Data Processing Based on Comparative Analysis and the Study of Consumers' Score on Cereal Under Different Influencing Factors

Weicong Chen[1, *, †],Haowei Yuan[2, *, †]
[*]weiconc1@uci.edu, [*]2368504234@qq.com

[1]Henry Samueli School of Engineering University of California Irvine Irvine, USA

[2]School of Computer and Information Science, School of Software Southwest university Chongqing, China

[†]These authors contributed equally.

Abstract—This research paper focuses on determining the combining effect of different factors that might affect customers' satisfaction scores on cereal products. As the demand for breakfast cereals has been increasing dramatically during the COVID-19 period, knowing the taste of customer is vital in order to increase sales for companiees. We started our research for a solution to each of the three problems which are crucial for sales increase. In our research, we are using data that contains different characteristics of certain cereal products correlated with their specific customer satisfaction score. Then we will use different methods to determine the correlation of each factor with customer satisfaction score and the correlation of combinations of factors with customer satisfaction score. The methods that we used include unary linear regression, multiple linear regression, and nonlinear regression. In this exploration and analysis, we used SPSS and Excel related software to analyze the data and build a model, and obtained the coefficient before each independent variable in the model, so as to obtain the model. After comparing the results we get using different methods, we came up with the optimal solution generated by multiple linear regression. From the optimal solution generated by multiple linear regression, companies will have a clear understanding of which nutrition ingredients are the ones that affect customer satisfaction the most, and they would be able to improve their cereal products in accordance.

Keywords-Cereals; Customer satisfaction; Data analysis; Comparative analysis

1 INTRODUCTION

1.1 Research Background

What do you eat for breakfast? We think this might be the first question that pops up in many people's mind when they wake up in the morning. There are so many different choices for

MSIEID 2022, December 09-11, Chongqing, People's Republic of China

DOI 10.4108/eai.9-12-2022.2327613

breakfast with people in a different countries. For example, people in the United States might prefer pancakes, and Cantonese might prefer steamed dim sum. However, one type of breakfast is becoming more and more common in families across the globe. That is cereal. According to Justin Fox, there was an abnormal increase in demand for cereal in the year of 2020. Fox also stated in his article that "before 2020, the biggest one-month percentage increase in cereals consumption was 5.6%, in July 1973. Last March it was 26.9%" [1]. An important reason for this drastic increase is the pandemic, which might last for quite a long time as for recent analysis. Since most people were suffering from a lockdown and restaurants were closed, cereal became a more popular option for breakfast.

1.2 Literature Review

Then why does cereal stand out and become a popular breakfast choice besides the effect of the pandemic? First of all, it is very easy to consume. Normally, we can pour milk into the cereal, making a simple and delicious breakfast. What's more, Williams pointed out several health benefits for cereal consumption, such as increasing the possibility to meet people's nutrition needs, lower rate of obesity, and "better satiety" [2]. This makes cereal a perfect fit for most people who usually go to school or work on a weekday. Imagine that someone is having a late start to the day and needs to hurry to work. There is not enough time for this person to cook a delicate breakfast and enjoy it. However, the person can instead eat a bowl of cereal by simply taking it out and soak it with milk. This can save people's time making breakfast and satisfy their daily nutrition needs, which is even more important than the time-saving benefit.

Since the cereal demand is increasing, the profit generated in the cereal market increases as well. However, there are so many different companies producing all kinds of cereal products. Some might focus on their flavors, and others might focus on nutrition contribution. According to Emilie Croisier, Jaimee Hughes, Stephanie Duncombe, and Sara Grafenauer, "Breakfast cereal options have expanded over the years to cater to a growing demand and are often marketed as a convenient source of nutrient-rich grain foods. According to the 2011–2012 National Nutrition and Physical Activity Survey (NNPAS), 36% of Australians (aged 2 years and above) consumed breakfast cereals, and the data suggests that a higher proportion of consumers are in the younger (2–8 years) and older (71 years and over) population groups" [3]. We can see that the data from the United States and the data from Australia suggest that cereal consumption is happening among a large proportion of the population.

There should be no doubt that the primary concern for any profitable business is profit. Then we can divide all the methods of maximizing profit into two basic categories: increasing sales and decreasing costs. In order to increase the sales of a product, the most important thing that we should know is what our customers prefer. Producing what consumers prefer is the key to making the products stand out in a greatly competitive market. While most research from the cereal production companies focuses on merely their products' flavor or nutrition components, our research will be based on a more comprehensive scope.

1.3 Research Framework

We will analyze the customer satisfaction score of a certain type of cereal based on many different factors such as nutrition component, weight, calories, and even its position on a store

shelf. We will do for our analysis to incorporate different types of methods, such as simple linear regression, multiple linear regression, and non-linear regression, to find out the relationship of different factors with the customer rating. We would first determine how every single factor is related to the rating. Then we would consider a combination of factors with a higher level of significance to form a model for companies to decide which factors affect the customer rating more significantly. With our model, companies would better understand what the consumers of cereal prefer and what they dislike. As a result, the cereal production companies can modify their product according to the crucial factors that we come up with in our model to increase their customer satisfaction and eventually increase their sales.

2 METHOD

In this method exploration, we mainly used comparative analysis method. In the selection of data, we selected the secondary data that had been collected and integrated, and the data came from 80 cereals on the official website of Kaggle data set. These datasets have been gathered and cleaned up by Petra Isenberg, Pierre Dragicevic and Yvonne Jansen. The original source can be found here. This dataset has been converted to CSV.

The dataset is based on consumer ratings of cereals based on different factors, such as the number of different nutrients and different manufacturers. The dataset provides a comprehensive overview of the major nutrients contained in the cereal, which can be understood as internal factors that affect consumers' ratings of the cereal. In contrast, other factors such as different manufacturers and the location of the cereal when sold can be called external factors. Because the nutrient content of internal factors can be expressed numerically, these data are more suitable for data processing. Of course, external factors should also be analyzed separately.

In order to find out the most suitable method for exploring this CSV data set, the comparative analysis method is used. In specific data processing methods, unary linear regression, multiple linear regression, and nonlinear regression are selected.

(1) Unary linear regression

Unary linear regression is a method to analyze the linear correlation with only one independent variable (independent variable X and dependent variable Y). The value of an economic index is often affected by many factors. If only one factor is the main and decisive factor, linear regression can be used for predictive analysis.

(2) Multiple linear regression

Multiple regression is used in regression analysis, if there are two or more independent variables, because problems in real life are complex, and multiple regression is usually more meaningful. Wang and Meng stated that “the fitting value obtained by multiple linear regression model has high precision and the predicted value is true and reliable. The calculated results are in good agreement with experimental results” [4].

(3) Nonlinear regression

If the dependent variable of the regression model is the function form of independent variable more than once, and the regression law is graphically represented as various curves of different shapes, nonlinear regression should be used.

Li once analyzed the effect of SPSS and other software on nonlinear regression [5]. In the data processing, we can use some similar data processing software such as excel or SPSS, the CSV data can be solved by the above several different methods for processing, finally to get the real factors that influence the score, and various factors influence the size of the final can put forward more valuable guidance to manufacturers, So that manufacturers can be in the relevant aspects of investment and practice, increasing enterprise competitiveness in the market.

3 RESULTS

As we have discussed above, we are now experiencing a global issue due to COVID-19. The virus has brought us problems with healthcare and has made people's lives much more inconvenient. In order to avoid getting infected by the virus, people tend to be more willing to stay at home and cook instead of dining in a restaurant. As a result, demands on products that could easily be prepared, such as breakfast cereal, has grown dramatically during the COVID-19 period. However, a market report on the breakfast cereal industry by Mordor Intelligence indicates that the pandemic is not the only reason for the growth in breakfast cereal demand.

According to Mordor Intelligence, "The growth of the market is mainly witnessed due to the changing food habits and influence of western culture on the dietary patterns of consumers, as it provides a convenient solution to readily accessible food that optimizes the ease of consumption without further preparation. Also, consumers' preference toward nutritious and healthy food regularly is driving the market" [6]. Whenever there is an increase in market demand, usually, an increase of market competition will follow. Hence, if a company wants to stay competitive in the market, one of the most important things it should keep in mind is customer satisfaction. If more customers are satisfied with the product, the demand for the product would certainly increase. Hence, if a company wants to improve its customer satisfaction level, it has to answer the following three problems.

3.1 What are the main factors affecting customers' satisfaction level on cereal products?

There are lots of factors that might affect customers' view on a certain breakfast cereal product. One important factor can be whether the product is beneficial for people's health. Even though breakfast cereals are similar to fast food, regarding its convenience, people who are consuming these products might still be intrigued about how healthy the product is. With the rapid development of technology and the massive improvement of life standards, people are increasingly concerned about their wellness. In an article from Food Quality and Preference, Mayara Lima pointed that "given a choice between a regular and a sugar-reduced product, consumers are expected to evaluate their characteristics and select the one that is associated with the most desirable outcome in terms of the goal they want to achieve" [7]. Nowadays, people are paying greater attention to their body shape, so many try to avoid food products containing a high proportion of sugar. Zero-sugar drinks and food products are becoming more popular than ever. However, to cater to the favourable taste of most customers,

a certain amount of sugar has to be considered while producing breakfast cereals. As a result, sugar level is another crucial factor that should be taken into account. Another health-related factor that customers are concerned about can be the protein level. This is a significant indicator of whether a certain breakfast cereal product is nutritious enough to satisfied people's nutrition demand for breakfast. If a person chooses to eat cereal for breakfast instead of other options such as eggs, bread, and meat, the protein level of the cereal has to satisfy the person's demand for breakfast because most people would not give up their health for mere convenience. In addition to these two health-related factors, the portion size of breakfast cereal products is another factor that can contribute to customers' satisfaction levels. When two breakfast cereal products do not have a significant variation in price or nutrition component, their portion sizes are critical factors that can impact customer satisfaction.

3.2 How strongly does each single factor relate to customers' satisfaction level?

Once we have identified several key factors, we should start to figure out how each factor affects customers' satisfaction. How can we determine this causal relationship quantitatively? Several methods can help. The simplest and relatively efficient way is to use a simple linear regression model. This model determines the linear relationship between two variables quantitatively, and it would output several indicators for how well these two variables are linearly correlated. An obvious drawback of this model is that not all relationship between two variables is perfectly linear and many might follow the other type of relationship, non-linear correlation. In the following discussion section, we will perform a simple linear regression analysis on a dataset about breakfast cereal provided by Chris Crawford from Kaggle. This dataset listed several characteristics of cereal products and their customer scores, which are the direct indicator of customer satisfaction. Once we have determined how well each factor is linearly correlated to customer score, we can then solve the next problem.

3.3 What is the optimal solution?

The last problem that we are going to deal with is the optimal combination of factors that a company should be focusing on to maximize customers' satisfaction level. `This is the ultimate and most complicated problem because we are going to deal with the combination of factors and figure out the effect they jointly have on a single outcome. With this problem, a multiple linear regression model is often more effective. It can generate how well a couple of independent variables are jointly correlated with the dependent variable and tell us how strongly each independent variable positively or negatively relates to the dependent variable. The result we get can help breakfast cereal production companies focus on the most critical factors when producing breakfast cereals. For example, suppose the result indicates that sugar and protein are two of the most significant factors affecting the customer score and sugar level is positively related to customer score while protein level is negatively related to customer score. In that case, a suggestion can be derived from this result indicating less sugar and more protein components should be included in the breakfast cereal during production.

These are the three main problems we are trying to solve. The discussion session would provide a detailed explanation of the methods we used to solve these questions and analyze the result we get from the dataset provided by Kaggle.

4 DISCUSSION

Wu obtained “the risk factors affecting the survival time of HCC patients by using different processing methods to process the data set analysis” [8]. To get as accurate as possible on the influence factors of cereal score and parts that manufacturers need to consider, we will get the results of the dataset through the comparative analysis of three methods and get a more suitable dataset of processing methods. It is concluded that the data of several data processing methods as applicable. The following is a process based on the analysis of the three methods. The values of the four dependent variables, mfr, type, weight, and cups, are not listed. The first two values are not universal, and the results of the latter two values are not influential factors in rating, whether they pass the R Square Value test or the P-value Value test.

4.1 Unary linear regression

When using unary linear regression, we can only get the unary linear fitting degree between the score and one influencing factor because unary linear regression is the simplest linear regression method. It mainly obtains the degree of fitting degree between the independent variable and the dependent variable through the value of R Square, which is between 0 and 1.R Square value closer to 1, it is proved that the independent variable and dependent variable of linear fitting degree is higher, the greater is the impact of the relationship between, below is obtained by the data analysis method of Excel spreadsheet R Square value between different influence factors and the score, the method to calculate score every time and the relationship between the different influence factors. As can be seen from the figure, the amounts of calories and sugars are the highest, followed by fiber and protein. The fitting degree of other influencing factors is low, so manufacturers should focus on cereal's calories and sugars content, followed by the amount of fiber and protein.

Table 1 R Square and Coefficient of Factors

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins
R Square	0.47523	0.2215	0.1675	0.161	0.3412	0.0027	0.57711	0.1445	0.0579
coefficient	-0.49701	6.0385	-5.7124	-0.0672	3.443	0.1709	-2.4008	0.0749	-0.1512

4.2 Multiple linear regression

Multiple linear regression is needed to analyze the dataset when using multiple linear regression, because there is more than one value of the dependent variable and multiple influencing factors jointly affect the score. Because the value of each dependent variable in the given data set has its special meaning, the prediction model is

$$Rating=\beta0+\beta1*calories+\beta2*protein+\beta3*fat+\beta4*sodium+\beta5*fiber+\beta6*carbo+\beta7*sugars+\beta8*potass+\beta9*vitamins+\beta10*shelf+\beta9*weight+\beta9*cups$$

The prediction through multiple linear regression is as follows:

Table 2 Prediction Result

Multiple R	1
R Square	1
Adjusted R Square	1
Standard Error	3.04E-07
Observation	77

When we get this data, we think the value of R Square is abnormal, so we use Python to divide the test set and training set by ourselves. The result is as follows:

```
y1_predict = estimator.predict(x_test)
print("predict: ", y1_predict)
error1 = mean_squared_error(y_test, y1_predict)
print("Root Mean Squared: ", error1)
rate = estimator.score(x_test, y_test)
print("the correct rate:", rate)
print("R2:",r2_score(y_test,y1_predict))
```

```
Root Mean Squared:  9.260612306570948e-14
the correct rate: 0.9999999999999992
R2: 0.9999999999999992
```

Figure 1 Test Process

The value of R Square is 1 due to one digit forward, which proves that this model is too perfect for this data set, which too small a data set may cause. Therefore, it is speculated that this prediction model will play a better role if the data set is larger. The table 3 confirms the model.

Table 3 Key Indicators Od Factors

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	54.92718	3.63E-07	1.51E+08	0
calories	-0.22272	5.66E-09	-3.9E+07	0
protein	3.273174	5.09E-08	64278081	0
fat	-1.69141	6.23E-08	-2.7E+07	0
sodium	-0.05449	4.96E-10	-1.1E+08	0
fiber	3.44348	4.31E-08	79919750	0
carbo	1.092451	1.74E-08	62683538	0

sugars	-0.7249	1.82E-08	-4E+07	0
potass	-0.03399	1.47E-09	-2.3E+07	0
vitamins	-0.05121	1.93E-09	-2.7E+07	0
shelf	-3.7E-08	5.28E-08	-0.70404	0.483962
weight	-4.3E-07	5.21E-07	-0.82566	0.412063
cups	1.38E-07	1.92E-07	0.716834	0.476084

Because in the prediction model, the confidence interval is 95%, and the P values of shelf, weight, and cups obtained from the figure are all greater than 0.05, so these three influencing factors can be omitted. The influence of each influencing factor can be determined by the absolute value of the coefficient of each influencing factor.

The final prediction model is:

*rating=54.92718-0.22272*calories+3.273174*protein-1.69141*fat0.05449*sodium+3.44348*fiber+1.092451*carbo-0.7249*sugars-0.03399*potass-0.05121*vitamins*

The multivariate linear model fits well for predicting the rating of cereal.

4.3 Nonlinear regression

The advantage of nonlinear regression is that it adopts iterative method to fit various complex curve models set by users: iterative method usually means stable results. The definition of residual is greatly expanded from the least square method, which means that the error measurement means are greatly enriched, we can use the weighted least square method, autoregression model and so on. Provides users with extremely powerful analysis capabilities, especially for laboratory data analysis.

Our prediction model is as follows:

*Rating=a * calories + b * protein + c * fat + d * sodium + e * fiber + f * carbo + g * sugars + h * potass + i * vitamins*

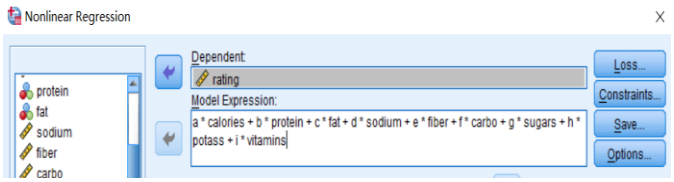

Figure 2 Nonlinear Model Construction

It can be seen from the iteration records that the iteration is terminated after 8 iterations and the optimal solution has been found. This method is to continuously substitute "parameter estimated value" into the "loss function" to solve, and the loss function adopts the minimum

"sum of squares of residuals". After 8 iterations, the sum of squares of residuals reaches the minimum value, at which point the optimal solution is found and the iteration is terminated.

Table 4 Iteration History

Iteration Number	Residual Sum of Squares	Parameter								
		a	b	c	d	e	f	g	h	i
1.0	155164.704	.000	.000	.000	.000	.000	.000	.000	.000	.000
1.1	30707.326	.043	1.968	1.129	.019	1.426	.363	.263	.036	.077
2.0	30707.326	.043	1.968	1.129	.019	1.426	.363	.263	.036	.077
2.1	5849.462	.113	5.114	-3.394	-.026	3.517	1.384	-.456	.006	-.067
3.0	5849.462	.113	5.114	-3.394	-.026	3.517	1.384	-.456	.006	-.067
3.1	4248.745	.138	5.260	-3.142	-.062	6.488	1.911	-.448	-.096	-.102
4.0	4248.745	.138	5.260	-3.142	-.062	6.488	1.911	-.448	-.096	-.102
4.1	4248.745	.138	5.260	-3.142	-.062	6.488	1.911	-.448	-.096	-.102

The prediction model can be obtained from parameter evaluation:

*Rating=0.138 * calories + 5.260 * protein – 3.142 * fat - 0.062 * sodium + 6.488 * fiber + 1.911 * carbo – 0.448 * sugars – 0.096 * potass – 0.102 * vitamins*

Table 5 Paramter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.138	.106	-.074	.350
b	5.260	1.294	2.679	7.842
c	-3.142	1.439	-6.014	-.270
d	-.062	.012	-.087	-.037
e	6.488	1.024	4.445	8.532
f	1.911	.429	1.054	2.767
g	-.448	.467	-1.379	.484
h	-.096	.036	-.168	-.024
i	-.102	.045	-.191	-.013

It can be seen from the ANOVA table that R square=0.717 and the fitting degree is 0.717, indicating that this model can explain 71% of the variation and has a high fitting degree.

Table 6 Prediction Result

Source	Sum of Squares	df	Mean Squares
Regression	150915.959	9	16768.440
Residual	4248.745	68	62.482
Uncorrected Total	155164.704	77	
Corrected Total	14996.800	76	

Dependent variable: rating

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .717.

In conclusion, in the three methods, the second multiple linear regression has the highest degree of fit, and manufacturers should focus on the amount of these nutrients, which together determine the score of cereal.

5 CONCLUSION

In this paper, we used three different regression methods to analyze the influencing factors of customers' ratings on cereal. Finally, we determined a good multiple linear regression model through comparative analysis and found that the model has a high linear fitting degree, which can be predicted.

In addition, this research is of great practical significance at present. Since the outbreak of COVID-19 in 2020, more people have chosen to stay at home, and consumers increasingly purchase these fast-food products. Zhang said that the epidemic had promoted the growth of fast-food sales [9], and these foods, like cereals, belong to the fast-food category. Therefore, relevant manufacturers should better understand which parts of cereal consumers value more to have greater competitiveness in the industry.

Similarly, this paper also has some limitations. We used the data set of the processed secondary data downloaded from Kaggle, rather than the first-hand data collected through our field research. Samuel Lefever stated that the “nonrandom nature” of secondary data might be a major limitation when using the data for research purposes [10]. In addition, there are some other methods not used in data processing, which can be improved.

In future research, we can go to supermarkets and other places for field investigation, collect first-hand data, and process and analyze data through various other methods to get a better prediction model.

REFERENCES

[1] Fox, J. (2021, February 24). The Cereal Industry Had a Very Weird Year. Bloomberg.com. https://www.bloomberg.com/opinion/articles/2021-02-24/beyond-grape-nuts-cereal-makers-had-a-very-weird-year.

[2] Williams, P. G. (2014). The benefits of breakfast Cereal consumption: A systematic review of the evidence base. Advances in Nutrition, 5(5). https://doi.org/10.3945/an.114.006247

[3] Croisier, E., Hughes, J., Duncombe, S., & Grafenauer, S. (2021). Back in time for BREAKFAST: An analysis of the changing breakfast cereal aisle. Nutrients, 13(2), 489. https://doi.org/10.3390/nu13020489

[4] Wang Huiwen, Meng Jie.Prediction modeling method of multiple linear regression. Journal of Beijing University of Aeronautics and Astronsutics, 2007, 33(004):500-504.

[5] Li Haikui, XIAO Yali, MIAO Jun.Analysis and evaluation of nonlinear regression function in Common Statistical Software [J]. Journal of Henan Agricultural University, 2003(02):200-204.

[6] Mordor Intelligence. (n.d.). Breakfast cereals Market: 2021 - 26: Industry Share, size, growth - Mordor intelligence. Breakfast Cereals Market | 2021 - 26 | Industry Share, Size, Growth - Mordor Intelligence. Retrieved September 14, 2021, from https://www.mordorintelligence.com/industry-reports/breakfast-cereals-market#faqs.

[7] Lima, M. (2019). It is not all about information! Sensory experience overrides the impact of nutrition information on consumers' choice of sugar-reduced drinks. Food Quality, 74, 1–9. https://doi.org/10.1016/j.foodqual.2018.12.013

[8] Wu Jianhu, et al." Multi-variable missing data processing methods and comparison of analysis results." Journal of Second Military Medical University 25.009(2004):1013-1016.

[9] Zhang FENGming. Survey report on frozen food consumption during COVID-19 [J]. Modernization of shopping malls (1):3.

[10] Lefever, S., Dal, M. and Matthíasdóttir, Á. (2007), Online data collection in academic research: advantages and limitations. British Journal of Educational Technology, 38: 574-582. https://doi.org/10.1111/j.1467-8535.2006.00638.x