A Predictive Analysis of Food Nutrient Density Project Proposal

David Berberena

Bellevue University

DSC 680 Applied Data Science

Amirfarrokh Iranitalab

September 1, 2024

A Predictive Analysis of Food Nutrient Density Project Proposal

Topic

After researching interesting topics, I have decided to conduct a predictive analysis of various food ingredients to forecast nutrient density. Many people are looking for the healthiest food in the world, and while no one ingredient checks all the boxes, those food ingredients that are the most nutrient-dense can lead consumers to a healthier lifestyle.

Business Problem

Food is one of the necessities of life, and the nutritional and dietary needs sector of the food industry focuses on providing the most informed recommendations on what foods help to stay healthy and what foods come with certain dietary consequences. An overabundance of junk food (or any food not consumed in moderation) in a person's diet may lead to medical issues, and nutritionists and dieticians need to have proven food regimens for these individuals to regain their health. To create these regimens, predictive analytics can help to identify how nutrient-dense a food ingredient is based on other factors such as the amount of proteins, sugars, carbohydrates, fats, vitamins, and minerals. For as long as doctors have been around, we have been told that many health complications can be solved with proper sleep, food, and exercise. Predicting the nutrient density of food ingredients so organizations like the United States Department of Agriculture and the Food and Drug Administration can provide the most accurate information to consumers through their daily dietary needs recommendations and other suggestions is paramount to ensuring their continued credibility. New foods created by companies also need to be sorted into specific food categories (junk foods, dietary supplements, powerhouse fruits and vegetables, pulses, etc.) by these organizations, and predicting nutrient density plays a hand in determining how a new food should be classified.

Datasets

Food data is extensive and can be found through many credible channels. The datasets I will use have been found on Kaggle, and originate from the Comprehensive Nutritional Food Database. I will be utilizing five datasets, all of which contain nutritional values of food such as food ingredient name, calorie count per one hundred grams, carbohydrates per one hundred grams (measured in grams), protein per one hundred grams, fats, vitamins measured in milligrams, minerals measured in milligrams, and nutrient density.

Methods

Initial analysis of the business problem and the data prompts exploratory data analysis of the datasets. Since all five datasets share the exact same variables, I will first join them in a Jupyter Notebook via Pandas's concatenation function. Then I plan to look for factors that strongly correlate with nutrient density by using a heatmap to display the correlation coefficients and their relationship to the nutrient density variable. I also would like to plot other charts such as bar charts and scatter plots to visualize the distribution of data points and to see what food ingredients are highest in each variable present in the dataset to see their relevance. After the exploratory data analysis and any other transformations made to the data are complete, I will be splitting the dataset into a training and test set at an 80/20 ratio and crafting a multiple linear regression model in the form of an ordinary least squares model using Python's predictive analytics libraries.

The data allows for both classification and regression analysis, yet the goal of this project, as it stands now, is to predict nutrient density provided the most significantly impactful variables present in the data. This lends itself to regression analysis, so the ordinary least squares model is a viable model choice. The intent is to then cross-validate and normalize the model using both k-fold cross-validation and Ridge regression (L2 Normalization) to output the most accurate regression model. The metrics that will be used to calculate the model's predictive performance are the RMSE (Root Mean Squared Error) and the R^2 value (coefficient of determination). The RMSE value will provide an answer to the difference in model predictions and actual values, while the R^2 value indicates the amount of variance in the data explained by the model. Should the model perform poorly, I will craft a stacking regressor containing a decision tree and a random forest regressor with gradient boosting to identify which model under which hyperparameters performs the best by attempting to maximize the R^2 value and minimize the RMSE.

Ethical Considerations

Ethics are always a concern when dealing with data; this research is no different. A slippery slope that is apparent in the implications of this study is that while the most nutrient-dense foods can be easily fleshed out of the data that has been found, not every individual should flock to the supermarkets in search of these foods to overconsume them. Dietary needs and restrictions are created based on the principle of a well-balanced diet, and while there are a few foods that are considered healthy to eat all the time, there are always underlying effects of eating too much of a single food. An example of this that comes to mind is the Brazil nut. While very healthy for many individuals, there are some who have tree nut allergies and are unable to eat them without serious medical consequences like anaphylaxis, hives, rashes, and other side effects. Everyone else who can consume the Brazil nut safely must do so in moderation and intermittently, as almost daily intake of a handful of these nuts can cause Selenium poisoning (again, too much of a good thing can be a bad thing).

This research is working towards the identification of nutrient-dense foods provided their dietary makeup across a standardized measurement (in the data's case one hundred grams) for nutritionists and dieticians to incorporate them into recommended food regimens for different groups of people needing different macro and micronutrients. The results of the study are to be viewed as data-backed recommendations based on food guidelines already established by the FDA and the USDA. If studies such as this were to be viewed as ironclad commands to change the way we consume food, junk foods such as potato chips and candies, and processed foods like shelf-stable desserts would be unavailable for purchase as they are mostly nutrient-insufficient. The USDA and FDA cannot force people to eat only nutrient-dense foods and cannot bar food manufacturing companies from producing, distributing, and selling nutrient-insufficient foods, so for the results of this study to be considered anything other than an avenue for dietary recommendation systems is rather questionable regarding data ethics.

Challenges/Issues

The analysis of the food nutrition data I have accessed may prove a challenge as there are many conclusions that may come from this study that need to be firmly established as recommendations as opposed to solid direction. The implications of the ordinary least squares model need to be properly shared so that the reader understands the intent to find the most impactful variables to nutrient density, while the graphical analysis of the data is simply to show foods that contain the highest amount of any particular macro or micronutrient, without telling the reader to consume these foods to change their dietary habits or focus on the most impactful factors to nutrient density to alter their diet. The story of the data must be navigated carefully, as many potential consequences could come from the lackadaisical communication of the analysis's outcome.

References

In addition to the Kaggle datasets derived from the Comprehensive Nutritional Food Database, other article sources have been found to fortify the study's methodology and overall meaning after the model evaluation stage. ScienceDirect's The American Journal of Clinical Nutrition has an issue that outlines the specifics of nutrient density and how it can be evaluated. A nutrition review within The National Library of Medicine has also been found to emphasize the recommendations made by the USDA and share the evolution of the Dietary Guide for Americans to include mention of nutrient-dense foods. The article links are below:

1. https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset
2. https://www.sciencedirect.com/science/article/pii/S0002916523050748
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6489166/