Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2010764

A evaluation system based on PRES algorithm

Summary

In this paper, we build a PRES product review evaluate system. This model can integrate the sales volume, evaluation and other parameters of a commodity on the market to score it, and take it as a standard to measure its success. In addition, it can calculate its market competitiveness in the future. In the process of statistical analysis, we summarize the common points of competitive products, which reveals the characteristics of excellent products and the shortcomings of failed products. We hope that our analysis can help sunshine evaluate and improve its product market competitiveness.

First, we preprocess the data. Our focus is on the analysis of text data. We use the TextRank algorithm to extract the keywords in the comments to reveal the product features and functions that users are most concerned about. On the other hand, we use nltk to conduct emotional analysis to further quantify the emotional tendency of the comments. These two factors are one of the important indicators for us to evaluate the success of commodities. Secondly, we analyze the data and make assumptions. We filter the data according to the integrity and usefulness of the data, and actively remove some outliers and missing data. After classifying the data according to different commodity types, we focus on product sales volume and star ratings, carry out data visualization analysis combining with the time axis, and speculate the relationship among reviews, stars and sales volume in the data, as well as the extent of change of sales volume and stars under the influence of time.

Based on the above two steps, we designed PRES (product review evaluation system). According to the user comments and star rating of the product, we have quantified the impact of user comments, combined with VP (verified purchase), VI (vine), Sr (star rating), HV (helpful votes) index multi-dimensional modeling. To find the most suitable model parameters, we use the genetic algorithm to optimize the curve parameters. Finally, we obtain a score index that can fit the sales situation very well, and the model can evaluate whether a product can have a good sales volume through time and scoring details, which can achieve 90% correlation with the overall market trend.

Subsequently,we use two classification models to predict whether the goods are successful or not. Here we consider the effect of time and market comprehensively, and the standard to measure the competitiveness of products is the difference between the growth rate of their sales volume and the average growth rate of sales volume of similar products. At first, we use the decision tree of Gini coefficient to classify, but the accuracy of the model is poor. So we turn to SVM(support vector machine) model. Although there are few kinds of products that we can use to analyze, the model can still achieve 70% classification accuracy, which proves that the product success can be predicted by comments and time analysis in big data.

Finally, the model is validated by sensitivity analysis. The satisfactory results enable us to apply the model to the actual situation to solve practical problems. In addition, we also identified keywords that can express a lot of information about products, such as born, fit, plastic, etc.

Keywords: Textrank; Data visualization; Product review evaluate system ; SVM; Genetic algorithm

Contents

1 Introduction **2**

1.1 Background 2

1.2 Our work 2

1.3 Assumptions and Notations 2

1.3.1 Assumptions 2

1.3.2 Notations 3

2 Data analysis **3**

2.1 Correlation analysis between sales-volume and star-rating 3

2.2 Correlation analysis between sales-volume and time 5

2.3 Correlation analysis between star-rating and time 6

3 Data preprocessing **8**

3.1 The importance of reviews 8

3.2 Keyword extraction method based on Textrank 8

3.3 Finding high quality evaluation words 10

4 Model construction **11**

4.1 Keyword mining and comment quantification 11

4.2 Identify data measures 12

4.3 Product growth rate predicting system 14

4.4 Chain reaction phenomenon 15

5 Validating the Model **16**

6 Strengths and weaknesses **17**

6.1 Strengths 17

6.2 Weaknesses 18

7 Letter **19**

Appendices **22**

1 Introduction

1.1 Background

In amazon's online marketplace, customers have the opportunity to rate and review purchased items. Ratings, also known as "stars", are used by users to indicate their satisfaction with a product by giving it a rating of one to five stars. Reviews are where users submit textual reviews to express more opinions and information about the product. Both ratings and reviews are an important factor in influencing other users' willingness to buy a product. In addition, other users can comment on the effectiveness of these reviews in assisting their own product purchase decisions, which is called helpfulness rating. Also, companies can use the data to analyze market and product functional design indicators to reveal whether the product has potential success in the market.

The sunshine company plans to market three new products, including microwave ovens, baby pacifiers and hair dryers. We were asked to provide an online sales strategy by analyzing several indicators such as existing market reviews and ratings, measuring and quantifying them, and measuring their relationships. In addition, give reasonable advice to the sunshine company on how to design the important features of the product to improve its attractiveness.

1.2 Our work

In this article, we divide our work into the following sections.

- Preprocess data, draw statistical charts and word clouds to carry out data analysis which based on visualization technology.
- Using algorithms for keyword analysis to calculate impact factor of review and combining product rating information and other indicators to establish an evaluation model to quantify users' evaluation of the product.
- Analyze sales trend and influencing factors to establish a prediction model of sales growth rate.
- Evaluate and improve the model and provide reliable suggestions for the products to be marketed by sunshine company based on the results.

1.3 Assumptions and Notations

1.3.1 Assumptions

In our model, we make the following assumptions.

- we assume that the product company will not experience a precipitous drop in sales due to bankruptcy or major product issues
- we assume that the external environment such as policy, logistics and natural disasters, will not change significantly, resulting in drastic changes in sales.
- we ignore the influence of other categories of products or industries on this product, that is, the product will not be affected by factors other than its own quality problems and other similar products.
- we assume that most consumers are rational consumers who can correctly analyze the quality of products and use other reviews to evaluate the value of products.

1.3.2 Notations

Notation	Meaning
I	The total score of a review
V_0	number of helpful voting
V_1	number of unhelpful voting
Q	quality information of a sentence
ρ	the emotionality of a sentence
R	Emotional quality factors
I	The impact of reviews on product sales
SPM	sales per month of a product
SV	Sales Volume
SR	Star Ratings

Table 1: notations

2 Data analysis

2.1 Correlation analysis between sales-volume and star-rating

Take the sales of microwave ovens for example. First, we examine the star_ranking which is a very obvious quantitative characteristics. We choose 20 products with middle sizes of sales volume. The products are divided by product_id, and are plotted from top to bottom in terms of sales. It can be seen from the chart 4 that:

1. Sales of different products vary, with the first product in the chart alone being six times more than the last product in the chart. In the whole data set, some products sell hundreds of sales, and some products sell as low as a single digit.

2. In general, there are several types of sales volume and star composition of each product.
   - high praise leading type, such as the No.12 product.
   - poor evaluation leading type, such as No.3 product.
   - equilibrium type, such as No.11 product.
   - bipolar type, such as No.8 product
3. Generally speaking, the sales volume of products is positively related to the acquisition rate of high star rating, which is very reasonable, but there are also exceptions. For example, No.3 products have a high 1-star rating, while No.12 products have the opposite. The reason behind it remains to be explored
4. 5-star and 1-star evaluation of most products account for a large proportion. In fact, in a total of 1615 evaluations, 1 star 402, 2 stars 112, 3 stars 134, 4 stars 300, and 5 stars 667, we can see it from [1]. This shows that buyers tend to be extreme in evaluation, which may mean that we need to set a certain weight to buffer when calculating stars as indicators, and also shows the importance of text analysis results for comments when measuring the final evaluation.

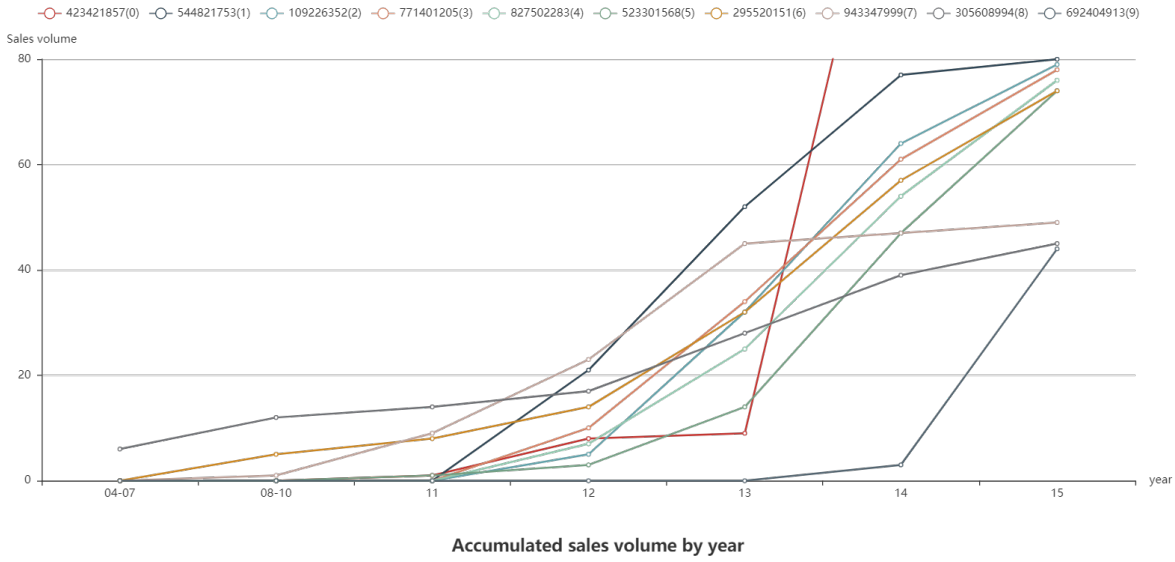
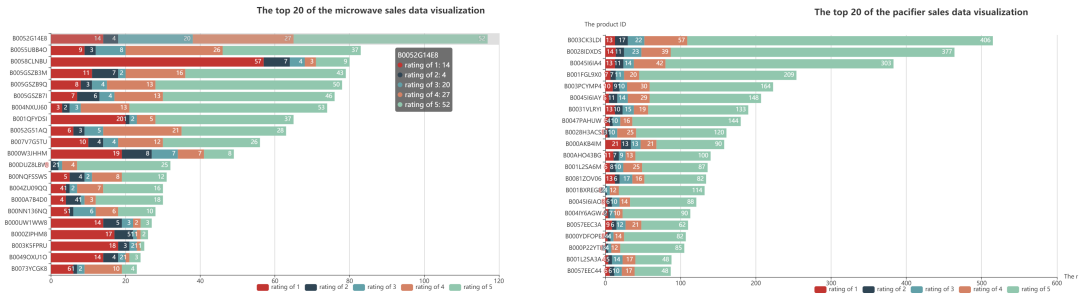


Figure 1: star rating composition

5. We compare the chart of microwave oven with the chart of baby's pacifier horizontally, and the composition of the two is quite different. In the case of infant pacifiers, 5 stars are highly praised, usually more than 60%, and 1 star is very few. This is related to the product category and value. The low price of infant pacifier and low product quality differentiation make it

less possible for buyers to give 1-star poor rating as a small object in life. It also shows that it is difficult to use a model to evaluate all three kinds of products.



(a) microwave oven (b) pacifier

Figure 2: accumulated SV-SR

2.2 Correlation analysis between sales-volume and time

We select the top 10 products in total sales and analyze the total sales figure 3. As the latest date of the data given is August 2015, the data in 2015 is incomplete and cannot fully reflect the total sales volume. What we have done is to expand the sales volume in 2015 to 1.5 times of the original according to the proportion of time interval. It can be seen from the figure that the sales of the products that can be ranked in the top 10 of the total sales volume show an upward trend with the growth of time, and they all conform to the law of slow rise first, then sharp rise before and after 2012, and then slow growth before and after 14 years.

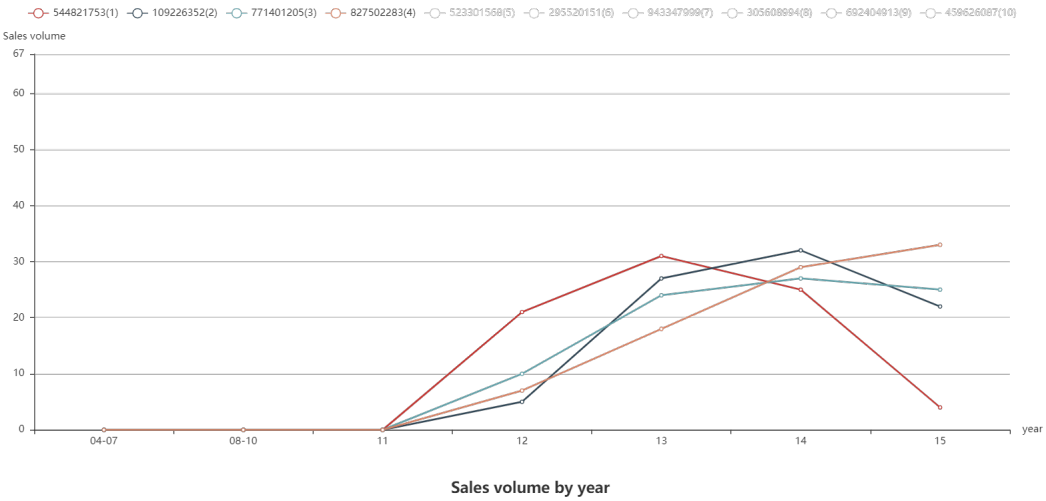
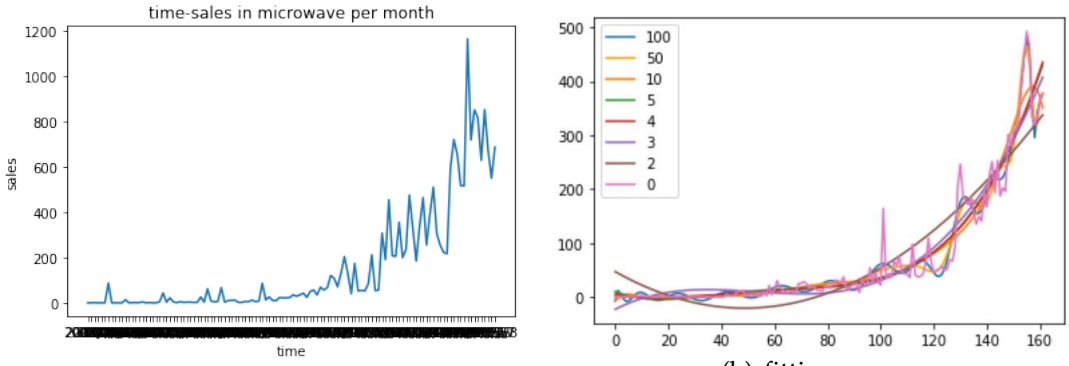


Figure 3: microwave oven SV-time

So when we analyze and judge the sales volume of a product, we can't simply

take whether the sales volume increases as the judgment index of a product. Because of the promotion of online shopping, the market is becoming larger and larger, more and more consumers are buying hairdryer on Amazon, so the sales growth is a very common phenomenon, and it is not easy to judge the product success through the sales growth.



(a) raw curve (b) fitting curve

Figure 4: microwave oven monthly SV

2.3 Correlation analysis between star-rating and time

We take the time as the x-axis, and take the proportion of different star numbers (hair_dryer) given by consumers at different time points as the y-axis to make this picture. Since 2003, the proportion of different stars has gradually stabilized and converged to a specific proportion.

At the beginning, due to the small sales volume, customers do not understand the products and other reasons, the proportion of different stars presents irregular situation. With the increase of sales volume, products are gradually understood by customers, target consumer groups are gradually determined, and the proportion of stars given by customers is also gradually determined. It can be seen that the proportion of 5 stars fluctuates around 60%, which is far higher than other stars, indicating that the consumer group is generally recognized for hair dryer products. The next steps are 4 stars, 1 stars and 3 stars. 2 stars accounts for the smallest proportion.

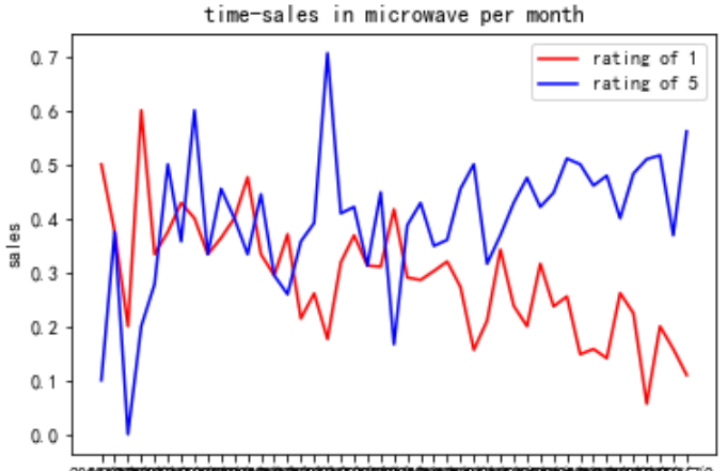
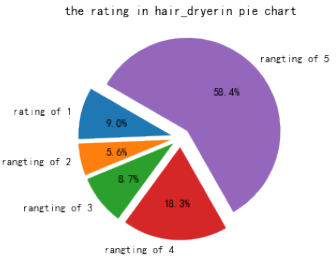
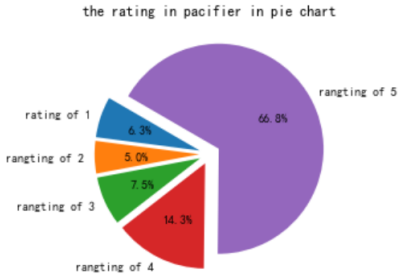


Figure 5: SR-time curve

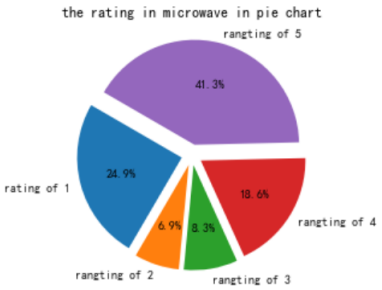
This pie chart is obtained by extracting the star rating data in 2015. The results are basically consistent with those of our curve analysis. We think that giving 5 stars is a good comment, giving 3-4 stars is a medium comment, giving 1-2 stars means a business trip comment. The positive rate is 58.4.



(a) hair dryer



(b) pacifier



(c) microwave oven

Figure 6: SR composition pie chart

3 Data preprocessing

3.1 The importance of reviews

We have 15 columns of data, but only part of them can be used in the analysis. For example, the marketplace is discarded because of its unique value "US" . In addition, there are some data information redundancy, such as product_title and product_id are two groups of almost one-to-one values (only very few cases exceptions ,which can be ignored), so we only keep one of them. Of course, the two most interesting things to look at when looking at all the data are the ratings and the reviews.

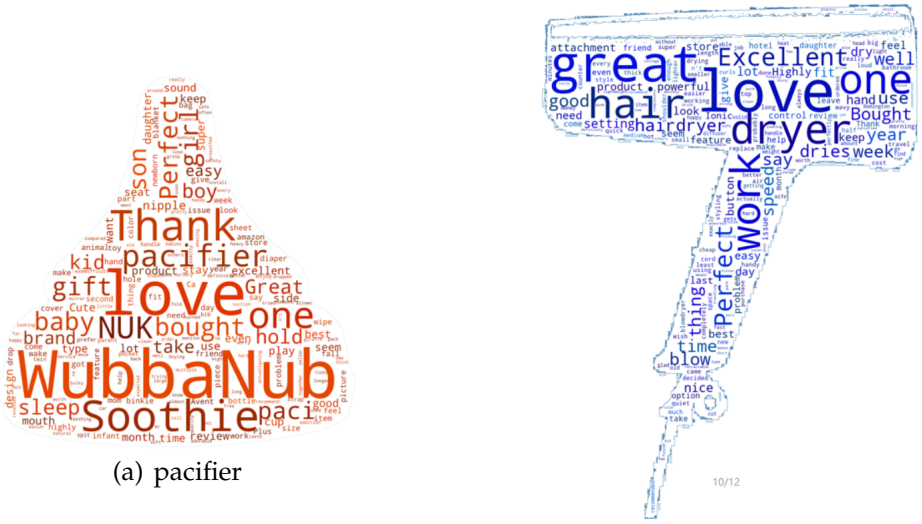
An evaluation of a commodity consists mainly of reviews and ratings. The rating is a discrete value of 1-5 stars, which directly reflects the buyer's satisfaction. Reviews are textual, not as intuitive as ratings, but that also means there is more information to be mined. In reality, when potential buyers enter the product details page, they almost always refer to the reviews of other buyers in the review section. A star rating can only tell potential buyers whether the product is good or bad, but cannot reveal the reason why the product is good or bad. However, specific reviews can have a big impact on buying decisions. For example, a five-star rating of a microwave oven with praise for its energy efficiency might prompt some hesitant environmentalists to make an immediate purchase. Complaints about noise in a 1-star rating of a hair dryer can lead some noise-conscious buyers to decide not to buy.

After analysis, we believe that the information in the reviews can be measured by two indicators, one is the degree of importance and the other is the degree of satisfaction. The two complement each other. The degree of importance reflects the importance of the evaluation in the eyes of potential buyers, while the degree of satisfaction reflects the degree of praise or criticism of the product.

Textrank algorithm was used to calculate the degree of importance, and nltk emotion analysis was used to calculate the degree of satisfaction. We will introduce these two algorithms in the following sections.

3.2 Keyword extraction method based on Textrank

There were a lot of reviews from users, and it contained a lot of information. These evaluations include valuable information, such as space-saving design or great size, which can provide us with suggestions on designing product, as well as information that is not explicitly helpful for improving the product, such as “good product” and “It works bad”. So we need to extract the keywords.



(a) pacifier

(b) hair dryer

Figure 7: word cloud

We used the Textrank algorithm to transform the text into a graph model and extract keywords from the ratings given. Each words in the reviews will serve as a vertex for graph. Then we will record the information of edge connections among all vertices. Here we set the windows size to 3, which means we check the adjacent 3 word to get the edge connections information.

Algorithm 1 Textrank algorithm

$n \leftarrow$number of word in text
$edge\_weight[n][n] \leftarrow \{0\}$
for each word w in text **do**
  for every another word v in text **do**
    if $adjecent(w, v)$ **then**
      $D \leftarrow distanct\ between\ w\ and\ v$
      $edge\_weight\,[w]\,[v] \leftarrow \frac{1}{D}$
    end if
  end for
end for
$\epsilon \leftarrow 0.0001$
$score[n] \leftarrow \{0\}$
repeat
  for every word w in text **do**
    score[i] $\leftarrow 1 - d$
    for every another word v in text **do**
      $score[i] \leftarrow score[i] + d \times [(\frac{weighted\_edge[i][j]}{inout[j]}) \times score[j])]$
    end for
  end for
until exist a score[i] changed more than ϵ

The score is iteratively updated until convergence. Here we get the convergent score through 33 iterations. On the basis of word score, we can calculate the score of the phrase. First, identify the phrase in the text, and then add up the score of the keyword in the phrase to get the score of the phrase. This algorithm is more likely to extract long phrases from text, because long phrases are more informative than words, and we are better able to extract the information we need from them. Below are some of the key words in the comments given by 5 stars microwave purchasers. We can see that small microwave ovens are the most concerned feature of consumers, which can be speculated to be related to people living in big cities facing smaller living areas

key word	score
great space saver	25.743746042251587
ikea wall cabinets	24.604225158753456
good deal great space-saving option	24.40525159239769
small corner space	21.745492219924927
great product thirty inch wide	21.708362460136414
small space	18.850110054016113
happy great stuff	18.573667466640472
great price	17.894643306732178
great size	16.759092807769775
counter top space	16.654603481292725
small sized unit	16.587942510843277
small kitchen	16.32162570953369

Table 2: key word score table

3.3 Finding high quality evaluation words

After calculation of TextRank algorithm, we obtained the keywords and key phrases of the five grade evaluation of three products. The data already reflect the features and design of the product that users care about, but there are some flaws in the results. For example, phrases such as "it works great" and "I like it" appear frequently, but do not provide us with information about the functional design of the product. To solve this problem, we use manual tagging to remove items that are not helpful for finding keywords that describe product features and functionality. As for the emotional tendency of the comment, whether it is praise or criticism, we completed it through the emotional analysis of NLTK.

(a) hair dryer (b) pacifier

(c) microwave oven

Figure 8: high quality word cloud

4 Model construction

4.1 Keyword mining and comment quantification

After the above processing, we get the most common words in all reviews. After word meaning analysis, we select the 600 words that can increase the importance of reviews and the corresponding importance for each product to construct a dictionary. We believe that these words represent the most concerned aspects of the product, which may be the function, design, quality, etc. of the product. When these words appear in a comment, they can attract the attention of other potential buyers and affect the score of the product in the hearts of potential buyers, thus affecting the purchase decision.

With this dictionary, we can quantify the impact of comments and emotional tendencies. We select a comment for analysis, and divide the comment into multiple sentences. We search each sentence by word to match the dictionary, and

calculate the total importance of a sentence. Then, the emotional analysis of the sentences is carried out, and a weight index of positive, negative and neutral emotions is analyzed as a measure of satisfaction. See the table below for these indicators. The results of all sentences are combined to get the total comment score, which takes the importance and satisfaction into account. The higher the score, the better the sales growth.

indicators	score
Total_weight	The importance of a sentence in a comment
Review_pos	Emotional positive index of a sentence in a comment
Pos_weight	$Review_pos \times Total_weight$
Review_neg	Emotional negative index of a sentence in a comment
Neg_weight	$Review_neg \times Total_weight$
Review_neu	Emotional neutral index of a sentence in a comment
Neu_weight	$Review_neu \times Total_weight$
compound	Generaltendency of sentence emotion

Table 3: sentiment analysis result

4.2 Identify data measures

First of all, we consider that the information contained in the text is the largest, and through visual analysis we find the difference relationship in the text, so we use textrank to extract keywords in the text in the data preprocessing, and use a set of scoring standards to quantify the useful information contained in the comments, and get the value Q. In addition, we also use emotion analysis to score the emotion of the comment as an important evaluation standard for sales indicators ρ.

$$\rho_{pos} = \frac{\sum_{i=1}^{n} \rho_{pos}^{(i)}}{n}$$

$$\rho_{neg} = \frac{\sum_{i=1}^{n} \rho_{neg}^{(i)}}{n}$$

$$\rho_{neu} = \frac{\sum_{i=1}^{n} \rho_{neu}^{(i)}}{n}$$

Then, according to each sentence, we get the product of the user's emotionality ρ and the quality information Q in each sentence, and we can get the quality information of different emotionality in each sentence.By summing up and averaging the emotional quality of each sentence in a comment, we can get the quality of the comment and the emotional tendency of the comment, which plays an important role in analyzing the quality of the product. In addition, considering the uneven number of comment sentences, it is necessary to consider the influence of sentence number on the quality in the calculation. When the sentence is too short, design a penalty coefficient $(1 - \frac{1}{n+1})$.Where n is the number of sentences

in a comment.Considering the possibility of high quality of long sentences, long sentence complement coefficient is introduced. $\ln(n+e)$,where n is the number of sentences in a comment. In conclusion, the formula is obtained through simplification

$$R_{pos} = \frac{\ln(n+e)\sum_{i=1}^{n} Q^{(i)}\rho_{pos}^{(i)}}{n+1}$$

$$R_{neg} = \frac{\ln(n+e)\sum_{i=1}^{n} Q^{(i)}\rho_{neg}^{(i)}}{n+1}$$

$$R_{neu} = \frac{\ln(n+e)\sum_{i=1}^{n} Q^{(i)}\rho_{neu}^{(i)}}{n+1}$$

Further, considering that the positive effect of positive emotion in high star is higher than that in low star, the product star and emotional tendency are introduced in the subsequent calculation to calculate the comprehensive impact of reviews on product sales in positive or negative emotions. Most importantly, considering the opposite of negative emotion and positive emotion, a coefficient C is added to neutral emotion. C is the overall emotion of comment from emotion analysis, which is used to guide neutral comment to the appropriate emotional quality score.

$$I_{pos} = S_r\rho_{pos}R_{pos}\epsilon_{pos}$$

$$I_{neg} = (5 - S_r)\,\rho_{neg}R_{neg}\epsilon_{neg}$$

$$I_{neu} = S_r\rho_{neu}R_{neu}\epsilon_{neu}C$$

$$I = I_{pos} - I_{neg} + I_{neu}$$

However, only analyzing comments is not enough. We consider that other variables, such as vine and helpful votes, can expand or reduce the effect of comments on products. In the aspect of considering voting, the voting coefficient V is determined jointly by the helpful voting v_0 and the unhelpful voting v_1. By designing a formula, the number of votes can be used as a coefficient to effectively evaluate the quality of the comment content.

$$V = \log(V_0 + e) - \log(\theta_0 V_1 + e)$$

In addition, we consider V_n and V_p. And also regard them as a coefficient that can evaluate the quality of evaluation. Considering the two sides of V_n, that is to say, when there is one star, there will be greater negative emotional resonance of users and when there are five stars, there will be greater positive emotional resonance of users. We regard V_n as a boolean variable, and it is linked with star level. However, customers who buy at a discount may have relatively less requirements for product quality. Therefore, we regard whether to buy at a discount as a factor affecting product sales. Finally, we fit a general trend according to the total sales volume of a class of products.In fact it's like a exponential function.

$$SPM = f(x), \quad f(x)\ is\ fit\ from\ the\ sales\ per\ month\ curve$$

All the parameters obtained from our analysis are taken as a coefficient and multiplied by a weight $Weight$. The formula described above is as follows.

$$Y = SPM \times Weight \times I(\alpha^V)(1 + \theta_1 \frac{(S_r - 3)}{2} V_n)(1 + \theta_2 V_p)$$

4.3 Product growth rate predicting system

In the previous analysis, we combined reviews, helpful_votes, and other parameters to get a text-based score for evaluation. Besides, the star rating given by consumers is the most direct indicator of the quality of a product, so we use the average rating as ratings-based measures.

Input:

We select the data of the products in the last six half years (three years), including the average of the comprehensive scores of consumers and the average rating as the input model of a product (2×6 numbers).

Output:

Whether the growth rate of a product in the same period is higher than the average market level.

Here, we use the SVM (support vector machine) model for classification learning. Take the 6 values as the input. In the output, 1 represents the potential successful product, while 0 represents the opposite. The label is regarded as a Boolean value, so the problem is changed into two classification problems.

we found that the data set of hairdryer has good data integrity and data stability, so we choose this data set for analysis. But on the other hand, there are too few kinds of hair dryer brands in the market. Through data cleaning, 56 groups of product data can be obtained finally. That is to say, the problem belongs to small data classification learning, and the effect of using small data machine learning classification algorithms is better. However, due to the small amount of data, the model may be unstable. We import the input values into the model and get the following results:

$$\begin{pmatrix} & False & True \\ False & 18 & 6 \\ True & 8 & 7 \end{pmatrix}$$

The result is derived from the confusion matrix, and the model has some instability with an accuracy of 67.8%. Although the value is not very high, it still proves the feasibility of predicting the success of the product by past star ratings and reviews through historical data analysis. And we think there will be better performance in larger data sets.

4.4 Chain reaction phenomenon

When analyzing the star-sales bar chart of microwave ovens, we found some interesting products. From the perspective of macroeconomics, the overall sales situation of the market is on the rise over time, and the annual increase is increasing. However, it was found that did not increase the annual average sales volume of the product product_ID B000UW1WW8 from 13 to 15 years and has maintained single-digit annual sales.

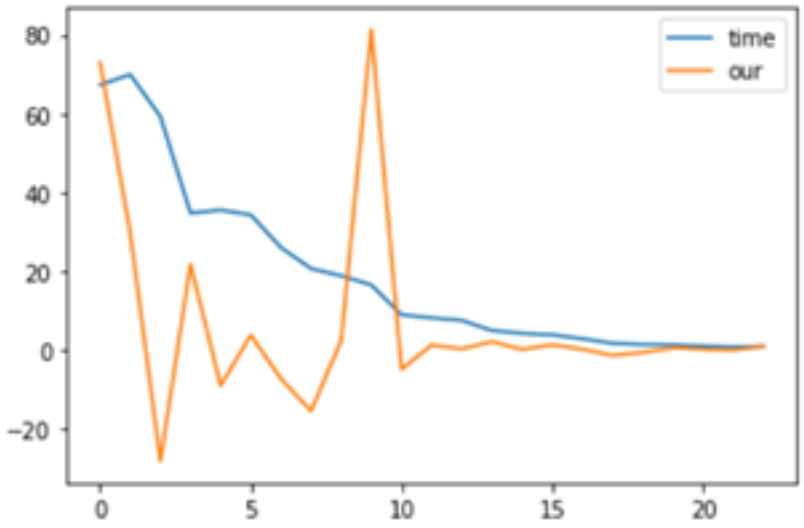


Figure 9: SV-time curve of B000UW1WW8

In order to analyze the reasons, we selected the product review data of product, of which product_ID is B000UW1WW8, and computed the data via our model. We found that the product scored very high from 2011 to 2013, which fully explained that the curve rises in 11 and 12 years matched the market curve better. But from early 14 to 15 years, the scores evaluated by the model began to fall sharply, and gradually fell below the average. This also explains why the time sales curve of the product between 2013 and 2015 began to decline after the sales volume reached its highest value.

We analyzed and looked for some reviews of this product. With the help of Excel tools, we could easily find that there were 4 low-star reviews back in 2008, and the reviews received a lot of helpful votes. Since people have also been more tended to consider it as low-star product. The product's average star rating from 2008 to 12 was only 2. With the passage of time, after the existence of some high star votes, I think that the product made up for the shortcomings through technical advancement, and then the high star evaluation appeared after the first high star, but the sales volume was still far below the market average curve due to the past low ratings. Therefore, this product was considered as a failed product because of the large number of low-star.

We have made this phenomenon a chain reaction reaction phenomenon, that is, due to the existence of Extreme emotions in a review, the overall star rating would be influenced. We also used the coefficient of extreme emotions and star ratings as an influence coefficient to modify the overall score in PRES.

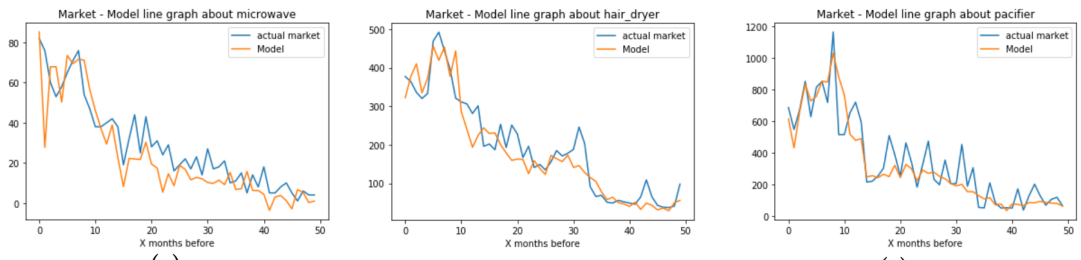
5 Validating the Model

In the pre,u ser reviews and star rating can be quantified, and product sales can be evaluated in combination with time. We use genetic algorithm to optimize the parameters. We need to fit all the coefficients and weights, which is a very large project. Fortunately, with the help of mathematical experience, a general parameter range can be determined so as to reduce to the best. Finally, our parameters are as follows:

category	theta0	theta1	theta2	alpha
pacifier	1.4335	0.1902	0.5494	1.6083
microwave_oven	1.2483	0.2414	0.3870	1.1418
hair_dryer	2.6520	0.1896	0.5420	1.9970
category	epsilon_neu	epsilon_pos	epsilon_neg	weight
pacifier	19.5033	36.9083	66.6574	0.0491
microwave_oven	22.8934	5.1082	176.7704	0.0496
hair_dryer	32.0008	76.3114	179.6366	0.0322

Table 4: parameters

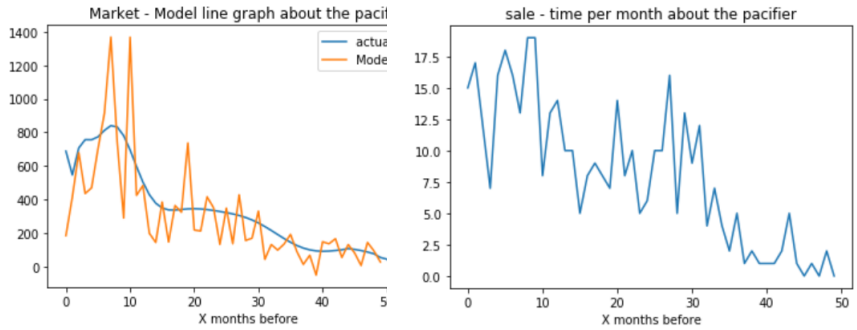
Through these three values combined with our time curve, we can score through user evaluation. We get an impact factor of the evaluation on the user evaluation at a certain time. The sum of all influencing factors in a range is the quantification of the monthly sales volume. In the time dimension, it can be seen that the influencing factors are related to the monthly sales volume, and the correlation coefficient is about 0.9 through calculation. This shows that our model can fit the market sales trend to a great extent.



(a) (b) (c)

Figure 10: fitting curve

It can be seen that our model can perfectly fit the market sales curve under the optimal result obtained by genetic algorithm. And when we analyze a single commodity, we can well fit the trend of sales of a single commodity. This is because we combine the positivity and negativity in the comments. Take a product at random. This time, choose a hair dryer with the product ID of b0028idxds. We can see that the overall monthly sales curve of this hair dryer rises, which is very similar to that of the market. They all rise first, reach the peak five months ago, and then decline slightly. And the model curve completely expresses this trend. Through calculation, the correlation coefficient between the model fitting curve and the actual monthly sales volume curve is 0.8, which is a good prediction of this data!



(a) (b)

Figure 11: fitting curve of pacifier

6 Strengths and weaknesses

6.1 Strengths

- We have fully considered all available data and possible situations, and provided lots of coefficients for the model to be adapted to a variety of products, not just hair dryers, microwave ovens and pacifiers.
- Due to the existence of dirty data, our data preprocessing is effective to remove the dirty data. Moreover, outliers have been modified sufficiently, such as interpolation and coefficient correction, in order to maximize data integrity.
- We used GA(genetic algorithm) so that the optimal value can be found.
- We provide a visual presentation of the results and a simple number of inputs and outputs, enabling people to quickly understand our model.
- We chose to build our own scoring system instead of simply using classical algorithms such as PCA(Principal components analysis) or AHP(Analytic Hierarchy Process). our model is more considerate and thoughtful

6.2 Weaknesses

- In predicting the success of the product, we only have very few product data to train, so it is possible to present a low prediction value.
- In the emotional analysis, we referenced the emotional analysis system, named NLTK, and we think moving it to evaluate the emotion in shopping review online may result in the loss of emotional value.

7 Letter

TEAM 2010764

March 9, 2020

Marketing Director

Sunshine Company

Dear director,

We are honored to be employed by your company to carry out this data analysis. After hard and detailed research, we can confidently return our research results to you. Using TextRank technology, we conduct text analysis and extract some keywords from the review and title of goods. These keywords can tell us what aspects of products consumers care about, and what characteristics of products they will buy better. The first is the review analysis, which helps to add popular features to your products and make them more popular with consumers. About microwave ovens, 30% of the top words are about the size characteristics of microwave ovens.

Small size, suitable for the small urban area of the kitchen of a small microwave oven is most popular with consumers. In addition, the convection device of a microwave oven, more cooking options are the other two things people are most concerned about. Among the poor reviews, the most complained by consumers is about safety. In my opinion, people feel dangerous, which means that products should make people feel reliable and safe. Circuit breakdown, poor quality products often cause overheating and other phenomena due to circuit problems. As for the nipple, people prefer to use it as a consumable, so the price factor plays a great role in it.

Products with high discounts and low prices are often the most popular. Second, parents are most concerned about quality. Products with taste or poor quality will have a bad reputation. In addition, because it's for babies, a lovely appearance is also a very concerned place for parents. As for hairdryers, customers prefer quiet products that do not harm hair and high-power products; they hate products that are heavy, easy to overheat or easy to break the folding handle. Therefore, light and quiet products will sell well. Through the analysis of the product title, you can know what kind of properties the product will sell better. As for microwave ovens, products with poor reputation often have the characteristics of "over the range" and "large", which shows that products with large space and high power are not popular. As for the pacifier, the description with "monkey", "girafee" and "puppet" will sell better, probably because the lovely animals make people have a good impression. With regard to hairdryers, those with descriptions of 'professional' and 'following' tend to sell poorly. And, black or purple hair dryers are more popular with consumers than white ones. Further,

your company can obtain the latest review data of other brand, and the review data can be used to train the latest PRES and TextRank. After all, it is too difficult to use data, from before 2015 and cannot accurately, to predict what the market will be like today, over a long period.

Your can import the model in combination with the test user comments, and simulate the real market situation, then evaluate and predict the market sales volume by PRES to obtain the market sales evaluation value. We believe that you will improve the essential aspect of your products, according to the key words we provide.

Through continuous product optimization, when the user comments can support model training, we can predict whether the product is successful or not with SVM. Through the verification of our two models, if your product is ultimately evaluated as a successful product, we believe that this product will succeed.

Thank you for your trust and pleasant cooperation. We look forward to the success of your products

References

[1] Singla,Zeenia,Randhawa,Sukhchandan,Jain and Sushma "Statistical and sentiment analysis of consumer product reviews ". BOOK,2017/07/01 10.1109/ICCCNT.2017.8203960 Publishing Company , 1984-1986.

[2] D. N. Devi, C. K. Kumar, and S. Prasad, "A feature based approach for sentiment analysis by using support vector machine," in Ad- vanced Computing (IACC), 2016 IEEE 6th International Conference on. IEEE, 2016, pp. 3–8. Addison-Wesley Publishing Company, 1986.

[3] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in LREc, vol. 10, no. 2010, 2010.

[4] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.

[5] Chong, Alain, Ngai, Eric, Ch'ng, Eugene, Li, Boying, Lee, Filbert. (2015). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. International Journal of Operations, Production Management.

[6] Archak, N., Ghose, A. and Ipeirotis, P.G. (2011), "Deriving the pricing power of product features by mining consumer reviews", Management Science: INFORMS, Vol. 57 No. 8, pp. 1485-1509.

[7] Baker, B.M. and Ayechew, M.A. (2003), "A genetic algorithm for the vehicle routing problem", Computers and Operations Research, Vol. 30 No. 5, pp. 787-800.

[8] Chen,S.F.S., Monroe,K.B.andLou,Y.C.(1998), "The effects offramingprice promotionmessageson consumers' perceptions and purchase intentions",Journal of Retailing,Vol.74 No.3,pp.353-372.

Appendices

Here are part of the diagrams that we drawn in the process of our work.
There are more diagrams to help us analyze the data, but they are not reflected in the paper

	star_rating	helpful_votes	unhelpful_votes	vine	verified_purchase	Total_pos	Weight_pos	Total_neu	Weight_neu	Total_neg	Weight_neg	compound
count	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000	11470.000000
mean	4.116042	2.179076	0.384220	0.015606	0.855362	0.241582	0.347559	0.710345	1.433713	0.038137	0.072968	0.279923
std	1.300333	14.241304	1.843085	0.123950	0.351751	0.204918	0.332535	0.200329	1.066324	0.070040	0.128517	0.274673
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.890000
25%	4.000000	0.000000	0.000000	0.000000	1.000000	0.100000	0.064960	0.630000	0.634944	0.000000	0.000000	0.090000
50%	5.000000	0.000000	0.000000	0.000000	1.000000	0.190000	0.280984	0.750000	1.294786	0.000000	0.000000	0.280000
75%	5.000000	1.000000	0.000000	0.000000	1.000000	0.330000	0.528127	0.840000	2.085638	0.050000	0.110468	0.470000
max	5.000000	499.000000	98.000000	1.000000	1.000000	1.000000	2.463679	1.000000	6.291684	1.000000	1.415696	0.980000

Figure 12: data diagram after NLP processing

	star_rating	y
star_rating	1.000000	0.309914
y	0.309914	1.000000

Figure 13: Correlation matrix

	review_id	star_rating	y	time
0	R9T1FE2ZX2X04	5	3.500000	8/31/2015
1	RE36JAD5V53PO	4	7.404579	8/31/2015
2	RIDHM8B7SCCV3	5	3.118876	8/31/2015
3	R14QGWPCHU9LSE	5	3.500000	8/31/2015
4	R35BHQJHXXJD59	4	35.255721	8/31/2015
...	...	...	...	...
11465	R2O50YNP83CG34	5	7.429425	8/21/2002
11466	R2JQPUYU65C4QD	1	1.384684	8/13/2002
11467	R3GO6L5PWBS0IW	5	6.374316	7/13/2002
11468	R3JMGN42OJCL97	5	12.310257	4/20/2002
11469	R2XM83JYE2KDE2	3	9.965606	3/2/2002

(a)

	star_rating	y
count	11470.000000	11470.000000
mean	4.116042	222.997723
std	1.300333	559.428273
min	1.000000	-5402.066168
25%	4.000000	0.000000
50%	5.000000	58.048867
75%	5.000000	294.017434
max	5.000000	8897.566633

(b)

Figure 14: data diagram after model calculation

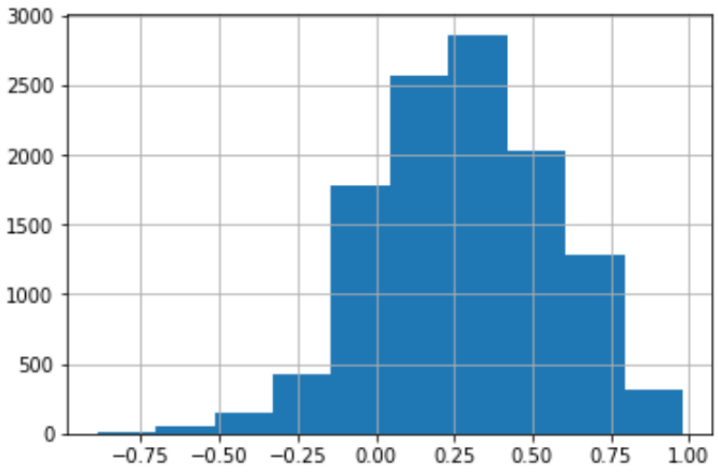


Figure 15: Comment emotion distribution map