Food Preferences: Predicting the Favorability of Foods Based on Nutrition

Beomsuk Seo
Data Science
University of California, San Diego
beseo@ucsd.edu

Abstract

It is generally agreed upon that cooking at home is a much more favorable alternative to buying pre-made food. As stated by the National Library of Medicine, “An intuitive style in eating decision-making, for example, basing decisions on one's gut feeling, has been related to a less healthy diet, whereas deliberately deciding what to eat, such as making plans about eating behavior, has been related to a healthier diet”(NIH). I propose that one’s eating preferences are related, or influenced by, the nutritional value of the food one chooses to eat. I will incorporate a machine learning model in order to confirm or deny this proposal, and explore a dataset gathered using listings of recipes and their reviews from food.com,

Introduction

Are we naturally attracted to healthy or unhealthy foods? Many would agree with the former, saying that they always strive to eat healthy and that it fulfills them. On the other hand, some people enjoy the pleasure of consuming an unhealthy meal. Whether it be someone who is seeking to improve their health or someone that is unaware of how to eat properly, all of us can learn more about our own eating habits.

But what defines healthy? Any meal that is low in calories? A ‘superfood’ that hits all of one’s micronutrients? The truth is that everybody has different health goals, and the foods we eat depend heavily on those goals. Nonetheless, there are multiple generalities and consensuses about what rules we should all follow, no matter what goals we may have. For example, one shouldn’t gain or lose weight too quickly, consume too much sugar, and that a home-cooked meal will generally be healthier than one bought at a fast food restaurant.

Many of us look to the internet for recipes like these. Along with the rise of rapid consummatory social media platforms (e.g. Youtube Shorts, Instagram Reels, Tiktok), a wave of “healthy eating”, meal prepping, and fitness took the world by storm. However, it is also common that what could be perceived as ‘healthy’ may not be at all, no matter how popular. In order to figure out whether or not a healthy recipe is a popular one, we look to Food.com, one of, if not the most extensive database of recipes online, along with reviews for these recipes.

Dataset

The data collected from Food.com consists of over two hundred thousand recipes posted online, along with over a million cumulated reviews from two hundred thousand users that have tried out those recipes. This data is split into two datasets, one containing recipe data, and the other containing information about the posted reviews. Within the recipes dataset, the 230,000 entries consist of how long the recipe takes, nutrition facts, ingredients, and so on. The reviews dataset comes with a rating out of five, as well as the review text.[1.1]

user_id	recipe_id	date	rating	review
38094	40893	2003-02-17	4	So simple, so del..

Figure 1.1 A sample entry of the reviews dataset

Interesting findings

The review and recipe datasets both have a few interesting findings. In the reviews dataset, we can see that a majority of the reviews indicate five stars, which can indicate that many only leave a review if they enjoyed that particular recipe.[1.2]

In the recipes dataset, there seem to be extreme outliers, although far and few between. For example, one recipe's data states that it takes 145 steps, and another takes 0.

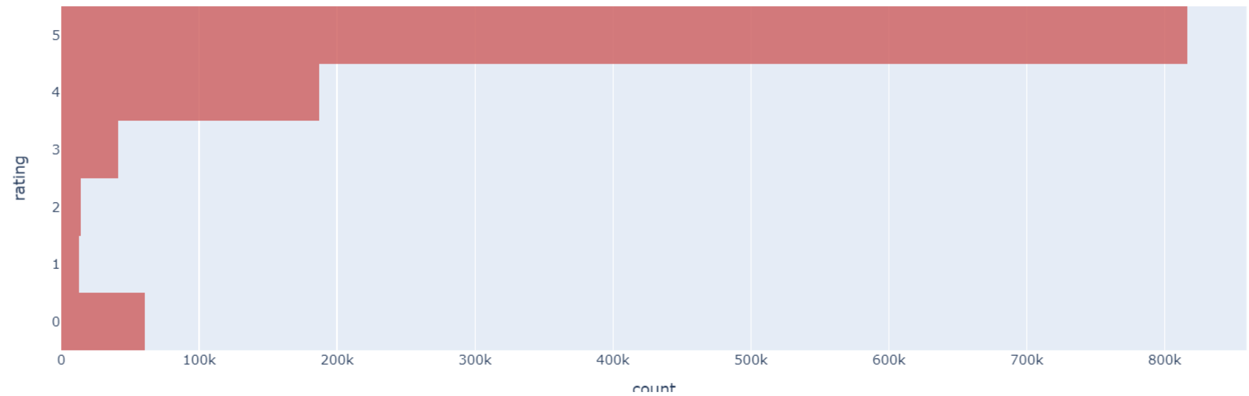

Figure 1.2 A histogram distribution of the 'rating' column in the reviews dataset Notes: most reviewers rate a recipe 5 stars, 1 star rating is least common to be rated.

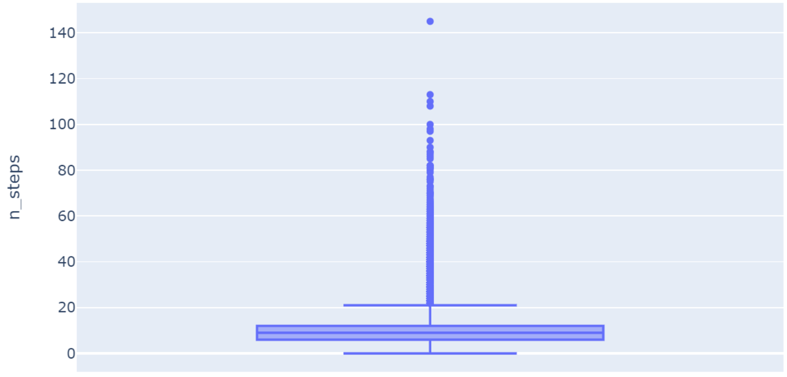

Figure 1.3 Box plot of "n_steps" column

As we can see, most of the data for n_steps lies between 0-20, and anything above 20 can be considered as outliers.

Predictive Task

In order to predict the preferential liking of a recipe based on its nutrition, we will test two models, one incorporating logistic regression and the other using linear regression, the reason being that the value we are trying to predict is the rating a review has based on the recipe, which is strictly categorical. In order to evaluate my model's accuracy, I will create a train-test split using the data, in order to not test the model on the same data that it was trained on. A relevant baseline that can be used for comparison is a dummy model that will always predict the rating of two and a half stars out of five. I will assess the validity of my model's predictions using the mean squared error (MSE), since the rating is on a scale of zero to five and we want to penalize those predictions that are too far off.

The features I will use are based on the nutrition facts of each recipe. The nutrition variable in the recipes dataset is formatted as such: [51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0]. Cross-referencing food.com and the dataset itself helps us figure out that these seemingly arbitrary numbers each represent a different nutrition fact, which I will label accordingly to represent the feature.[2.1] While processing our data, we will "flatten" the nutrition variable so we can use each individual value as a feature, leading us to our model representation.

calories	total fat	sugar	sodium	protein	Saturated fat	carbohydrates
51.5	0.0	13.0	0.0	2.0	0.0	4.0

Figure 2.1 A sample entry of the 'nutrition' column in the recipes dataset (Note: other than the calories value, all the other labeled values represent the percentage daily value)

Model

In the beginning I had planned to use a logistic regressor, using a balanced weight classifier. The strength of this model was that since our rating is a categorical variable, we can easily fit the model to the data we have. However, I constantly ran into issues with its accuracy. The data was not overfitted, indicated by the fact that even when tested on the data the model was trained on, its accuracy remained below 20%. Therefore, I ended up using a linear regression model instead.[3.1] The strength of this model is its simplicity and works fast, especially with fitting the multitude of features. The weakness, on the other hand, is that we would receive continuous results: instead of predicting 4 stars, we would get 4.446, which is a problem because of the lack of continuity in the original data. The accuracy on this model was similar, but I found a different way of measuring the accuracy that would test well on our data. The model accounts for the continuity by using a rounding system for every rating score (e.g. 4.4 = 4.0, 2.6 ≃ 3.0).

$$rating(review, recipe) \simeq \theta_0 + \theta_1 * [calories] + \theta_2 * [fat] + \theta_3 * [sugar] + \theta_4 * [sodium] + \ ...$$

Figure 3.1 Format for our linear regression model

Literature

The dataset used for this research comes from recipes and ratings listed on food.com, collected by Julian McAuley and team. It was used to generate personalized recipes based on historical user preferences, which is cited below. Some similar datasets that have been studied in the past include the Fastfood nutrition dataset and numerous food preferences surveys. Currently scientists conduct studies using exploratory data analysis, along with a multitude of machine learning techniques including regressors, classifiers, and decision trees in order to study this type of data. The conclusions from existing work are similar to my own findings: the nutrition of a food does not affect one's food preferences nor their dietary intake.

Results and Conclusions

Using a logistic regression model unexpectedly performed quite poorly. However, after switching to linear regression and testing accuracy on ranges instead of exact numbers proved to have a much higher accuracy.[5.1] Other than the Mean Squared Error, I assessed the accuracy of the model by testing the predictions in different ranges: whether or not the prediction was correctly between certain ranges (e.g. 4 or 5 stars, 3/4/5 stars, 2/4/5 stars). I believe that the model was more successful than the past models because of the way I tested the validity: with only 5 possible stars, the accuracy expectedly goes up when testing a wider range. This leads me to conclude that the nutrition of a certain food is not likely to affect one's given rating of that recipe. This is drawn from the mean squared error of the improved model, as well as the lack of success with the logistic regression model, which was the more logical model to use for this type of problem.

Mean Squared Error (MSE)	>= 4 stars (4 or 5)	>= 3 stars (3, 4, 5)	>= 2 stars (2, 3, 4, 5)
1.5874415499931862	0.88785589	0.92377	0.935787

Figure 5.1 Model Results (Note: other than the MSE, the values are accuracy ratios out of 1.)

Citations

Generating Personalized Recipes from Historical User Preferences, Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley, EMNLP, 2019

Sproesser G, Aulbach M, Gültzow T, König LM. Do nutrition knowledge, food preferences, and habit strength moderate the association between preference for intuition and deliberation in eating decision-making and dietary intake? Appl Psychol Health Well Being. 2023 Aug;15(3):957-982. doi: 10.1111/aphw.12419. Epub 2022 Dec 7. PMID: 36478397.