots of 5s which was the chosen root node for the tree.

	Actual				
Predicted	1	2	3	4	5
1	3	0	3	2	2
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	249	480	847	1469	3724

Figure 26

```
> #Confusion Matrix or a contingency table for the train data
> print(tab.A.trn)
> #Error Rate
> 1-sum(diag(tab.A.trn))/sum(tab.A.trn)
[1] 0.4502872
> #Accuracy
> sum(diag(tab.A.trn))/sum(tab.A.trn)
[1] 0.5497128
```

Figure 27

Three trees A, B, C, were created and each were tested and pruned to maximize their performance.

Decision Tree A: Determining a 5 Rating

Sentiment was the biggest determinant of a rating of 5 with the cut off at sentiment below 6. Between 6 and 10 sentiment, a 5-star rating was determined by the department name. Dresses, tops, and trends are less tolerant of low sentiment and still receiving a 5 rating. While sentiment is correlated with rating, reviews are likely to have words with less positive sentiment reflecting a comparative product or difficulty with sizing a 5 -rated garment. This tree achieved an accuracy rate of 55%.   See Figures 26 & 27 for the confusion matrix and accuracy calculation.

Five-Star Rating Decision Tree

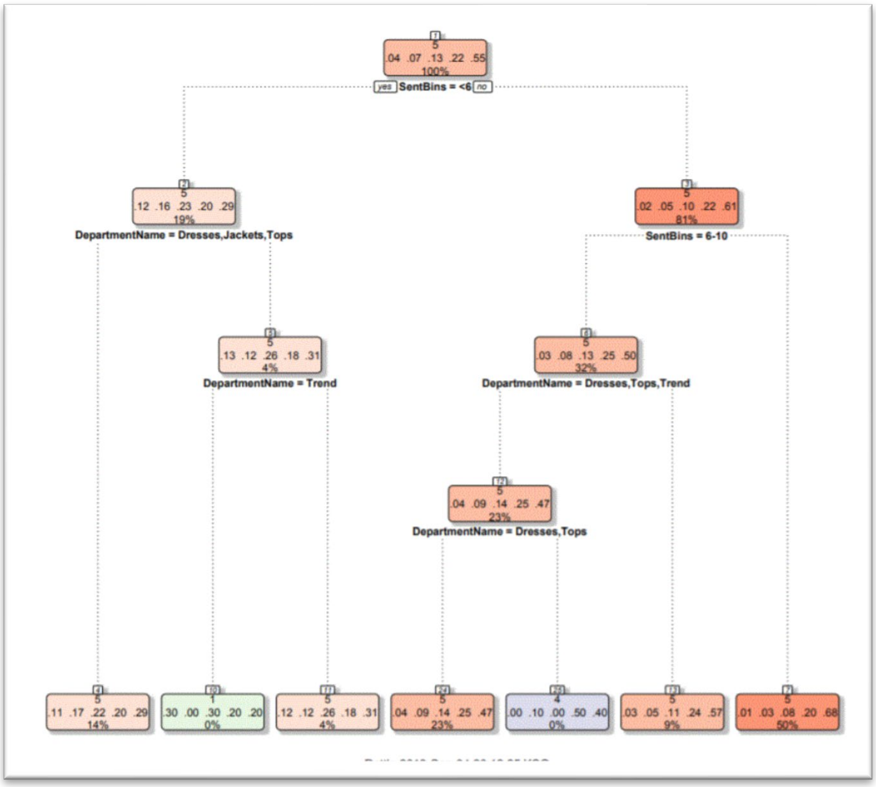
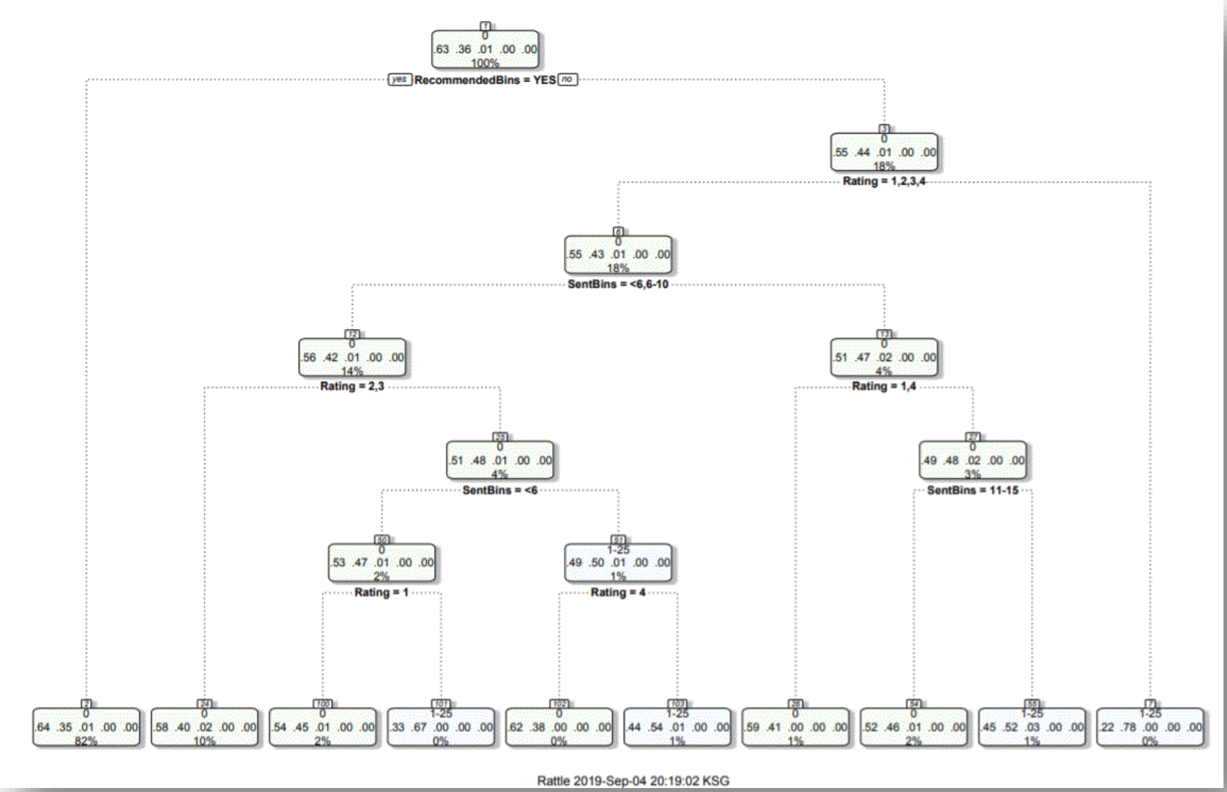


Figure 28

<u>Decision Tree B: Determining the Type of Review Receives the Most Positive Feedback</u>

This second tree was narrowed to just include rating, recommendation, and sentiment score. Without pruning, the tree was detailed, but not clear enough to present to a business audience. Here is the tree using the pruning function rpart control to the level “control = rpart.control(cp = .0009).” Positive feedback is resulting from the branches to the left. For example, 82 percent of the cases with positive feedback came with a recommendation.

Positive Feedback Decision Tree



Rattle 2019-Sep-04 20:19:02 KSG

Figure 29

With some further pruning this tree becomes clearer. This turned out to be the most helpful of the three trees and it had a higher accuracy rate of 63% and its featured in the results section below.

Here is the confusion matrix for decision tree B.

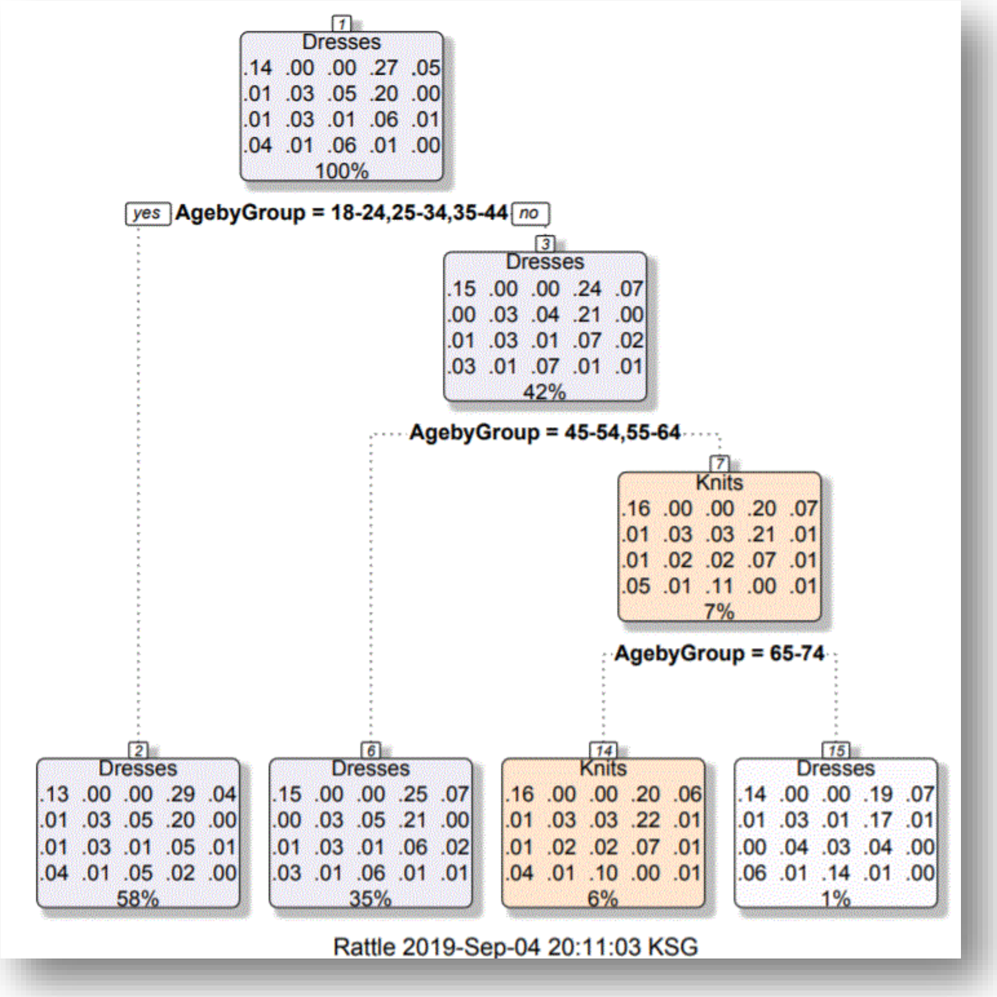
	Actual						
Predicted	0	1-25	26-50	51-75	76-100	101-125	125+
0	4173	2352	59	7	6	0	0
1-25	83	105	4	0	0	0	0
26-50	0	0	0	0	0	0	0
51-75	0	0	0	0	0	0	0
76-100	0	0	0	0	0	0	0
101-125	0	0	0	0	0	0	0
125+	0	0	0	0	0	0	0

Table 4

Decision Tree C: Are Age and Purchased Product Related

A powerful marketing tool for some popular e-commerce sites is collaborative filtering. This allows prioritization of real time product promotion in the process of making a purchase. “Purchasers of X also purchased product Y”. Collaborative filtering of products to enhance the cross-sell and up-sell opportunities is facilitated using a customer identifier and association rule mining of products. This trees purpose was to search for a common descriptor of purchasers such as demographics that could suit this purpose thus allowing for the collaborative filtering in absence of a customerID in this dataset. The conclusion was that collaborative filtering was not possible for decision tree analysis and here is why.

There were very few differences in class by age, or age by clothing class and the error level was 73%. This was the case for larger and pruned trees. There was some definition with knits, perhaps more often purchased by the young and the old, <45 years and > 64 years. It is interesting to let the decision tree parse all the data but with accuracy at just 27% and little information beyond this the null hypothesis proved to be true. There is not enough accuracy in defining which garments are purchased according to age to do and effective association rule mining and collaborative filtering of garments.



Rattle 2019-Sep-04 20:11:03 KSG

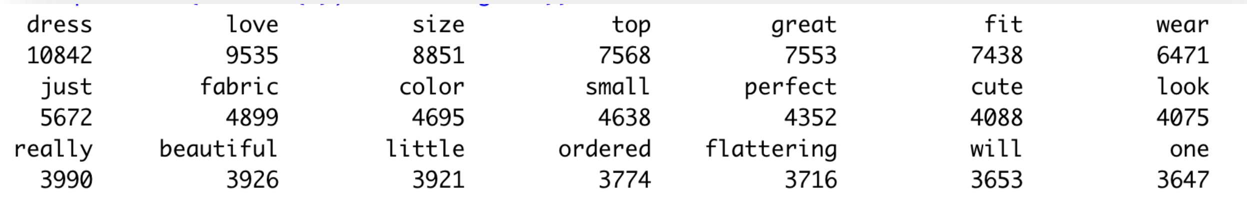
Figure 30

Results

Model 1: Text Mining & Sentiment Results

The results of the text mining found words grouped in the following categories:

- Product - dress, top
- Opinion about the product - love, great, small, perfect, cute, beautiful, little, flattering
- Action words related to the product - wear, ordered



```
 dress          love          size           top         great           fit          wear
 10842          9535          8851          7568          7553          7438          6471
  just        fabric         color         small       perfect          cute          look
  5672          4899          4695          4638          4352          4088          4075
really     beautiful        little       ordered    flattering          will           one
  3990          3926          3921          3774          3716          3653          3647
```

Figure 31

The sentiment scores showed the expected correlation between the Ratings and the Sentiment Score. For example, the ratings of 1 had a mean Sentiment Score of 3.9 while a rating of 5 had a mean Sentiment Score of 12.7.   A more unexpected result was the Ratings of 1 with some high sentiment scores and a Rating of 5 with some low sentiment scores as shown in Figure 32.

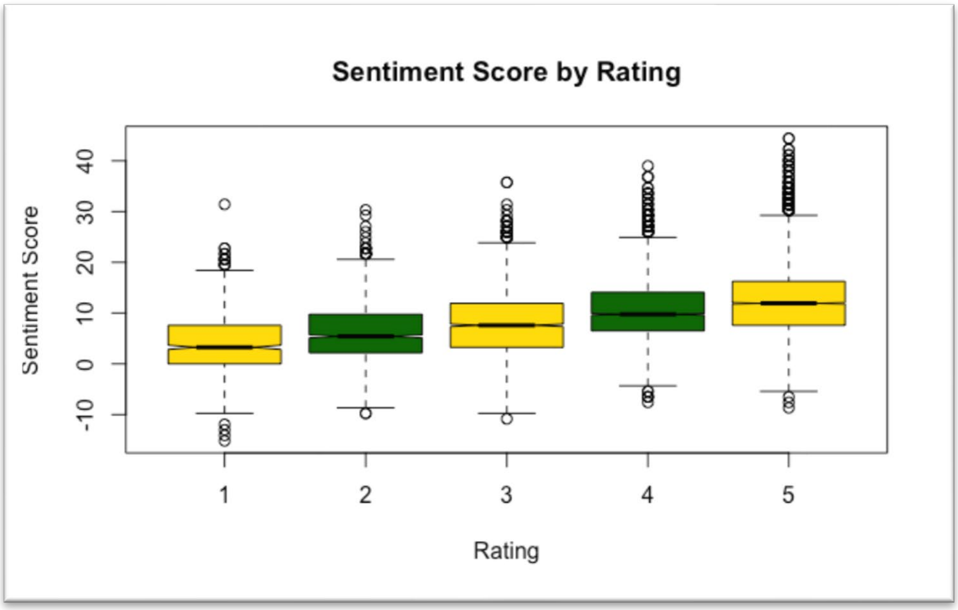
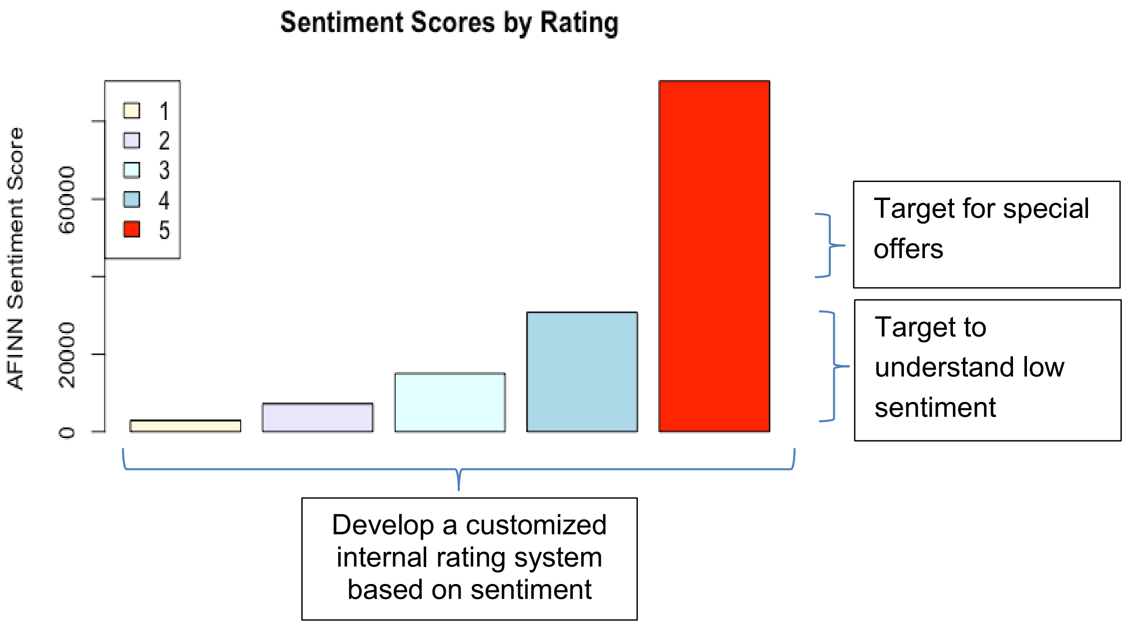


Figure 32

A deeper investigation into this data shows how this information could be used by the retailer to better understand:

1. Why consumers would give a rating of 5 and then give a product review that is negative? This could provide insights on hidden customer service issues where additional marketing to this person could result in more positive reviews.
2. The sentiment score data could be utilized to create a customer internal scoring that should provide additional insights compared to the more subjective 1-5 ratings from consumers. Figure 33 shows an example of how customers with a sentiment score of 5 could be segmented to market to customers with low/medium/high sentiment scores.
3. The text of those customers with a rating of 3 could be mined in more detail to better understand how to resolve issues like fit, quality and product selection which may be the reason this group is ambivalent about recommending a product.
4. A deeper investigation into the text of Ratings 1 and 2 may reveal true issues experienced by customers but could also be fake reviews inserted by an unscrupulous competitor. New techniques such as classification using transformers (BERT, XLNet) which could be applied to better understand how many potential fake reviews are included in this dataset.



Target for special offers

Target to understand low sentiment

Develop a customized internal rating system based on sentiment

Figure 33

Model 2: Association Rule Mining Results

Tuned Dataset – Run 1:

In the initial dataset's final tuned run, the primary finding was that a Rating = 5 was strongly associated with Recommended = YES. This held to be true across classes and age groups. Figure 34 shows the top 19 strongest rules from the initial dataset.

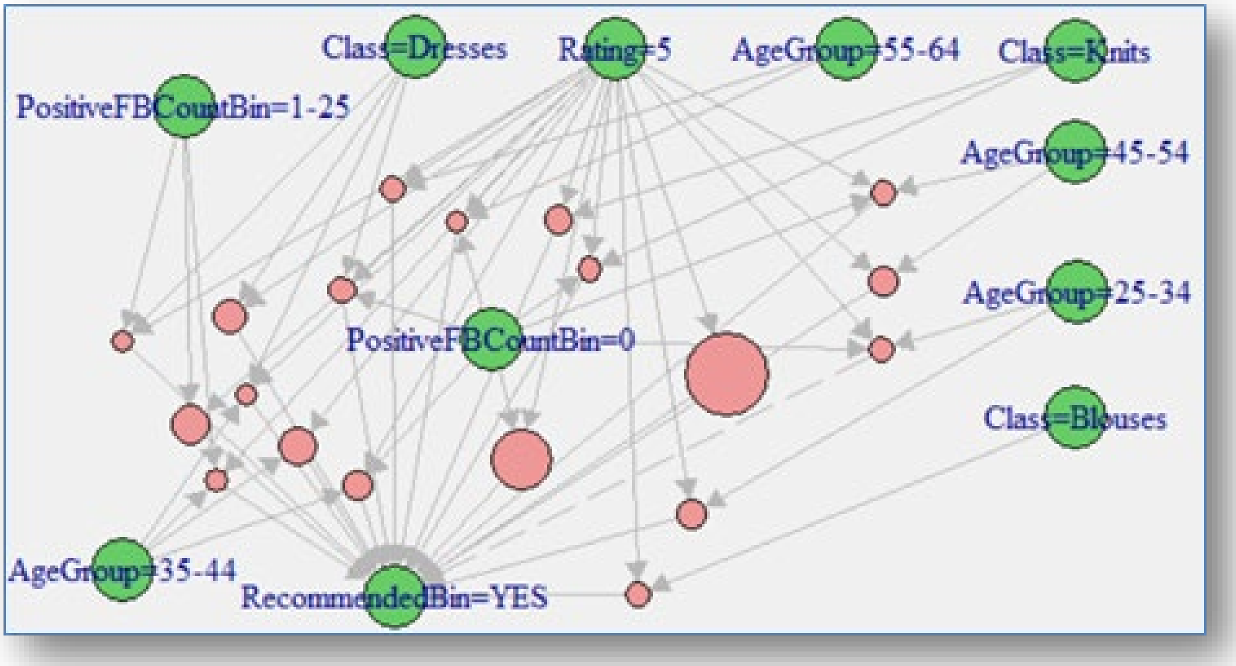


Figure 34

Figure 35 below shows the Relative Attribute Frequency of the attributes in the 19 strongest rules initially mined from the dataset.

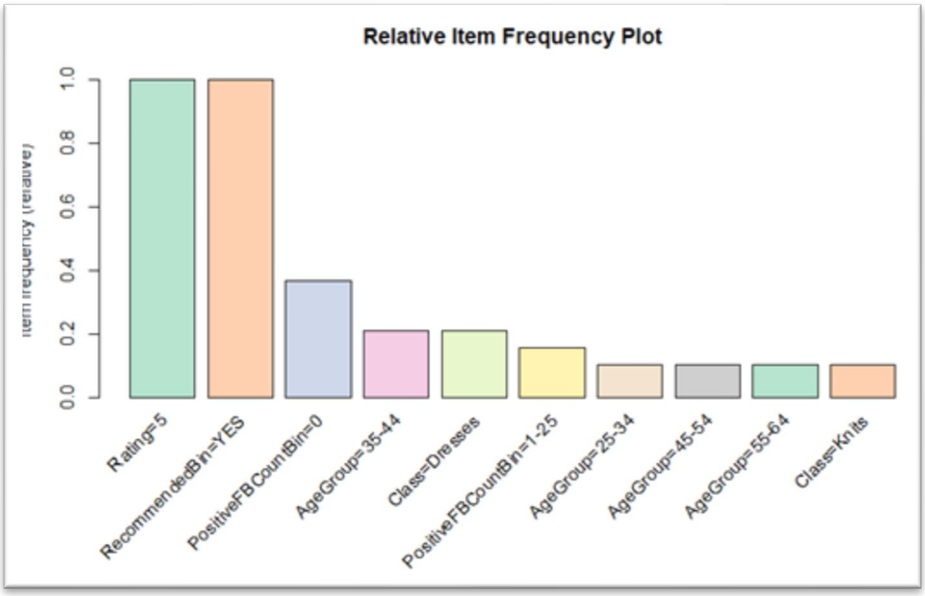
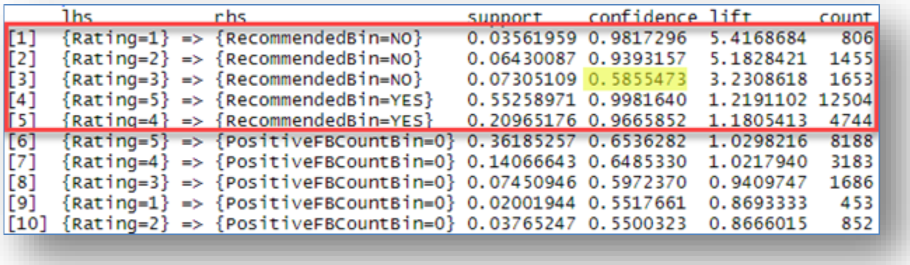


Figure 35

Secondary Runs – Rating Associations, Positive Feedback Counts, and Recommended = NO:

Rating Associations

As the initial dataset provided a wealth of information related to records with Rating = 5, Ratings 1-4 were investigated further. Unsurprisingly, Ratings of 1 & 2 are strongly associated with Recommended = NO with a confidence level between 93-98%. Interestingly, a Rating of 3 is associated with Recommended = NO, but with a lower confidence level of only 58.6%. A Rating = 4 is also unsurprisingly associated with Recommended = YES with a confidence level of 96.7% and support of over 20%.

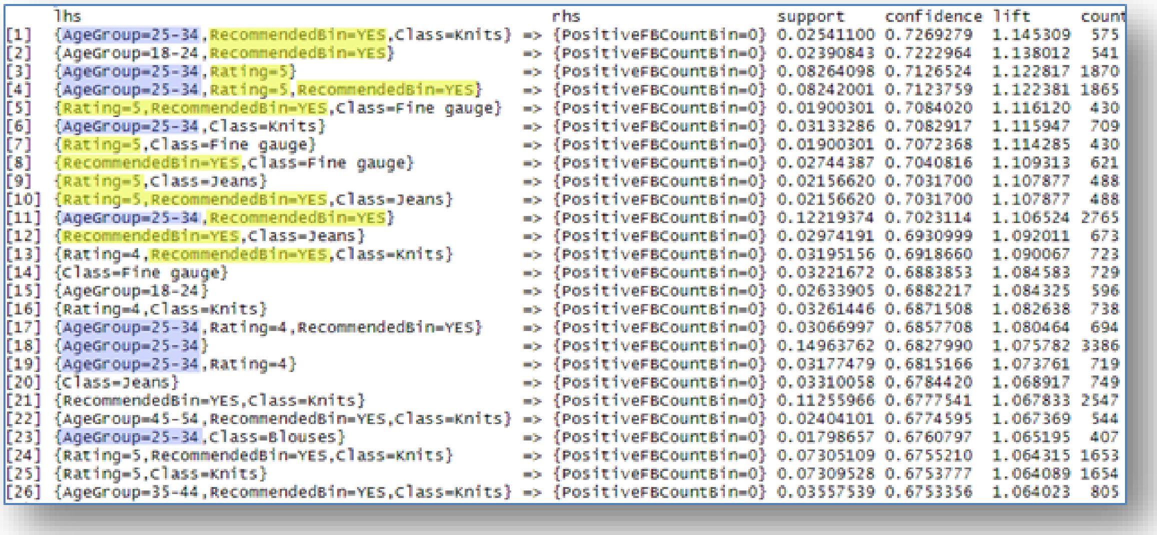


```
     lhs            rhs                     support    confidence lift      count
[1]  {Rating=1} => {RecommendedBin=NO}      0.03561959 0.9817296  5.4168684   806
[2]  {Rating=2} => {RecommendedBin=NO}      0.06430087 0.9393157  5.1828421  1455
[3]  {Rating=3} => {RecommendedBin=NO}      0.07305109 0.5855473  3.2308618  1653
[4]  {Rating=5} => {RecommendedBin=YES}     0.55258971 0.9981640  1.2191102 12504
[5]  {Rating=4} => {RecommendedBin=YES}     0.20965176 0.9665852  1.1805413  4744
[6]  {Rating=5} => {PositiveFBCountBin=0}   0.36185257 0.6536282  1.0298216  8188
[7]  {Rating=4} => {PositiveFBCountBin=0}   0.14066643 0.6485330  1.0217940  3183
[8]  {Rating=3} => {PositiveFBCountBin=0}   0.07450946 0.5972370  0.9409747  1686
[9]  {Rating=1} => {PositiveFBCountBin=0}   0.02001944 0.5517661  0.8693333   453
[10] {Rating=2} => {PositiveFBCountBin=0}   0.03765247 0.5500323  0.8666015   852
```

Figure 36

Positive Feedback Count Associations

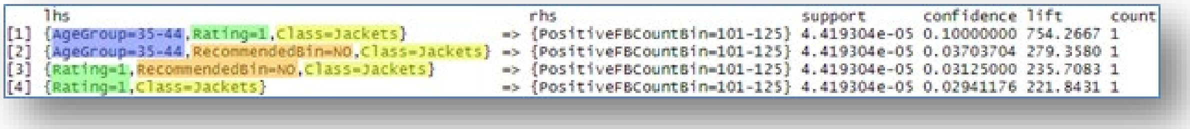
PositiveFeedbackBin = 0



```
     lhs                                                  rhs                      support    confidence lift     count
[1]  {AgeGroup=25-34,RecommendedBin=YES,Class=Knits}    => {PositiveFBCountBin=0} 0.02541100 0.7269279 1.145309  575
[2]  {AgeGroup=18-24,RecommendedBin=YES}                => {PositiveFBCountBin=0} 0.02390843 0.7222964 1.138012  541
[3]  {AgeGroup=25-34,Rating=5}                          => {PositiveFBCountBin=0} 0.08264098 0.7126524 1.122817 1870
[4]  {AgeGroup=25-34,Rating=5,RecommendedBin=YES}       => {PositiveFBCountBin=0} 0.08242001 0.7123759 1.122381 1865
[5]  {Rating=5,RecommendedBin=YES,Class=Fine gauge}     => {PositiveFBCountBin=0} 0.01900301 0.7084020 1.116120  430
[6]  {AgeGroup=25-34,Class=Knits}                       => {PositiveFBCountBin=0} 0.03133286 0.7082917 1.115947  709
[7]  {Rating=5,Class=Fine gauge}                        => {PositiveFBCountBin=0} 0.01900301 0.7072368 1.114285  430
[8]  {RecommendedBin=YES,Class=Fine gauge}              => {PositiveFBCountBin=0} 0.02744387 0.7040816 1.109313  621
[9]  {Rating=5,Class=Jeans}                             => {PositiveFBCountBin=0} 0.02156620 0.7031700 1.107877  488
[10] {Rating=5,RecommendedBin=YES,Class=Jeans}          => {PositiveFBCountBin=0} 0.02156620 0.7031700 1.107877  488
[11] {AgeGroup=25-34,RecommendedBin=YES}                => {PositiveFBCountBin=0} 0.12219374 0.7023114 1.106524 2765
[12] {RecommendedBin=YES,Class=Jeans}                   => {PositiveFBCountBin=0} 0.02974191 0.6930999 1.092011  673
[13] {Rating=4,RecommendedBin=YES,Class=Knits}          => {PositiveFBCountBin=0} 0.03195156 0.6918660 1.090067  723
[14] {Class=Fine gauge}                                 => {PositiveFBCountBin=0} 0.03221672 0.6883853 1.084583  729
[15] {AgeGroup=18-24}                                   => {PositiveFBCountBin=0} 0.02633905 0.6882217 1.084325  596
[16] {Rating=4,Class=Knits}                             => {PositiveFBCountBin=0} 0.03261446 0.6871508 1.082638  738
[17] {AgeGroup=25-34,Rating=4,RecommendedBin=YES}       => {PositiveFBCountBin=0} 0.03066997 0.6857708 1.080464  694
[18] {AgeGroup=25-34}                                   => {PositiveFBCountBin=0} 0.14963762 0.6827990 1.075782 3386
[19] {AgeGroup=25-34,Rating=4}                          => {PositiveFBCountBin=0} 0.03177479 0.6815166 1.073761  719
[20] {Class=Jeans}                                      => {PositiveFBCountBin=0} 0.03310058 0.6784420 1.068917  749
[21] {RecommendedBin=YES,Class=Knits}                   => {PositiveFBCountBin=0} 0.11255966 0.6777541 1.067833 2547
[22] {AgeGroup=45-54,RecommendedBin=YES,Class=Knits}    => {PositiveFBCountBin=0} 0.02404101 0.6774595 1.067369  544
[23] {AgeGroup=25-34,Class=Blouses}                     => {PositiveFBCountBin=0} 0.01798657 0.6760797 1.065195  407
[24] {Rating=5,RecommendedBin=YES,Class=Knits}          => {PositiveFBCountBin=0} 0.07305109 0.6755210 1.064315 1653
[25] {Rating=5,Class=Knits}                             => {PositiveFBCountBin=0} 0.07309528 0.6753777 1.064089 1654
[26] {AgeGroup=35-44,RecommendedBin=YES,Class=Knits}    => {PositiveFBCountBin=0} 0.03557539 0.6753356 1.064023  805
```

Figure 37

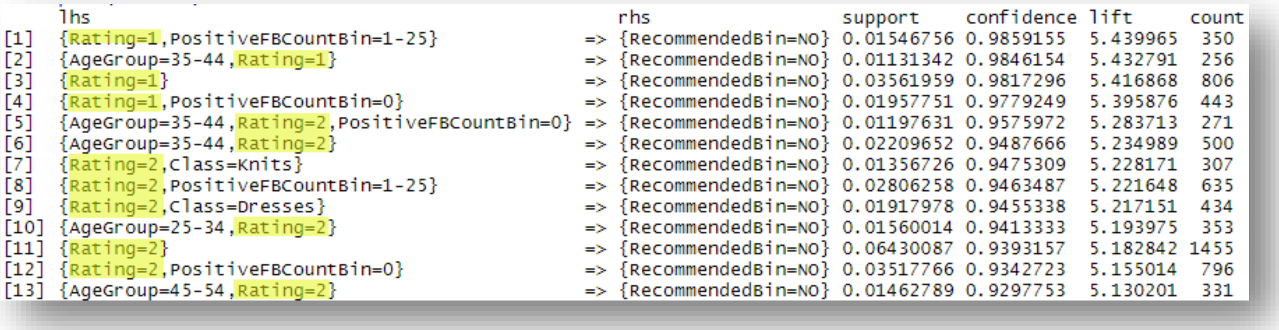
PositiveFeedbackBin = 101-125 Associations



```
    lhs                                                  rhs                          support      confidence lift     count
[1] {AgeGroup=35-44,Rating=1,Class=Jackets}           => {PositiveFBCountBin=101-125} 4.419304e-05 0.10000000 754.2667 1
[2] {AgeGroup=35-44,RecommendedBin=NO,Class=Jackets}  => {PositiveFBCountBin=101-125} 4.419304e-05 0.03703704 279.3580 1
[3] {Rating=1,RecommendedBin=NO,Class=Jackets}        => {PositiveFBCountBin=101-125} 4.419304e-05 0.03125000 235.7083 1
[4] {Rating=1,Class=Jackets}                          => {PositiveFBCountBin=101-125} 4.419304e-05 0.02941176 221.8431 1
```

Figure 38

Recommended = NO Associations



```
     lhs                                                  rhs                  support    confidence lift     count
[1]  {Rating=1,PositiveFBCountBin=1-25}                => {RecommendedBin=NO} 0.01546756 0.9859155  5.439965  350
[2]  {AgeGroup=35-44,Rating=1}                         => {RecommendedBin=NO} 0.01131342 0.9846154  5.432791  256
[3]  {Rating=1}                                        => {RecommendedBin=NO} 0.03561959 0.9817296  5.416868  806
[4]  {Rating=1,PositiveFBCountBin=0}                   => {RecommendedBin=NO} 0.01957751 0.9779249  5.395876  443
[5]  {AgeGroup=35-44,Rating=2,PositiveFBCountBin=0}    => {RecommendedBin=NO} 0.01197631 0.9575972  5.283713  271
[6]  {AgeGroup=35-44,Rating=2}                         => {RecommendedBin=NO} 0.02209652 0.9487666  5.234989  500
[7]  {Rating=2,Class=Knits}                            => {RecommendedBin=NO} 0.01356726 0.9475309  5.228171  307
[8]  {Rating=2,PositiveFBCountBin=1-25}                => {RecommendedBin=NO} 0.02806258 0.9463487  5.221648  635
[9]  {Rating=2,Class=Dresses}                          => {RecommendedBin=NO} 0.01917978 0.9455338  5.217151  434
[10] {AgeGroup=25-34,Rating=2}                         => {RecommendedBin=NO} 0.01560014 0.9413333  5.193975  353
[11] {Rating=2}                                        => {RecommendedBin=NO} 0.06430087 0.9393157  5.182842 1455
[12] {Rating=2,PositiveFBCountBin=0}                   => {RecommendedBin=NO} 0.03517766 0.9342723  5.155014  796
[13] {AgeGroup=45-54,Rating=2}                         => {RecommendedBin=NO} 0.01462789 0.9297753  5.130201  331
```

Figure 39

Model 3: K-Means Results

The K-Means analysis clustered the review records into 5 groups by Rating. Figure 40 below shows the clusters had a great deal of overlap, indicating the k-means clustering of ratings by age, recommended, positive feedback count and sentiment was not distinctly successful in identifying variances in attributes contributing to reviews. The addition of other attributes such as size, or frequency of purchase may lead to more distinct clustering results. This was not performed within the scope of this effort.

From the graph it is apparent cluster 3 (the Rating 3 group) was very broadly distributed. This appears to support the conclusions identified in the Association Rules Mining that Rating = raitings of 3 are more middle of road with coverage in broader areas. This indicates the greatest variance lies within the Rating = 3 cluster group.

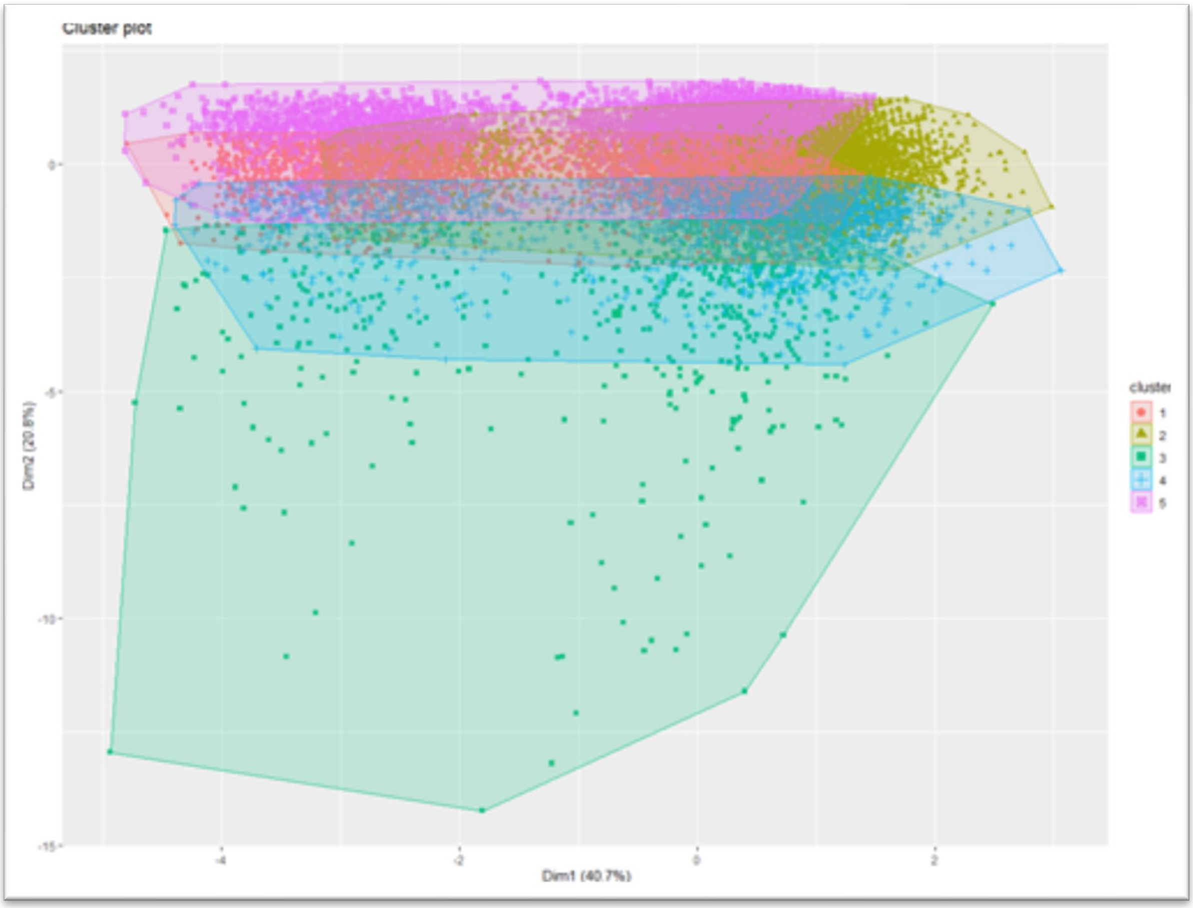
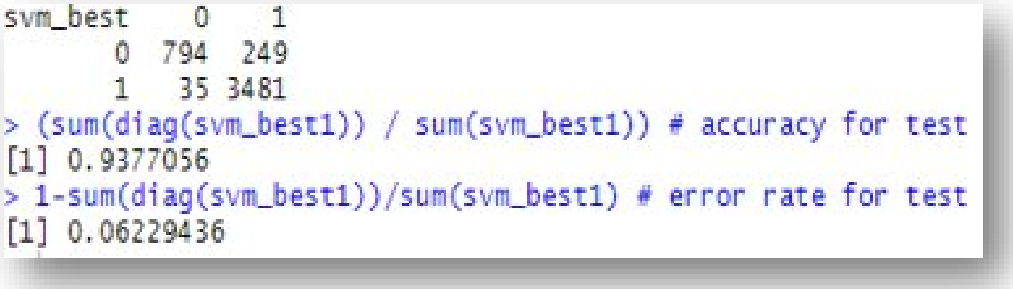


Figure 40

Model 4: Support Vector Machine Results



```
svm_best    0    1
       0  794  249
       1   35 3481
> (sum(diag(svm_best1)) / sum(svm_best1)) # accuracy for test
[1] 0.9377056
> 1-sum(diag(svm_best1))/sum(svm_best1) # error rate for test
[1] 0.06229436
```

Figure 41

Figure 41 presents the best Support Vector Machine Results for determining if a product will be recommended. This best result is captured by a linear SVM and uses the default cost of 1. What is unique about e-commerce dataset is that the almost all the SVMs ignore cost. What is meant by that is there must be extreme cost values of 10,00 and above before the SVM would get a different accuracy score. Generally, the higher the cost the more accurate a SVM will become, however the dataset just contains 22,000 total records. There are hundreds of millions of these transactions occurring each year. Thus, the cost was left at its default setting to avoid over-training.

Over-training occurs when a model is only useful for the data it was trained on. Meaning that although there could be a much higher score for SVM it is better to leave it as is, where it already has over 93.77% accuracy.

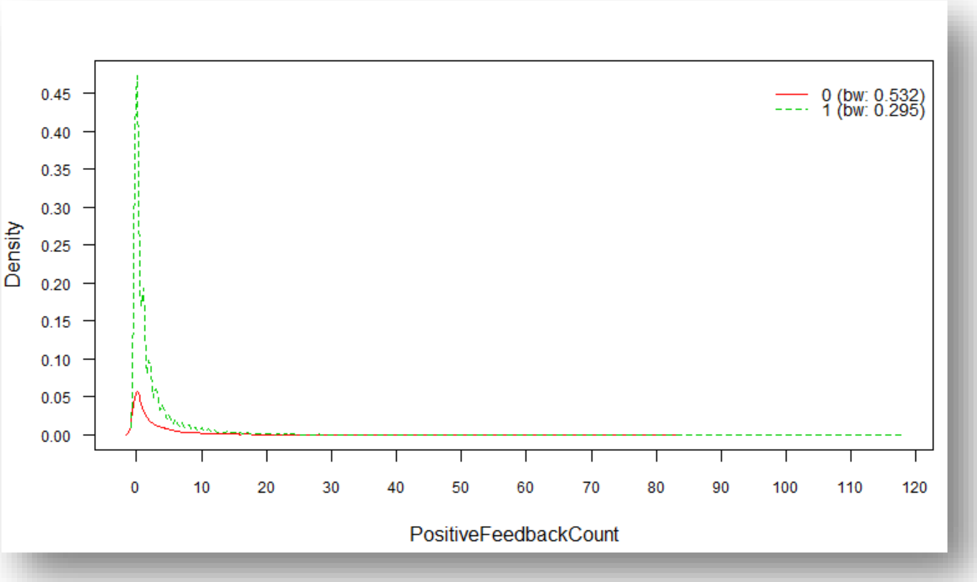


Figure 42

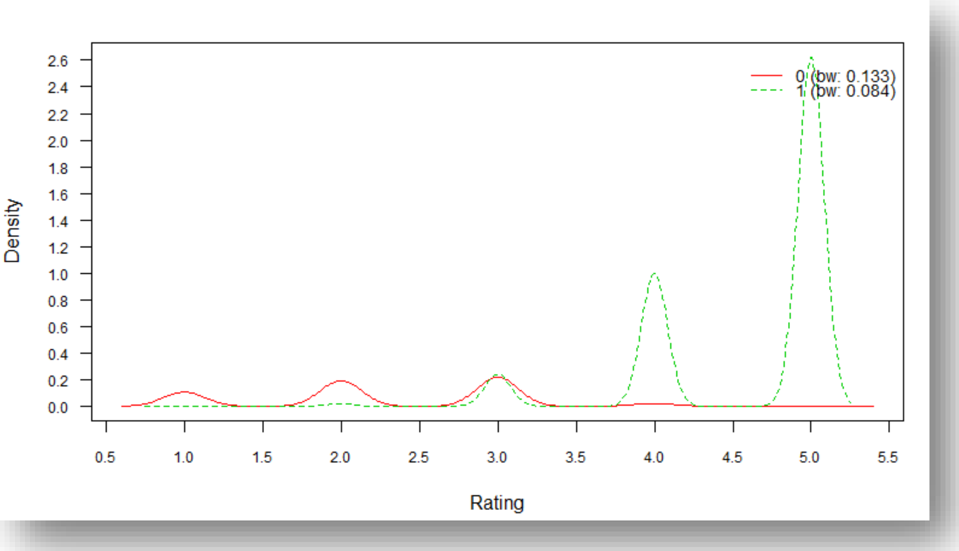


Figure 43

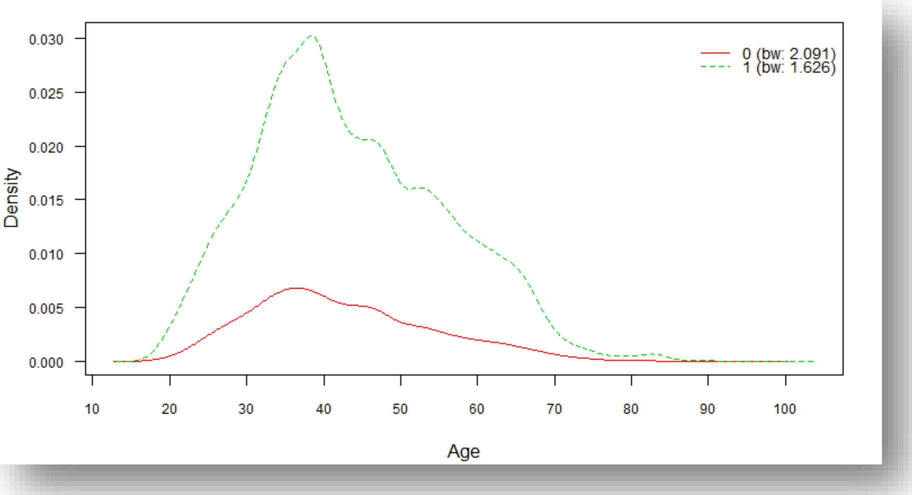


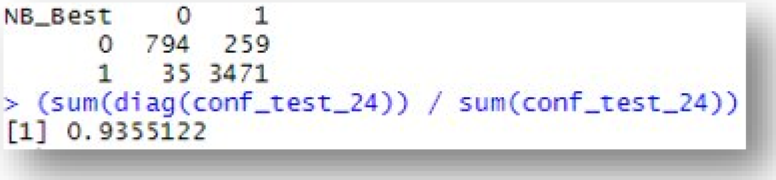
Figure 44

Figure 42 – Figure 44 are density plots from the Support Vector Machine's result. Each figure displays a lot about e-commerce and provides a valuable strategy. For example, Figure 42 displays that for non-recommended reviews, it is rare for them to get high numbers of positive feedback. However; as discussed in the introduction 86% of people will avoid making a purchase if the product has negative reviews. Thus, a great strategy would be to address these negative reviews as soon as a pattern is emerging.

Figure 43 demonstrates the breakdown of recommendations. The story it provides is that at rating 4 and 5 are almost always guaranteed to recommend the product and 1 & 2 ratings are almost always going to not recommended the products. But reviews with a rating of 3 is an almost 50/50. This is where good customer support should first be focused. The individuals who rated 3 are one of the best opportunities to potentially gain additional customers who support the brand. Amazon is a mega e-commerce store, that sales a wide arrange of items. At one point, Amazon was Walmart's competitor, and currently there is no competition Amazon reigns supreme. How did Amazon accomplish this, it was not because they had a niche market, Walmart overlapped with items and vendors. The key factor for Amazon's success is their policy on returns. If you receive a broken product, or after trying it on, Amazon will still allow the item to be returned. While Walmart has a much more stringent and time-consuming process.   In summary, customer service and brand loyalty generally win the day, and focusing on the group who rates the products with a 3, will have the fastest return on investment for good customer service.

Figure 44 is showing the density model by age for the customers. The graph has a right tail skewed with majority of the consumers on the right-hand side. At the age of 40, there is a significant change in the model where folks are more likely to recommend a product. This indicates that there could be different customer bases. Thus, tailoring the online experience for individuals who are under 40 and above 40 could be a valid strategy to increase sales or even help control the amount of negative reviews for a product.

Model 5: Naïve Bayes Results



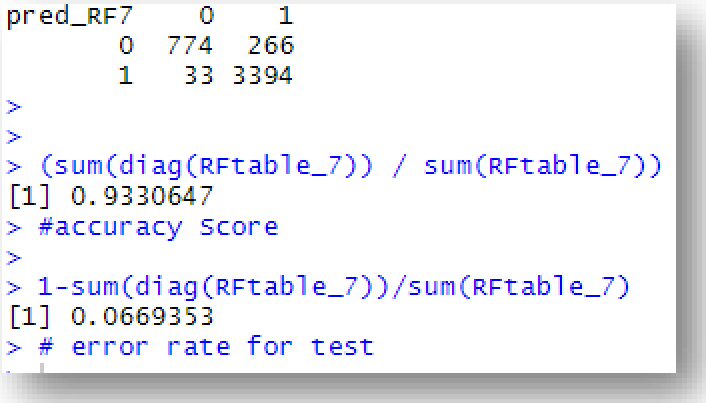
```
NB_Best    0    1
      0  794  259
      1   35 3471
> (sum(diag(conf_test_24)) / sum(conf_test_24))
[1] 0.9355122
```

Figure 45

Figure 45 is showing the best Naïve Bayes model. The NB model is a linear model, where kernel is false and laplace is set to 1. The model is also using the numeric dataset as that removes independent variable correlations and values that do not have an impact of if the product is recommended or not.

As seen in Figure 45, Naïve Bayes preforms almost as well as SVM. It is only 2 thousandths of a decimal of a point away from SVM. Or in other words, for the same set it only guessed 1o more wrong then the best SVM Model. Also, like the SVM model the errors generally occur in assuming that some of the not recommended would be recommended.   Most of the confusion is when a person gives a rating of 4 for the product but would not recommend the product. But much like SVM, Naïve Bayes also highly weights ratings and age for whether it is expecting a product to get a recommendation. The focus on ratings also helps reinforce that if a product has a lower then expected rating it is crucial to act immediately.

Model 6: Random Forest Results



```
pred_RF7     0    1
       0   774  266
       1    33 3394
>
>
> (sum(diag(RFtable_7)) / sum(RFtable_7))
[1] 0.9330647
> #accuracy Score
>
> 1-sum(diag(RFtable_7))/sum(RFtable_7)
[1] 0.0669353
> # error rate for test
```

Figure 46

Figure 46 is displaying the best Random Forest model. It uses the numeric dataset like NB and SVM. Random Forest preformed the best when it uses the over 200 tress and default maturity. RF preformed close to NB and SVM. All three models key off rating for whether a product would be recommended.

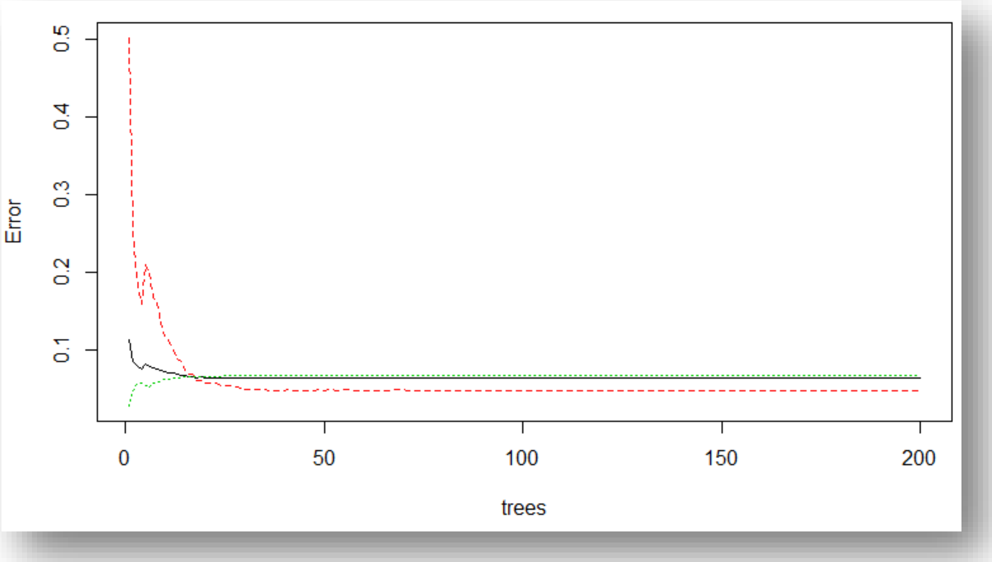


Figure 47

Figure 47 is demonstrating how the error rate drastically drops to less than 10% by 50 trees. As stated in the model the more trees given to a RF model the more accurate it will predict. But that after 200 trees there is a steep drop-off of returns for effort.

What this signifies that there might be other data points that are not considered in the dataset that can explain why a person would give an item a rating of 5 and then not recommended the item to others. If more time was allowed it would be suggested to match up the common errors back to the ratings to see if that provides insight, or if this behavior is something that cannot be fully explained in the data captured. In summary there are some human behaviors may not be quantifiable or determined with accuracy.

Model 7: Decision Tree Results

Decision trees took a different approach and set out to test three theories. The results of two of the three hypotheses supported the null hypothesis. In other words, tree A and C were not able to confirm/predict a rating based on clothing item or department. However, decision tree B did provide directional prediction on Positive Feedback.

Just 37% of all reviews received a Positive Feedback indicator which is the label for this analysis. Yet some of those reviews received more positive feedback. The positive feedback count data were organized into the following buckets for the DT analysis.

Positive Feedback Count Distribution

Positive Feedback Count	# Reviews
0	4256
1-25	2457
26-50	63
51-75	7
76-100	6

Table 5

Of the 2,500 reviews received one or more indicators, 8/10 times it was a review that recommended a product suggesting that getting approval from a past purchaser is particularly helpful. Almost 20% of the time a positive review can happen with no recommendation and review of 4 or lower. However, in these cases is 75% likely to be associated with negative sentiment in terms of the text reviews themselves. Sentiment scores were bucketized into the following categories as shown in Figure 48.

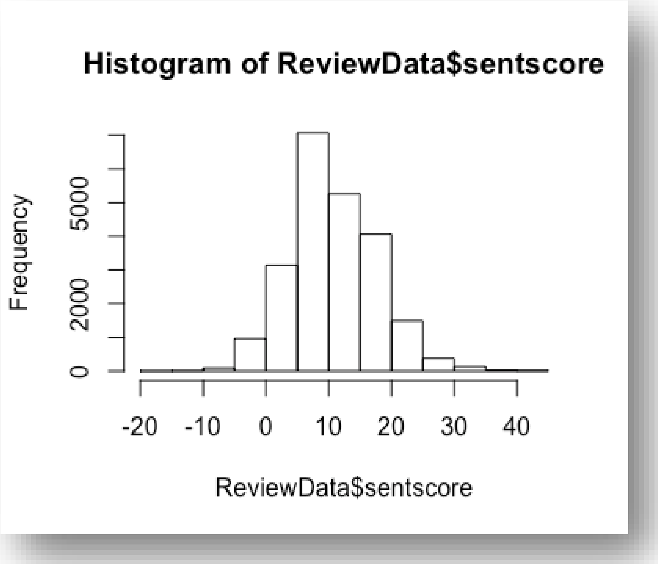


Figure 48

The decision tree demonstrates that the most helpful reviews are the ones that either let the shopper know details or concerns about a product or provides reasoning for the recommendation from previous purchasers. The decision tree data confirms that the most helpful reviews are giving a shopper a comfort from other shoppers, or specific feedback about why their experience wasn't 5-star.

In reading the tree, the low sentiment score branches ladder up to positive feedback on the occasions where a product is not recommended and received a lower rating. Otherwise 82% of the positive feedback is from reviews that include a recommendation.

Positive Feedback Decision Tree

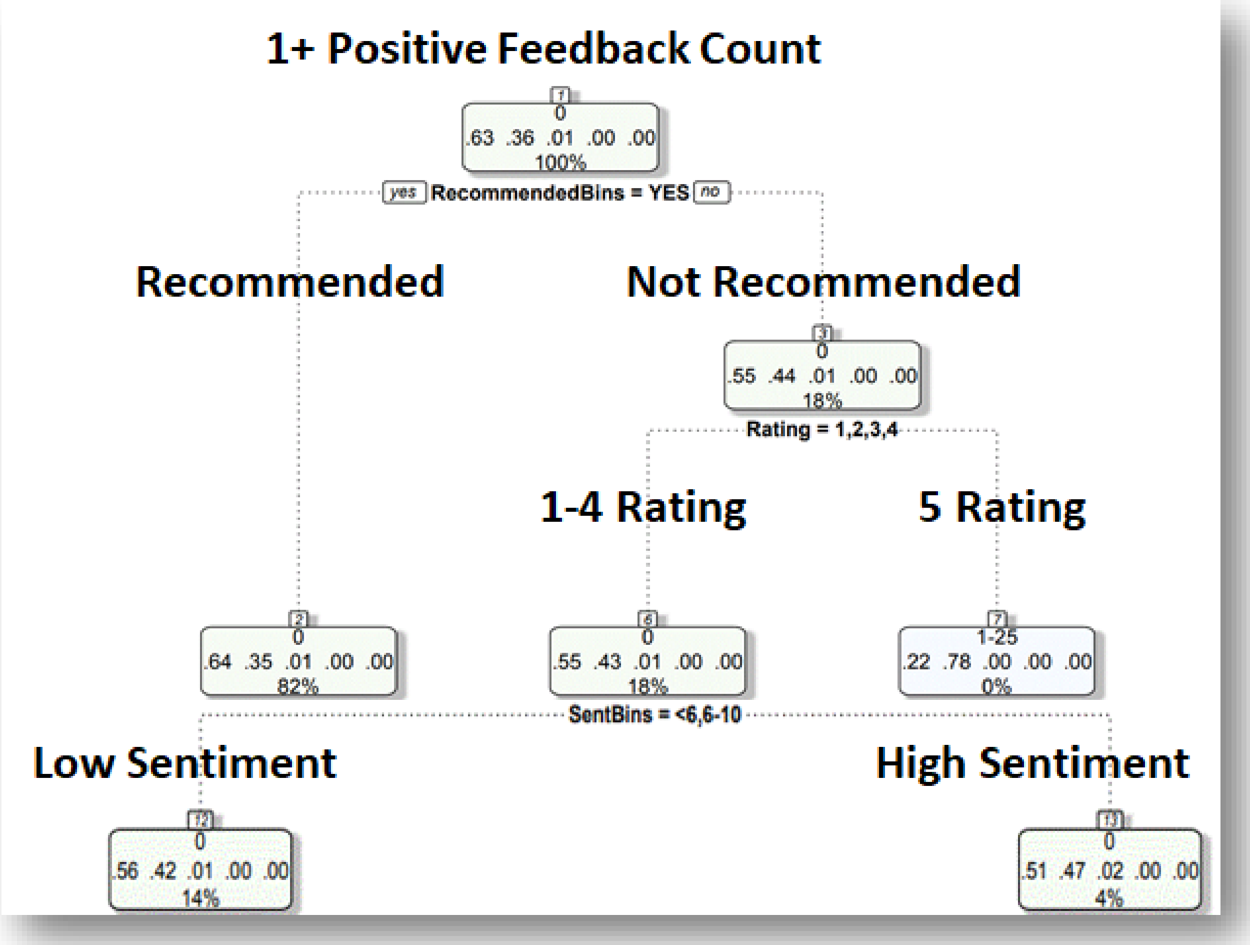


Figure 49

Conclusions

Over the past decade, the way women's clothing is purchased has seen a dramatic shift away from "brick and mortar" stores to online e-commerce. With this change, new challenges for retailers increase since customers are not able to try items on before purchasing. Online product reviews do provide a convenient way for consumers to read other shoppers opinions about a product prior to making a buying decision. These reviews also provide a wealth of great information for retailers to use to ensure they have the right products that fit great and provide a seamless purchasing experience for the customer.

The analysis of the 23,486 reviews for this study provided very helpful insights to help a retailer better understand the customer and respond to her/his needs. This data was from one retailer; however, the methods could be applied to any type of online reviews. While all the reviews are online, the impacts are not isolated to online e-commerce sales. As mentioned in the Introduction, brick and mortar sales are dramatically impacted by online reviews as well.

Below is a summary of the short-term recommendations for this women's e-commerce site:

1. Develop a segmented marketing plan for customers over the age of 40 and under the age of 40. Patterns were clear that these two age demographics buy differently and should be marketed to in a different way.
2. Target the top tier of sentiment scores with special offers and early releases of the newest and most expensive products. This group typically provides a 5 rating but more importantly, they write reviews that are more positive which could influence other customers to buy.
3. Understand why some customers rate the products as a 5 but have low sentiment scores. Do hidden customer service or product issues exist? Are sizes just a little off for a product they absolutely love after receiving a replacement for size?
4. Target customer service improvement for customers with product ratings of 3 to convince more to recommend a product.
5. Analyze the Ratings of 1 or 2 to explore the issues that cause low reviews that negatively influence other buyers. If a problem has been addressed, respond to the review with a description of the remedy.

In addition to the steps above that can be implemented immediately, a more detailed analysis of these suggested changes should be considered:

1. Implement a verified review process to ensure that fake reviews are minimized.
2. Implement customer identification and purchase tracking.
3. After implementing the verified user process and customer identification tracking, utilize this information to provide collaborative filtering of products to enhance the cross-sell and up-sell opportunities.

With the change in how women's clothing is purchased, retailers must become even more responsive to customer needs by following the short/long term recommendations above. Compared to "brick and mortar" purchases, a shopper can have multiple windows from different sites open to compare the variables that are most important to them such as fit, quality, price and shopping experience. Likewise, even in retail stores shoppers will use smart phones and other devices to evaluate reviews and prices of products before making purchases from a store. Online reviews and company responses to those reviews both have a strong influence on establishing the reputation of products as well as the reputation of the companies.

A retailer who isn't continuously innovating can quickly lose customers or worse. Examples of companies who have filed for bankruptcy or gone out of business include Bonton, Barney's, Nine West and many others.[viii] The nimble companies who can harness the power of data to transform their business and provide better customer service will continue to gain market share as the demand for women's fashion doesn't show signs of slowing in the foreseeable future.

[i] Why Are Reviews Important? (2019, March 3007). Retrieved from https://planetmarketing.com/blog/why-are-reviews-important/#1b

[ii] Meena, Satish. “Forrester.” *Forrester Analytics: Online Fashion Retail Forecast, 2017 To 2022 (Global)*, 2018, www.forrester.com/report/Forrester+Analytics+Online+Fashion+Retail+Forecast+2017+To+2022+Global/-/E-RES145235.

[iii] Yu, Bei. (2019). Week 8 Slides: KNN, Support Vector Machines, and Ensemble Learning.

[iv] Yu, Bei. (2019). Week 8 Slides: KNN, Support Vector Machines, and Ensemble Learning.

[v] Yu, Bei. (2019). Week 8 Slides: KNN, Support Vector Machines, and Ensemble Learning.

[vi] Yu, Bei. (2019). Week 8 Slides: KNN, Support Vector Machines, and Ensemble Learning.

[vii] Brid, Rajesh S. “Decision Trees-A Simple Way to Visualize a Decision.” *Medium*, GreyAtom, 26 Oct. 2018, medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb.

[viii] Editor. “Retail Woes: A Running List of Fashion Bankruptcies.” *The Fashion Law*, The Fashion Law, 6 Aug. 2019, www.thefashionlaw.com/home/retail-woes-a-bankruptcy-timeline.