Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2011014

A Probabilistic Model of Relationship Between Product's Characteristic Factors and Star Rating based on Data Mining and Bayesian Network Model

The Internet is changing every aspect of our lives, and the rapid development of e-commerce is due to this. With the rise of this industry, how to seize the opportunities to create more wealth should be our common focus. Here's how we've been able to mine the secrets behind the data to make more money for sunshine.

The idea is to use the Bayesian model, which will help us make a better prediction of unknown relationships between factors. We expect to establish a Bayesian network model of customer satisfaction influencing factors based on extracted characteristic factors with satisfaction as the target and other feature factors as input variables. In this way, the correlation among factors and the conditional probability of the child node under the parent node event can be obtained, and the ranking table of the influence factors on the satisfaction degree of goods can be output. In addition, we expect to make an optimization based on the Bayesian network model, so as to be able to more accurately find the relationship between several variables and solve the specific problems.

After data cleaning, we screened out the high-frequency words in the text comments of each type of products through the word frequency analyzer. And we found the keywords of the reviews of three types of products. Through the establishment of the emotional dictionary and the processing of the database to achieve the quantitative process of the text comments. Using Modeler, we built a Bayesian network to measure the relationship between stars, comments, and helpful votes. We obtained the correlation between the characteristics of each product and their impact on the reputation of the product. Through the establishment of the number of comments based on the time series and the average star evaluation line graph, we found that star evaluation has a certain positive influence on the content and sales of the text review. Based on these, we gave Sunshine Company the design and sale strategy of these three products.

Contents

1 Introduction

1.1 Statement of the Problem

In the past decade we have witnessed the rapid development of e-commerce, thanks to the increasing maturity of Internet technology. The business model of e-commerce is becoming more and more common in our daily life. In the United States, Amazon is the pioneer of e-commerce. In addition to the description of the product, we can also see consumers' comments on the product in the dazzling product page of amazon. This is an essential feature of almost every major online store, not just Amazon.

The value of these review and rating data mainly lies in three aspects. Firstly, it can give consumers an opportunity to evaluate the product and provide them with a platform to express their opinions. Secondly, it can give consumers who are still hesitating to buy a reference, which may be different from the description of the attributes of the goods on the product page. In addition, the data can help businesses successfully make and sell products. Sunshine Company attaches great importance to the value behind the data, so we hope that our team can identify and analyze the valuable data, which will help them design the product and determine the sales strategy.

We now have data on hair dryers, microwaves and pacifiers, which includes product ratings, reviews and more. Through reasonable screening, classification and analysis of the data, we established a relevant model to address the needs of sunshine company.

2 Assumptions and Notation

2.1 Assumptions

In our model, we make the following assumptions:

- We assume that the relationship between the variables will not change over time.
- We assume that amazon didn't tamper with the data.

2.2 Notation

Table 2-1: The notation of the Model

Notation	Meaning
i	The amount of the reviews
R	Product reputation
S	Star rating
RS	Review score
PO	Power characteristic score of hair dryer
DU	Durability characteristic score of hair dryer
PR	Price characteristic score of hair dryer
CO	Comfort characteristic score of hair dryer
AP	Appearance characteristic score of hair dryer

3 Data Processing

3.1 Data Cleaning

Prior to further analysis of the data, we found that there was data redundancy in the given file. Therefore, we need to process the given data.

First, we delete the useless attributes in the table, which include the marketplace and product_category attributes, because the data is the same, so it is not meaningful for correlation analysis.

Second, we filter the content of the product_title attribute in each table. We import the pacifier.tsv file into the MySQL database and use the SQL statement to remove the

data from the product_title property that does not contain the word "pacifier", and the other two tables do the same. This enables us to obtain more accurate results later.

Table 3-1: Examples of Useless and Invalid Data on the Pacifier Table

marketplace	product_title	product_category
US	aden by aden + anais sleeping bag	Baby
US	baby Einstein	Baby

We summarize the remaining data and the results are shown in the following table:

Table 3-2: Summary of Collected Information

Product	Amount	Number of Favorable Comments
Hair Dryer	11460	85.41%
Pacifier	12074	88.51%
Microwave	1599	68.61%

3.2 Keywords Extraction

After obtaining valid data, we wrote a word frequency analyzer in Java. This program allows us to extract the frequently used words in the review_headline and the review_body that represent the consumer's comments on the quality of the product. From these keywords we can know which aspects consumers are most concerned about. In addition, extracting these keywords will be helpful for the following analysis of commodity characteristic factors.

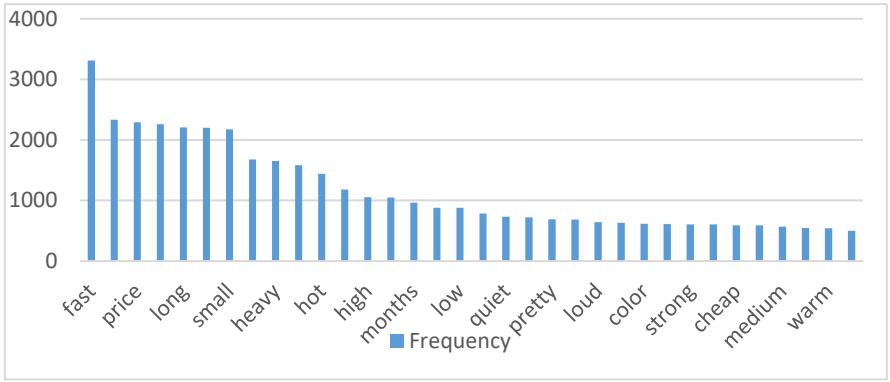


Figure 3-1: The Keywords Extracted from the Hair Dryer Reviews

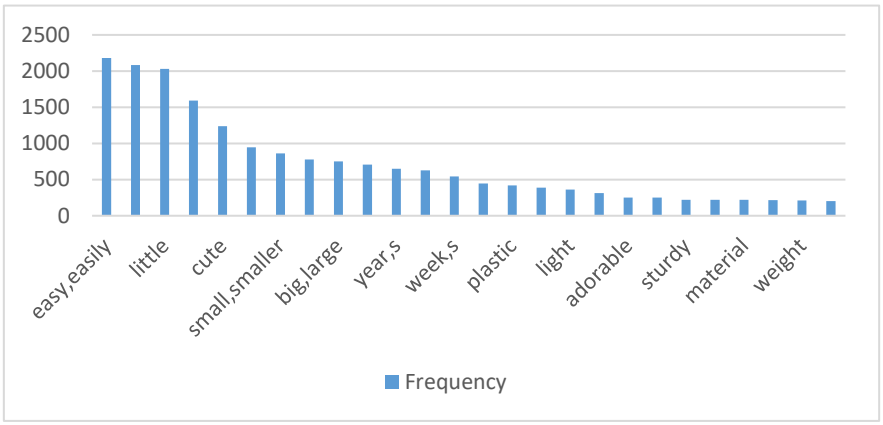


Figure 3-2: The Keywords Extracted from the Pacifier Reviews

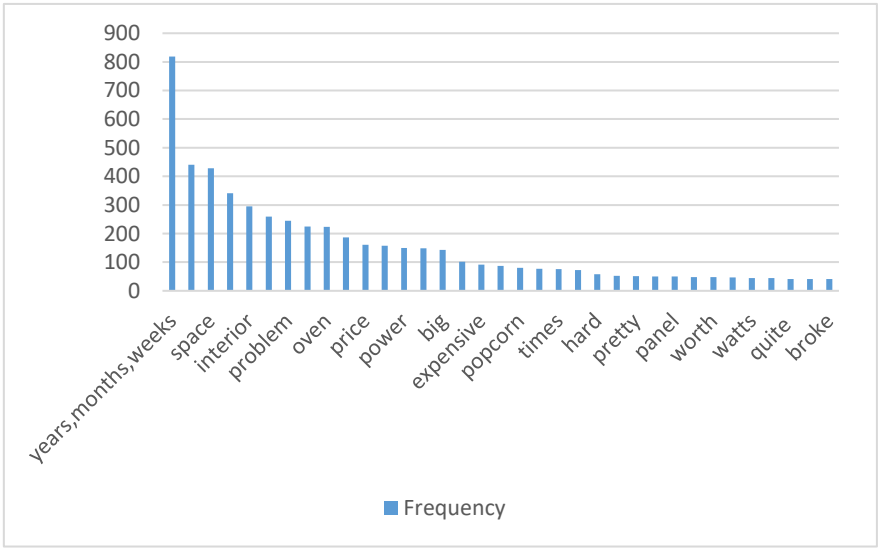


Figure 3-3: The Keywords Extracted from the Microwave Reviews

3.3 The Summary of the Characteristic Factors of the Review

After merging the synonyms in the keyword, we extract the characteristic values of different products. The results are as follows

Table 3-3: The Summary of Product Characteristics Factors

Product	Characteristic Factors
Hair Dryer	Power, Appearance, Durability, Comfort, Price
Microwave	Function, Quality, Price, After-sales Service, Volume
Pacifier	Appearance, Health, Texture, Quality, Size

3.4 Build the Emotional Dictionary and Score the Text Data

Next, we manually build the emotional dictionary. In the case of the hair dryer, we quantified each review with an adjective describing the main characteristic factor. Positive words scored +1, negative words scored -1, and the rest scored 0.

Table 3-4: The Example of the Score Table of the Hair Dryer

Characteristic Factors	Words	Score
Power	Powerful, Fast, Strong	+1
	Weak	-1
Appearance	Beautiful, Suit	+1
	Ugly	-1

Durability	Long, Durable	+1
	Short	-1
Comfort	Comfortable, Easeful	+1
	Uncomfortable	-1
Price	Cheap, Inexpensive	+1
	Expensive	-1

4 Bayesian Network Model of the Star Ratings, the Reviews, and the Helpfulness Ratings

Based on the quantified data, modeler software is used to import the data into the software in Excel format. After data filtering and type selection, build the Bayesian network model of each product. Rank the importance of each characteristic factor and the conditional probability of each node.

4.1 Model Overview

Bayesian networks can represent a set of random variables and their conditional dependencies by a directed acyclic graph. Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian networks aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors.

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} P\left(A_k \middle| \bigcap_{j=1}^{k-1} A_j\right)$$

$$P(x_1, x_2, \ldots, x_n) \propto P(y) \prod_{i=1}^{m} P(x_i|y)$$

Currently we are not sure about the relationship between star ratings, reviews, and helpfulness ratings. After the establishment of the Bayesian network model, the relationship between the factors and the future prediction can be obtained through the data processing. Next, we introduce the establishment of our model by taking the hair dryer product as an example. The results are as follows:

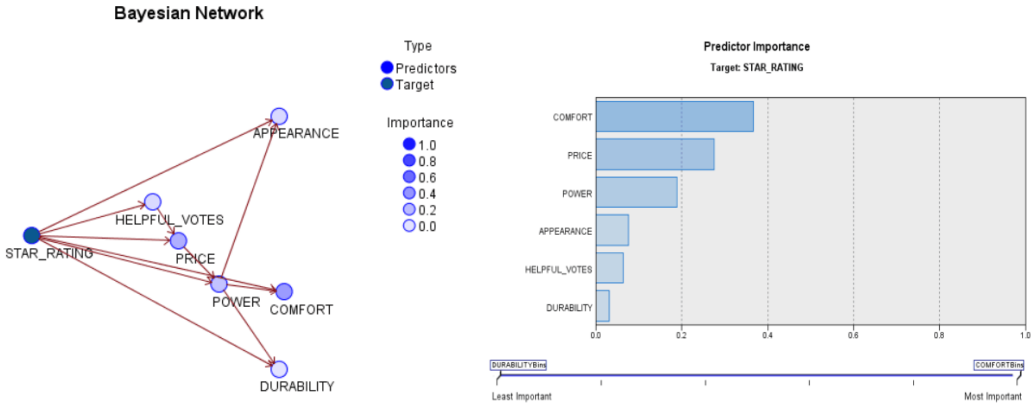


Figure 4-1: The Relationship Between the Factors Figure 4-2: The Predictor Importance

In the Bayesian network model, we set the star rating as the target, with the remaining blue circles representing the remaining input variables. The darker the color of the circle, the more important the variables it represents in the model. In this Bayesian network structure diagram, the star rating is used as the target to connect with the other 6 variables through arrows, which indicates that consumers' satisfaction is closely related to other variables, and there is a certain correlation between each input variable. The figure above reflects the importance of each factor, with the Comfort at the top and the durability at the bottom.

Table 4-1: Conditional Probabilities of Star Rating

Probability				
1	2	3	4	5
0.07	0.06	0.10	0.20	0.57

As can be seen from the conditional probability table of the overall star evaluation node, the highest probability score is 5, and the occurrence probability is 0.57, which

indicates that the overall satisfaction of the buyers to the hair dryer is relatively high.

4.2 Model Specific Analysis

Table 4-2: Conditional Probabilities of Helpful Votes

Parents	Probability				
Star Rating	<=64	64~128	128~192	192~256	>256
1	0.99	0.01	0.00	0.00	0.00
2	0.99	0.01	0.00	0.00	0.00
3	0.99	0.01	0.00	0.00	0.00
4	1.00	0.00	0.00	0.00	0.00
5	0.99	0.00	0.00	0.00	0.00

The score is the parent node of the number of helpful votes. It can be seen from the table that when evaluating 1-5 stars, the number of helpful votes is distributed in the case of ≤64, and the probability approaches 1. This shows that the size of the helpful votes does not affect the probability of others writing biased comments.

Table 4-3: Conditional Probabilities of Price

Parents		Probability		
Star Rating	Helpful Votes	<=-0.4	-0.4~0.4	>0.4
1	<=64	0.24	0.68	0.08
1	64~128	0.00	0.50	0.50
1	128~192	0.00	1.00	0.00
1	>256	0.00	1.00	0.00
2	<=64	0.15	0.80	0.05
2	64~128	0.50	0.00	0.50

3	<=64	0.13	0.80	0.07
3	64~128	0.33	0.33	0.33
3	>256	0.00	1.00	0.00
4	<=64	0.07	0.87	0.05
4	64~128	0.00	1.00	0.00
4	128~192	1.00	0.00	0.00
4	192~256	0.00	1.00	0.00
5	<=64	0.07	0.86	0.07
5	64~128	0.18	0.73	0.09
5	128~192	0.00	1.00	0.00
5	192~256	0.33	0.17	0.50
5	>256	0.00	1.00	0.00

The star rating and the number of helpful votes are the parent nodes of the price. It can be seen from the table that the conditional probability distribution is around 0, indicating that consumers are relatively satisfied with the price. The conditional probability distribution of reviews with more votes was more dispersed, indicating that consumers expected more biased price reviews.

Table 4-4: Conditional Probabilities of Durability

Parents		Probability		
Star Rating	Power	<=0.4	-0.4~0.4	>0.4
1	<=-0.4	0.00	0.78	0.22
1	-0.4~0.4	0.24	0.42	0.34
1	>0.4	0.13	0.60	0.27
2	<=-0.4	0.20	0.60	0.20

2	-0.4~0.4	0.21	0.31	0.48
2	>0.4	0.08	0.78	0.14
3	<=-0.4	0.21	0.64	0.14
3	-0.4~0.4	0.21	0.34	0.45
3	>0.4	0.10	0.65	0.25
4	<=-0.4	0.25	0.50	0.25
4	-0.4~0.4	0.14	0.44	0.43
4	>0.4	0.09	0.74	0.17
5	<=-0.4	0.00	0.85	0.15
5	-0.4~0.4	0.10	0.47	0.43
5	>0.4	0.04	0.79	0.17

It can be seen from the table that the scores with durability node condition probability greater than 0.5 are concentrated in -0.4~0.4. When the overall score is 4 and 5, and the price score is less than or equal to -0.4, the favorable rating of durability is 0.66 and 0.70 respectively, which indicates that consumers are generally satisfied with the durability of hair dryers on the market. There is room for improvement in durability.

Table 4-5: Conditional Probabilities of Power

Parents		Probability		
Star Rating	Price	<=0.4	-0.4~0.4	>0.4
1	<=-0.4	0.01	0.88	0.11
1	-0.4~0.4	0.04	0.72	0.24
1	>0.4	0.00	0.88	0.12
2	<=-0.4	0.00	0.88	0.12
2	-0.4~0.4	0.04	0.64	0.32

2	>0.4	0.07	0.79	0.14
3	<=-0.4	0.03	0.76	0.21
3	-0.4~0.4	0.03	0.54	0.43
3	>0.4	0.06	0.74	0.19
4	<=-0.4	0.00	0.69	0.31
4	-0.4~0.4	0.02	0.54	0.44
4	>0.4	0.00	0.79	0.21
5	<=-0.4	0.01	0.75	0.24
5	-0.4~0.4	0.01	0.45	0.54
5	>0.4	0.00	0.69	0.31

Scores with conditional probability of power node greater than 0.5 are concentrated at more than -0.4. When the overall rating is 4-5 and the durability score effect is equal to -0.4, the probability of power score of >0.4 is 0.46 and 0.42 respectively, which indicates that the customer is satisfied with the power of the hair dryer in the market.

Table 4-6: Conditional Probabilities of Appearance

Parents		Probability		
Star Rating	Power	<=-0.4	-0.4~0.4	>0.4
1	<=-0.4	0.00	1.00	0.00
1	-0.4~0.4	0.00	0.98	0.02
1	>0.4	0.00	0.90	0.10
2	<=-0.4	0.00	0.90	0.10
2	-0.4~0.4	0.01	0.99	0.00
2	>0.4	0.00	0.97	0.03
3	<=-0.4	0.00	1.00	0.00

3	-0.4~0.4	0.01	0.98	0.01
3	>0.4	0.00	0.99	0.01
4	<=-0.4	0.00	1.00	0.00
4	-0.4~0.4	0.01	0.98	0.02
4	>0.4	0.00	0.99	0.01
5	<=-0.4	0.00	1.00	0.00
5	-0.4~0.4	0.00	0.96	0.04
5	>0.4	0.00	0.99	0.01

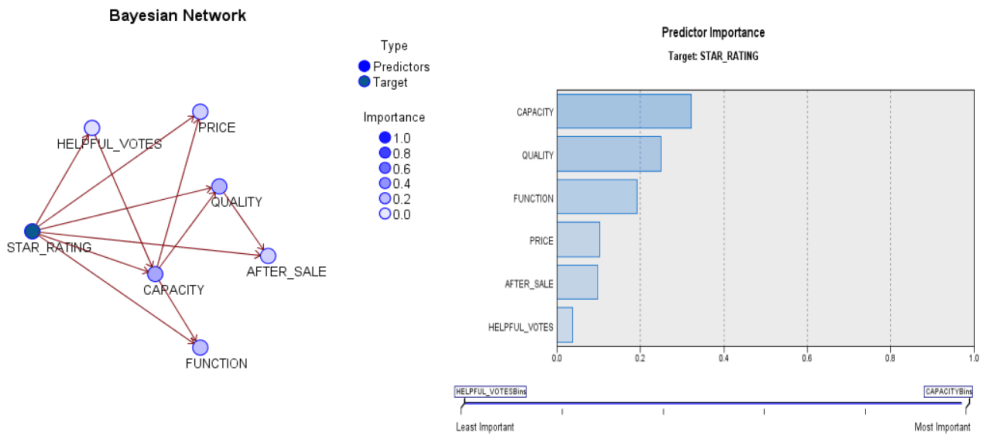
128

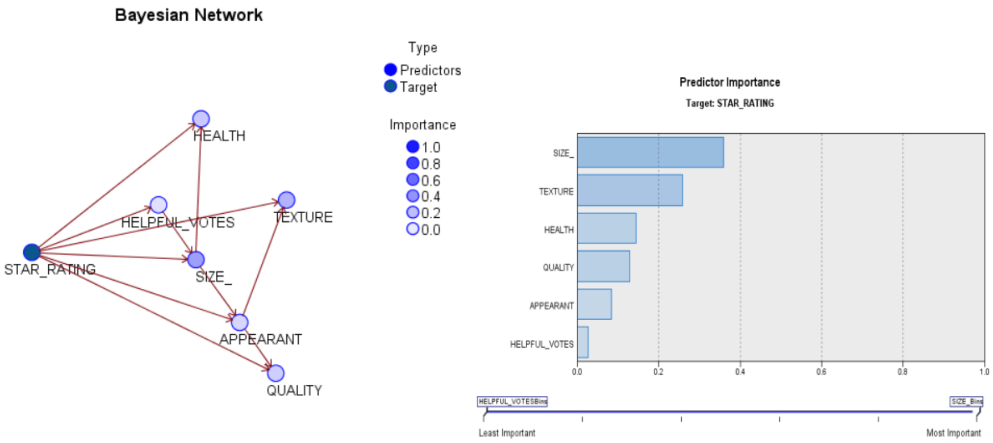
129 Table 4-7: Conditional Probability of Comfort

Parents		Probability		
Star Rating	Power	<=-0.4	-0.4~0.4	>0.4
1	<=-0.4	0.11	0.89	0.00
1	-0.4~0.4	0.15	0.80	0.05
1	>0.4	0.08	0.90	0.02
2	<=-0.4	0.00	1.00	0.00
2	-0.4~0.4	0.18	0.76	0.06
2	>0.4	0.05	0.92	0.03
3	<=-0.4	0.14	0.86	0.00
3	-0.4~0.4	0.16	0.73	0.11
3	>0.4	0.04	0.93	0.03
4	<=-0.4	0.08	0.75	0.17
4	-0.4~0.4	0.14	0.59	0.28

4	>0.4	0.04	0.86	0.09
5	<=-0.4	0.15	0.85	0.00
5	-0.4~0.4	0.07	0.62	0.30
5	>0.4	0.03	0.88	0.09

Star rating and power is the parent of appearance and comfort. It can be seen from this table that the score of appearance nodes is concentrated in -0.4~0.4, and the probability approaches to 1 almost everywhere. This shows that the customer is ok with the appearance of the hair dryer, but wants the appearance of the hair dryer to be better. When the score is 1-3, the comfort score is almost all concentrated in -0.4~0.4. It can be seen that when the customer is not satisfied with the whole hair dryer, the comfort level of the hair dryer is satisfactory. When the overall rating is 4-5, and the power score is greater than -0.4, the probability of comfort score greater than 0.4 increases. This shows that the improvement of power has a positive impact on the comfort of customers when using the hair dryer.





141

142 Figure 4-3: The Bayesian Network Structure of Microwave Oven

143 Figure 4-4: The Predictor Importance of Microwave Oven

144 Figure 4-6:The Predictor Importance of Pacifier

145 Figure 4-5: The Bayesian Network Structure of Pacifier

146 Table 4-8: The Probabilities of Star Rating of Microwave Oven

Probability				
1	2	3	4	5
0.26	0.08	0.10	0.20	0.35

147

148 Table 4-9: The Probabilities of Star Rating of Pacifier

Probability				
1	2	3	4	5
0.05	0.06	0.09	0.18	0.61

According to the results, consumers are most concerned about the microwave oven function, quality and price, followed by after-sales service, and finally volume. Consumers are satisfied with the quality, function and price of the current microwave oven on the market, after-sales service to be improved but still relatively satisfied, and volume is a point of consumer dissatisfaction.Overall, the microwave's Star Rating isn't bad, but consumers aren't very happy. There is a lot of room for improving word of mouth.

Similarly, the factors influencing consumers' satisfaction with the pacifier were in order of appearance, health, texture, quality and size. Consumers were satisfied with appearance, health and texture, but there was room for improvement in texture and quality, and consumers were least satisfied with the size of the nipple. Finally, the overall satisfaction of the pacifier was relatively high.

5 Solving Specific Problems

5.1 Measurement of Product Reputation Based on Time Series

Taking hair dryer as an example, we selected the product with the highest sales volume. To calculate the product's reputation, we listed the following formula：

$$RS = (\overline{PO} + \overline{DU} + \overline{PR} + \overline{CO} + \overline{AP}) \div 2 + 2.5$$

$$R = 0.2 \times \mathrm{S} + 0.8 \times RS$$

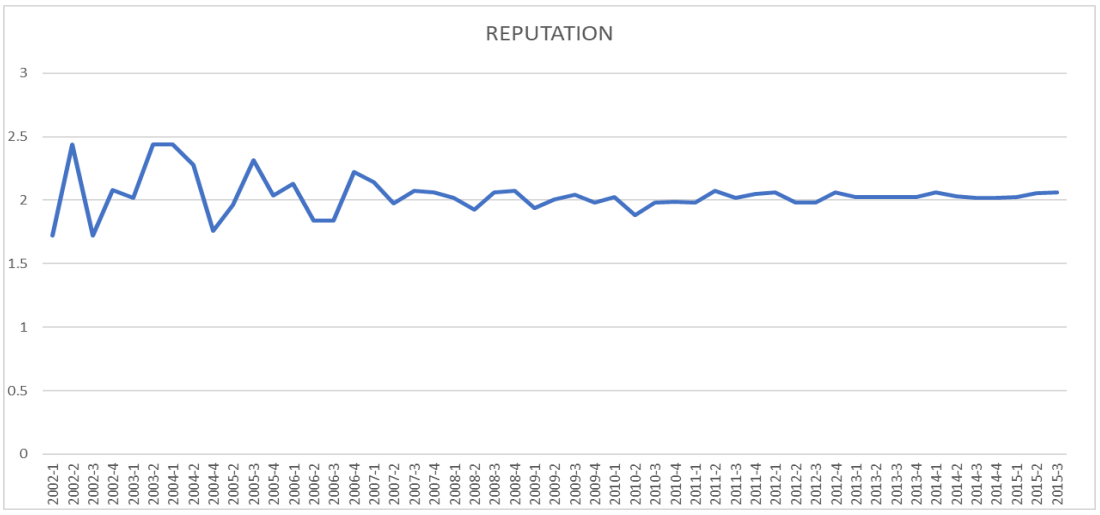
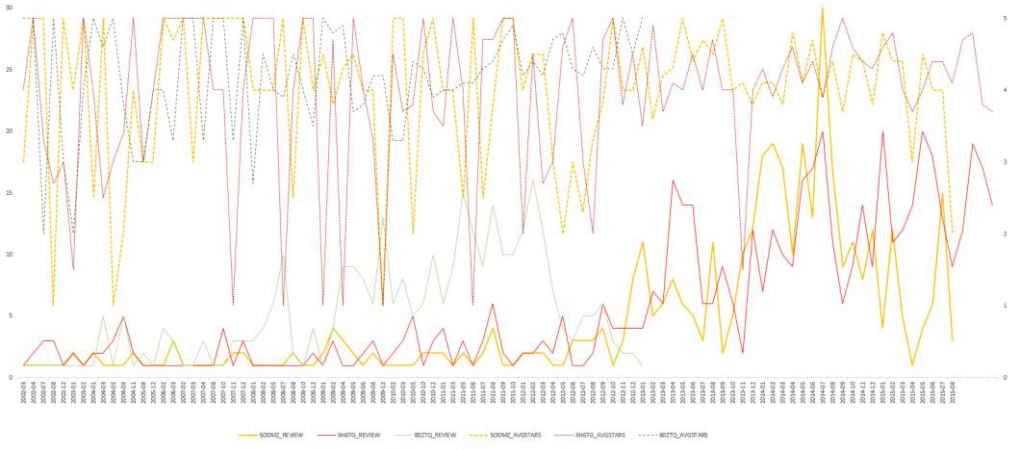


Figure 5-1: Time-based Changes in Product Reputation

As can be seen from the figure, there are some fluctuations in the reputation of the product when they were first launched. With the continuous improvement of the product characteristics over time, the reputation of the product will tend to be stable.

5.2 Correlation Between the Star Rating and the Number of Reviews

We extracted the review data of the top three products of hair dryer sales. The Star rating and the number of reviews are shown as follows according to time:

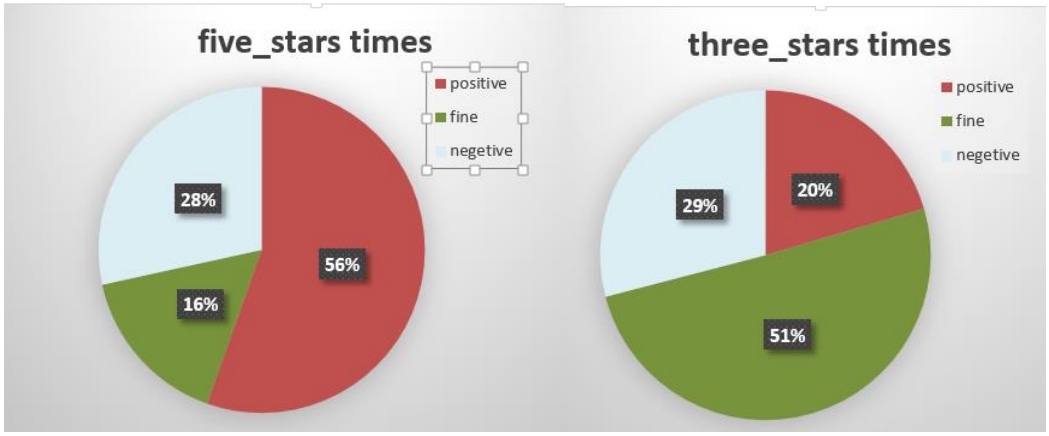


It can be seen from the figure that Star Rating is correlated with the number of reviews. In general, the number of reviews fluctuates with the change of average Star Rating,

and the trend of change is generally the same. It indicates that the two are positively correlated, and the effect has a certain lag.

5.3 The influence of characteristic emotion words based on text reviews on Star Rating

We extracted the data with Star Rating of 1, 3 and 5, and analyzed their reviews in the database with emotion words. The following figures are obtained:



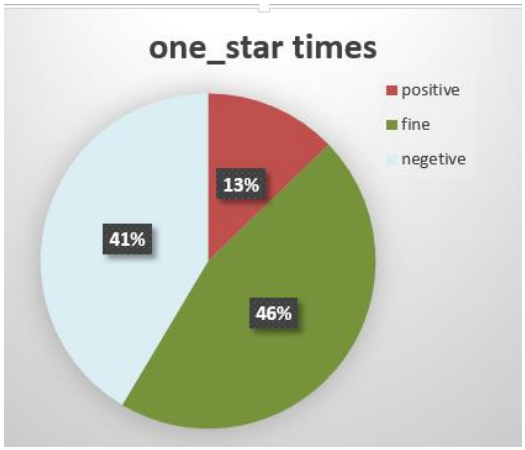


Figure 5-2: Category Distribution of Emotion Words in Five-star Reviews

Figure 5-3: Category Distribution of Emotion Words in Three-star Reviews

Figure 5-4: Category Distribution of Emotion Words in One-star Reviews

It is obvious from the figures that the distribution of emotional words in different Star

Rating is different. It can be said that the text-based content reviews are related to the rating levels.

6 Model Strengths and Weaknesses

While the model we built has some advantages, we also found some disadvantages in the process of applying it.

6.1 Strengths

We chose the Bayesian model to solve the main problem, which helped us to make a better prediction of the unknown relationship between the factors. This study is based on online review data to measure the composition of consumers' satisfaction with different products, and uses the high-frequency keywords in the reviews and the emotional attributes of the keywords to describe the product concerns. On this basis, we established an emotion dictionary, set up quantified scoring criteria and digitized online comments. Finally, the obtained data was imported into the software to establish a Bayesian network model for satisfaction analysis. Data processing method is one of the advantages of this study. The modeling steps are clear and concise.

6.2 Weaknesses

Limited data and incomplete samples. We studied only representative product data. The insufficient sample size of the data may lead to the lack of characteristic factors, thus affecting the accuracy of the satisfaction model. At the same time, in the process of quantitative rating, due to the different contents of reviews and the incomplete content of the emotion dictionary we established, online reviews did not completely include the extracted characteristic factors, which resulted in the loss of some data, thus affecting the accuracy of the prediction results.

7 Conclusion

We conduct data cleaning on the review information of three kinds of commodities,

extract the characteristic factors in the online review and establish the emotion dictionary. Then, based on the effective information of the quantitative scoring rules, the data was used to establish the bayesian network model of influencing factors of satisfaction. Our conclusions are as follows:

1. According to the results we obtained, it is found that different products have different factors affecting consumer's satisfaction. The major factors affecting the hair dryer are power, appearance, Comfort, Price. Function, quality, price, after-sales service and volume are the main factors affecting the microwave oven. The main factors influencing the pacifier are appearance, health, texture, quality and size. We recommend that companies that make these types of products focus their design efforts on these areas.

2.After analyzing the specific text content of the review, we find that the emotional tendency of the review has a certain influence on the sales volume of the product. For companies, this is an idea that can be used. That is, by improving the product's star rating to improve the product heat.

Due to time and data limitations, we still have a lot of room for improvement in the modeling process. If we have more time and data, we can analyze three different types of products of each type separately as the target variable to obtain the characteristic analysis of each type. In this way, a more accurate Bayesian network structure can be established and a more accurate product feature evaluation scheme can be determined, so as to help the company more accurately understand the needs of consumers and design products with higher satisfaction.

8 Letter to the Marketing Director of Sunshine Company

Dear Sir/Madam,

In response to your questions and requests about the design and sales strategy for the next three products， we are writing to inform you of our work.

We used the data you gave us to carry out relevant analysis and got many results. As

for the design strategy of the product, we will give you different suggestions for different products. First, the hair dryer. We recommend that you focus on power, appearance, appearance, comfort, and price. These factors have a greater impact on consumer satisfaction. Function, quality, price, after-sales service, volume are the important factors for a microwave oven. We suggest that you design the microwave oven with a larger capacity, which should help you to compete with your competitors. Finally, about the baby pacifier, you should focus on these aspects of the design appearance, health, texture, quality, size. Currently, consumers are less satisfied with the size of the existing pacifier on the market.

Then there is the sales strategy of the product. We found that the Star Rating of the product has a certain relationship with the content of review. We suggest that your sales department should pay more attention to the Star Rating of products to avoid its low value. Keeping the Star Rating at a high level will help sales. When a product is first launched, its reputation, the content of ratings and reviews, can fluctuate wildly. This is normal in the short term. However, you need to pay close attention to these fluctuations and make continuous improvement in subsequent products, so that the product can have a stable good reputation and high sales in the future.

Thank you again for choosing our team to finish this work. We do hope those strategies will be helpful to you.

Yours, sincerely

Team#2011014

9 References

[1] Jiang, HM. Predicting future importance of product features based on online customer reviews[J]. Journal of mechanical design. 2017, 139(1 11：1 163 一 l 173

[2] Greaves, F. Use of sentiment analysis for capturing patient experience from free-text comments posted online[J]. Journal of medical internet research, 2013, 15

(11):154—159

[3] Caballero, D; Caro, A; Avila, M. New fractal features and data mining to determine food quality based on mri[J]. IEEE LATIN AMERICA TRANSACTIONS, 2017, 15(9): 1777—1784

10 Appendix: wordcount.java

```
package wordcount;

import java.awt.List;
import java.io.BufferedReader;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.util.ArrayList;
import java.util.Collections;
import java.util.Comparator;
import java.util.Map;
import java.util.Map.Entry;
import java.util.TreeMap;

public class wordcount {

    public static void main(String[] args) throws IOException {
```

```
298         // TODO Auto-generated method stub
299         BufferedReader br = new BufferedReader(new FileReader("E://test.txt"));
300         ArrayList<String> lists =new ArrayList<String>();
301         String readLine =null;
302         while((readLine=br.readLine())!=null){
303             String[] wordsArr1=readLine.split("[^a-zA-Z]");
304                 for (String word : wordsArr1) {
305                         if(word.length() != 0){
306                             lists.add(word);
307                         }
308                 }
309         }
310         br.close();
311         Map<String,Integer>wordsCount = new TreeMap<String,Integer>();
312         for(String li:lists){
313             if(wordsCount.get(li)!=null){
314                     wordsCount.put(li, wordsCount.get(li)+1);
315         }else{
316             wordsCount.put(li, 1);
317         }
318         }
319         SortMap(wordsCount);
320         System.out.println("sdfsfs");
321     }
```

```
322
323         private static void SortMap(Map<String, Integer> oldmap) {
324
325                 ArrayList<Map.Entry<String,Integer>> list = new ArrayList<Map.En-
326     try<String,Integer>>(oldmap.entrySet());
327
328                 Collections.sort(list,new Comparator<Map.Entry<String,Integer>>(){
329                         @Override
330                         public int compare(Entry<String, Integer> o1, Entry<String, Integer>
331     o2) {
332                                 return o2.getValue() - o1.getValue();
333                         }
334                 });
335                 for(int i = 0; i<list.size(); i++){
336                         System.out.println(list.get(i).getKey()+ ": " +list.get(i).getValue());
337                 }
338         }
339
340
341 }
```