Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2018884

2020 MCM/ICM Summary Sheet

A Wealth of Data

Summary

Nowadays, purchasing through the Internet has became a popular shopping trend. The Amazon marketplace platform provides an opportunity for customers to purchase with rating products and rating reviews of products. Through the establishment of the models and data analysis, we hope to be able to derive meaningful quantitative and qualitative patterns, relationships,measures and parameters between data such as star ratings, product ratings from user reviews and helpfulness rating of product which sold in the online market and the product's reputation and sale status. This analysis uses the models in order to assist the company to develop the product sales strategy, supplier production plan and new product design plan and identify potential necessary design functions or features to help the company to optimize the product. The users' reviews, star ratings and the helpfulness ratings of the reviews are quantified for mathematical modeling based on the time model. We're using the mathematical methods such as multiple regression analysis and curve fitting in order to obtain the model and measurement method finally.

The steps and methods of the model analysis and establishment are as follows. Firstly, quantified the users' reviews for the products, helpfulness ratings and star ratings by the sentiment analysis and word frequency statistics. Explore the possible indicators Influence the future trend of the product and implied associations among the users' reviews, helpfulness and star ratings by counting the distribution of them. After that, explore the influence degree of each factor on product by means of regression fitting the indexes preselected. Finally, explore the relationship functions between these variables by fitting the index such as sales status, users' reviews, helpfulness ratings and star ratings. Search the possible problems and looking for suggestions through the word frequency statistics and pos tagging of the data.

Through the model and analysis mentioned above, we obtained the relationship among the product reputation, product sale status, users' reviews, helpfulness ratings and star ratings. We came conclusions that the helpfulness ratings based on the users' reviews benefits best for the company to track, the rating level of the products are closely related to text-based quality descriptors. In addition, we determined the measurement pattern for potentially successful or failed products.

MCM Team 2018884

P.O.Box XXXXXX

China

March 9, 2020
America
Sunshine Company

Dear Director:

Hello, the Marketing Director of Sunshine Company! We are the MCM team 2018884. We are very concerned about the sales and development of your company's products. The data shows that the three types of products designed by your company have their advantages and disadvantages. In our opinion, the best sales strategy is: through careful analysis of user feedback, starting from the reality, find the entry point for improvement and innovation of the product, improve the product's own hard power, and add strong word-of-mouth marketing to enhance the product's reputation. The analysis of this strategy is as follows.

After getting the three data files provided by Sunshine's data center, we carefully browsed the data files which include review, product-id, star-rating, help-votes, total-votes, vine and review date. Then ,we adopted the method of emotion analysis to quantify the text reviews, so as to conduct scientific and effective analysis of the large number of comments contained in them. After that, a series of data combination and analysis were carried out by means of mathematical modeling, and it was found that time-based patterns in these data interact in ways that will help the company craft successful products.

First, we processed the data of the three types of products comprehensively, and then produced a series of images. After further comprehensive and detailed analysis of multiple images, we found that:

1. The public's comments on the product are relatively objective and reasonable, and most scoring products by reviews is of great help to the product evaluation.

2. The reviewers rated the three categories of products in each star, and the 3 points were the highest point of the lines whose general trend is very similar. This indicates that most people who participated in the review believe that if the score interval is set at [0,6), the score of these three products is mainly distributed at the position of 3.

Secondly, we classified the data and inserted time clues, and refined the analysis into each time period to make the results more accurate and convincing. In order to prove that reviews can be used to predict the sales conditions of online markets, and find out the potential success or failure of products after online sales, we use regression analysis to compare the effects of various factors on purchases, and find that in most cases, the impact of staged reviews and total reviews on periodical sales is greater than other variables. In addition, in the process of analyzing the trend of the number of stars and the image

of the change of scoring products by reviews over time, it is found that the customer's comments are hardly affected by the change of star rating. On the contrary, we conducted word frequency statistics on reviews, established a multiple linear regression standardization model, and obtained multiple linear regression equation for three types of products. After finding the mathematical relationship between review texts and stars, we found that specific quality descriptors of text-based reviews are strongly associated with rating levels. Therefore, we believe that it can take the method based on statistics and analysis reviews and supplemented by other factors to track the sales status of the online marketplace. And three equations obtained can be used to identify potential factors that lead to success or failure of a product after it is sold online.

In order to be able to find out the reason of users don't rate highly on these three types of products, we found out the negative words that appear more frequently in the reviews of the three types of products ,which represent the parts of the products need improvement through analysis and statistics of the word frequency of the reviews. For example, this type of hair dryer products are old, heavy, short and thick, this type of microwave products are small, limited, old and noisy and this type of pacifier products are old and hard. These aspects are also potentially important design features that would enhance product desirability.

In order to better understand the product reputation which is expressed by the quantity of the products purchased after it is sold online, we draw the quantity of products purchased and the changes in star ratings, reviews and helpfulness ratings in a line chart. Through the analysis of the image, it can be seen that the sales volume and the helpfulness ratings show a clear inverse relationship, that is when the helpfulness ratings of reviews is reduced, the purchase amount of the product increases. This clearly reflects the rise in product reputation, this clearly reflects the rise in product reputation, which is also in line with the status quo of online marketplace.

During the above analysis, we used the following methods:

1. The method of combining local analysis and global analysis is not easy to miss valid information, and it can make complex situations more concise and clear.

2. All solutions are supported by meaningful data and mathematical evidence to enhance the credibility of the conclusions reached.

Therefore, we provide our solution to your company, and we are confident that we can solve your company's problems in online marketplace.

Respectfully

MCM Team 2018884

Contents

1 Introduction

1.1 Issues Statement

Sunshine Company is planning to introduce and sell three new products in the online marketplace: a microwave oven, a baby pacifier and a hairdryer. Our group will use the data of other competing products provided and assist the company to find the best online sales strategies for these three products, identify the important design functions or features, increase the favorable conditions, desirability and competitiveness of products.

1.2 Problem Statement and Model Description

Our group performed these following tasks in order to determine sales strategy and identify important design features that have not yet been implemented to enhance product desirability.

1.Draw figures individually and in pairs for the other competing products of microwave, baby pacifier and hair dryer with indexes users' reviews, helpfulness ratings and star ratings. Intent to describe and analyze the phenomenon of online sales status from the figures.

2.Establish a model of the text's emotional tendency and objective degree through word frequency statistics and sentiment analysis, so as to show the score of the products through reviews.

3.Establish a multivariate standardized regression model for company to track the sales of the three products in the online market better.

2 Model and Analysis Methods

2.1 Data Processing

2.1.1 Processing of Reviews

We processed and analyzed the users' reviews by sentiment analysis. We extract the aspects from review headline and content, adjust the sentiment of the review and then aggregate the score to perform the polarity and subjectivity. The polarity float within range from -1 to 1, a high score means the review is positive, in contrast a low score means the review is negative. The subjectivity float from 0 to 1, a high score means the review is relatively subjective, in contrast a low score means the review is relatively subjective. We construct a binary function for reviews by using the result polarity p and subjectivity a of the review's sentiment analysis in order to represents the product score through this review objectively. We defined the result z of the function as the user ratings for products via reviews, we called it review score under. The review of user consists of z_1 and z_2, z_1 means the review score of the review headline and the z_2 means the review score of the review body. We set the ratio of review headline at 20% and the ratio of the review body at 80%. In order to make the representation of the score more reasonable, we subtract the minimum value of the review score as a whole. The score float from 0 to 6.

The functional relations we used in the above process include:

$$z_i = \frac{p}{|p|} \cdot \frac{x^{|p|}}{y\,(1+a)} \quad (i = 1, 2 \quad x, y \in R)$$

$$z' = 20\% \cdot z_1 + 80\% \cdot z_2$$

$$z = z' - \min z'$$

2.1.2 Processing of Helpfulness Ratings

We used the ratio of helpful votes and total votes to reflect the helpfulness ratings. We removed the records whose total votes equal to 0. It can be considered that when the total number of votes is high, the helpfulness rating of the review is trustworthy, and the helpfulness rating of relatively large number of votes is similar; and when the total number of votes is small, the trustworthiness is low, and a relatively small number of votes has a relatively large difference in the effectiveness of helpfulness rating. As a result, we defined a parameter α as representing trustworthiness and the value of it mapped in [ln(0.9),ln(1,1)]. Simultaneously, we designed two mapped functions to avoid there are too many comments with the highest total votes affecting most of the help levels of trust and causing data to be too dense.

The functional relations we used in the above process include:

$$h = \frac{v}{t}$$

$$\alpha = \begin{cases} e^{\left(\frac{\left(\overline{h}-\min h\right)}{\ln(1)-ln(0.9)}\cdot(h-\min h)\right)} & h < \overline{h} \\ e^{\left(\frac{\left(\max h-\overline{h}\right)}{\ln(1.1)-ln(1)}\cdot(\max h-h)\right)} & h \geqslant \overline{h} \end{cases}$$

$$H = \alpha \cdot h$$

2.1.3 Processing of Star Ratings

We count the total number of reviews and the cumulative value of star ratings in a certain period of time, the ratio of the cumulative value to the total number is the average of the star ratings in the period. For example: the average star ratings of the product to the current time node is represented by the average star ratings of the time period from the listing of the product to the beginning of the time period to the current time node.

2.2 Model Foundation

2.2.1 Model One: Image Analysis

The number of star ratings is summarized into a column map according to the classification of total star ratings, purchased star ratings, and Vine Voice star ratings. From the image, we can clearly see the

approximate number of star ratings and trend of star ratings of purchasers and Vine Voice, which are relatively more significant for the evaluation of products.

Plot the number of reviews scores into a column map to see the distribution of reviews at each level easily.

Helpfulness rating is defined as a percentage of the ratio of helpful votes to total votes. So, helpfulness rating is within [0,100]. Helpfulness rating is 0 if the review is not helpful for the product, the larger the helpfulness rating is, the more helpful the review is for the product. Draw a line graph of each value of the helpfulness rating, so that you can see the approximate number and overall trend of each helpfulness rating. In addition, we think that the reviews with a total votes of 0 are meaningless, and we have eliminated them in the process of data processing.

2.2.2 Model Two: Multiple Regression Analysis

In order to identify data measures based on ratings and reviews that are most informative for Sunshine company to track. We selected several products, conducted data statistics based on time, counted the sales volume in a time period as the dependent variable, and used indicators such as star_ratings, scores of review_headline, scores of review_content, and helpfulness in the time period as independent variables to establish multiple linear regression model by least square method. At the same time, we also standardized the data, that is, the original data is subtracted from its mean value, and then divided by the standard deviation of the variable to calculate the new variable value. The regression equation formed by the new variable becomes the standardized regression equation.

Significance of regression equation parameters:

·P: P is the probability that the "invalid hypothesis" holds.

(See the appendix for the statistics of each product sub-category)

3 Analysis

3.1 The First Question

3.1.1 Within

1. Star Ratings

The image counts the total number of reviews, the number of purchased reviews, and the number of Vine Voice reviews according to star ratings.

As can be seen from the figure of star rating of microwave, in the three classifications, the distribution of star ratings from 2 to 5 is very similar and shows an increasing effect. The number of star ratings of 5 is much larger than the number of other stars. The distribution of star ratings from 2 to 5 is very similar and shows an increasing effect. The number of star ratings of 5 is much larger than the number of other stars.

Possibility of causing problems in the image: 1. 1 star is only lower than 5 stars in the total number of reviews: there is a phenomenon of malicious reviews, or products purchased from offline stores or other factors cannot satisfy customers and go online Store reviews. 2. The number of one-star ratings is only lower than the four-star ratings : There is indeed a degree of polarization in the evaluation of products.

The star rating of hair-dryer graph is compared with the microwave graph. In the total number of reviews, the number of 1-star reviews is greatly reduced, so that the star classification distribution rules of the three classification methods are very similar and show increasing effects. It shows that for the public, this product has better comprehensive performance and has its own advantages.

As can be seen from the image of star ratings of pacifier, most buyers rated the product as 5 stars, and the number of people who rated 5 stars was much larger than the number of other stars. On the one hand, it can be confirmed that the overall performance of the product is really high-quality, and on the other hand, it can not be ruled out that a deliberate 5-star review occurs.

2. Reviews Score

These images count the distribution of the total number of reviews according to the rounded down review scores.

Comparing the graphs of the reviews scores of the three products, it can be seen that the distribution of review scores of the three products is roughly the same. You only need to reasonably analyze the review scores of one of the products. It can be seen from the graph of reviews score of pacifier that the number of reviews scores 0 and 5 is very small, the number of 1 and 2 is relatively small, the number of 4 is large, and the number of 3 is the most. This shows that most people participating in the review believe that the best score for pacifier is 3, and they have a moderately high attitude to the product. These three products still need to be improved and improved in terms of product performance.

3. Helpfulness Ratings

The graph ranks the number of comments at different levels according to help levels.

As can be seen from the image of helpfulness ratings of microwave, there are still some reviews that are less helpful for the product, while most reviews are 90% helpful for the product, and other reviews are helpful for the product focus on 20% -80%. This shows that most reviews are helpful and that the help levels are somewhat continuous.

Similar to the image of helpfulness ratings of microwave, most reviews are nearly 100

Compared to the image of helpfulness ratings of microwave, the total number of reviews increased significantly, and most reviews are nearly 100% helpful for the product. The number of reviews which are 0% helpful for the product has also greatly increased, indicating that people have a greater diversity of views on the product, and the possibility of deliberate reviews is not excluded.

3.1.2 Between

1. Star and Helpfulness

These images show the distribution of reviews helpfulness ratings under different stars.

It can be seen from the three pictures that no matter which product, which star rating, the helpfulness ratings are distributed more in a high range. This shows that the reviewers are random and the reviews of products are relatively objective and reasonable. That is to say, all possible assumptions about reviews on image issues when analyzing a single image are completely ruled out.

2. Star and reviews

The images show the number of every review whose score is corresponded to each star rating. And these images count the distribution of the total number of reviews according to the rounded down review scores.

Contrast observing the three products‘ reviews score images distribution find that the general trend of the images of the three products under every star is very similar. It can be seen from the image of star ratings and review scores that as the star rating increases, the low-score part of the reviews gradually decrease, and the high-score part of the reviews gradually increase. But most people who give reviews under each star rating gives the products 3 points. This shows that most people participating in the review believe that the best score for this product is 3 points, so the product is in the middle position. For microwave oven, 1-star product reviews scored with 3 points are still a lot, which indicates that 1-star reviews are still very objective. This shows that most people participating in the review believe that the best score for this product is 3 points, so they have a moderately high attitude to the product. For microwave oven, 1-star product reviews scored with 3 points are still a lot, which indicates that 1-star reviews are still very objective. The occurrence of this contradiction cannot be ruled out due to the lack of performance of a certain part of the product or the star rating attitude of other participants affecting their star ratings.

3. Reviews and Usefulness

The image shows the distribution of review scores at different helpfulness ratings. And these images count the distribution of the total number of reviews according to the rounded off review scores.

Judging from the image of the review score and helpfulness ratings of the three products, except for reviews with zero helpfulness rating, the distribution of review scores is clustered in a range of 40% -100%, and the score is concentrated in [3,5]. This shows that most of the review scores are of great help to the assessment of the product. From the product reviews, you can see what aspects of the products need to be improved and what the advantages of the product are.

3.2 The Second Question (a)

In order to identify data measures based on ratings and reviews that are most informative for Sunshine Company to track. First, we counted the total number of reviews for each product id, and selected the products with the most reviews from each of the three product categories. We used a time period as a unit to count the sales volume, star ratings, scores of review headline, scores of review content, helpfulness, the total number of reviews, vine, and the cumulative star ratings cumulative value, scores of review headline cumulative value, scores of review content cumulative value and helpfulness cumulative value by time period, and made a record form. We use the significance test

method and through a simple analysis of the data, we choose to use the quantity of sale as the dependent variable and the remaining variables as the independent variables to establish a multiple regression standardized model.

Multiple regression equation: $y = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + \beta \ \ (\alpha_i \in R)$

Variable Declaration:

Compare and contrast the five selected tables. From the beta values, it can be seen that the absolute value of the standardized regression coefficient of Stage Total is greater than the values of the remaining independent variables in most cases, which indicates that given in other independent variables , the total number of comments in a unit time period has a great impact on the sales volume of the product during the time period. The value of the Coefficient and Beta of the Review Title Score and Review Content Score is, in most cases, greater than the values of the other independent variables except Stage Total, which indicates that it is more beneficial to track the value of the Review Score accumulated to the current cycle by time period Analyze product sales in the online market.

3.3 The Second Question (b)

Divides product star rating, user review score, and help level according to time period, and uses the corresponding data for statistical analysis. The data uses cumulative product purchases (total cumulative value and cumulative value during the period) product ratings, product review ratings (Total Score, Reviews headline scores, and Reviews body scores) and help level.

Variable Explanation:

These pictures are the relationship between the cumulative sales of the product, Star Ratings, Reviews Scores, and Helpfulness, and are about the relationship between the sales volume of the product during, time period, Star ratings, Reviews Scores, and Helpfulness.

According to the analysis of the image of the microwave product B0055UBB4O, within a certain period of time, while the helpfulness of user reviews declines, the product sales, star ratings and review ratings increase.

According to the images of the hair dryer and the Pacifier, in a certain period of time, there is always a corresponding relationship between the helpfulness of the user reviews and the product ratings and product star ratings of the reviews. It can be explained that in a certain period of time, the decline of the helpfulness of the user reviews can predict the improvement of product reputation in part. In the image of B001FGL9X0, it can be found that after the review help level rises, the purchase volume drops.

Similarly, the increase in sales data also has the opposite relationship with the helpfulness. When the helpfulness of the review decreases, the sales volume of the product is increasing.

Therefore, it can be considered that when the helpfulness of product reviews decreases due to an increase in the number of purchases, people will more often think that the quality of the product is higher and its reputation is on the rise.

3.4 The Second Question (c)

First conduct word frequency statistics and sentiment analysis on the review title and review text of each product to get the score of the review title and review text. (Perform a sentiment analysis on each comment to obtain a positive and negative relative value p in the range (-1,1) and a subjective and objective degree value a in the range (0,1). And calculate the score of user reviews) Then score the help level of each product and finally get a form.

The fields are: star rating of each review, score of each review title, score of each review content, help level.

Please see the attachment for the form.

In order to determine the linear combination of text-based and rating-based metrics, we chose star ratings as the dependent variable and the remaining three as independent variables to establish a multivariate linear regression normalization model. (Set Star Ratings to y, ReviewTitleScore to X1, ReviewContentScore to X2, Helpfulness to X3, and constant term to b)

Hair_Dryer

Then the multiple regression equation can be obtained as:

$y = 0.136703x_1 + 0.751379x_2 + 0.648387x_3 + 2.669121$

Microwave

Then the multiple regression equation can be obtained as:
$y = 0.504711x_1 + 0.713557x_2 + 0.318163x_3 + 2.175574$
Pacifier

Then the multiple regression equation can be obtained as:
$y = 0.17388x_1 + 0.047955x_2 + 1.021438x_3 + 3.837628$

3.5 The Second Question (d)

These images include the distribution of stars according to time and the star ratings, reviews and helpfulness ratings of products.

We map the changes in the number of star ratings for each sub-product of the three categories of products over time and statistics. We select sub-products' images of the quantity trend of each star whose star trend has characteristics and helpful for analyzing the topic to analyze, and map a image that the review scores change over time. We have selected that :a sub-product of microwave oven's every star's number has increased at a similar rate, a sub-product of microwave oven's one star's number has increased extremely fast and other star's number has increased slow and three sub-products of three products' five star's number has increased extremely fast and other star's number has increased slow. Analyzing these three situations is easy to judge whether customers are more likely to write some type of review after seeing star ratings or not.

The first situation that every star‘s number has increased steadily:

Through the images that the review scores change over time, it can be seen that the trend line of the review title score and the review content score are relatively steady. So it can be seen as floating around or nearing a certain value near a certain value.

The second situation that low star‘s number has increased fast:

Through the images that the review scores change over time, it can be seen that the line trend is very similar to the steady growth image of each star.

The third situation that high star‘s number has increased fast:

Similarly, the line trend of the review score is still very close to the image that every star‘s number has increased steadily.

From the images and analysis of the three cases, it can be seen that after seeing the change in star ratings, customers did not write more reviews due to the number of high (low) stars. The attitude of review is still the attitude of most reviewers to this category of products. So we can conclude that customers are hardly affected by star ratings when they comment.

3.6 The Second Question (e)

In order to determine whether the specific quality descriptor based on text reviews is closely related to the scoring level. We quantified the descriptors based on the text comment, and obtained a score after the sentiment analysis of the comment title and comment content of each product. Then we give a certain weight to them (the review title accounts for 20%, and the review content accounts for 80%) For comprehensive scoring. Therefore, four quantitative indicators are obtained. We use the Star ratings of each product as the grade. We use the Star ratings score as the dependent variable, and the other four as the independent variables to establish a multivariate linear regression normalization model level of the product. Then, according to the absolute value of the standardized regression coefficient, the degree of influence of the independent variable on the dependent variable is determined, that is, the degree of close correlation.

Please see the attachment for the processed data table.

Please see the attachment for the word frequency statistics of specific quality descriptors for each comment text.

Variable Declaration:

The regression results are analyzed as follows:

Microwave

Known from the table: For products such as microwave, the absolute value of beta coefficients of TitlePolarity and ContentPolarity are relatively large, which indicating that both the title of the review and the specific descriptor of the content have a lot to do with the rating.

Hair_Dryer

Known from the table: For products such as hair dryer, the absolute value of beta coefficients of TitlePolarity and ReviewScore are relatively large, which indicating that both the title of the review and the specific descriptor of the content have a lot to do with the rating.

Pacifier

Known from the table: For products such as hair dryer, the absolute value of beta coefficients of TitlePolarity and ReviewScore are relatively large, which indicating that both the title of the review and the specific descriptor of the content have a lot to do with the rating.

In summary, no matter which data set, the specific quality descriptors of text-based reviews of the three types of products have a close relationship with the rating level of the products.