

Predicting food choice with machine learning

Working paper

Dawn Liu Holford¹ & Robert Komara

¹University of Essex, dliuxi@essex.ac.uk

This work was supported by a grant from the UK Economic and Social Research Council to DLH (grant reference ES/V011901/1)

Abstract

We investigated if we can accurately predict the different types of choices consumers make given certain food label information. The dataset featured 4620 food choices from 154 participants who chose in each experimental trial the healthiest option out of six options based on the nutritional information given. Using an auto machine learning software (h2o.ai), we trained a set of different algorithms to make binary choice predictions for whether participants would choose a particular type of option in the choice set, for example, if the least calorific option in the set would be selected. Predictive accuracy was generally >90% and the type of food and distinctiveness of the choice options contributed most strongly to predictions. We discuss how combining machine learning predictions with statistical testing could help us understand food choice decisions within its informational context.

Researchers, public health bodies, companies, and even consumers themselves have expended much effort to encourage people to make healthier food choices (e.g., (Wartella et al., 2010b; World Health Organization, 2021)). One way that national authorities have tried to inform consumers about the nutritional value of their choices is by mandating food to carry labels with nutrition information (Storcksdieck genannt Bonsmann et al., 2010; van den Wijngaart, 2002). However, measuring the effectiveness of this information is challenging, partly because determining the healthiest choice is complex and could vary for different individuals and food types (Food Standards Agency, 2008; Guthrie et al., 2015; Scarborough et al., 2007). It is thus informative to understand how food labels could inform different types of choices (e.g., minimising calories vs. minimising fat) and, indeed, what other elements—be they inherent characteristics or external factors—affect the ability of food labels to guide participants towards making the healthiest choice out of the many options they may encounter when buying food.

Modelling food choice and the role of machine learning

Analysing what shapes food choice relies on *modelling*. One begins with an intuition—based on from theory or empirical observation—about the factors that generate differences in what people choose. In a simple model, one might expect that attitudes affect people's propensity to choose healthier choices; that is, if an individual is more positive about healthy eating, that individual should be more likely to choose the healthiest option than an individual who has a less positive attitude. This model can then be tested through the analysis of real data, typically assessing the variation observed in that data and how it can be attributed to the model factors. In our example, a researcher might assess the correlation between attitudes and choice, finding in the data that more positive attitudes are related to more selections of the healthiest choice. She might then assess whether this relationship is *significant*—most commonly operationalised as a less than 5% probability of finding such a

result if a relationship did not exist, or " $p < .05$ ". This type of analysis—null hypothesis significance testing ("NHST")—compares the observed data distribution to a distribution that assumes the model is non-existent (hence the term "null") and is the most prevalent in psychology and the behavioural sciences (Bakker & Wicherts, 2011; Hubbard & Ryan, 2000; Yarkoni & Westfall, 2017).

Increasingly, alternative methods to model behavioural data have been championed. For example, a Bayesian approach begins by encoding prior assumptions in the model (rather than the NHST) that is subsequently updated with incoming data to generate a posterior data distribution (Schönbrodt et al., 2017). Based on this data distribution, one can then quantify the likelihood of a proposed model (e.g., that attitudes and choice are related) to the likelihood of the null model. The Bayesian method therefore offers model comparison to understand which model best fits—and thus explains—the data. Crucially, like with NHST, the focus is on *explaining* the data with the model, so there is a preference for models that minimise noise in the sample.

In theory, one might expect explainability to be the goal of the investigation, since the variables that explain people's behaviour should also predict their future behaviour. However, in practice, the tools in the researcher's toolbox often sacrifice predictability from explainability because statistical analysis is prone to overfitting a model to the existing data (Yarkoni & Westfall, 2017). One can easily test any number of models with many variable combinations on a dataset until one finds the best fit (i.e., the one that reduces noise, or "error"). However, whether that model will then best explain a *new* sample is a different question---and one to which the answer is often "no". Therefore, it is informative to consider an alternative approach to analysis: examining the *predictive ability* of the model, that is, that is, whether the model and its included variables (also known as "features") can accurately predict future outcomes. It is from this angle that we approach our investigation of healthy

food choice. We look at how accurately (and precisely) we can predict, based on known information about the individual and the circumstances of their choices, whether a consumer would make a certain choice.

Prediction is primarily the goal in the field of machine learning. In contrast to the model-fitting approach (minimising variance), machine learning seeks to prioritise the model that minimises prediction error. This is typically done by training and testing the model on different datasets (Yarkoni & Westfall, 2017). This can be done through splitting the collected data into two samples, where the test (or "hold-out") sample approximates out-of-sample data. Of course, splitting the data results in fewer observations for training, which can be problematic without a large dataset. Alternatively, one can "fold" the dataset such that in one train/test cycle, one fold is the training and the other the test set, with the cycle repeated for as many folds as one sets (Yarkoni & Westfall, 2017). Leveraging the techniques and principles of machine learning thus provides a good way to assess whether a model could predict behaviour, rather than just explain it (see (Yarkoni & Westfall, 2017), for a summary of the core concepts of machine learning and how they can improve traditional psychological approaches to analysis). Furthermore, while some have argued that this generates the opposite problem---predictability without explainability (McGovern et al., 2019) this actually depends on the complexity of the model used. A model with fewer variables or a straightforward decision-making process is clearer to interpret than one with many variables and interactions. Moreover, even with complex models, the machine learning literature suggests ways to quantify the contributions of variables in the models, for example by comparing the model predictions with the variable to predictions without it (Lundberg & Lee, 2017; Rodríguez-Pérez & Bajorath, 2020).

Despite the benefits machine learning can offer to psychological research, it is only in recent years that machine learning has been applied as a novel method of investigation in the

field. Within the area of food-related behaviour, there has been work investigating whether machines can judge nutrition better than humans (Rokicki et al., 2018), predict preferences for certain food types (Yu & Fu, 2020), and create healthier recommendations that align with health outcomes (Elsweiler et al., 2017; G. Mitchell et al., 2021; Panaretos et al., 2018). Gandhi et al. (in press) applied machine learning methodology to predicting healthiness judgements of food based on either the foods' nutritional information (e.g., amounts of fat, sugar, salt etc.), the associations of those food items with commonly used words, or both. Their models found that healthiness ratings for foods were indeed predictable—with accuracy rates of up to 91%.

A few studies have also looked at some factors that predict actual food choice. For example, (Dalenberg et al., 2014) investigated how the emotions evoked by food predicted food choice, complementing an analysis using mixed-effects models with a "leave-one-out" cross-validation method that folded the data as many times as the number of observations—effectively making isolated predictions for each observation ($n = 123$ in their case). Similarly, (Verwaeren et al., 2019) followed this approach (combining statistical inference with cross-validation) to assess predictive accuracy for their models of children's food choice based on sensory characteristics of the food (from $n = 149$ children). (Elsweiler et al., 2017) used a machine learning approach to predict whether participants would select one recipe over a similar one (in a pair) based on characteristics of the recipe's title, images, ingredients, ratings from other users, and nutritional content. Their modelling included both cross-validation and out of sample testing, and reported 64-66% accuracy in predicting which recipe would be selected among just over 1,100 observations from around 100 participants. So far, this growing literature indicates that machine learning can be a promising tool to investigate food choice behaviours.

Despite these promising indications, research in this area is still limited, especially compared to the wealth of studies in the wider food choice literature. An area within this literature that has received much attention is the use of nutrition labels, where investigations have spanned decades, including many systematic reviews about whether nutrition labels facilitate food choice (e.g., (Campos et al., 2011; Cowburn & Stockley, 2005; Grunert & Wills, 2007; Hersey et al., 2013; Hieke & Taylor, 2012; Soederberg Miller & Cassady, 2015) However, to our knowledge, machine learning has yet to be employed in these investigations. We therefore add to this body of work by leveraging machine learning as a tool to understand how people determine the healthiest food item in a choice set from their nutrition labels.

Nutrition labels and food choice

Nutrition labels have been widely endorsed as a tool to improve the healthiness of consumer diets (OECD & Publishing, 2008), with the majority of Western nations mandating some form of nutrition labels (Storcksdieck genannt Bonsmann et al., 2010). An estimated 80% of food products in Europe and the UK additionally adopt front-of-package (FOP) labels, which are posited to facilitate consumers' visual access to comprehensiveness information about key nutritional qualities of a food (Storcksdieck genannt Bonsmann et al., 2010) and increase the chances of this information being used to judge food healthiness (Wartella et al., 2010a). There is, however, scant evidence that these labels have brought down obesity rates as they were intended to (Storcksdieck genannt Bonsmann & Wills, 2012). One issue is of course whether people use labels; in general, people *purport* to use labels and find them more useful (Campos et al., 2011; Cowburn & Stockley, 2005) but actual usage is substantially less (Cowburn & Stockley, 2005; Higginson et al., 2002). The other questions is whether people use the labels *effectively*; here, empirical evidence is also conflicting. A number of studies reported that participants misunderstood labels (Graham &

Mohr, 2014; Liu et al., 2019; Mackey & Metz, 2009) or misjudged food nutrition based on the information (Levy et al., 2000). Some studies reported that participants could not reliably select the healthiest option in a set given nutrition labels (Gorton et al., 2008). Yet others reported that most of their participants successfully used nutrition labels to pick the healthiest option (Barreiro-Hurle et al., 2010; Grunert et al., 2010). This may of course be a case of what type of label was tested, since some reviews have found that FOP labels with colour-coding and text (compared to those without) facilitate selection of healthier choices (Cowburn & Stockley, 2005; Hersey et al., 2013).

A different problem may be that in practice, it is not that simple to judge which option is the healthiest. In certain cases, if one option stochastically dominates all others (e.g., if the option is lower in fat, sugar *and* salt than all the others), it is clear this is the best. However, one will more commonly encounter situations where options perform best on one attribute but worst on another (e.g., lowest in fat, highest in sugar). One way to select may be to identify the option that is best on average across the attributes. Alternatively, one might also consider one attribute to be superior to the others (e.g., lowest in sugar). Such a strategy, known as "take the best", is a shortcut "heuristic" that people often employ (Scheibehenne et al., 2007). It may well be a rational and appropriate criterion due to an individual's personal circumstance (e.g., specific medical conditions). It is also worth bearing in mind that even experts do not necessarily agree on what is the healthiest option. In a study where dieticians and nutritionists in the UK were asked to select the healthier of two real food options (with nutrition labels), the experts only reached $\geq 80\%$ agreement for the healthier product in 60% of the pairs (Food Standards Agency, 2008). (Scarborough et al., 2007) also found that nutrition professionals varied highly in their categorisation of the healthiness of some foods. Therefore, it is perhaps advisable to consider when consumers would make the best choice as determined by several different criteria rather than one single measure.

One barrier to systematically assessing choices against each criteria is the danger of inflated p -values inherent in testing data multiple times. Using the traditional analytical approach of NHST, if one tests five different criteria for healthiness on the same choices, there is at least a 23% chance¹ of finding a significant result. One could of course collect five different new datasets systematically to avoid this, however this requires additional time and resources. We propose that the advantages offered by machine learning, described above, present an opportunity to exploratorily analyse food choice data from different angles without "the tyranny of p -values" (Stang et al., 2010).

Present research

The rest of this paper is structured as follows. We first describe the food choice dataset, how it was obtained, and how choices were scored. Then, we then describe the models (including their predictor variables) and how they were trained and tested. We then present the modelling results: (i) the performance of the models when assessing choices for different healthiness criteria and (ii) the contribution strength of each predictor variable included in the models. Here, we include as well a test of the data using typical Null Hypothesis Significance Testing in a general linear model (GLM). The GLM was also assessed for fit and predictive accuracy and precision. This allowed as to compare the use of tree-based machine learning methods against the more commonly used GLM. Finally, we discuss the results in terms of relative model performance and the overall findings of our analyses.

Model data

About the dataset

¹ The familywise error rate, based on the typical $\alpha = .05$ significance level for just five comparisons = $1-(1-0.05)^5$.

The dataset consists of 4620 choices made by 154 participants (30 per participant). For each choice, participants chose the healthiest option out of a set of 6 randomised choice options, each depicted by a food label.

In each trial, these 6 options depicted varieties of a type of food (taken from the dataset of real foods that was used in the expert study described above; FSA, 2008). The different categories were: ready meals, yoghurts, sandwiches, crisps, soups, and breakfast cereals.

Each label showed energy, fat, sugar, and salt content for the food, with traffic light bands applied to fat, sugar, and salt (green, amber, red). %RI values were also provided as numerical figures, rounded to the nearest whole number.

Choice predictions

For each choice trial, participants' choices could be classified according to whether the chosen option aligned with the option in the given choice set that²:

- (i) Minimised the equally weighted average of %RI provided by each of the nutrients
- (ii) Minimised the equally weighted average across traffic light band categories for each of the nutrients (i.e., whether the option is likely to fall under red, amber, and green on average)
- (iii) For each of the following: Fat, Sugar, Salt, Energy³;
- (iv) Minimised the %RI for that nutrient
- (v) Minimised the traffic banding for that nutrient

For each of these, we ran the models to ascertain what predicted choices that corresponded to that classification.

Models

² Note that in each choice set, one option might correspond to more than one of the classifications above (e.g., a choice that minimised salt might also minimise sugar).

³ As traffic light banding was not available for energy, we only included energy under the minimising %RI classification.

Scripts and datasets to run the models are available at the following link:

<https://github.com/dlholf/foodml>

Data splits

As our dataset contained sufficient observations ($n = 4620$), we elected to use a hold-out sample (our test set), with cross-validation of the initial models run on the training set using five validation folds. We split the data 80/20 into a training and test set⁴. For the cross-validation folds, we used balanced resampling in the training set when there were imbalances of $> 80/20$ ratio in occurrences of the target. For example, 91% (4204) of choices corresponded with the option that minimised the average traffic light band on the label vs. 9% (416) did not. In contrast, 57% (2643) of choices corresponded with the option that minimised the average %RI of the nutrients vs. 43% (1977) did not.

Model algorithms

Using `h2o.ai`⁵, we trained a set of 20 base models (each tuned within the `h2o.automl()` function) to predict each choice alignment. Each model used 5 cross-validation folds within the training set, and was subsequently tested on a hold-out test sample.

Each model used the following predictors:

- Type of food (breakfast cereal, soup, ready meal, yoghurt, crisps, or sandwiches, each dummy coded)
- Time taken to make choice
- Demographics and individual characteristics:
- Attitudes to healthy eating
- Frequency of nutrition label use

⁴ One exception was for choices that minimised sugar TFL bands, because of a massive class imbalance, which meant an 80/20 training/test split would have left < 5 items for test prediction. A 60/40 split was used in this case.

⁵ <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html>

- Factor considered important in determining food healthiness (reducing calories, fat, sugar, salt, or all, each dummy coded)
- Age
- Gender
- Number of other criteria the choice also aligned with (e.g., that option might be the best for minimising energy _and_ fat)
- Distance (in original units) between the best and second best option in the choice set

Model selection (out of 20 automl base models)

For each choice alignment variable (the DV), we selected the best performing model from the 20 automl base models. Model selection was determined by inspection of the confusion matrix of predictions in the test sets, seeking to minimise classification error rates across both types of correct classification (true positives and true negatives), but also prioritise minimising false positives over false negatives. For each DV, the results reported are based on the final selected model.

Analysing model features

We analysed the contributions of individual predictors to the models using:

1. A trained surrogate model. This extracted the predictions made for the dataset (train and test sets) by each model and trained a single decision tree on the predictions. We then plotted the trees and their nodes, which are split based on relevant predictors. This gave a picture of the predictive variable (and its critical value) that led to resultant classification probabilities.
2. Shapley's (SHAP) values of each predictor, which assigns each predictor an importance value for a particular prediction (Lundberg & Lee, 2017). The SHAP values compare what a model predicts with and without a predictor to determine that predictor's contribution (Paris, 2020). These help us to identify whether predictors are making strong positive (i.e., 'yes') or negative (i.e., 'no') contributions in the model (Rodríguez-Pérez & Bajorath, 2020).

The SHAP values can be visualised in two ways:

- (i) Plotting the SHAP values for each predictor shows the change in log odds for each individual data point in the sample (with its unique predictor value).
- (ii) Partial dependency plots (PDP) for individual predictors show a finer-grained evaluation of the change in SHAP values (i.e., the predictive contributions) across different predictor values.

Results

Model fit and performance

Statistics for each of the classification models are presented in Table 1. We report the type of model used (in general, these were Gradient Boosted Machines, which are decision tree models), the AUC, log loss, mean per class error, mean square error, and three harmonic means (f) of precision and recall (Brownlee, 2020), along with the rate of false positives and false negatives and the total error rate observed when applying the model to the test set.

The mean f1 is the ratio of precision (percentage of correct predictions—minimising false positives) to recall (percentage of correct predictions for the positive class—minimising false negatives). The mean f0.5 calculates the ratio with more weight on precision and less on recall (i.e., more importance on minimising false positives), while the mean f2 calculates the ratio with less weight on precision and more on recall (i.e., more importance on minimising false negatives).

Model features and explainability

We assessed the sign and magnitude of how the SHAP values were correlated with the actual values observed for each predictor variable in the data. Tables 2 and 3 report for each model the sign of the predictor and the relative magnitude of the correlation (as a ranking against other predictors in the model, where 1 reflects the most correlated predictor).

Table 1.

Statistics for the best classification model for each DV

DV	Model type	AUC	Log loss	Mean per class error	MSE	Mean f1	Mean f2	Mean f0.5	False positive rate	False negative rate	Total error rate
<i>%RI classification models</i>											
Minimise sugar %RI	Gradient Boosted Machine	1.00	0.08	1.75%	0.02	0.71	0.73	0.73	7.01%	9.89%	7.90%
Minimise fat %RI	Gradient Boosted Machine	0.99	0.16	6.18%	0.04	0.72	0.76	0.72	5.76%	9.79%	7.03%
Minimise energy %RI	Extreme Gradient Boosted Machine	0.99	0.14	4.53%	0.04	0.73	0.75	0.73	5.29%	14.5%	8.56%
Minimise salt %RI	Gradient Boosted Machine	0.96	0.28	12.1%	0.09	0.62	0.66	0.63	14.4%	18.1%	15.6%
Minimise average %RI	Gradient Boosted Machine	1.00	0.10	2.25%	0.02	0.80	0.78	0.84	7.51%	3.16%	4.94%
<i>TFL band classification models</i>											
Minimise sugar traffic light band	Gradient Boosted Machine	1.00	1.22	0%	0.42	0.67	0.63	0.75	40%	0.05%	0.16%
Minimise fat traffic light band	Extreme Gradient Boosted Machine	1.00	0.02	6.48%	0.01	0.77	0.72	0.86	0%	0.68%	0.66%
Minimise salt traffic light band	Gradient Boosted Machine	0.99	0.36	2.93%	0.13	0.84	0.80	0.89	6.25%	4.05%	4.28%
Minimise average TFL band	Gradient Boosted Machine	1.00	0.24	0.04%	0.07	0.84	0.79	0.90	0%	0.36%	0.33%

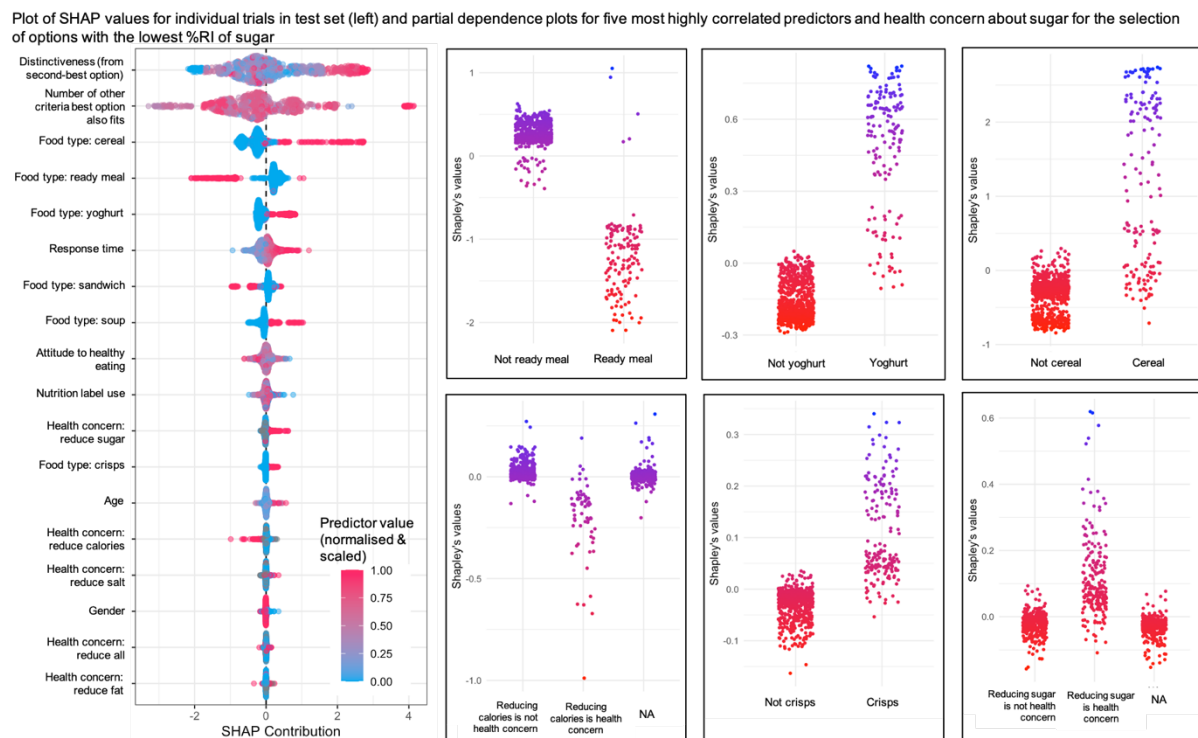
Table 2.

Rank and sign of predictor variables in the %RI prediction models

Predictor	<u>%RI Sugar</u>		<u>%RI Fat</u>		<u>%RI Energy</u>		<u>%RI Salt</u>		<u>Average %RI</u>		Mean rank
	Sign	Rank	Sign	Rank	Sign	Rank	Sign	Rank	Sign	Rank	
Food type: ready meal	-	1	+	13	-	1	+	4	+	2	4.2
Food type: yoghurt	+	2	-	17	-	5	+	2	-	6	6.4
Food type: cereal	+	3	-	2	+	14	-	1	+	1	4.2
Food type: crisps	+	5	+	1	+	10	-	13	-	7	7.2
Food type: soup	+	6	+	4	-	13	+	18	-	5	9.2
Food type: sandwich	-	7	-	3	+	3	-	3	+	3	3.8
Health concern: reduce sugar	+	9	-	6	-	12	+	8	+	12	9.4
Health concern: reduce fat	+	14	+	10	+	7	-	9	+	14	10.8
Health concern: reduce calories	-	4	+	8	+	4	-	11	-	8	7
Health concern: reduce salt	+	15	-	5	-	2	+	7	+	16	9
Health concern: reduce all	-	17	+	7	+	6	-	14	+	18	12.4
Age	+	11	+	9	-	18	-	10	-	13	12.2
Gender	-	12	-	15	-	9	+	5	-	15	11.2
Response time	+	8	-	16	-	17	+	17	+	10	13.6
Attitude to healthy eating	-	16	+	14	+	15	-	16	-	17	15.6
Nutrition label use	-	18	-	18	-	11	-	15	+	11	14.6
Number of other classifications	+	13	+	11	+	8	+	19	+	4	11
Distance from second-best option	+	10	+	12	+	16	+	6	+	9	10.6
Number of other options that were equally the best	NA	NA	NA	NA	NA	NA	-	12	NA	NA	NA

Plotting the SHAP values against the actual values of each predictor, as shown in the example for minimising %RI of sugar in Figure 1, illustrate the relative contributions of the predictors to the models.

Figure 1. SHAP values and partial dependence plots for predictors of whether the option with the lowest %RI of sugar would be selected.



Plotting a surrogate tree for each model allowed us to corroborate how the predictors were used in classification by a single decision tree. For example, as shown in Figure 2, the surrogate tree for minimising %RI of sugar shows the importance of yoghurts, cereals, soups, and how distinct the lowest-sugar option was from the other options in making predictions.

Health concerns—i.e., whether the participant viewed reducing sugar, fat, calories, salt, or all of them as the most important in determining food healthiness—also did not consistently contribute to predicting selections in the same direction. For example, concern

about reducing fat tended to predict the selection of lowest sugar, fat, energy, and overall %RI options, however it tended to predict non-selection of lowest fat %RI options.

Other individual characteristics such as age, gender, attitudes, frequency of nutrition label use, and the speed of responses, were generally ranked lower in terms of how well they correlated with SHAP values, indicating a less linear relationship with selection. Again, none of these variables had a consistent sign of correlation with SHAP values across all the models, with each predicting selection for some products and non-selection for others.

This tells us that there is clearly a trade-off to be made between the various products; and the same variables that help one to select the best option by one criteria may not be helpful with a different criteria.

Overall, food type tended to consistently contribute to predictions, showing clear correlations between the predictor values and SHAP values. The sign of this correlation was never consistent across all models, however—for example, sandwiches, which most consistently contributed to predictions, predicted the selection of lowest energy and overall %RI options. However, they predicted *non*-selection of the other lowest options (sugar, fat, and salt).

Figure 2. Surrogate decision tree for the prediction of selecting an option that is the lowest in %RI of sugar in the set.

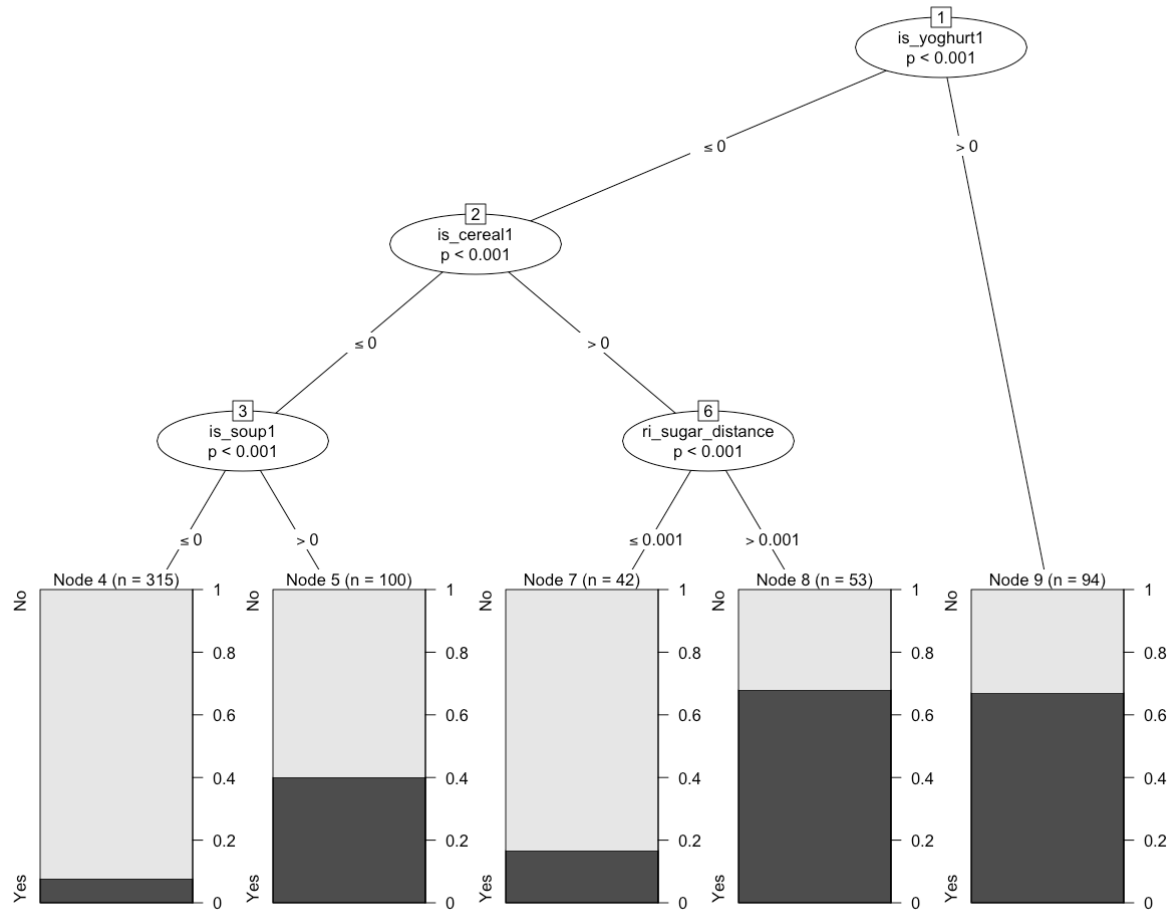


Table 3

Rank and sign of predictor variables in the TFL prediction models

Predictor	<u>TFL Sugar</u>		<u>TFL Fat</u>		<u>TFL Salt</u>		<u>Average TFL</u>		<u>Mean</u>
	Sign	Rank	Sign	Rank	Sign	Rank	Sign	Rank	<u>rank</u> Sign
Food type: ready meal	-	4	-	6	-	5	-	12	6.75
Food type: yoghurt	-	5	NA	NA	+	2	+	3	3.33
Food type: cereal	+	3	NA	NA	+	6	+	2	3.67
Food type: crisps	-	7	NA	NA	+	3	+	4	4.67
Food type: soup	-	10	NA	NA	+	4	+	5	6.33
Food type: sandwich	+	1	-	8	-	11	-	7	6.75
Health concern: reduce sugar	+	12	-	3	-	19	-	8	9
Health concern: reduce fat	-	8	+	1	-	12	+	14	8.75
Health concern: reduce calories	+	9	-	2	+	9	+	9	7.25
Health concern: reduce salt	-	14	-	4	+	10	+	11	9.75
Health concern: reduce all	+	19	+	7	+	15	+	13	13.5
Age	+	18	-	13	-	13	-	17	15.25
Gender	-	11	+	14	+	14	-	15	13.5
Response time	+	17	-	9	+	16	+	19	15.25
Attitude to healthy eating	+	15	-	12	-	17	+	18	15.5
Nutrition label use	-	13	+	10	-	18	-	16	14.25
Number of other classifications	+	6	+	11	+	8	+	10	8.75
Distance from second-best option	-	2	NA	NA	-	1	-	1	1.33
Number of other options that were equally the best	+	16	+	5	+	7	+	6	9.25

Contrast with GLM

As part of the `h2o.automl()` process, we were also able to identify models that used basic GLM (a typical statistical model that identifies the best parameters, or coefficients, that fits an equation that includes the predictors to the data). For each of the models, we compared predictions under the GLM vs. best machine learning model and examined which predictors offered better explanations of the data.

Overall, the GLMs performed worse. For example, the GLM for reducing %RI sugar had a 23% error rate in predictions (vs. 8% for the GBM model) and harmonic mean of precision/recall of 0.52 vs. 0.71 (meaning a lower false positive and false negative rate in the GBM model). The only instance in which a GLM outperformed the tree-based model in predictive ability was in reducing false positives for the model minimising sugar traffic light banding (20% false positives vs. 40%)—this was a case where the class imbalance was so high (>99%) that we should in any case be wary of overfitting.

Table 4.

Statistics for the GLM classification models

DV	AUC	Log loss	Mean per class error	MSE	Mean f1	Mean f2	Mean f0.5	False positive rate	False negative rate	Total error rate
%RI classification models										
Minimise sugar %RI	0.82	0.47	23.2%	0.15	0.52	0.58	0.50	22.5%	24.7%	23.2%
Minimise fat %RI	0.83	0.47	20.8%	0.15	0.53	0.59	0.52	23.5%	19.6%	22.3%
Minimise energy %RI	0.88	0.40	21.1%	0.13	0.60	0.63	0.62	33.3%	13.5%	26.2 %
Minimise salt %RI	0.80	0.50	26.6%	0.17	0.52	0.58	0.50	29.4%	26.3%	28.4%
Minimise average %RI	0.92	0.36	12.6%	0.11	0.75	0.75	0.79	11.0%	10.4%	10.6%
TFL band classification models										
Minimise sugar traffic light band	1.00	0.003	21.4%	< .001	0.91	0.87	0.96	20%	0.05%	0.11%
Minimise fat traffic light band	1.00	0.03	5.33%	0.01	0.86	0.81	0.93	28%	0.34%	1.10%
Minimise salt traffic light band	0.98	0.11	19.1%	0.04	0.88	0.85	0.93	12.5%	3.19%	4.17%

Explainability via NHST

We complemented our machine learning models with null hypothesis significance testing using mixed-effects models using the lme4 package in R. These models included random intercepts for participants in order to account for repeated trials in the data. For multicategorical variables (food type and health concerns), we used as the reference class the category that least contributed to predictions.

While automl was able to run GLM-based classification models, the imbalances of certain categorical variables posed a problem for the model specifications for a mixed-effect GLM resulting in models that failed to converge for models with selections that minimised TFL values. We therefore confined our mixed-effects analyses to models with selections minimising %RI.

Table 5 shows the odds ratios (and their respective *p*-values) obtained for each variable in the mixed-effects GLMs for the five %RI-minimising models. As found with the machine learning models, the variables did not have the same sign across all models, further supporting that different variables will be of predictive importance depending on *which* nutrient one wishes to minimise in one's diet—and the necessary trade-off in decision-making needed.

Another similarity with the machine learning models was that regardless of which criteria was used to score the healthiest option, food type tended to have consistently large odds ratios in predicting whether the best option was selected—in some cases, even greater than the distinctiveness of that option and whether it was the best by more criteria (both of which one would sensibly expect to be consistent predictors of best-option selection). This underscores the importance of considering how different food types might affect how people use food labels to judge their healthiness.

The contribution of what people were most concerned about in determining healthiness was comparably small, with each concern about 1.5 times on average more likely than the least concern to predict best-option selection (vs. not).

Finally, demographic variables and individual differences in attitudes and nutrition label use had virtually no role in predicting whether participants would select the best option for most of the criteria. (The exception was gender, where females were 1.6 times more likely than males to pick the option with lowest %RI of salt as the healthiest.)

Overall, these patterns align well on the whole with what we observed with the machine learning insights.

Table 5.

Odds ratios and significance of predictor variables in the %RI GLMs

Predictor	%RI Sugar		%RI Fat		%RI Energy		%RI Salt		Average %RI		Mean absolute OR
	OR	<i>p</i>	OR	<i>p</i>	OR	<i>p</i>	OR	<i>p</i>	OR	<i>p</i>	
Food type: ready meal	0.29	< .001	1.40	.090	0.05	< .001	0.02	.001	14.74	< .001	17.91
Food type: yoghurt	7.37	< .001	NA	NA	0.20	< .001	27.59	< .001	0.12	< .001	12.07
Food type: cereal	7.24	< .001	1.82	.001	NA	NA	0.66	.460	3.45	< .001	3.51
Food type: crisps	0.30	< .001	4.22	< .001	1.13	.464	4.60	.002	NA	NA	3.32
Food type: soup	1.15	.317	2.21	< .001	0.63	< .001	NA	NA	0.80	.221	1.55
Food type: sandwich	NA	NA	0.20	< .001	9.26	< .001	0.14	< .001	6.40	< .001	6.95
Health concern: reduce sugar	1.64	.016	0.53	.010	NA	NA	1.24	.274	0.96	.827	1.45
Health concern: reduce fat	1.24	.398	NA	NA	1.43	.122	0.73	.202	1.22	.326	1.31
Health concern: reduce calories	0.57	.060	1.04	.188	2.89	< .001	0.70	.204	0.85	.457	1.66
Health concern: reduce salt	1.32	.285	0.36	.001	0.86	.531	1.44	.091	0.86	.460	1.57
Health concern: reduce all	NA	NA	0.73	.188	1.50	.030	NA	NA	NA	NA	1.43
Age	1.16	.204	1.00	.971	0.86	.165	1.05	.679	1.13	.191	1.04
Gender	1.01	.944	0.74	.113	0.75	.108	1.62	.010	0.85	.290	1.30
Response time	1.06	.193	0.98	.667	0.96	.432	0.97	.434	1.06	.271	1.04
Attitude to healthy eating	1.02	.869	1.02	.819	1.03	.686	0.98	.787	0.97	.640	1.02
Nutrition label use	0.99	.892	1.00	.999	0.95	.570	1.04	.713	1.08	.326	1.04
Number of other classifications	1.28	< .001	3.06	< .001	6.35	< .001	1.10	.093	27.61	< .001	7.88
Distance from second-best option	1.27	.028	2.86	< .001	1.18	.197	11.64	< .001	1.99	< .001	3.79
Number of other options that were equally the best	NA	NA	NA	NA	NA	NA	1.44	.343	NA	NA	NA

Note: Odds ratio (OR) < 1 indicate that the variable is _less_ predictive of selecting the option for the relevant column. Mean OR is calculated as

the mean _magnitude_ (i.e., using the positive odds, i.e., the inverse of ORs < 1).

Discussion

Using an automated machine learning software (h2o.ai), we trained a set of 20 tree-based model algorithms to predict in a dataset of 4,620 choice observations whether an individual picked the healthiest option (out of 6). We operationalised the healthiest option in different ways, running a separate set of models to predict whether a participant's chosen healthiest option was one that minimised:

- (i) the equally weighted average of %RI provided by each of the nutrients
- (ii) the level of %RI contributions each for fat, sugar, salt, or energy
- (iii) the equally weighted average across traffic light band categories
- (iv) the traffic light band category each for fat, sugar, or salt

Overall, the models were able to predict whether participants picked the healthiest choice in >90% of instances (except for %RI of salt, where the total error rate was approximately 16%). For the majority of models, the false positive rate (i.e., predicting a healthiest choice when it was not the case) remained low (<10%), with acceptable false negative rates as well (i.e., failing to predict when someone picked the healthiest choice). The models were tested using cross-validation folds within the training sample (80% of the data) and again on a 20% holdout sample. Therefore, the error rates reported, which are from the holdout sample, reflect an ability to predict in a completely unseen dataset which healthy choices would be made.

These results are comparable to those obtained in a few previous studies where machine learning techniques were able to predict participants' choice of food based on food-evoked emotions (50-80% accuracy; (Dalenberg et al., 2014), sensory food characteristics (69% accuracy; (Verwaeren et al., 2019), and recipe characteristics (64-66% accuracy; (Elsweiler et al., 2017). We therefore add to the growing base of evidence about the strong

potential for machine learning to predict—and subsequently understand—food choice behaviours.

However, performance did vary across the different healthiness criteria we tested for in our modelling. Within the same food label, there were many different ways one could determine the food's healthiness: based on the numerical %RI value that indicates the contribution of 4 possible nutrients (sugar, fat, energy, salt), or the traffic light band colour given to 3 of the nutrients (sugar, fat, salt). For the %RI values, which are more precise, options could be differentiated on the basis of smaller differences, and a single option was most likely to be identified as the best for each nutrient (and more so for their average %RI). These %RI-based differences between options can often be small, to the point that one may wonder if they make any real difference to consumers. Past work has suggested that small differences in nutritional content still matter and can drive selection of one product over another (Miller et al., 2015). In our data, the model based on %RI differences had prediction rates of 84% (for selecting options with lowest salt %RI) to 95% (for selecting options with the lowest average %RI). This suggests that there are indeed variables in the the food information environment that can help us understand when consumers identify these finely-differentiated options as being healthiest.

In contrast to %RI information, for the traffic light bands, which capture a broader range of values within them, there was often more than a single option within the choice set that would be the best for a particular nutrient. The data prior to modelling suggested that the traffic light bands were to some extent effective: large proportions of participants selected the best choice under each of the traffic light band criteria (in the most extreme case, by 997:1 for sugar). Unfortunately, this resulted in sample imbalances that made it harder to predict when participants might fail to identify this best choice. In such cases, the model may learn to predict this is the default choice. For instance, predicting the healthiness of choices based on

sugar traffic light banding (i.e., whether the "green" sugar option was selected) had an extremely high false positive rate (40%). Interestingly, despite also suffering similar imbalances in the sample, predictive accuracy for the other traffic light band criteria (fat, salt, average band rating) suffered much less from false positives. As such, there appear to be predictive variables that highlight when people select foods with labels that are "green" on average, or for fat or salt, but not sugar.

What explains choices?

One critique of machine learning models has been that there is low explainability, i.e., it is difficult to identify *why* the model predicted people would choose that option. However, it is possible to place a value on how much each predictor variable in the model contributes to a prediction (i.e., the SHAP value: a smaller SHAP indicates the variable is less important to the prediction, whereas a larger positive SHAP indicates the variable is more important for predicting a choice will be selected, and a larger negative SHAP indicates the variable is more important for predicting that the choice will *not* be selected). We correlated the SHAP values with the predictor values to determine the directional association between predictor and predicted choice. We were also able to assess the probability of a certain prediction based on different predictor values (i.e., through a surrogate model). We also ran a robustness check using GLM regression models to check the effects of the same predictors, which broadly corroborated with our machine learning analysis. Altogether, these methods allowed us to identify the features of the food information environment that might explain predictions.

A key finding from this analysis was that people likely use different healthiness criteria depending on the type of food they purchase and, to a lesser extent, what nutrients they perceived as detrimental to health. Type of food was constantly an important contributor to predictions, but how it aided predictions varied across the healthiness criteria. For example, yoghurts and cereals were positively associated to predicting lowest-sugar choices,

but these tended to predict higher fat choices. In contrast, lowest-fat choices were better predicted for crisps and soups, but these same foods tended to predict choices that were higher in average nutrient contribution. This suggests that low sugar may be a consumer criteria for picking the healthiest yoghurt and cereal, whereas low fat has more importance in selecting the healthiest crisps. Crucially, the fact that the direction of prediction varies for different foods reflects a trade-off people are required to make, where picking the lowest sugar product often means forgoing the lowest fat one. Indeed, this trade-off was visible from how participants' concerns about which nutrient should determine healthiness. As would be expected, these concerns positively predicted the respective healthiest choices—i.e., one's belief that reducing sugar is most important to determining health increases the prediction that one would select the lowest sugar option as the healthiest choice. However, believing that reducing sugar was most important also contributed to a *less* likely selection of the lowest %RI fat option; similarly for other nutrient reduction priorities. There was no one criteria (even best average reduction of all nutrients) that was also positively associated with lowering all of the possible %RIs, highlighting the complexity of making a food healthiness decision within a given food information environment.

The variation in predictive explainability among foods—even for selecting the lowest average %RI option—also suggests that how one judges healthiness is likely not a simple matter of prioritising one nutrient in the decision. Indeed, concerns about individual nutrients were generally less strongly correlated with predictive contributions. Only salt and energy/calorie concerns were among the top five strongest contributors to any model's predictions—and in fact, concern about reducing salt was ranked higher in predicting higher energy and fat choices than in predicting lowest salt choices. This could mean that participants' beliefs about the healthfulness of nutrients did not strongly guide their choices—or more likely, the beliefs were more complex than stated, and varied depending on the foods.

Our findings here align with other recent work that people's existing knowledge about different food items is better able to predict how healthy they judge those foods to be, over and above the nutritional content of the food (although nutritional information can be helpful on top of that existing knowledge). For example, Gandhi et al. (in press) found that a model trying to predict food healthiness judgements by using linguistic representations of people's knowledge about the food items fared better (by about 12% greater predictive accuracy) than a model that simply used nutritional information to predict judgements—although combining both types of information in the modelling improved predictive accuracy.

Contrary to what one might expect, other individual characteristics of participants such as demographics (age, gender), attitudes to healthy eating, and nutrition label use frequency had less importance in the models' predictions. Gender was important to the choice of lowest-salt products by %RI, where female participants were predicted to select these more than male ones, but this was the only instance in which demographic differences ranked among the top 5 predictors used by the models. Altogether this suggests a complex relationship between individual difference variables and the propensity to choose certain food options. However, the range of individual difference variables in our sample was fairly limited, since there would only have been 154 unique individuals in the dataset. With a larger and more diverse sample, it is still possible that these characteristics might take on more importance for predictions. However, we did already see that the direction of the predictive contributions from each variable varied depending on which criterion was modelled, and we would expect this to still be the case in a more diverse dataset. In other words, we might predict older participants to select lowest sugar and lowest fat options more, but lowest calorie options less.

Limitations

While our work offers greater insight on how machine learning methods could be leveraged to better understand the complexities involved in deciding whether something is healthy, it does have some limitations. First, although our sample sizes were comparable to other initial studies in machine learning and food choice (e.g., repeated observations from 100-200 participants; (Dalenberg et al., 2014; Elswailer et al., 2017; Verwaeren et al., 2019); Gandhi et al., in press), these are still small datasets (totaling observations in the thousands) compared to many datasets in the wider machine learning field (millions of observations, (Galeano & Peña, 2019; Gudivada et al., 2015); potentially billions, (Sh. Hajirahimova & S. Aliyeva, 2017)) and even some areas of psychology (Yarkoni & Westfall, 2017). Larger datasets are naturally beneficial in avoiding model overfitting to a specific dataset, since it becomes less likely for the model to learn patterns unique to the training data (Yarkoni & Westfall, 2017). While the crossfold validation and out-of-sample testing procedures we used for model predictions provide some measure of protection against overfitting, larger samples remain a better way to improve model generalisation and balance issues with under or overfitting.

Second, our participants were drawn from a predominantly "WEIRD" (Henrich et al., 2010) undergraduate student population, which may not be fully representative of the wider population. Our work presents more evidence for how machine learning methods could be applied to understanding food choice, but much more analysis of data from a wider population is needed to verify the insights we obtained from our limited data. This limitation cannot be understated: the danger of encoding biases from the training dataset into algorithms has been exemplified by high-profile cases where biased algorithms result in biased real-world decisions, often to great detriment to already-disadvantaged minorities (Cathy O'Neil, 2016; Eubanks, 2018).

Conclusion

As the application of machine learning methods to behavioural questions is still in its infancy, we believe our findings on a small scale are of value to shaping future research questions and investigations. We hope that this report may act as early evidence and validation for researchers looking to enlist machine learning methods in understanding the complexities of nutrition understanding and food choice.

References

- Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3), 666–678. <https://doi.org/10.3758/s13428-011-0089-5>
- Barreiro-Hurle, J., Gracia, A., & de-Magistris, T. (2010). Does nutrition information on food products lead to healthier food choices? . *Food Policy*, 35(3), 221–229. <https://doi.org/10.1016/j.foodpol.2009.12.006>
- Brownlee, J. (2020). *A Gentle Introduction to the Fbeta-Measure for Machine Learning*.
- Campos, S., Doxey, J., & Hammond, D. (2011). Nutrition labels on pre-packaged foods: a systematic review. *Public Health Nutrition*, 14(8), 1496–1506. <http://www.ncbi.nlm.nih.gov/pubmed/21241532>
- Cathy O’Neil. (2016). *Weapons of Math Destruction*.
- Cowburn, G., & Stockley, L. (2005). Consumer understanding and use of nutrition labelling: a systematic review. *Public Health Nutrition*, 8(1), 21–28. <http://www.ncbi.nlm.nih.gov/pubmed/15705241>
- Dalenberg, J. R., Gutjar, S., ter Horst, G. J., de Graaf, K., Renken, R. J., & Jager, G. (2014). Evoked Emotions Predict Food Choice. *PLOS ONE*, 9(12), e115388. <https://doi.org/10.1371/journal.pone.0115388>
- Elsweiler, D., Trattner, C., & Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. *SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 575–584. <https://doi.org/10.1145/3077136.3080826>
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*.

- Food Standards Agency. (2008). *Defining the correct answers: Professional nutritionists and dietitians assessment of the relative healthiness of foods used within the main survey for the independent signpost evaluation study* (F. S. A. Food Standards Agency, Ed.). Food Standards Agency, FSA.
- G. Mitchell, E., M. Heitkemper, E., Burgermaster, M., E. Levine, M., Miao, Y., L. Hwang, M., M. Desai, P., Cassells, A., N. Tobin, J., G. Tabak, E., J. Albers, D., M. Smaldone, A., & Mamykina, L. (2021). *From Reflection to Action: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations*. 1–17.
<https://doi.org/10.1145/3411764.3445555>
- Galeano, P., & Peña, D. (2019). Data science, big data and statistics. *TEST 2019* 28:2, 28(2), 289–329. <https://doi.org/10.1007/S11749-019-00651-9>
- Graham, D. J., & Mohr, G. S. (2014). When zero is greater than one: consumer misinterpretations of nutrition labels. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association*, 33(12), 1579–1587. <http://www.ncbi.nlm.nih.gov/pubmed/24707845>
- Grunert, K. G., Fernández-Celemín, L., Wills, J. M., Storcksdieck genannt Bonsmann, S., & Nureeva, L. (2010). Use and understanding of nutrition information on food labels in six European countries . *Zeitschrift Fur Gesundheitswissenschaften*, 18(3), 261–277.
<https://doi.org/10.1007/s10389-009-0307-0>
- Grunert, K. G., & Wills, J. M. (2007). A review of European research on consumer response to nutrition information on food labels. *Journal of Public Health*, 15, 385–399.
<https://doi.org/10.1007/s10389-007-0101-9>
- Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). *Big data: Promises and problems*. <https://rebootingcomputing.ieee.org/images/files/pdf/mco2015030020.pdf>

- Guthrie, J., Mancino, L., & Lin, C.-T. J. (2015). Nudging Consumers toward Better Food Choices: Policy Approaches to Changing Food Consumption Behaviors. *Psychology & Marketing*, 32(5), 501–511. <https://doi.org/https://doi.org/10.1002/mar.20795>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29. <https://doi.org/10.1038/466029a>
- Hersey, J. C., Wohlgenant, K. C., Arsenault, J. E., Kosa, K. M., & Muth, M. K. (2013). Effects of front-of-package and shelf nutrition labeling systems on consumers. *Nutrition Reviews*, 71(1), 1–14. <http://www.ncbi.nlm.nih.gov/pubmed/23282247>
- Hieke, S., & Taylor, C. R. (2012). A critical review of the literature on nutritional labeling. *Journal of Consumer Affairs*, 46(1), 120–156. <https://doi.org/10.1111/j.1745-6606.2011.01219.x>
- Higginson, C. S., Kirk, T. R., Rayner, M. J., & Draper, S. (2002). How do consumers use nutrition label information? *Nutrition & Food Science*, 32(4), 145–152. <https://doi.org/10.1108/00346650210436253>
- Hubbard, R., & Ryan, P. A. (2000). Statistical Significance with Comments by Editors of Marketing Journals: The Historical Growth of Statistical Significance Testing in Psychology—and its Future Prospects. *Educational and Psychological Measurement*, 60(5), 661–681. <https://doi.org/10.1177/0013164400605001>
- Levy, L., Patterson, R. E., Kristal, A. R., & Li, S. S. (2000). How well do consumers understand percentage daily value on food labels? *American Journal of Health Promotion*, 14(3), 157–160. <https://doi.org/10.4278/0890-1171-14.3.157>
- Liu, D., Juanchich, M., Sirota, M., & Orbell, S. (2019). People overestimate verbal quantities of nutrients on nutrition labels. *Food Quality and Preference*, 78, 103739. <https://doi.org/10.1016/j.foodqual.2019.103739>

- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. v Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.
- <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>
- Mackey, M. A., & Metz, M. (2009). Ease of reading of mandatory information on Canadian food product labels. *International Journal of Consumer Studies*, 33(4), 369–381.
- <https://doi.org/10.1111/j.1470-6431.2009.00787.x>
- McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100(11), 2175–2199. <https://doi.org/10.1175/BAMS-D-18-0195.1>
- Miller, L. M. S., Cassady, D. L., Beckett, L. A., Applegate, E. A., Wilson, M. D., Gibson, T. N., & Ellwood, K. (2015). Misunderstanding of Front-Of-Package Nutrition Information on US Food Products. *PloS One*, 10(4), e0125306.
- <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4414362&tool=pmcentrez&rendertype=abstract>
- OECD, & Publishing, O. (2008). *Promoting sustainable consumption – good practices in OECD countries*. .
- Panaretos, D., Koloveryou, E., Dimopoulos, A. C., Kouli, G.-M., Vamvakari, M., Tzavelas, G., Pitsavos, C., & Panagiotakos, D. B. (2018). A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the ATTICA study. *British Journal of Nutrition*, 120(3), 326–334. <https://doi.org/DOI: 10.1017/S0007114518001150>

- Paris, U. L. (2020). *Push the limits of explainability — an ultimate guide to SHAP library*.
- Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34(10), 1013–1026.
<https://doi.org/10.1007/s10822-020-00314-0>
- Rokicki, M., Trattner, C., & Herder, E. (2018). The Impact of Recipe Features, Social Cues and Demographics on Estimating the Healthiness of Online Recipes. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, Issue 1).
<https://www.nhs.uk/Livewell/Goodfood/Pag>
- Scarborough, P., Rayner, M., Stockley, L., & Black, A. (2007). Nutrition professionals' perception of the 'healthiness' of individual foods. *Public Health Nutr*, 10(4), 346–353.
<https://doi.org/10.1017/S1368980007666683>
- Scheibehenne, B., Miesler, L., & Todd, P. M. (2007). Fast and frugal food choices: Uncovering individual decision heuristics. *Appetite*, 49(3), 578–589.
<https://doi.org/10.1016/j.appet.2007.03.224>
- Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339. <https://doi.org/10.1037/met0000061>
- Sh. Hajirahimova, M., & S. Aliyeva, A. (2017). About Big Data Measurement Methodologies and Indicators. *International Journal of Modern Education and Computer Science*, 9(10), 1–9. <https://doi.org/10.5815/ijmecs.2017.10.01>
- Soederberg Miller, L. M., & Cassady, D. L. (2015). The effects of nutrition knowledge on food label use. A review of the literature. *Appetite*, 92, 207–216.
<https://doi.org/10.1016/j.appet.2015.05.029>

- Stang, A., Poole, C., & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, 25(4), 225–230. <https://doi.org/10.1007/s10654-010-9440-x>
- Storcksdieck genannt Bonsmann, S., Fernández Celemin, L., Larrañaga, A., Egger, S., Wills, J. M., Hodgkins, C., & Raats, M. M. (2010). Penetration of nutrition information on food labels across the EU-27 plus Turkey. *European Journal of Clinical Nutrition*, 64(12), 1379–1385. <https://doi.org/10.1038/ejcn.2010.179>
- Storcksdieck genannt Bonsmann, S., & Wills, J. M. (2012). Nutrition labeling to prevent obesity: Reviewing the evidence from Europe. *Current Obesity Research*, 1(3), 134–140. <https://doi.org/10.1007/s13679-012-0020-0>
- van den Wijngaart, A. W. (2002). Nutrition labelling: Purpose, scientific issues and challenges. *Asia Pac J Clin Nutr*, 11(2), S68-71. <https://doi.org/10.1046/j.1440-6047.2002.00001.x>
- Verwaeren, J., Gellynck, X., Lagast, S., & Schouteten, J. J. (2019). Predicting children's food choice using check-all-that-apply questions. *Journal of Sensory Studies*, 34(1), e12471. <https://doi.org/https://doi.org/10.1111/joss.12471>
- Wartella, E. A., Lichtenstein, A. H., & Boon, C. S. (2010a). *Examination of front-of-package nutrition rating systems and symbols*. National Academies Press. <https://doi.org/10.17226/12957>
- Wartella, E. A., Lichtenstein, A. H., & Boon, C. S. (2010b). History of nutrition labelling. In E. A. Wartella, A. H. Lichtenstein, & C. S. Boon (Eds.), *Front-of-package nutrition rating systems and symbols: Phase I report*. National Academies Press.
- World Health Organization. (2021). *Action framework for developing and implementing public food procurement and service policies for a healthy diet*.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology:

Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100–

1122. <https://doi.org/10.1177/1745691617693393>

Yu, P., & Fu, M. (2020). *TPJF: Machine Learning Based Intelligent Prediction of Preference*

for Japanese Food; TPJF: Machine Learning Based Intelligent Prediction of Preference

for Japanese Food. <https://doi.org/10.1145/3441250.3441273>