STAT 542 Final Project: Analysis of Open Food Data

Liqian Ma (lma11), Shubo Sun (shubos2) and Zhaohong Wang (zw59)

5/06/2021

Project description and summary

In this project, we would like to take a deeper look into the open food data. This dataset is from Kaggle, and it is originally posted on the openfoodfacts website. We separated our project into two parts. In the first part, we tried to perform clustering analysis using the nutrition data. From the clustering result, we tried to understand if the labels/categories of the food and nutritional clusters are related. In the second part, we performed text analysis and extract the top 150 most common ingredients of the food. We employed classification algorithms to predict the nutrition scores of the food.

We wanted to achieve two goals in this project. In the first part, we selected carbohydrates_100g, sugars_100g, fiber_100g, proteins_100g, sodium_100g, trans.fat_100g, cholesterol_100g these 7 variables to perform unsupervised learning to look for relationships of these 7 common nutritions in the food dataset. Also, we categorized the food into 3 categories, candies, chips and diary by their product name. In order to validate our clustering results, we compared the clusters with the 3 categories of food. To perform unsupervised learning, we used K-means, K-medoids and SOM methods. K-means and K-medoids clustering, compared to SOM, could better identify the 3 categories of the food. In the second part, we performed supervised learning using the text information from the open food dataset. The response variable is nutrition score, and based on the value of the score, we separated this variable into 2 levels, low nutrition and high nutrition. We then used 7 classification algorithms (Logistic Regression, SVM, KNN, LDA, QDA, AdaBoost and Random Forest) and compared the results. The best performing model we chose had greater than 85% accuracy in predicting the nutrition levels using the test dataset.

Literature Review

Open Food Facts

We read the paper "Discriminating nutritional quality of foods using the 5-Color nutrition label in the French food market: consistency with nutritional recommendations"[1]. We believe what we need to do in our project is similar and this paper can be used as a good reference. The goal of this paper was to create a 5-color nutrition label for foods on French market based on the nutrients to discriminate nutritional quality of foods. The nutrition label was the compared to French nutritional recommendations.

In their analysis, foods are categorized into different groups, such as 'Fruit and Vegetables', 'Cereals and potatoes', 'Meat, Fish and Eggs'. The authors used the Food Standards Agency nutrient profiling system (FSA) to assess the nutrition scores. Sugar, fat, and sodium contribute to positive scores and fruits, vegetables, fibers, and proteins would decrease the FSA score. Then they use cutoffs were determined to split all the food into 5 different categories (5-Color nutrition label) according to the FSA score.

The results of this study showed that food categories as determined by the 5-color nutrition label were consistent with French recommendations. Most of the fruits and vegetables were labeled as green category, which means more healthy. On the other hand, most of the sugar snacks were grouped into red or pink categories, which are less healthy. By introducing this 5-color nutrition label, general public may find easier to make healthy food choice decisions.

Text data analysis

In the article "Text Mining in Organizational Research", authors introduced basic concepts regarding text mining[2]. Specifically, text analysis generally comprises three steps: (1) text preprocessing, (2) application of TM operations, and (3) postprocessing. During text preprocessing, it is important to clean the text (e.g., lowering the cases, removing stopwords). Then the text is generally transformed into a matrix structure, termed document-by-term matrix. This structure would allow text analysts to obtain word frequencies, which can be then used to construct covariates. In our analysis, we learned from these ideas to obtain word

[1] Julia C, Ducrot P, Peneau S. et al. (2015) Discriminating nutritional quality of foods using the 5-Color nutrition label in the French food market: consistency with nutritional recommendations. Nutr J. 14:100. doi: https://doi.org/10.1186/s12937-015-0090-4

[2] Kobayashi B, Mol ST, Berkers HA, et al. (2017) Text Mining in Organizational Research. Organ. Res. Methods. 21:733-765. doi: https://doi.org/10.1177/1094428117722619

frequency table and construct dummy variables. Due to the presence of amount of vocabularies, commonly the document term matrix could be very sparse, and hence impose high computation cost. To overcome this problem, we only used the top 150 common words. After preprocessing, clustering and supervised learning algorithms can be applied to the reconstructed data and post processing requires that the validity of the model to be evaluated.

Another paper we read about text analysis is "How We Do Things With Words: Analyzing Text as Social and Cultural Data"[3]. In this paper, authors described their experiences in analyzing text data, which involed rich social and cultural concepts. There are three goals that the authors wanted to achieve. First, they discussed key issues that everyone may encounter when trying to do text analysis. Second, they provided some important questions that may be helpful to guide work in this area. The final goal was to help interdisciplinary collaborations. At the beginning of the paper, the authors covered data acquisition, compilation and metadata incorporation. For our project, the data were already provided, we did not need to do this step. However, it is important to be aware of the ethical concerns surrouding digital data acquisition and the value of data labeling, which could provide context of additional information regarding the text documents.

Then, operationalization methods were explained in the article. For example, the authors talked about how to do the data preprocessing (e.g., tokenization) and model selection (e.g., supervised learning, topic modeling). Compared to supervised learning, the authors discussed that topic modeling is especially suited for insight-driven analysis due to high interpretability. The authors next emphasized on model validation, because it is important that the models are reliable and provide consistent results. Usually, for text analysis models, the models are evaluated by comparing the model result with the real labels. In this case, human-generated labels need to be very accurate, or the result could be meaningless. We also conducted model validation in our project to make sure the reliability of our models.

Data Processing and Summary Statistics

Data processing for unsupervised learning

The data processing for the unsupervised learning comprises 3 steps: (1). Select and use only the most commonly labeled nutrient facts including carbohydrates, sugars, fiber, proteins, sodium, trans fat and cholesterol as predictors. (2). Construct categorical response variable by identifying keywords in the product name variable. Specifically, the products with keywords "chocolate" or "candy" in the product names and without another 21 keywords such as "wafer", "cupcakes", "cookie" or "brownie" were categorized as "candies". The products with keyword "chips" in the product names and without "cookies" or "chocolates" were categorized as "chips". The products with keywords "milk" or "cheese" in the product names and without another 12 keywords related to "sweets" such as "chocolate", "candy" or "duds" were categorized as "diary". (3). Remove all the uncategorized observations and observations with missing values. After these steps, 22782 obeservations were left for further analyses. For clustering analysis, 70% of these observations are used as training data and the rest 30% testing data.

Data processing for supervised learning

For supervised learning, text information was used as predictors for the nutrition scores (nutrition.score.fr_100g). Only observations from the United States were kept. To reduce the computational demand, a random subset of 15000 complete observations was sampled from all observations of the United States. Because the variable additives and ingredients_text essentially contain the same information, only the variable additives is used to generate a text document matrix by isolating individual words or phrases of certain patterns. Specifically for each observations the additive phrases with only 1 hyphenation (e.g., word-word, word-number) were kept, while those with other patterns for instance having 2 hyphenations (e.g., word-word-word) were excluded. The text document matrix is a 11147x15000 matrix with columns representing the observations and the each row representing the individual extracted words/phrases from all observations. Subsequently from the word document matrix, a word/phrase frequency table is then generated by counting the number of times each word/phrase appeared in total across all observations. We extracted

[3] Nguyen D, Liakata M, DeDeo S, et al. (2020) How We Do Things With Words: Analyzing Text as Social and Cultural Data. Front. Artif. Intell. 3:62. doi: 10.3389/frai.2020.00062

the top 150 most frequent additive words/phrases and construct each of them as a dummy variable to be used as predictors for the response variable nutrition score. The response variable was then constructed from the variable nutrition.score.fr_100g to be a 2-level categorical variable (greater than 10 and smaller than or equal to a score of 10, with 10 being the median of all scores). Similar to the unsupervised learning, 70% of the 15000 observations are used as training data and the rest testing data.

Summary Statistics

The original dataset contains 356027 rows and 163 columns. In the data, the packaging information mainly includes the product tags with both product names and brand tag names to indicate the food categories, the places of origin, the additives and ingredients information to categorize the food, and the serving size with information related to the number of nutrient substance and additives.

After data processing, the data used for unsupervised learning contains 22782 observations and 7 variables related to nutrients, carbohydrates_100g, sugars_100g, fiber_100g, proteins_100g, sodium_100g, trans.fat_100g, cholesterol_100g, and 1 constructed food category variable (**Table 1**). The summary statistics of these variables are shown in **Table 2**.

Table 1: Frequency of each reconstructed food category

category	freq
candies	8216
chips	3449
diary	11117

Table 2: Summary statistics of nutrient variables

nutrient	mean	median	min	max
carbohydrates_100g	36.9003415	43.480	0	100.000
cholesterol_100g	0.0311417	0.009	0	62.500
fiber_100g	2.3352398	0.400	0	83.300
proteins_100g	10.2113305	7.140	0	100.000
sodium_100g	0.5028593	0.321	0	708.333
sugars_100g	19.6554495	5.000	0	100.000
trans.fat_100g	0.0782130	0.000	0	38.460

The data used for supervised learning contains 15000 observations and 150 predicting variables of additives/ingredients, and 1 response variable nutrition level. **Table 3** show the top 10 frequent additives/ingredients, respectively. **Table 4** presents the frequency of nutrition levels.

Table 3: Top 10 most frequent additives/ingredients

words	freqs
salt	12242
sugar	9791
oil	8674
flour	7978
water	7496
milk	5397
wheat-flour	3968
syrup	3836

words	freqs
powder	3778
starch	3237

Table 4: Frequency of each nutrition level

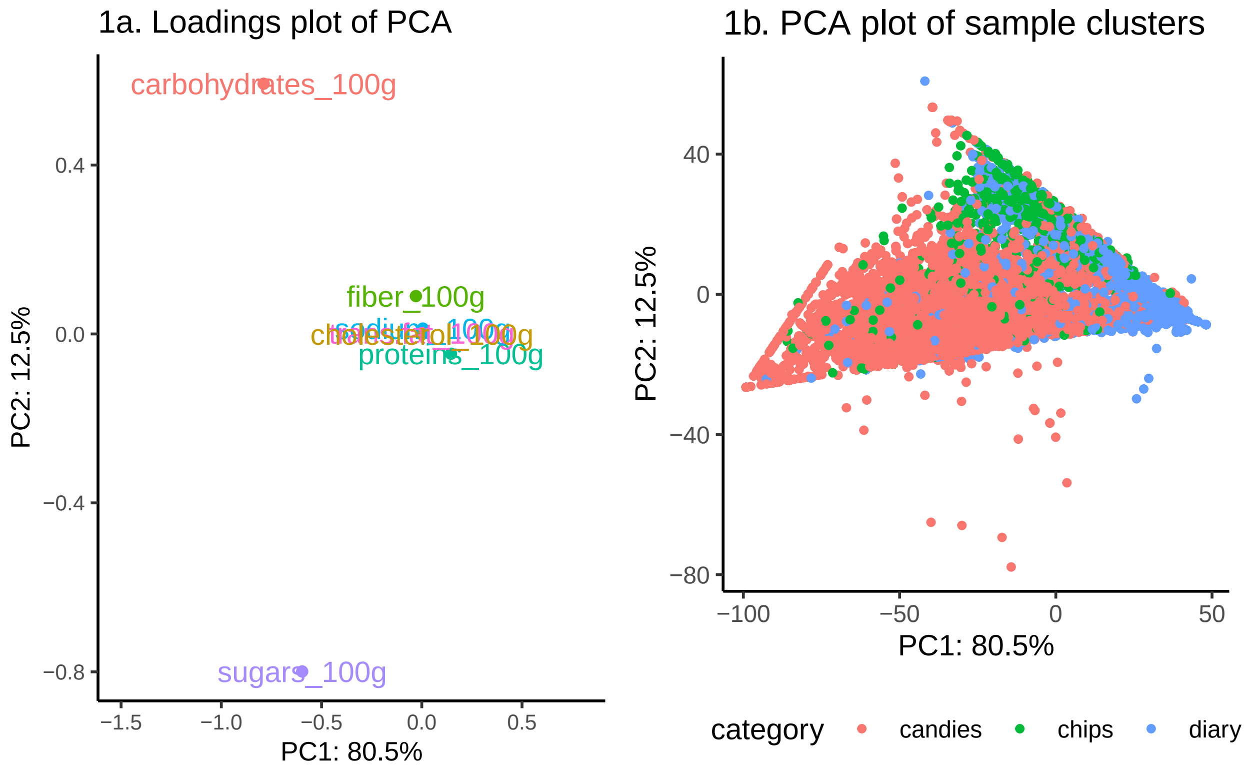
nutrition level	freq
less than 10	7785
more than 10	7215

Clustering analysis of food nutrition

In this section, we are interested in understanding the heterogeneity of nutrient profile of distinct food categories (diary, candies and chips) that are produced in the United States. That is, whether diary, candies and chips have distinct enough nutrient values (i.e., carbohydrates, sugars, fiber, proteins, sodium, trans fat and cholesterol) to be clustered into three different clusters. First, to visualize the data in a lower dimensionality, we performed dimensionality reduction via principal component analysis (PCA). The result shows that the first principal component could explain over 80% of the variation in the original data, and the second component could explain about 12.5%, so the first two components explained over 90% of the variation in total (**Fig 1**). The first principal component has positive eigenvalue in protein, sodium and cholsterol with negative eigenvalue value in the remaining nutrient contents (**Fig 1a**). The second principal component has positive eigenvalue value in carbohydrates, fiber, sodium and trans fat, and the carbohydrates variable has predominant contribution to the second principal component than the other nutrient contents (**Fig 1a**). **Fig 1b** shows some level of separation of all three types of food. Diary has more positive value in PC1, and intuitively diary products tends to contain more protein, sodium and cholsterol; candies has more negative value for PC1, so candies-related food will expectedly contain more sugars.

Then, we proceeded to cluster analysis using K-mean, K-medoids and unsupervised self-organizing map to investigate whether the common nutrient contents would be able to cluster data observations into three groups. The clustering results were then compared and validated using the constructed food category label.

Fig 1. PCA plot of all observations



K-means clustering

K-means clustering is initiated from randomly spliting a dataset into K different subsets/clusters, and randomly assign the observations in each cluster a corresponding cluster label. This algorithm then iteratively calculates the cluster mean vectors and re-assign each observation to the closest cluster mean, until the cluster assignment no longer changes.

On the 70% of training data, we performed K-means clustering with $k = 3$, because we categorized the food into 3 categories using their product names. We also set 20 random initializations to reduce the chance of being stuck at a local minimum. The clustering results on training data are shown in **Table 5**: most of the observations with food category candies were clustered into cluster 1, diary into cluster 2 and chips into cluster 3, even though there are also a relatively big number of candies that were categorized into cluster 3. By voting majority, we re-named the first cluster of K-mean to be candies, the second diary and the third chips. We subsequently calculated the distances between each observation in the testing data and the cluster mean vector of the re-named clusters. Individual observations were then assigned to the corresponding cluster when the distance between the observation and cluster mean is the smallest. The prediction error using the test dataset is approximately 17% (**Table 6**).

Table 5: K-means clustering result compared to food categories

train clusters	candies	chips	diary
1	4284	104	145
2	339	65	6754
3	1092	2249	915

Table 6: Training and testing error of K-means clustering

training error	testing error
0.167	0.172

K-medoids clustering

K-medoids is an alternative version of K-means. Unlike K-means that calcualtes the cluster mean vectors, K-medoids searches for the one observation that minimizes the distance to all others in the same cluster, and this observation is used as the cluster center, akin to the median. For this reason, K-medoids could be more robust than K-means against potential outliers.

On the 70% of training data, we performed K-medoids clustering with $k = 3$. The clustering results on training data are shown in **Table 7**: quite similar to the K-means results (**Table 5**), most of the observations with food category candies were clustered into cluster 1, diary into cluster 2 and chips into cluster 3. However, K-medoids seems to cluster more observations with food category candies into cluster 1 and less into cluster 3, compared to K-means clustering. By voting majority, we again re-named the first cluster of K-medoids to be candies, the second diary and the third chips. We subsequently calculated the distances between each observation in the testing data and the cluster center vector of the re-named clusters. Individual observations were then assigned to the corresponding cluster when the distance between the observation and cluster center is the smallest. Both training error and prediction error using the test dataset are better than K-means at approximately 14% (**Table 8**).

Table 7: K-medoids clustering result compared to food categories

train clusters	candies	chips	diary
1	4682	135	170
2	370	65	6764
3	663	2218	880

Table 8: Training and testing error of K-medoids clustering

training error	testing error
0.143	0.147

Self-organizing map

Unlike K-means, the cluster means of self-organizing map (SOM) have geometric relationships. Additionally, it does not use all the observations at once, rather it includes the observations one-by-one. Therefore, the cluster mean is updated as observations being added into the model.

We built an unsupervised SOM model by specifying a 10x10 grid of cluster means. **Fig 2** shows that some clusters contain a large fraction of certain features, and some clusters have very small fractions of almost all features. For example, the right half clusters contain a high amount of carbohydrates and sugar in general, while the clusters at the lower left corner contain generally higher amount of proteins. Also, clusters with a large proportion of proteins tend to have few other nutrients. Overall, we could see that each cluster is mainly represented by one or two nutrients. We then performed hierachical clustering on the codebook vectors from the SOM model and cut the tree to generate 3 clusters (**Fig 3**). The clusters were then assigned to each observation in the training data (**Table 9**). Unfortunately, SOM followed by hierarchical clustering was not able to distingush the three food categories based on the nutrient profiles.

Overall, K-means, K-medoids and SOM + hierarchical clustering were all able to identify certain underlying patterns of the data represented by 7 common nutrients: carbohydrates, sugars, fiber, proteins, sodium, trans fat and cholesterol. To investigate if different composition of these nutrients would imply certain food categories, we constructed 3 food categories from the product name variable to be diary, candies and chips. When $k = 3$, K-means and K-medoids clustering were both able to find 3 clusters using the nutrient covariates that distingush the food categories. After assigning labels to the clusters according to voting majority, only 14% of the samples were misclassified in both training and testing set with K-medoids clustering. However, even though SOM was also able to detect certain underlying patterns of the observations using the nutrients information, those distinctions were not as closely related to the food categories of those observations.

Fig 2. SOM Codes Plot

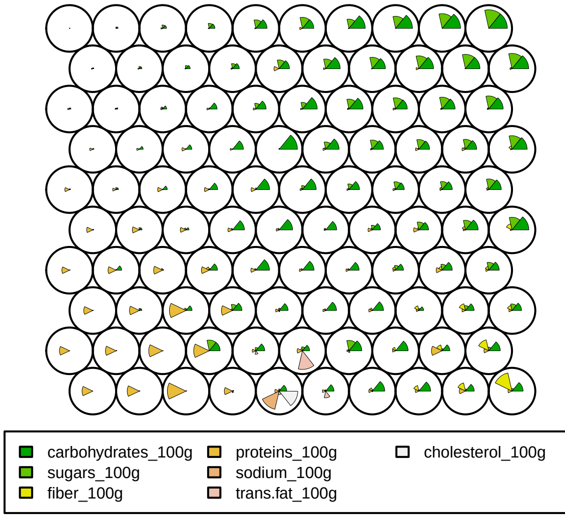


Fig 3. Hierachical clustering of codebook vectors

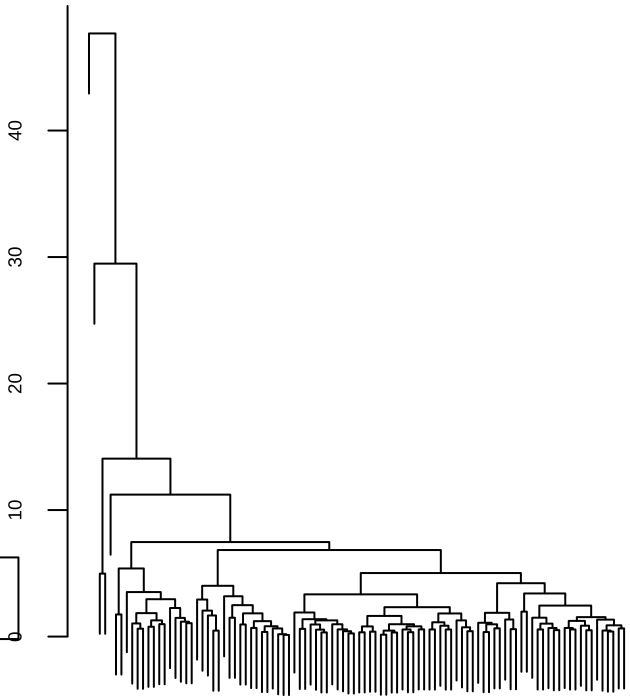


Table 9: SOM clustering result after hierachical clustering compared to food categories

train clusters	candies	chips	diary
1	5712	2418	7806
2	0	0	3
3	3	0	5

Supervised Learning of Additives/Ingredients - Classification

To investigate the relationship between the French Nutrition Scores and the ingredients/additives of each product, supervised learning for classification is a good choice. In this section, we are interested in predicting the nutrition scores of each product to be either "greater than 10" or "smaller than or equal to 10". Hence our response Y_i for each observation is binary.

After text preprocessing, we selected the 150 most frequent additives/ingredients as the predictors. Each of the additives/ingredients was treated as a dummy variable such that if the product include this ingredient/additive, and then value of this observation for this ingredient/additive is TRUE, otherwise it would be FALSE. If we define $X = (X_1, X_2, \ldots, X_n)^T$ to be set of the observations, and Z_i to represent the probability that the additive/ingredient name of X_i contains j_{th} additive/ingredient we selected, then the observed value of each entry will be

$$X_{ij} = \begin{cases} TRUE & Z_i = 1 \\ FALSE & Z_i = 0 \end{cases}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, 150$.

We trained and tuned the parameters (if applicable) of 7 classfication algorithms, Logistic Regression, SVM, KNN, LDA, QDA, AdaBoost and Random Forest, using the training data, and predicted the response (nutrient score) using the testing data.

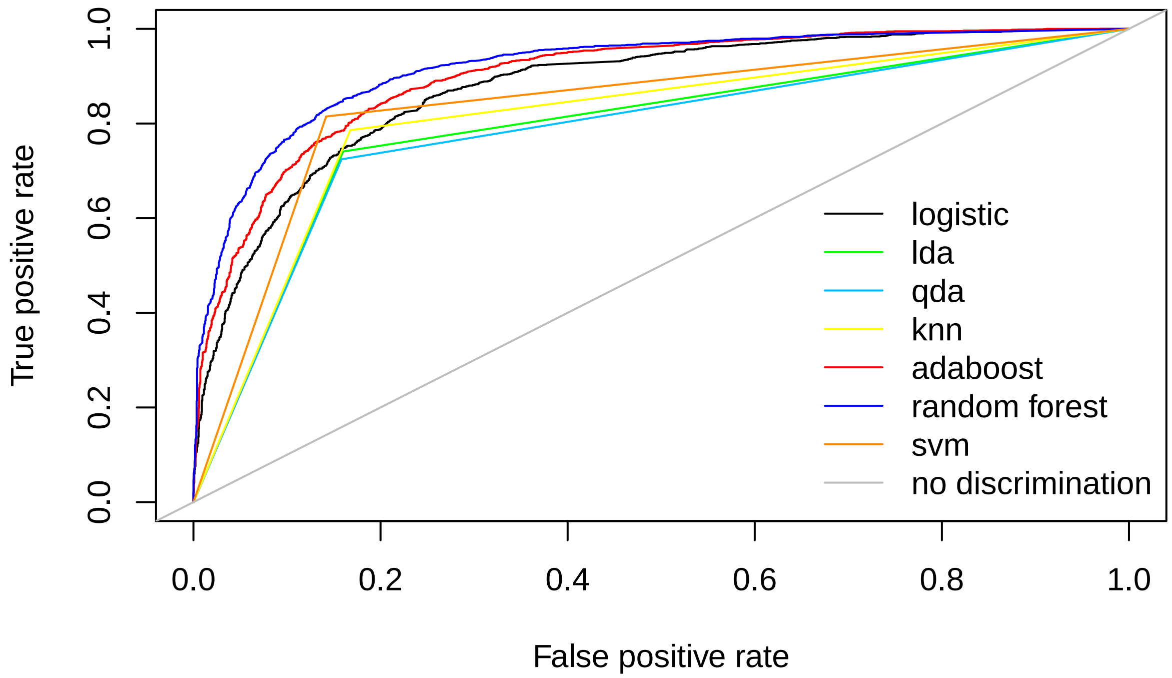
For the algorithms being used, SVM was tuned with 3-fold cross validation and the best tuning parameter γ was 0.0067. KNN was tuned on a grid of 13 K with 3-fold cross validation to select the best performing $K = 1$. For AdaBoost, we set the shrinkage parameter δ to be 0.1, and tuned the number of trees to be optimal at 1224. For Random Forest, we also set the shrinkage parameter δ to be 0.1, and tuned the numbers of trees and `mtry` within the RandomForest function. The optimal number of trees for random forest was 1297, so we used 1300 trees. A 3-fold cross validation used to tune $mtry = 10, 12, 15, 20, 30, 35, 40, 50$, and at $mtry = 12$ it achieved the best prediction accuracy. The summary of the tuning parameters and the resulted optimal parameters are listed in **Table 10**.

Table 10: Summary of the results of 7 classification models

method	tuning parameter	best parameter	testing error
Logistic Regression	NA	NA	0.227
SVM	γ, 3-fold CV	$\gamma = 0.0067$	0.161
KNN	K, 3-fold CV	K = 1	0.180
LDA	NA	NA	0.202
QDA	NA	NA	0.216
AdaBoost	n_tree, $\delta = 0.1$	n_tree = 1224	0.189
Random Forest	n_tree, mtry, $\delta = 0.1$	n_tree = 1297, mtry = 12	0.156

A ROC curve shows the performance of a classification model at all classification thresholds on the training data, while area under the ROC (AUC) measures the two-dimensional area under the ROC curve, which ranges from 0 to 1. The line no discrimination represents AUC of 0.5, and a classifier with AUC of 0.5 would not able to distinguish between classes. All the classifier tested here had AUC greater than 0.5 and smaller than 1, meaning that they are capable of distinguishing between classes. ROC curve generally shows a tradeoff between the sensitivity and specificity. The classifier with greater true positive rate and smaller false positive rate should be preferred. **Fig 4** shows that compared to other classification methods, the tuned Random Forest has the greatest AUC value. **Table 10** also shows that the Random Forest model has the best prediction accuracy on the testing data.

Fig 4. ROC curve of classification models



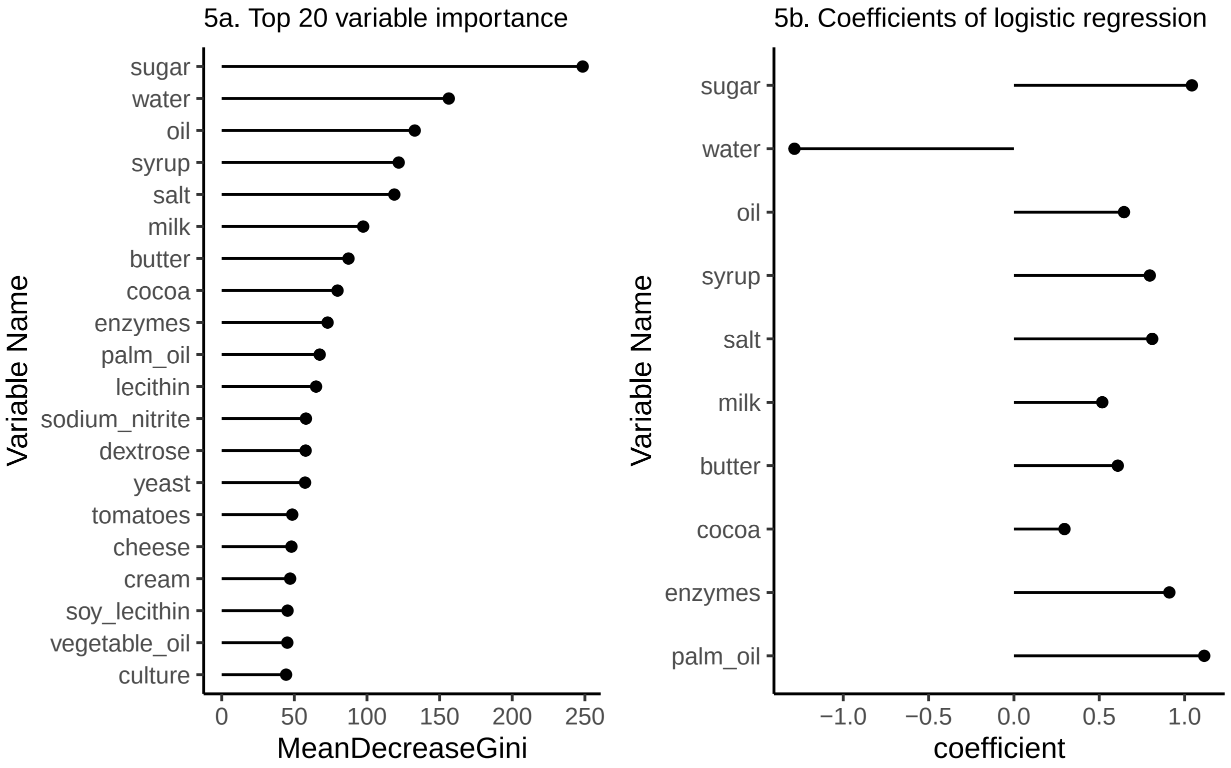
Final Classification Model: Random Forest

Based on previous analysis of data, the Random Forest model fitted using 1297 trees with the shrinkage estimator $\delta = 0.1$ and $mtry = 12$ has the smallest classification errors, so we treated this as our final model to classify the nutrient scores of food produced in the United States.

From **Fig 5a** by Random Forest on the training data, the two most important ingredients/additives that could influence the nutrient scores are sugar and water. Other relatively important variables are syrup, oil, salt, milk, and butter. This result is not surprising because most of food in the supermarkets directly include these substance or include some other ingredients that are consisted of these substances. The rest of the ingredients have approximately the same importance. However, one disadvantage of Random Forest is that it cannot directly indicate the relationship between the predictors and the response, so a Logistic regression was applied to find some possible relations.

The top 10 most important ingredients of the training data were plotted in **Fig 5b**. Since the response is binary, a positive coefficient would show a greater chance for a randomly selected food to have nutrient score greater than 10. Water has very negative coefficient, meaning food containing water with a large proportion tends to have nutrient score smaller than or equal to 10. The other coefficients are positive, so the food containing these ingredients tends to have a nutrient score greater than 10. Combining the results from **Fig 5a** and **Fig 5b** could make it easier for customers to understand the roles of each ingredient to the overall nutrient scores of the food, and better choose the food they may want. Additionally, these results could also help the supermarkets to make informed decisions to meet different demands of different customers based on the nutrient scores. Ultimately, this will also allow food industries and supermarkets to decide what types of food should be produced more or sold more. In summary, the classification error rate is about 16% for this model, and considering the large sample size, our random forest model could classify the observations with a great accuracy.

Fig 5. Variable importance and coefficients


5a. Top 20 variable importance

5b. Coefficients of logistic regression

Conclusion

For unsupervised learning, we reconstructed the observations into three food categories (candies, chips and diaries) and found patterns of the dataset through clustering design. Detailedly, clusters were built based on the sugars, carbohydrates, proteins, trans fat, fiber, sodium and cholesterol of each food. K-means and K-medoids clustering found three clusters highly related to the food categories and yielded low prediction errors. PCA also showed some patterns of the three categories: diary food tends to contain more protein, sodium and cholesterol, and candies-related food tends to contain more sugars. Hierarchical clustering and unsupervised SOM failed to distinguish the three food categories, but another underlying pattern was found that the food with more proteins tends to include very few of other nutrients simultaneously.

For supervised learning on classifying the nutrient scores into either "greater than 10" or "smaller than or equal to 10", Logistic regression, SVM, LDA, QDA, KNN, Adaboost and Random Forest were used to compare their performance of prediction accuracy. Using the best tuning parameters, Random Forest model achieved the greatest AUC value. As a result, we used random forest as our final classification model. From the result of this classification model, sugar and water were the most important factors predicting the nutrient score of food based on the dummy variables created from 150 most frequent additives/ingredients. The coefficient plot from Logistic regression corresponds to the importance of variables in the classification model, and further indicates the relationships between the nutrient and each of 150 most frequent additives/ingredients.

References

1. Julia C, Ducrot P, Peneau S. et al. (2015) Discriminating nutritional quality of foods using the 5-Color nutrition label in the French food market: consistency with nutritional recommendations. Nutr J. 14:100. doi: https://doi.org/10.1186/s12937-015-0090-4

2. Nguyen D, Liakata M, DeDeo S, et al. (2020) How We Do Things With Words: Analyzing Text as Social and Cultural Data. Front. Artif. Intell. 3:62. doi: 10.3389/frai.2020.00062

3. Kobayashi B, Mol ST, Berkers HA, et al. (2017) Text Mining in Organizational Research. Organ. Res. Methods. 21:733-765. doi: https://doi.org/10.1177/1094428117722619