A Predictive Analysis of Food Nutrient Density

David Berberena

Bellevue University

DSC 680 Applied Data Science

Amirfarrokh Iranitalab

September 22, 2024

A Predictive Analysis of Food Nutrient Density Draft Paper

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. A nutrient-dense food can be defined as one that supplies relatively more nutrients relative to their caloric value (Drewnowski & Fulgoni, 2014). 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.

Background/History

Food has existed on the planet longer than humans have, and with the introduction and subsequent progression of technology throughout time, food has been analyzed in many different ways. From dieticians and nutritionists to chefs and food critics, food is studied for various reasons. Recently, America has been pursuing food for enhanced dieting, weight loss regimens, and increased knowledge. Our country is known as one of the most obese countries in the world, and for those who wish to educate themselves on how they can avoid being part of this statistic, one of the most common things individuals do is to count the calories they ingest. This is done in an attempt to stave off being overweight or obese, yet there are many other nutrients food provides that people are starting to pay attention to. Different types of diets like the keto diet are designed to avoid certain nutrients altogether in an effort to lose weight quickly, yet other diets are meant to maximize other nutrients to gain muscle or energy for competition. This study can directly guide individuals who are looking to incorporate or exclude certain nutrients by showcasing which nutrients are most important in predicting nutrition density and which foods can be rotated into one's diet to maximize any given nutrient intake. It also provides nutrition professionals with information that helps for them to recommend more nutrient-dense foods to their clients and the general public so as to contribute to the obesity epidemic in the country positively.

Data Explanation

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 (Dey, 2024). I will be utilizing five datasets, all of which contain nutritional values of

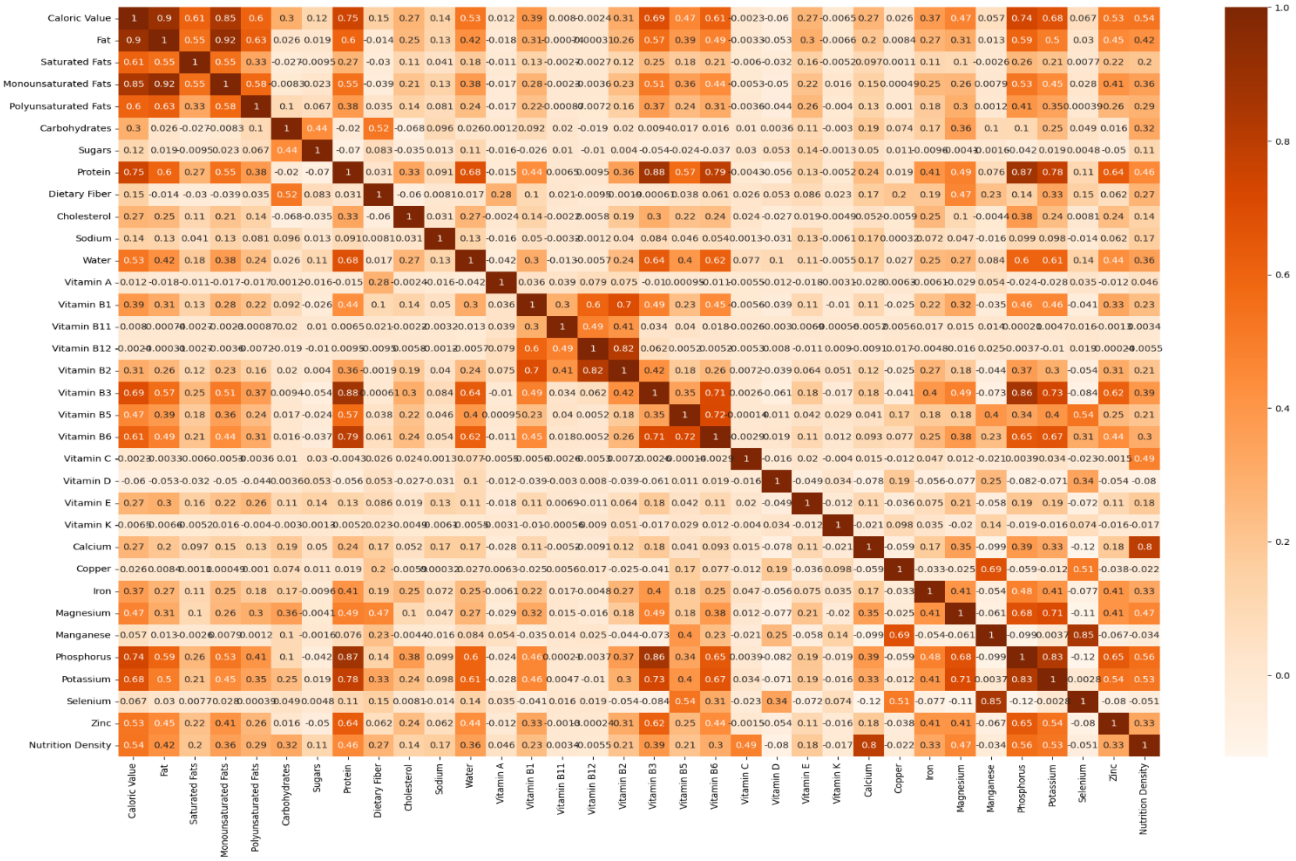
food such as food ingredient name, calorie count per one hundred grams, carbohydrates per one hundred grams (measured in grams), protein per one hundred grams, fats, vitamins measured in milligrams, minerals measured in milligrams, and nutrient density. Initial analysis of the business problem and the data prompts some preparation of the datasets. Since all five datasets share the same variables, I have joined them in a Jupyter Notebook via Pandas's concatenation function. Once this was completed, I had very minimal transformations left to perform on the combined dataset. I only needed to remove irrelevant columns, change the data types of the variable observations, and capitalize the food ingredients for better readability. Now the dataset is ready for the exploratory data analysis phase.

Questions that I believe end users would be interested in knowing about the study are vast, but there are a few key ideas that come to mind that may benefit the general public the most:

1. What food ingredient has the most calories?
2. What food is the most nutrient-dense?
3. Which nutrients are highly correlated with nutrition density?
4. What nutrition density measure do most food ingredients fall between?
5. What are the top ten foods rich in protein?
6. Are any vitamins and minerals impactful to nutrition density?
7. How important is calcium to overall nutrition density?
8. How can we use this study's results in my everyday life?
9. Do we have to eat the foods that are found to be the most nutrient-dense?
10. Who are the results of this study directly impacting?

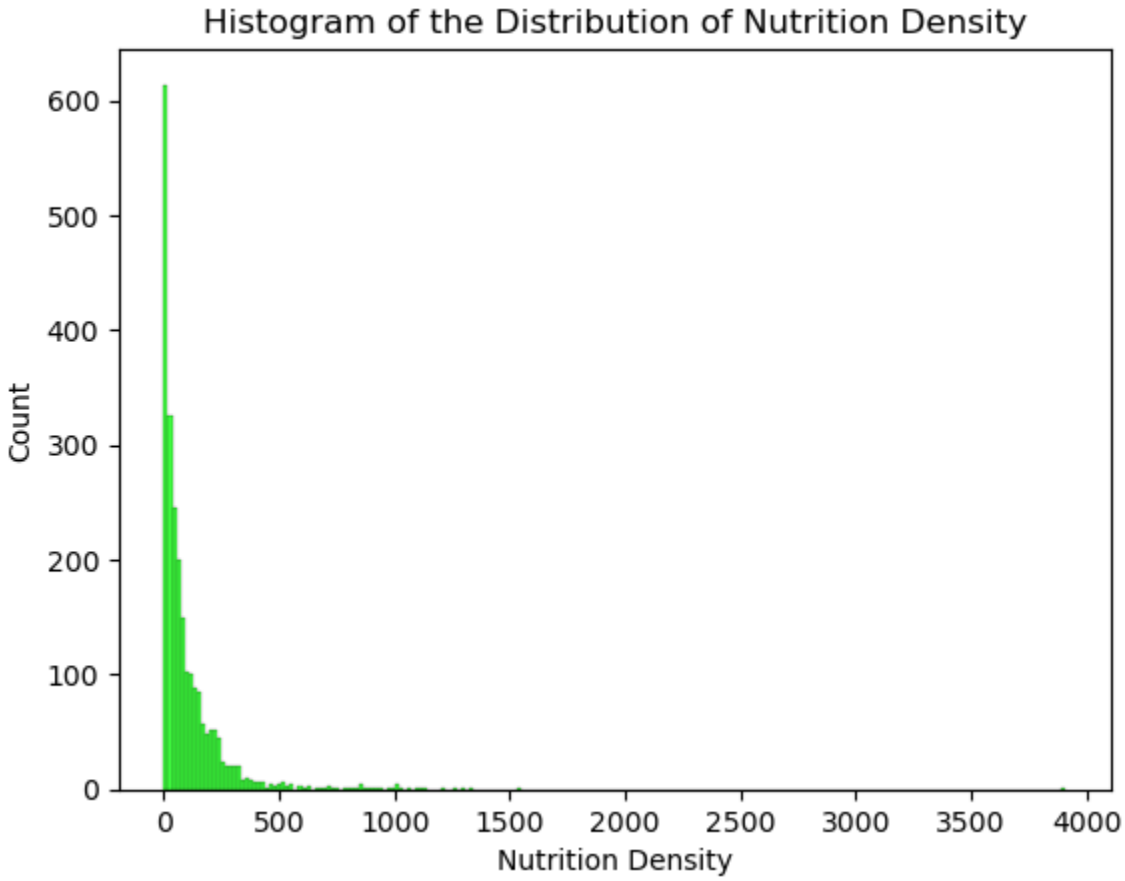
Methods

To begin the EDA phase, I wished 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. There were many variables within the dataset, so the heatmap yielded a cumbersome visualization, which is shown below.



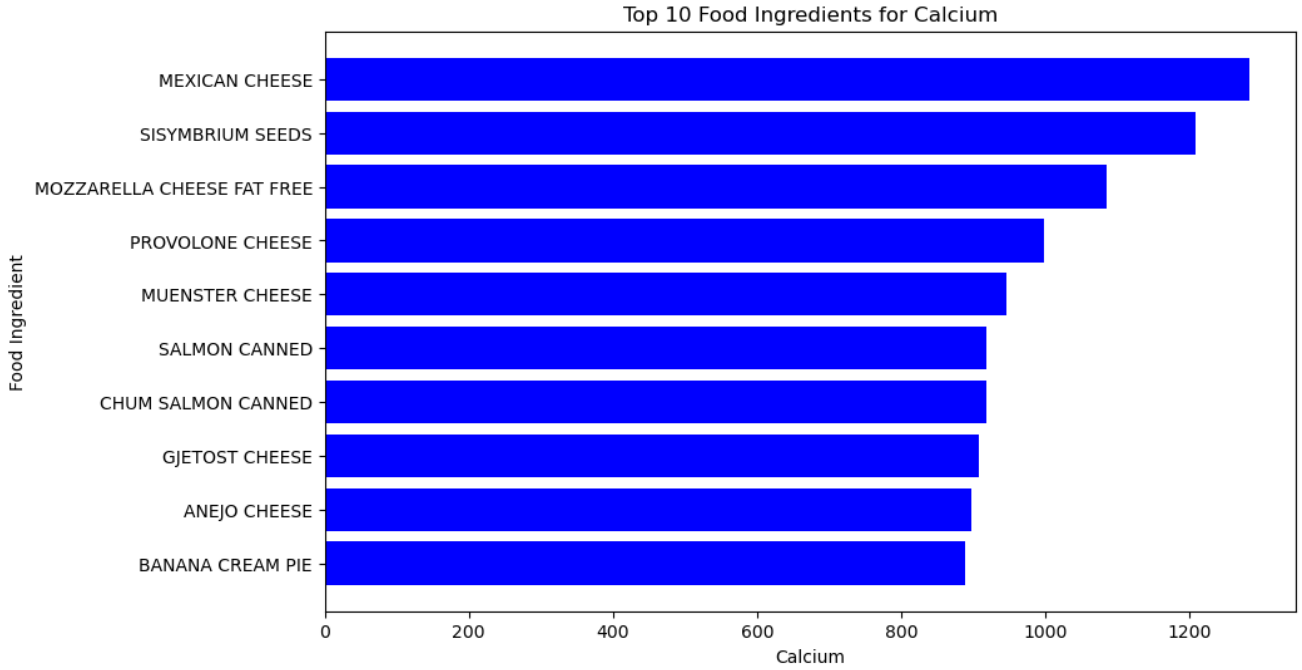
After reviewing the heatmap, the top five most highly positive correlated variables to nutrition density were revealed to be Calcium, Phosphorus, Caloric Value, Potassium, and Vitamin C, with Calcium having the closest correlation coefficient (0.796) to 1.

For each of the variables within the dataset, I wanted to see the data distribution, prompting me to create histograms for each variable. What was interesting to see was that every variable had a negative exponential distribution including the outcome variable of nutrition density. Below is my output for the nutrition density histogram.



This histogram shows that out of the almost 2400 observations, over 95% of the observations have a nutrition density score of 500 or less. This distribution is important to see as exponential distributions lend themselves to regression models.

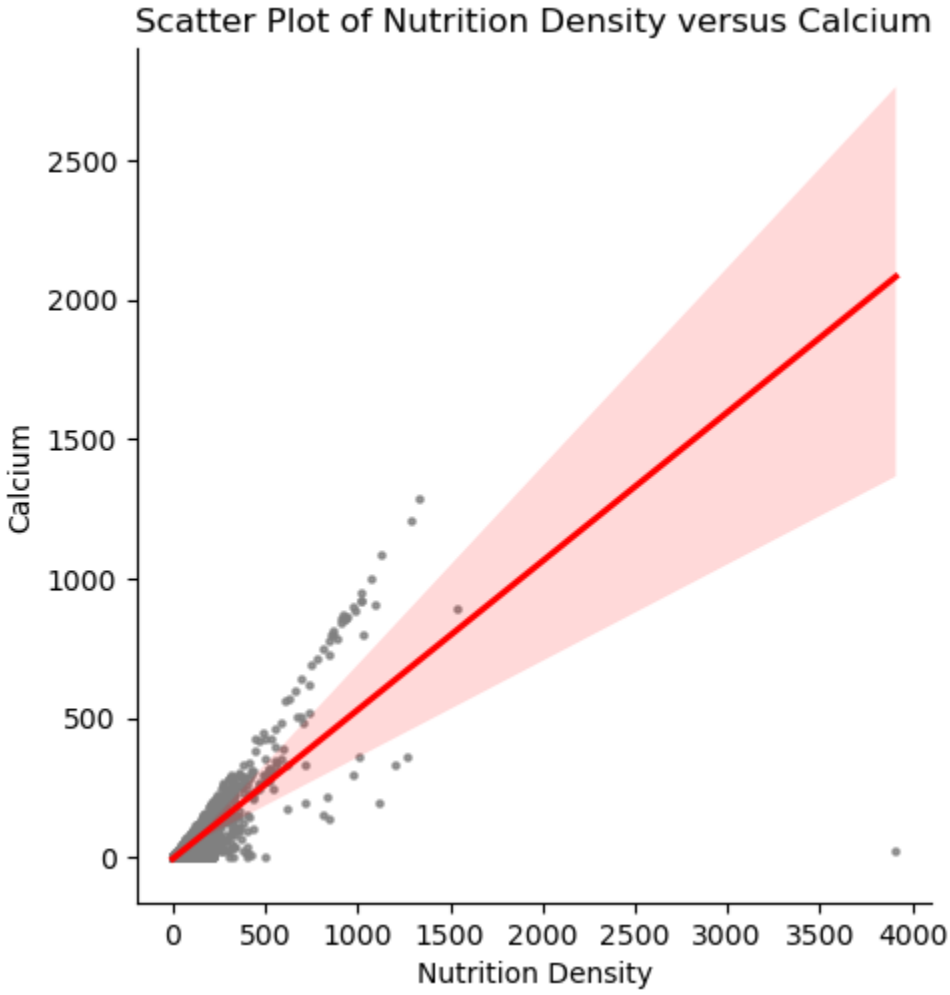
Within my code file, the sheer volume of output had become apparent as the time it took to generate the histograms was very long. To combat this, I extracted the top twenty highest positively correlated variables to continue the EDA process. Here is where I wanted to use bar charts to display the top ten food ingredients for each variable as shown in the data. You can see here how the bar chart came to look for the most positively correlated Calcium variable.



The dataset describes the Calcium variable as being measured in milligrams per 100 grams of the food ingredient, meaning that the most calcium-rich food in the dataset (Mexican cheese) contains over 1200 milligrams of calcium per 100 grams.

The final element of exploratory data analysis that I wished to pursue was the relationship between the top twenty highest positively correlated variables and the nutrition density outcome variable as seen on a scatter plot. I made a scatter plot for each of the twenty

variables with a trend line to help visualize the observations of each variable in relation to nutrition density values. Below is the scatter plot for the Calcium variable.



It is clear from this visualization that the relationship and many of the other scatter plots that were created that a linear relationship exists, yet we can see from the large cluster of data points close to 0 on both axes that the exponential distribution we saw earlier in the EDA phase is apparent here in the scatter plot as well. The scatter plot also does a good job of showcasing any outlier data points, which we can see clearly in the above plot that one does exist, yet many more exist across the other scatter plots that were generated.

Model Creation and Analysis

Now that the exploratory data analysis and other transformations made to the data have been completed, I have split the dataset into training and test sets at an 80/20 ratio to craft 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.

I have created an OLS wrapper function that creates a pipeline to standardize the data using a Standard Scaler, run the ordinary least squares model, and utilize a grid search that cross-validates and normalizes the model using both k-fold cross-validation and Ridge regression (L2 Normalization) to output the most accurate regression model. The metrics that have been 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 provides the 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. Here is the output of one of the OLS cross-validation folds and the best estimator model with its hyperparameters found below.

OLS Regression Results

Dep. Variable:	Nutrition Density	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	7.959e+08
Date:	Fri, 13 Sep 2024	Prob (F-statistic):	0.00
Time:	04:22:28	Log-Likelihood:	2930.5
No. Observations:	1533	AIC:	-5793.
Df Residuals:	1499	BIC:	-5612.
Df Model:	33		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	105.1559	0.001	1.14e+05	0.000	105.154	105.158
x1	-0.0004	0.012	-0.033	0.973	-0.024	0.023
x2	29.1762	0.006	4726.062	0.000	29.164	29.188
x3	0.0002	0.003	0.079	0.937	-0.006	0.006
x4	0.0008	0.003	0.285	0.776	-0.005	0.006
x5	-0.0006	0.001	-0.483	0.629	-0.003	0.002
x6	27.5388	0.004	7454.634	0.000	27.532	27.546
x7	0.0008	0.001	0.710	0.478	-0.001	0.003
x8	34.0200	0.005	6772.480	0.000	34.010	34.030
x9	5.4590	0.002	3478.389	0.000	5.456	5.462
x10	0.0002	0.001	0.208	0.835	-0.002	0.002
x11	-0.0002	0.001	-0.217	0.828	-0.002	0.002
x12	0.0011	0.002	0.701	0.483	-0.002	0.004
x13	12.6743	0.001	1.08e+04	0.000	12.672	12.677
x14	-0.0008	0.002	-0.480	0.631	-0.004	0.003
x15	-9.269e-05	0.001	-0.087	0.931	-0.002	0.002
x16	-0.0007	0.004	-0.158	0.875	-0.009	0.008
x17	0.0012	0.004	0.274	0.784	-0.008	0.010
x18	0.0021	0.003	0.716	0.474	-0.004	0.008
x19	0.0009	0.002	0.535	0.593	-0.003	0.004
x20	8.83e-05	0.002	0.038	0.969	-0.004	0.005
x21	26.4332	0.001	2.67e+04	0.000	26.431	26.435
x22	0.0004	0.001	0.347	0.728	-0.002	0.002
x23	-0.0002	0.001	-0.178	0.859	-0.002	0.002
x24	-0.0022	0.001	-1.846	0.065	-0.005	0.000
x25	111.6601	0.001	9.88e+04	0.000	111.658	111.662
x26	0.0016	0.002	0.968	0.333	-0.002	0.005
x27	4.3115	0.001	3680.297	0.000	4.309	4.314
x28	0.0014	0.002	0.780	0.435	-0.002	0.005
x29	0.0007	0.003	0.293	0.770	-0.004	0.006
x30	-0.0032	0.004	-0.917	0.359	-0.010	0.004
x31	-0.0012	0.002	-0.544	0.586	-0.006	0.003
x32	-0.0013	0.002	-0.542	0.588	-0.006	0.003
x33	-0.0004	0.002	-0.194	0.846	-0.004	0.003

Omnibus:	142.247	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	874.648
Skew:	0.147	Prob(JB):	1.18e-190
Kurtosis:	6.689	Cond. No.	47.4

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Best estimator: Pipeline(steps=[('scale', StandardScaler()),
('model', Ridge(alpha=0.0029470517025518097, max_iter=10000))])
R-squared: 0.999999961641675
RMSE: 0.032870538744169224

Based on this output of the OLS model, we can see that it is overfitted by the perfect R-squared value of 1.000 and the equally perfect Prob (F-Statistic) metric of 0.000. The best OLS estimator's performance metrics are also indicative of overfitting as R-squared is almost 1.000 and the RMSE is almost zero, meaning that the model's predictions are almost exactly the same as the actual observations. This fact rules out OLS as a viable model choice, which takes me to

the stacking regressor method I have chosen as a backup consisting of a decision tree regressor, a random forest regressor, and a gradient boosting regressor.

The stacking regressor was made using a similar function to the OLS model function, with the same cross-validation technique and the best estimator declaration. Each model was generated on their own and was also incorporated into the stacking regressor model, allowing me to see how a decision tree model or a random forest regressor performed independently of the other models. The output for this model generation is displayed below.

```
Best performing base estimators and their performance metrics:

Decision Tree Best Estimator:

{'ccp_alpha': 0.0, 'criterion': 'squared_error', 'max_depth': None, 'max_features': None, 'max_leaf_nodes': None, 'min_impu
rity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'random_state': 123,
'splitter': 'best'}

Performance Metrics for Decision Tree:

     RMSE: 41.49383966411136
     R-squared: 0.9388757928138893

Feature importances:

     Calcium: 0.5451
     Vitamin C: 0.2703
     Caloric Value: 0.1070
     Phosphorus: 0.0321
     Dietary Fiber: 0.0120

Random Forest Best Estimator:

{'bootstrap': True, 'ccp_alpha': 0.0, 'criterion': 'squared_error', 'max_depth': None, 'max_features': 1.0, 'max_leaf_node
s': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fra
ction_leaf': 0.0, 'n_estimators': 200, 'n_jobs': None, 'oob_score': False, 'random_state': 123, 'verbose': 0, 'warm_start':
False}

Performance Metrics for Random Forest:

     RMSE: 34.89358787335249
     R-squared: 0.9567747767624285

Feature importances:

     Calcium: 0.5793
     Vitamin C: 0.2414
     Caloric Value: 0.0850
     Saturated Fats: 0.0178
     Carbohydrates: 0.0106

Gradient Boosting Best Estimator:

{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'squared_error',
'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_sam
ples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_iter_no_change': None, 'random_state': 123, 'subsa
mple': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}

Performance Metrics for Gradient Boosting:

     RMSE: 27.04256135158072
     R-squared: 0.9740377912504109

Feature importances:

     Calcium: 0.5559
     Vitamin C: 0.2746
     Caloric Value: 0.0980
     Phosphorus: 0.0160
     Potassium: 0.0068

Stacking regressor performance metrics:

     RMSE: 24.650970805242217
     R-squared: 0.978426826070024
```

The stacking regressor proved much more viable in modeling the food nutrition data, with each model within the stacking regressor performing better than the previous OLS model choice. With the RSME values rather low in regards to the nutrition density variable's observations and the R-squared values explaining a large percentage of variance within the data as shown by their close proximity to 1, the stacking regressor is the best model choice out of the five models crafted (OLS with Ridge normalization, decision tree, random forest, gradient boosting, snd stacking regressor). This can be confirmed with the very high R-squared value of 0.978 and the low RMSE value of 24.65 as it relates to the nutrition density variable. Looking at the factors that were considered important features to the creation of these models and the outcome variable of nutrition density allows the definition of three consistent variables across the decision tree, random forest, and gradient boosting base models: Calcium, Vitamin C, and Caloric Value. Checking back with the heatmap created in the EDA stage confirmed that some of these important features also had the highest positively correlated coefficients, allowing dieticians and nutritionists to begin recommending foods high in those vitamins, minerals, and nutrients.

Recommendations/Implementation Plan

With the model performance results in, we can now identify actionable insights for end users to act on should they wish to do so. As the implications of this study are more personal than business, the general population can benefit almost immediately from the information gleaned from the EDA and model generation stages of the study. The recommendations that will be outlined are from an educational standpoint based on the limited data procured for this

academic study and are not to be confused with concrete directives. Individual food diets are unique to each person and the health benefits they gain from their regimen are completely up to them. The major recommendations that can be shared from this study are found below.

1. Calcium, Phosphorus, and Caloric Value are the three most highly positively correlated values to nutrition density, so consuming foods high in these nutrients may allow individuals to increase the nutrition density they receive from their food diets.
2. Calcium, Vitamin C, and Caloric Value are the three most impactful nutritional values to nutrition density according to the stacking regressor ensemble model and the individual models that make up the ensemble model, so these macro and micronutrients are to be considered when maximizing or minimizing nutrition density.
3. Eating only the most nutrient-dense food ingredients shown in the data is not considered maintaining a balanced diet since the number of food ingredients that are above a given nutrition density threshold (say a score of 500) decreases drastically (by about 95% for the 500 nutrition density threshold) based on the negative exponential distribution of the data.

Going beyond the immediate results gleaned from the study, having identified a viable model for deployment opens up the opportunity for the model to be trained on more accurate data from various other credible sources such as the Food and Drug Administration and the United States Department of Agriculture. Now before anything like that would happen, reporting the results obtained and alerting local nutritionists and dieticians to its implications would be crucial so experienced professionals could weigh in on any adjustments that need to be made before they could start incorporating the predictive powers of the stacking regressor

that has been created into their work. Recommendations could then be gleaned from the deployed model by these nutrition experts to enhance their dietary recommendations to their clients, allowing the above recommendations to be implemented in the real world.

Assumptions

The OLS model that was created for this research looked to be a viable model choice based on the intended information to be gleaned from the model, yet the assumptions of the linear regression model did not line up with the distribution of the data. The nutrition density outcome variable as shown above in the histogram displayed an exponential distribution, which goes against the linear regression model's assumption that the data is linearly related. Before the final OLS model output, I tried to log-transform the outcome variable to fit into the linear assumptions the OLS model perpetuates, yet even while adding a small constant to prevent the log-transformation of non-positive values, the OLS model output showcased many NaN values and the performance metrics all indicated severe underfitting of the data when exponentiating the predictions.

Another assumption that was made that could be addressed in the future is the presence and impact of outlier data points. These outliers as seen in the scatter plots crafted in the EDA phase are most likely the result of a combination of erroneous data entry and food ingredients that exhibit high nutritional content of only a few nutrients. While the data still lends itself to an exponential distribution, removing these outliers could potentially normalize the data to where the log-transformed nutrition density variable would have been properly modeled by the OLS model. Superfoods like almonds, kale, salmon, and others that contain high

amounts of specific nutrients are definitely natural outliers among the many food ingredients, yet the data showcased many more outlier data points that bring the dataset's validity into question.

Challenges/Limitations

The analysis of the food nutrition data I have accessed has proven 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. Both the USDA and the Department of Health and Social Services (DHHS) are in charge of publishing the Dietary Guidelines for Americans (DGA), which has been established as a source of advice and not coercion for Americans to choose a healthier way to eat. The DGA that was published in 2015 specifically advises people to "Focus on variety, nutrient density, and amount" and "Follow a healthy eating pattern across the lifespan" (Drewnowski et al., 2019). The implications of the stacking regressor 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.

A limitation made evident in my research is the computational effort made to properly visualize and model the food nutrition data. While the study was successful in crafting a viable

model, the time and memory used to do so placed a strain on my local device. Studies such as these involving thousands of data points across numerous variables and the plotting of each variable multiple ways calls for big data technologies such as the Hadoop infrastructure. Hadoop Distributed File System (HDFS) and Apache Hive come to mind regarding software that can handle big data tasks such as parsing through large datasets like the one used in this study to quickly identify trends, patterns, and specific values while storing the data in other directories not directly tied to the memory and storage of the local device.

Ethical Assessment

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).

Another issue of ethical concern is that while this data comes from a reputable source, there are still many erroneous or inflated values within the data that cannot and should not be viewed as true, especially as this study is being done for educational gain and not real-world application. While the methodology may be somewhat synonymous with professional research, the data acquired should not be taken too seriously as some food ingredients and nutrients may be over or understated. 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. Organizations like 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 in our capitalist country, 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.

Future Uses/Additional Applications

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 of a food ingredient) for nutritionists and dieticians to incorporate them into recommended food regimens for different groups of people needing different macro and micronutrients. With more accurate data and enhanced predictive modeling techniques, 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. Should the results of similar research

be published for the general public, people can use the information to consciously change their dietary habits however they see fit. Identifying nutrient-dense foods and impactful nutrients to nutrition density combined with smartphone health applications currently in use today like WeightWatchers and Yuka gives the end user the power to harness their dietary needs in the palm of their own hands.

Conclusion

After addressing the many implications of the results found in the research performed, it can be said that nutrition density is a measure that takes all micro and macronutrients into account. A few of these nutrients were more impactful than others, yet with more accurate data we can see the stacking regressor model train and learn even more about the factors that are most significant to nutrition density. While the current results can answer the data-related questions posed above, the application this study has to dieticians, nutritionists, food scientists, and food organizations like the FDA and the USDA is of significant importance and impact for the general public along with companies concerned with food nutrition.

Appendix/References

1. Drewnowski, A., & Fulgoni, V. L. (2014). Nutrient density: Principles and evaluation tools. *The American Journal of Clinical Nutrition*, *99*(5), 1223S-1228S. https://doi.org/10.3945/ajcn.113.073395.
2. Drewnowski, A., Dwyer, J., King, J. C., & Weaver, C. M. (2019). A proposed nutrient density score that includes food groups and nutrients to better align with dietary guidance. *Nutrition Reviews*, *77*(6), 404-416. https://doi.org/10.1093/nutrit/nuz002
3. Utsav Dey. (2024). Food Nutrition Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8820139