"Predicting Food Groups from Nutritional Profiles using Machine Learning Classification Models"

Deborah L. Young

DSC 680: Applied Data Science

Business Problem

This project addresses the need for an automated system to categorize foods into their appropriate food groups based on nutritional content. The existing manual process is not only time-intensive but also susceptible to human error. Automating this process could substantially enhance efficiency, accuracy, and be beneficial in numerous applications such as inventory management, recipe curation, and nutrition education initiatives.

Background/History

Drawing on the MyFoodData database, which is informed by the USDA's FoodData Central, this project's cornerstone is to train machine learning models capable of predicting food groups from their nutritional profiles. Historically, the task of food categorization has been manual, subject to errors and inconsistencies. With the advent of machine learning, there's a potential to streamline this process by training models to predict food groups from extensive nutritional profiles, thus revolutionizing inventory management and educational approaches. This project seeks to leverage the advances in machine learning to automate and refine this process.

Data Explanation

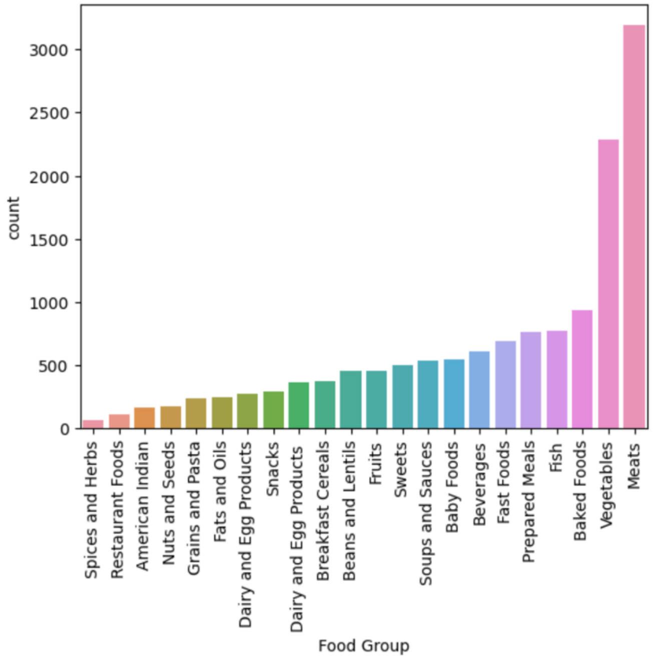
The dataset includes various features, such as macronutrient content (carbohydrates, fats, proteins), vitamins, minerals, and other nutritional values. However, it appears that there are several variables that are alternate names for the same thing, and several for different serving sizes, which might need consolidation or removal for better analysis and modeling.

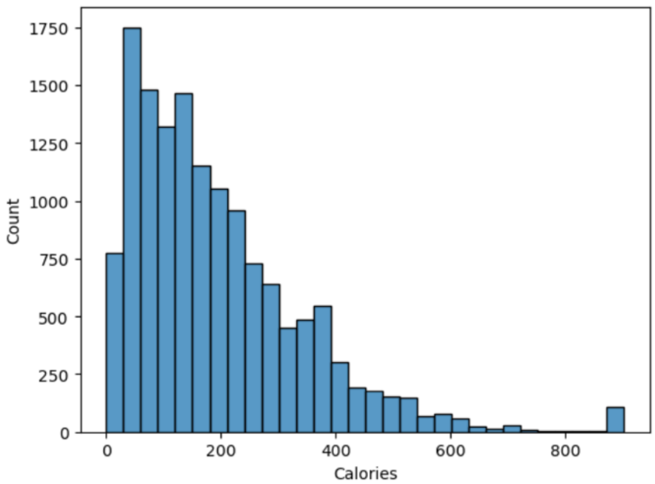
Methods

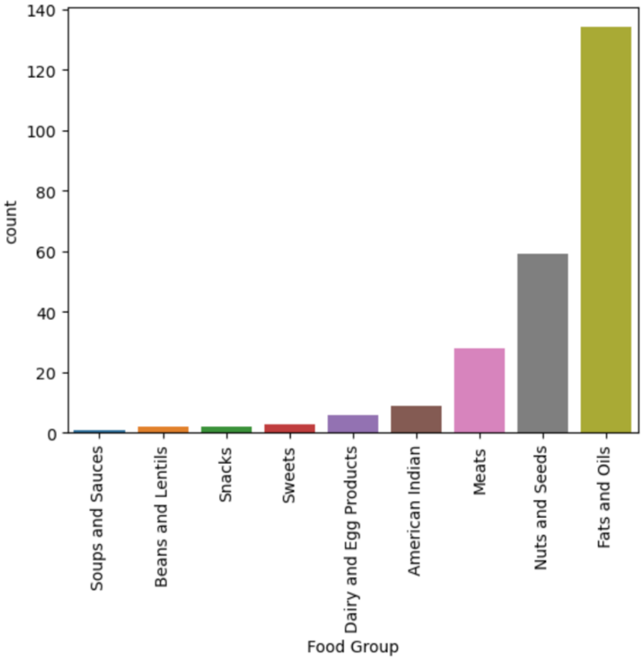
The methodological framework of the project involved the following stages:

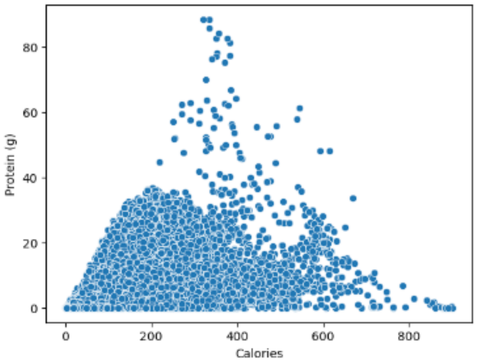
Exploratory Data Analysis (EDA): The goal of this is to assess data quality, visualize distributions, and clean data. The dataset consists of comprehensive nutritional facts for various food items. The descriptive statistics suggest that the dataset covers 14,164 items with an average of 194 calories per serving, with the caloric content given in weight (grams). The data analysis included univariate analysis to examine the distribution of variables like calorie content and food groups, as well as bivariate analysis to explore relationships between multiple variables, such as the proportion of macronutrients against caloric value and food group classifications.

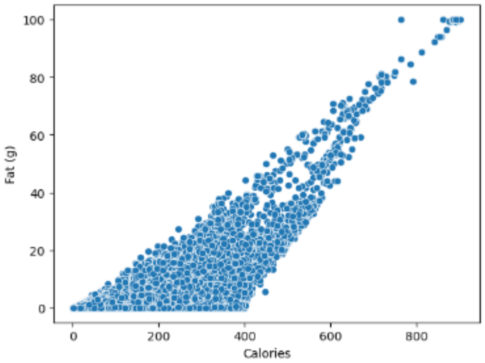
The data underwent thorough cleaning, which included removing features with excessive missing values. A detailed exploration and cleaning process were conducted on the dataset, with variables with over 80% missing values were removed. Components with trace amounts were converted from NA values to 0 based on subject matter indicating this was appropriate because a null value in a nutrient category represents that they are not present or are present in trace amounts.

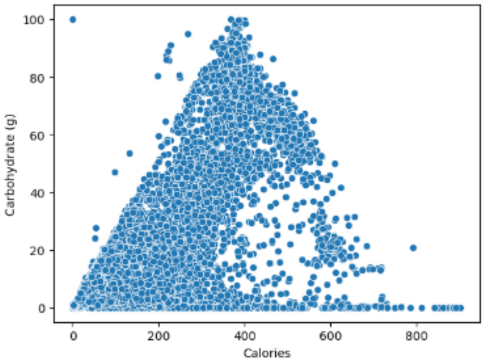





count
0
20
40
60
80
100
120
140
Soups and Sauces
Beans and Lentils
Snacks
Sweets
Dairy and Egg Products
American Indian
Meats
Nuts and Seeds
Fats and Oils
Food Group

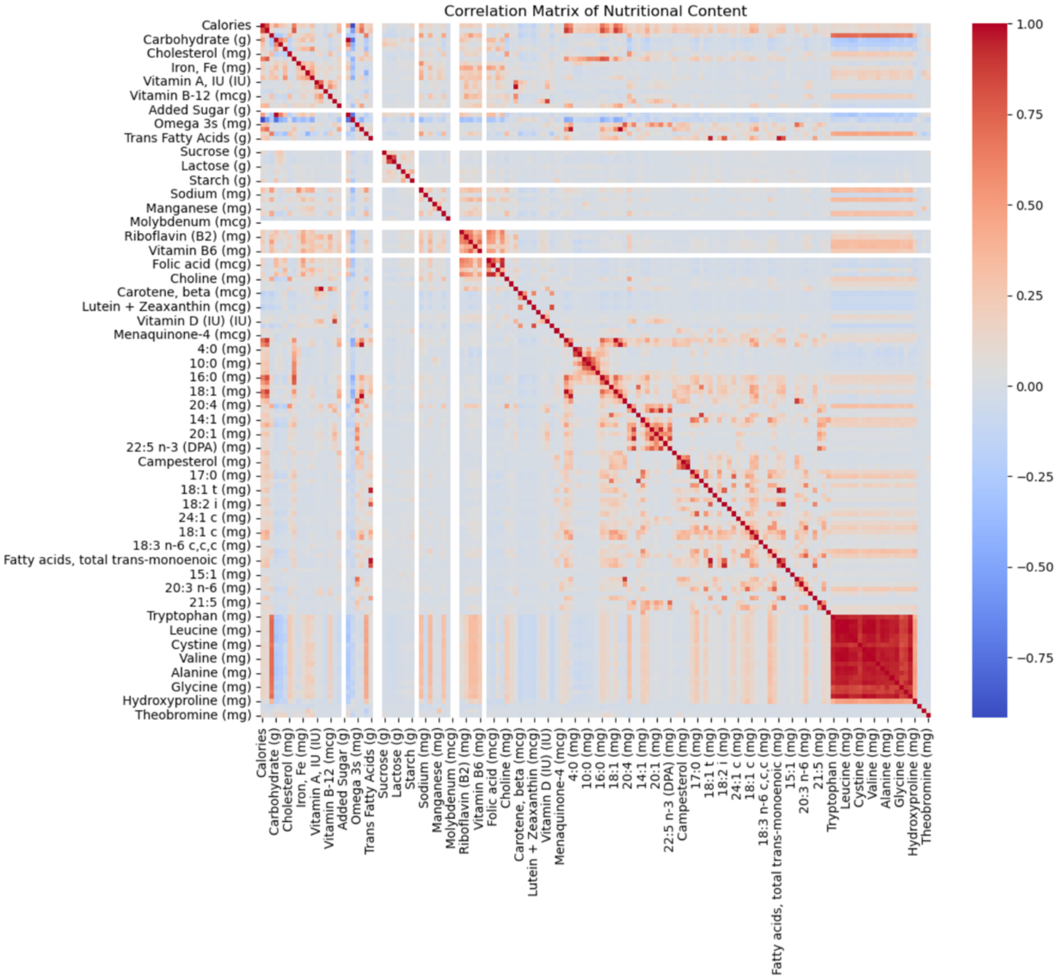

Protein (g)
0
20
40
60
80
0
200
400
600
800
Calories


Fat (g)
0
20
40
60
80
100
0
200
400
600
800
Calories


Carbohydrate (g)
0
20
40
60
80
100
0
200
400
600
800
Calories

Feature Selection/Dimensionality Reduction: For this imperative process, correlation matrices were utilized as well as other techniques to identify the most relevant features for the classification task. As the correlation matrix is challenging to decipher, this was accomplished by precising coding that printed extensive lists of preferred variables and removed based on their prevalence in the dataset. Where counts were nearly equal, null values in original dataset were referenced for preference.

Low variance features were eliminated due to their limited contribution potential to the predictive models. The remaining dataset was carefully prepared for the machine learning models to ensure accurate classification outcomes. Features with a high degree of correlation but low total counts were also excluded.



Modeling: Overall, the dataset is prepared by removing irrelevant features and converting null values to zeros where appropriate, setting the stage for further modeling tasks. The data was then ready to be encoded or scaled/normalized as needed for machine learning models, which will be used to predict the "Food Group" based on the nutritional profiles of the food items.

The models designated for this classification task were supervised learning models - Random Forest and Gradient Boosting. The data were divided 60/20/20 for training, testing, and validation. Techniques like SMOTE and RandomUnderSampler were utilized to address and mitigate issues arising from class imbalance within the data.

Evaluation: Models were evaluated using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness. The model performance metrics show the effectiveness of each model in classifying the food groups based on their nutritional content. Here's a summary of what these metrics indicate:

- Accuracy is the overall correctness of the model across all classes, showing how often the model correctly predicts the food group.
- Precision for each class indicates how many of the items predicted as belonging to that class actually do, highlighting the model's ability to avoid false positives.
- Recall (or sensitivity) shows how many of the actual items of each class were correctly identified, emphasizing the model's capability to find all relevant cases.
- F1-score is the harmonic mean of precision and recall, providing a balance between them for each class, useful when there's an uneven class distribution.

Analysis

Random Forest and Gradient Boosting models were used in this project. The performance of each model was rigorously evaluated using a variety of metrics. The results illustrate how the models perform differently across various food groups, with some groups being easier to predict than others. Models like RandomForest and GradientBoosting show high overall accuracy but have varying precision and recall across different food groups, indicating that some groups are more challenging to classify correctly due to similarities in nutritional content or imbalances in the data. The use of SMOTE and RandomUnderSampler techniques to address class imbalance affects precision, recall, and F1-scores, generally improving recall at the expense of precision, which is evident in the slight changes in these metrics between the original and resampled datasets.

The analysis also uncovered notable correlations between specific nutrients and food groups, which informed the feature selection process. The analysis highlighted the uneven distribution of caloric content and the challenge of high-calorie foods in categorization. Bivariate analysis explored macronutrient proportions against caloric values. The RandomForest with balanced class weights and oversampling of the minority class (using SMOTE) exhibited strong classification performance. This is evident from the accuracy and the weighted average scores for precision, recall, and f1-score.

The RandomForest model with these settings achieved:

- An accuracy of 0.90
- A macro average precision of 0.86
- A macro average recall of 0.84
- A macro average f1-score of 0.85
- A weighted average precision of 0.90
- A weighted average recall of 0.90
- A weighted average f1-score of 0.90

These results are better than those achieved by the RandomForest without oversampling, the GradientBoostingClassifier on the original data, and the GradientBoostingClassifier with oversampled minority class, and also outperformed the models with undersampling of the majority class.

Conclusion

The project concludes that machine learning models are highly effective in categorizing foods based on nutritional content. Random Forest, in conjunction with appropriate dimensionality reduction and data resampling techniques, exhibited superior performance and holds promise for real-world application. This model was not only able to correctly classify the food items with high accuracy but also maintained a good balance between precision and recall across different classes, which suggests that it was not only predicting the most frequent classes correctly but was also effective across the less frequent ones. The success of the model suggests significant improvements over manual categorization methods.

Assumptions and Limitations

The project assumes the accuracy and comprehensiveness of the MyFoodData dataset. A notable limitation includes the challenge of handling high-dimensional data and ensuring that the models generalize well to unseen data.

Challenges

Key challenges encompassed managing the high dimensionality of the dataset, computational demands, and ensuring the robustness and fairness of the models across diverse food groups.

Future Uses/Additional Applications

The potential applications for these models are vast and include improving inventory management systems, aiding in diet planning, and enhancing nutritional education programs.

Recommendations and Implementation Plan

It is recommended to refine the model through enhanced feature selection techniques and to explore additional machine learning algorithms. The implementation plan suggests a phased integration of the model into real-world systems, beginning with a pilot program to evaluate practical applicability.

Ethical Assessment

In the realm of machine learning projects, ethical considerations are paramount. They encompass ensuring fairness and transparency, safeguarding against biases in data and algorithms, maintaining privacy and security of data, and adhering to regulatory and ethical standards. Ensuring that models do not perpetuate or amplify existing social biases and are transparent in their functioning and decision-making processes is essential. An ethical ML project should also provide clear documentation and accountability measures. The aim is to complement human expertise, enhance decision-making, and improve efficiency without infringing on individual rights or reinforcing harmful stereotypes.

The ethical assessment of this project centers on maintaining accuracy and avoiding cultural biases in the dataset. We prioritize transparency and aim to complement, not replace, human expertise in nutrition. Challenges include ensuring the machine learning model does not inadvertently perpetuate biases or misrepresent certain food groups, which could misguide dietary recommendations. The model's deployment will be carefully monitored to align with ethical standards and will be adjusted based on feedback to mitigate any unintended consequences.

References

FoodData Central. (n.d.). Retrieved January 16, 2024, from https://fdc.nal.usda.gov/download-datasets.html

Mahadevan, M. (2022, July 31). Step-by-Step Exploratory Data Analysis (EDA) using Python. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/

Nutrition Facts Database Spreadsheet. (n.d.). Myfooddata. Retrieved January 16, 2024, from https://tools.myfooddata.com/nutrition-facts-database-spreadsheet.php

OpenAI. (2024). ChatGPT [Large language model]. https://chat.openai.com

What is Exploratory Data Analysis ? (2021, July 22). *GeeksforGeeks*. https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

10 Questions an Audience May Have

1. *How does the manual process of food categorization currently work, and what are its drawbacks?*

   The manual process involves individuals assigning food items to categories based on their nutritional profile. It is time-consuming and subject to human error, which automation seeks to address.

2. *How do you handle variables that are essentially duplicates in the dataset?*

   Variables that are alternate names for the same thing or that pertain to different serving sizes are consolidated or removed to streamline the dataset for more effective analysis and modeling.

3. *What machine learning parameters were used in this project?*

   Random Forest and Gradient Boosting models were primarily used, with techniques like SMOTE and RandomUnderSampler to address class imbalance.

4. *What is SMOTE, and why is it used in modeling?*

   SMOTE is a technique for oversampling minority classes in a dataset to create a balanced class distribution, which helps in improving model performance for underrepresented classes.

5. *Can you explain the importance of feature selection in this project?*

   Feature selection is crucial because it helps in identifying the most relevant features that contribute to the predictive power of the models, which is particularly important in high-dimensional datasets.

6. *How do you plan to implement these models in a real-world setting?*

   A phased integration is suggested, starting with a pilot program to assess practical applicability and refining the model based on real-world feedback.

7. *How accurate is the data from MyFoodData, and what assumptions are made about it?*

The project assumes the data is accurate and comprehensive. However, as with any dataset, there is a possibility of errors or omissions.

8. *What potential applications do these models have beyond inventory management?*

   They could be used in diet planning, nutritional education programs, and possibly in health-related research and public health initiatives.

9. *What recommendations do you have for future research or application of these models?*

   Future research should explore additional machine learning algorithms, refine feature selection techniques, and possibly expand the dataset to enhance diversity and representativeness.

10. *How do you ensure the model doesn't replace human expertise in nutrition?*

    The model is designed to support and enhance human decision-making, not replace it. It provides a tool for experts to use in conjunction with their knowledge and judgment.