Exploring the effect of Aquatic products diet on our "health" using data mining technique

By: Xirui Guan, Wenrui Jiang

Abstract

The purpose of this research is to create a method to find the connection between the aquatic products intake with the "healthy" diet habit. Data is getting from the Center of Disease Control and Prevention and Foods and Drugs Administration. Healthy diet habits were divided into several particular diet habits, for example, "balance of protein in total energy", "Calcium intake", and "Fiber intake" etc. Each diet habit is a response variable that has 3 classes: "Low", "Balance", and "High" by their corresponding range. The predicted variables are different groups of aquatic products; they are "refresh water fish", "sea fish", "shellfish", and "Shrimp and Crab". Each group is the sum of the total number of products the respondent ate during the last 30 days. Four different models are used to analyze the data in this research: "Logistic Regression", "Random Forest", "Naive Bayes Classifier", and "Support Vector Machine". In conclusion, the outcome of the models suggests that the data set does not have enough evidence to support our claim that aquatic products have a strong relationship with elements' diet habits.

Introduction

Our body needs various different nutritions and energy in order to maintain our daily basic operation. The main way that our body receives these nutrients is through eating, intake of nutrition from various different foods. This is very hard to tell the amount of different nutrients

are intake from each meal. It may cause an unbalance of nutrient intake and damage our body. Therefore, it is an important task to find the range of the amount of nutrients humans need.

There are two major nutrient groups: Energy and trace element. There are three different variables in energy intake: Protein, Carbohydrates, and Fat. The commonly used unit to calculate the energy is Calories, not only the total amount of Calories intake is important for our health, but also the proportion of the protein, carbohydrates and fat in total energy intake. The recommended proportion of protein for adults is 10% ~ 15% and 1 gram of protein can supply 4 Calories. The recommended proportion of carbohydrates is 55% ~ 75% for adults a gram of carbohydrates can supply 4 Calories. The recommended proportion of fat is 15%-30% and 1 gram of fat can supply 9 Calories. (*Nutrient and Health - Energy and Protein*)

There are many trace elements humans need to intake each day. It will be too complex and unnecessary to search each individual trace element. Therefore, the research will focus on those trace elements that are related with the aquatic products. The trace elements chosen in this research are calcium, zinc, potassium, and iron. (The Energy contained in shellfish and sea food_baiduzhidao) Furthermore, there exist some trace elements that do not relate with the aquatic product that can be used as a control group. Dietary fiber is a trace element that is not contained in most aquatic products. (“Dietary fiber: Basics of healthy eating”)

The recommended calcium intake amount is 800 mg to an upper bound of 2500 mg per day. (“Calcium and Calcium Supplements: Achieving the Right Balance”) The recommended zinc intake amount is 8mg ~ 22mg per day (*Office of Dietary Supplements - Zinc*). The recommended daily iron intake from 16 mg to upper bound of 45 mg.(National Academies Press (US). The recommended daily potassium is 2000 mg ~ 3500 mg. (*Office of Dietary Supplements*

- *Potassium*) These dietary information will be the range that separates the variable into different classes.

Background/Related Work (data clean)

This data comes from the Center of Disease Control and Prevention and Foods and Drugs Administration data collection from 2005 to 2014. The database is in csv format and is calculated based on person-times, with a total of 50,965 people and 549 variables. This also represents over fifty thousand rows and 500 columns of data. The data includes four aspects, personal information, body calorie intake, body rare element intake and aquatic product intake. Due to too much missing data and the fact that many people filled in incomplete information or had data that was obviously unrealistic, we deleted more than 50% of the columns and rows missing from the data. And some specific outliers are deleted, for example, all data are 0. This means that the provider of this data did not eat anything during the day, including drinking water, which is obviously unreasonable

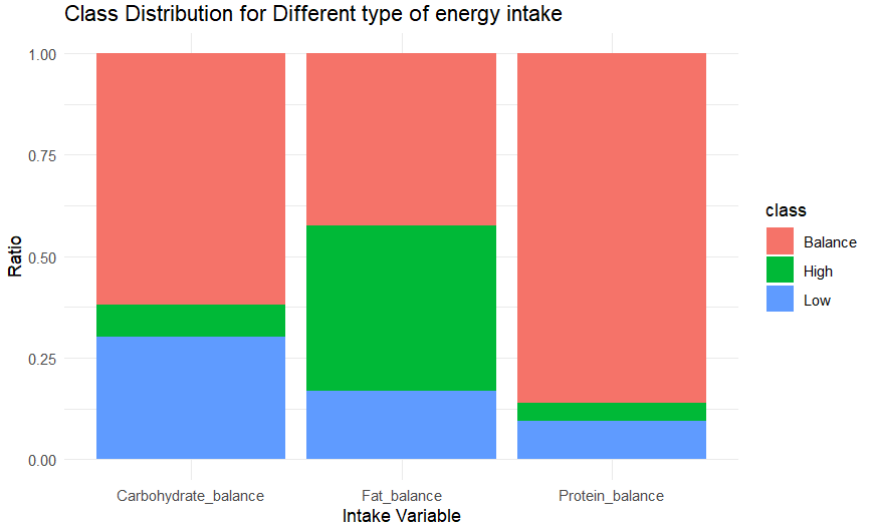
After initial data cleaning, our data was reduced to 24,682 rows and 100 columns. Although it has been reduced a lot, it is still too large and the complexity of building the model will be very high. Therefore, we selected several specific dependent variables according to the objectives of this study and incorporated the types of seafood products to reduce the complexity of the model. Finally, after integration, we got 4 independent variables, shellfish, shrimp and crab, marine fish, and freshwater fish. These four independent variables represent the number of times a person eats seafood every thirty days.

```
cleaned_df.SEQN   shellfish        shrimp_crab        seafish        freshwaterfish
Min.   :31129   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.:43615   1st Qu.: 0.0000   1st Qu.: 1.000   1st Qu.: 0.000   1st Qu.: 0.000
Median :56529   Median : 0.0000   Median : 2.000   Median : 0.000   Median : 0.000
Mean   :56673   Mean   : 0.6767   Mean   : 2.132   Mean   : 1.367   Mean   : 1.195
3rd Qu.:68702   3rd Qu.: 1.0000   3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.: 2.000
Max.   :83729   Max.   :53.0000   Max.   :44.000   Max.   :38.000   Max.   :32.000
```

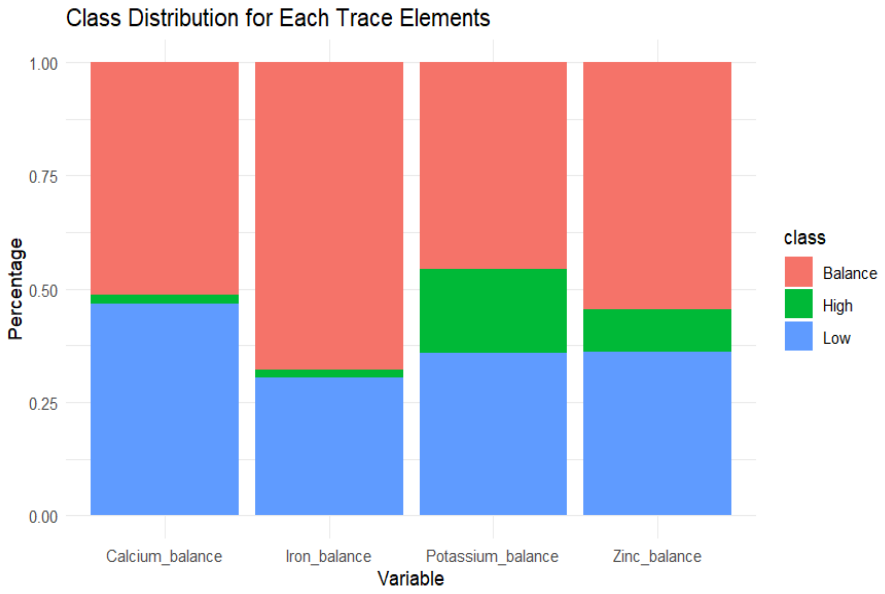
.

You can see the changing range of the four variables from the figure above. shellfish is from 0 to 53, shrimp_crab is from 0 to 44, seafish is from 0 to 38, and freshwater fish is from 0 to 32. The distribution of all variables is right-skewed.

Based on the human intake data, we selected 7 dependent variables for analysis. Since the 7 data have differences and correlations between units, we converted them from numeric types to categorical types based on health indicators. In the classification type, there are three groups: "High", "Balance" and "Low". They represent the level of intake for health recommendations respectively.



The above picture shows the comparison of carbohydrate, fat and protein intake respectively. It can be seen that except for the fat intake of most people, which is too high, the carbohydrate intake and protein intake of most people are balanced.



This is about the intake of trace elements. We selected four trace elements that are highly related to aquatic products for analysis, namely calcium, iron, potassium, and zinc. It can be seen that in terms of trace elements, the intake of most Americans is between balanced and too low, and only a few people have excessive intake.

Based on the above 7 categorical variables and 4 numeric variables, we established different models on fishery production levels for comparison.

Methodology

I. Brief

In order to study the relationship between aquatic products and human nutrition intake, we selected more dependent variables. This was also to compare the models of each dependent variable and observe the differences. Therefore, we established four models for each dependent variable, namely: logistic regression, random forest, Naive Bayes and SVM models. The following model will only use some variables as an example, and the complete data is reflected in the code.

II. Logistic Regression

```
Coefficients:
     (Intercept)    shellfish  shrimp_crab     seafish freshwaterfish
High   -3.212808  0.002250558  0.035868070  0.04551710     0.07012005
Low    -2.078405 -0.061072411 -0.008699556 -0.01241575    -0.04751114
```

The figure above shows the multinomial logistic regression model that describes the relationship between different aquatic products to the protein intake balance. In a multinomial logistic regression model, each level of response variable (except the reference level) has its own intercept term and coefficient of predict variable. In this case if the respondent ate 1 more seafish, it will increase the chance of he/she being on a high protein diet by log-odd 0.046 and decrease the chance of low protein diet by log-odd 0.012. The intercept describes the log-odd of the level when all predicted variables are 0 Overall, the model has an accuracy of 0.867. The confusion matrix of the model suggests the model predicts most test data as a Balance protein diet. This happens because of the unbalanced data, most classes in the data set are Balance protein diets. This is not helpful for the research, by adjusting the class weight for each level of class we might fix this problem. Class weight of class means the attention that a class has in the model. If the class weight is high, the class receives more attention from the model, when the model analyzes the data, it will be focused. On the other hand, a small value class weight will decrease the attention of that class, during the analyzes, it will be "ignored". The goal is to balance the weight of each class, under this principle, the higher amount level will assign a smaller weight, and lesser amount level will assign a larger weight, apply the formula:

```
          Reference
Prediction Balance High  Low
   Balance    3460  154  377
   High          1    0    0
   Low           0    0    0
```

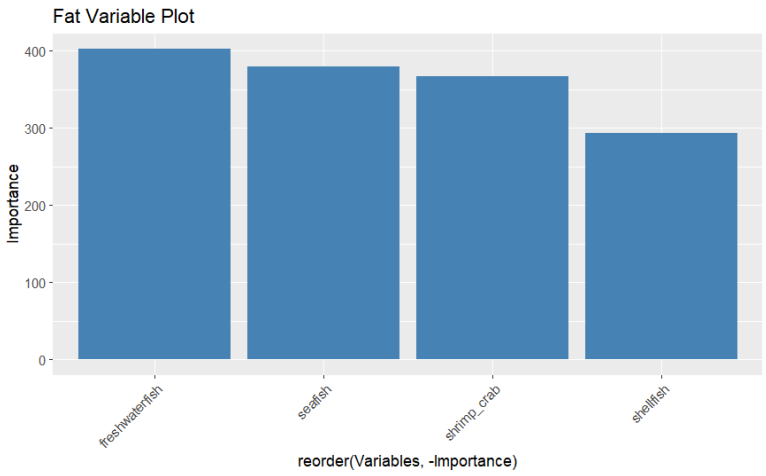
class weight = (total amount of observation) / (3 * the proportion of the class in the data set).

After inserting the class weight parameter into the model, we can see the change of prediction result as the confusion matrix shown at right.

Prediction	Balance	High	Low
Balance	168	9	21
High	1109	70	85
Low	2184	75	271

The predicted amount was from Balance level shift to Low level and High level. It seems the performance of the model (accuracy: 0.12) is lower than the previous, when class weight is not included.

III. Random forest



The above plot is a comparison of important variables in the random forest model of Fat intake. We can see that the importance gaps of all variables are not decreasing and are declining gently. The variable of highest importance, that is, the variable with the highest correlation with Fat in the random forest model, is the number of freshwater fish meals.

Prediction	Balance	High	Low
Balance	12	72	17
High	688	2471	619
Low	21	70	22

Next is the confusion matrix of the fat random forest model. It can be seen that the model has the highest prediction accuracy in the range of High, and predicts most of the data as High. The total accuracy of this model is 0.6275, which is higher than 0.6, which means that the model shows a partial connection between the independent variables and the dependent variables. But the Kappa value is -0.0032. A negative kappa value indicates that the model performs poorly on other data which may be balanced data. In general, this model can only guarantee the feasibility of this data, but its use on other data may cause a decrease in accuracy.

IV. Naive Bayes (W)

Another model that is used in research is the Naive Bayes Classifier. Naive Bayes Classifier based on the Bayes theorem, assumes all the features are conditionally independent given the class label. The classifier classifies the data by choosing the highest probability of the label of the text. Each variable contributes to the model by class-conditional probabilities:

	Reference		
Prediction	Balance	High	Low
Balance	548	250	351
High	177	98	76
Low	1122	370	1000

$$P(X1,X2,...Xn \mid Y) = P(X1|Y)\ P(X2|Y)\ \ldots P(Xn|P).$$

An example of the variable of the Naive Bayes Classifier in the research is shown in the figure. This is the result of Naive Bayes classifier classifying the diet habit of Potassium. The effect of each predicted variable in on the outcome of prediction in this model are:

```
Conditional probabilities:
        shellfish
Y            [,1]     [,2]
  Balance 0.6772370 1.670591
  High    1.0319220 2.540295
  Low     0.5149033 1.658469

        shrimp_crab
Y            [,1]     [,2]
  Balance 2.147801 2.251541
  High    2.396505 2.573173
  Low     1.976120 2.291201

        seafish
Y            [,1]     [,2]
  Balance 1.432097 2.562183
  High    1.774194 3.019726
  Low     1.096740 2.330940

        freshwaterfish
Y            [,1]     [,2]
  Balance 1.237970 2.200112
  High    1.609543 2.699162
  Low     0.906920 1.819060
```

As we can see above, each figure shows the conditional probability of each predicted variable on each response level. The first column shows the mean value of distribution of each predicted

variable and the second column shows the standard deviation of the distribution of each predicted variable. Each level of the response variable's probability is independently calculated, the model will choose the level that has the highest probability to classify the observation. The individual predicted variable on response variable is by this formula. P(X = x | Y = y) = (sqrt (2 * pi * (standard deviation)_y)) ^ (-1) * exp(-((x-mean)^2) / (2*(standard deviation)*y)). For example P(X = seafish | Y = Balance) = (sqrt (2 * pi * (2.56)^2)^(-1) * exp (- (seafish-1.43)^2 / (2*(2.56)^2))) .

V. SVM

For the SVM model, this is the model that took the longest in this analysis. There are three types of SVM models, linear SVM, polynomial SVM and Gaussian kernel SVM. For this analysis, we chose Gaussian kernel SVM because our dependent variable is categorical data. The following is the analysis result of carbohydrate modeling for four independent variables.

Confusion Matrix and Statistics

	Reference		
Prediction	Balance	High	Low
Balance	1637	228	663
High	28	4	16
Low	816	84	516

It can be seen that the total accuracy of this model is moderate, which is 0.5403. As can be seen in the above table, most of the guessed data are balanced and are more in line with the original data of the dependent variable. This model does not compare the Kappa value because the accuracy of the model is no longer enough to indicate the inevitable connection between the model and the data, especially for the classification model, so there is no need for further testing.

Evaluations

Summarize the models made for each response variable, some models have clear and relatively high accuracy, but the main reason is because of the unbalanced data set. This is also supported by the situation when we insert the class weight into the model, the accuracy of the model heavily drops, moreover the kappa value of the model also suggests the models' high accuracy is based on the unbalanced data set. In the case of a relatively balanced data set, the accuracy of the models is not high enough to support the existence of a strong relationship between the aquatic product and diet habits; most models' accuracy are less than 0.6.

	Reference		
Prediction	Balance	High	Low
Balance	1678	71	608
High	4	1	1
Low	1024	36	569

Furthermore, we also build models for the control group diet fiber. The model for predicting diet fiber also has 0.56 accuracy, it is very close to the result of other elements. This shows that the data set does not support the strong relationship between aquatic products and nutrition diet habits.

Conclusion

According to the model results, the data set does not support our hypothesis that the number of aquatic product intakes does not decisively affect the nutrient balance of each individual's daily intake. Although some models have relatively high accuracy, their kappa values indicate that this occurs very frequently because the data is unbalanced. None of these models can achieve high accuracy on new data with the same classification. We did not find any

strong relationship between seafood and protein and fat dietary habits. Trace elements (calcium, zinc, potassium) may have a moderate relationship with aquatic products. In this study, nonlinear models performed no better than linear models. In addition, we also used some trace elements that are not found in aquatic products for modeling comparison and found that some irrelevant trace elements also showed this moderate relationship. This is a very high value for the trace element model in this experiment. Therefore, based on the research data and model, we determined that this data set does not represent the relationship between the number of times each person consumes water products per month and balanced nutritional intake.

Citation

1.*Nutrient and Health - Energy and Protein*. 16 Nov. 2018, www.cfs.gov.hk/english/multimedia/multimedia_pub/multimedia_pub_fsf_29_02.html.

2.“Calcium and Calcium Supplements: Achieving the Right Balance.” *Mayo Clinic*, 1 Nov. 2022, www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/calcium-supplements/art-20047097.

3.*Office of Dietary Supplements - Zinc*. ods.od.nih.gov/factsheets/Zinc-HealthProfessional.

4.*Office of Dietary Supplements - Potassium*. ods.od.nih.gov/factsheets/Potassium-Consumer.

5.National Academies Press (US). “Iron.” *Dietary Reference Intakes for Vitamin a, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc - NCBI Bookshelf*, 2001, www.ncbi.nlm.nih.gov/books/NBK222309.

6.海鲜贝壳都含什么能量_百度知道. zhidao.baidu.com/question/1371902264175687819.html.

7.“膳食纤维：健康饮食的基础.” *Mayo Clinic*, 4 Nov. 2022, www.mayoclinic.org/zh-hans/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/fiber/art-20043983#:~:text=%E8%86%B3%E9%A3%9F%E7%BA%A4%E7%BB%B4%E5%8F%88%E7%A7%B0%E7%A7%B0,%E4%BA%8E%E6%B0%B4%E7%9A%84%E4%B8%8D%E6%BA%B6%E6%80%A7%E7%BA%A4%E7%BB%B4%E3%80%82.

R code note

1. Put the data sets into the R working directory (it make be set in Rstudio Session)
2. Run chunks one by one following the order.