Food Policy 131 (2025) 102826

Contents lists available at ScienceDirect

Food Policy

journal homepage: www.elsevier.com/locate/foodpol

Machine learning on national shopping data reliably estimates childhood obesity prevalence and socio-economic deprivation

Gavin Long [a], Georgiana Nica-Avram [a], John Harvey [a,*], Evgeniya Lukinova [a], Roberto Mansilla [a], Simon Welham [b], Gregor Engelmann [a], Elizabeth Dolan [a], Kuzivakwashe Makokoro [a], Michelle Thomas [b], Edward Powell [c], James Goulding [a]

[a] N/LAB, University of Nottingham, Nottingham NG8 1BB, UK
[b] School of Biosciences, The University of Nottingham, Sutton Bonington Campus, LE12 5RD, UK
[c] The Co-operative Group, One Angel Square Manchester, M60 0AG, UK

ARTICLE INFO

Keywords:
Deprivation
Obesity
Machine learning
Dietary Monitoring
Digital Footprints
Food Security

ABSTRACT

Deprivation pushes people to choose cheap, calorie-dense foods instead of nutritious but expensive alternatives. Diseases, such as obesity, cardiovascular disease, and diabetes, resulting from these poor dietary choices place a significant burden on public health systems. Measuring nutritional insecurity is difficult to achieve at scale and so the ability to study the relationship between nutritional outcomes and deprivation at a national level is very challenging. This makes it difficult to understand the effect of new policies or track changes over time. To address this challenge, we develop a machine learning approach using massive anonymised transactional data (4 million members and 2.5 billion transactions) in partnership with the retailer The Co-operative Group UK. We engineer a series of variables related to obesogenic diets, including a new measure called 'Calorie-oriented purchasing'. These variables help illustrate how large-scale transactional data can discriminate between neighbourhoods most affected by deprivation and childhood obesity. Through comparative assessment of machine learning approaches, we find better performance from tree-based models (Random Forest, XGBoost) with the best-achieving accuracy of 0.88 for predicting deprivation and an accuracy of 0.79 for childhood obesity. Calorie-oriented purchasing emerges as a robust predictor of deprivation and childhood obesity at the census area level. Results show this approach can help summarise nutritional insecurity, and support its spatio-temporal monitoring. We conclude with policy implications and recommend retailers adopt new measures for measuring national nutrition insecurity.

1. Introduction

The Academy for Medical Sciences has reported an 'appalling decline' in children's health in the UK, partly due to soaring obesity rates (PA Media, 2024). Childhood obesity, characterised by the WHO as "one of the most serious public health challenges of the 21st century" (World Health Organization, 2020), is a major and growing issue in the UK and globally (World Health Organization and others, 2016). Recent statistics from England show 10% childhood obesity at age 5, increasing to 23.4% by age 11 (Baker, 2022). While age 5 rates have remained relatively stable, obesity at age 11 has risen dramatically since 1990, along with associated health impacts like childhood type 2 diabetes (Davies, 2019), which disproportionally affect children living in the most deprived areas (The Food Foundation, 2023). The UK government aims to halve childhood obesity by 2030, but concerns persist about the effectiveness of this goal without more interventionist policies (Knai et al., 2016).

The links between nutrition, health and deprivation are well established (Tarasuk et al., 2013; Cetateanu and Jones, 2014; Howard Wilsher et al., 2016; Amin et al., 2021). Recent research has highlighted poor diet and experiences of food insecurity as important factors in lower life expectancy (The Food Foundation, 2023), compromised health and wellbeing (Tarasuk et al., 2013; Amin et al., 2021), increased levels of childhood obesity (Howard Wilsher et al., 2016; Cetateanu and Jones, 2014) and essential micronutrients deficiencies (Mansilla et al., 2024; Harvey et al., 2023). In the UK, Cetateanu and Jones (2014) and Howard Wilsher et al. (2016) used geographic analysis and parametric methods on large cross-sectional samples to examine these relationships and found positive associations between the presence of unhealthy food outlets and sales of unhealthy food and childhood obesity, and with area deprivation. As

* Corresponding author.
E-mail address: john.harvey@nottingham.ac.uk (J. Harvey).

https://doi.org/10.1016/j.foodpol.2025.102826
Received 14 February 2024; Received in revised form 6 February 2025; Accepted 10 February 2025
Available online 27 February 2025

these studies demonstrate, the causes of poor health are multifactorial, but the ranking of factors contributing to these findings remains poorly understood. This makes identifying areas most at risk resource-intensive. Implementing policies to reduce poverty and ensure optimal health requires current, accurate assessments of a population's health, deprivation, and nutrition. This paper investigates whether large-scale grocery purchase data can provide new deprivation measures and health insights to inform such policies.

In this study, we pursue two objectives. First, to evaluate the effectiveness of using supermarket food shopping data to predict area-level deprivation; and second, to explore how insights derived from food purchases offer additional information on childhood obesity, which can thus be used to inform policies focused on food provision and/or deprivation, such as the English Indices of Multiple Deprivation (Ministry of Housing, Communities & Local Government, 2019). These objectives revolve around two research questions:

RQ1 — Can large-scale transactional data on grocery purchases predict both deprivation and childhood obesity at the neighbourhood level?

RQ2 — What features derived from transactional data are most strongly correlated with deprivation and childhood obesity, and how might these metrics inform national food policy?

Deprivation metrics are typically applied to specific populations or areas, incorporating socioeconomic and/or health factors. For example, the Townsend Index relies on Unemployment, Non-car ownership, Non-home ownership, and Household overcrowding (Townsend et al., 2023). While numerous other deprivation metrics exist (Morris and Carstairs, 1991), we focus on England's Indices of Multiple Deprivation (Ministry of Housing, Communities & Local Government, 2019, IMD) — the predominant area-level deprivation measure in the UK. This index defines deprivation as a composite measure of societal and economic disadvantages at the local level, specifically at the geographic level of 'Lower-layer Super Output Areas' (henceforth 'LSOAs'[1]). The IMD incorporates seven domains to calculate a deprivation score that encompasses economic, social, and environmental factors. This score is commonly used to allocate resources for addressing socioeconomic inequalities. However, the IMD's complexity and data requirements mean it is only updated every 4–5 years, with the latest version released in 2019. Recent UK government obesity reports (Baker, 2022) clearly highlight the relationship between obesity and deprivation, as measured by the IMD, with levels of childhood obesity in the most deprived areas being approximately double those in the least deprived areas. The primary contribution of this work is therefore to demonstrate how transaction-level data can be used to predict area-level deprivation and childhood obesity.

Measures of diet are not included in the calculation of the IMD. This omission likely stems from a reliance on expensive and limited-coverage surveys to assess nutrition. For example, the UK's National Diet and Nutrition Survey (NDNS), like the US National Health and Nutrition Examination Survey, depends exclusively on self-reported dietary data. However, self-reported nutritional intake faces methodological challenges like selection bias and recall issues among participants (Kipnis et al., 2002). The complexity and cost of survey-based approaches make it difficult to link geographically focused diet insights with deprivation metrics. Alternative methods for assessing nutritional intake, such as through national-level shopping data, could potentially help predict area-level deprivation and health outcomes. This would support the development of new, rapidly updatable and multidimensional measures of deprivation. Moreover, as food sold in retail environments directly contributes to diet-related disease, such national sales data could illuminate the link between diet and health outcomes among populations most at risk, like children of school-going age with obesity.

As supermarkets are the predominant source of household food in high-income countries, grocery purchase records can offer valuable, rapid and localised insights into the diets of large groups of people (Timmins et al., 2018; Jenneson et al., 2022). There is precedent in re-purposing commercial data to estimate food and nutrient intake (Brimblecombe et al., 2013); monitor nutrient availability and population health (Aiello et al., 2019, 2020; Badruddoza et al., 2023); and evaluate the impact of interventions or policies (Andreyeva and Tripp, 2016; Schwartz et al., 2017; Amin et al., 2021; Berger et al., 2021) at population level.

These large-scale and granular data sources also enable novel data-driven approaches to hypothesis testing to be adopted in health research (Mooney and Pejaver, 2018), especially via the use of Random Forests models (Amin et al., 2021; Bannister and Botta, 2021; Badruddoza et al., 2023; Villacis et al., 2023). Studies using machine learning (ML) with grocery purchase records have shown the potential of these methods over traditional econometric models in predicting deprivation, as further discussed in Section 2.1.3. They also address the inconclusiveness of prior research in this area due to reliance on 'small population samples drawn from large urban areas'; much of the evidence comes from the USA, where, compared to the UK, the makeup of food environments may be different due to neighbourhood design and segregation (Cetateanu and Jones, 2014, p.68). Additionally, there is a reliance on the location and density of food outlets as proxies to diet without a clear understanding of their relationship with diet (Howard Wilsher et al., 2016).

There remains a significant need to improve testing of the generalisability of ML methods on new, unseen data, and to enhance the explainability of these methods by identifying key variables that drive the predictions, with methods such as Model Class Reliance (Smith et al., 2020; Dolan et al., 2023). Notably, previous work by Bannister and Botta (2021) used Tesco sales data from 1.6 million London customers in 2015 to estimate area deprivation levels via Random Forests models. However, the study was limited to Greater London, which had a restricted set of nutritional features and limited insight into variable importance measures. Our research seeks to build upon existing work by instead utilising national grocery purchase data from a different retailer, and expanding the range of food categories and associated nutritional features in the analysis. Additionally, we expand on predicting IMD deprivation by exploring measures of variable importance for the relationships between deprivation, diet, and childhood obesity.

To address the issues described above, we collaborated with The Co-operative Group (here referred to as the Co-op or The Co-op), a major UK food retailer, and obtained 30 months' worth of shopping transaction data for their members. Employing anonymised geolocation data provided by the retailer, we linked these transactions to LSOAs in the UK, census areas defined by the Office for National Statistics (ONS). We build upon previous research by introducing three metrics related to the obesogenicity of food purchases: Calorie-oriented purchasing (COP), Calorific Density (CD), and Obesogenic Potential (OP). We use grocery shopping data and the associated nutritional content to develop predictive models, and estimate area deprivation levels and childhood obesity rates. We then apply ML explainability tools to identify and compare key variables influencing these models.

The results of this method provide a three-fold contribution. First, we establish the viability of using aggregated food purchase transactions to predict area-level deprivation and childhood obesity. These rapidly updatable data sources and predictions can be used to augment deprivation statistics that are otherwise rarely revised. Second, we highlight insights into the relationship between diet, deprivation, and childhood obesity (specifically through calorie-oriented purchasing), whilst recognising deprivation as a challenging confounding factor in the intricate relationship between diet and health outcomes. Third, we use these insights to draw implications that could enhance national deprivation metrics and ensure that funding support is being deployed as efficiently and effectively as possible.

[1] For a full list of the acronyms used in this paper refer to Appendix A

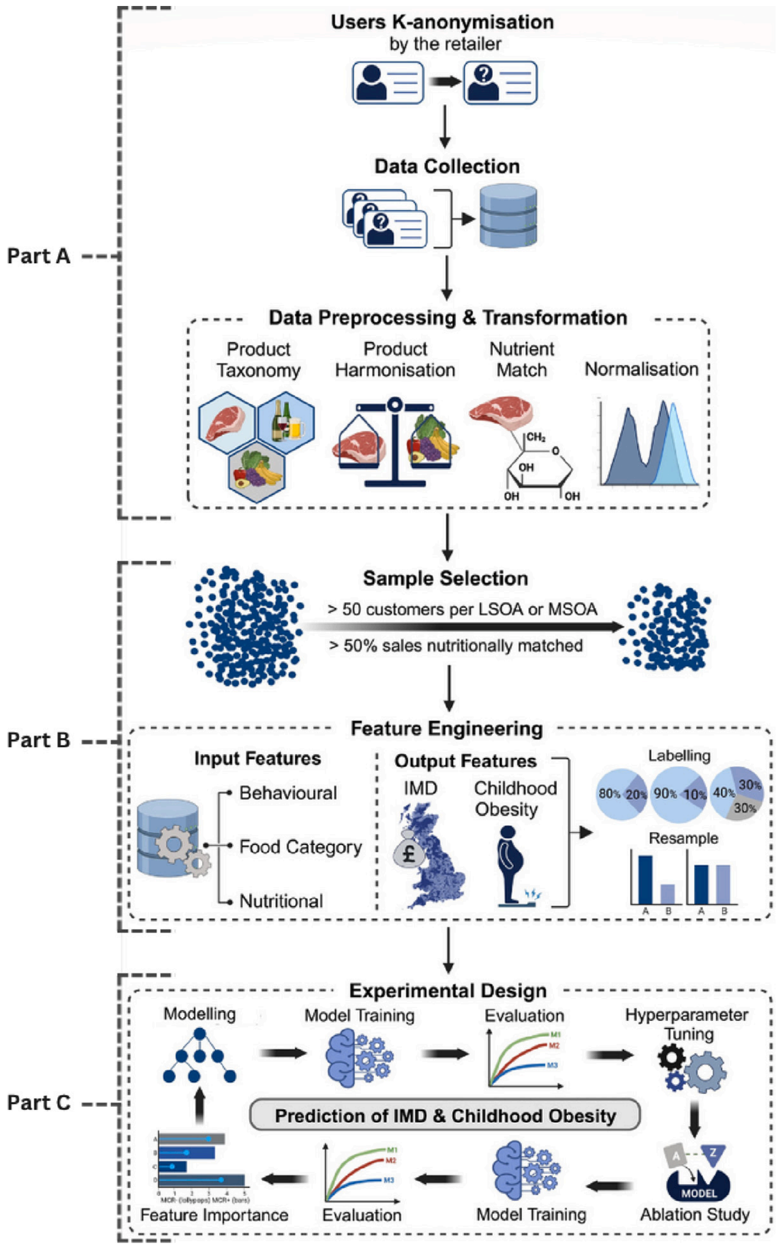

Fig. 1. The workflow diagram outlines the overall methodology used in this project. It illustrates each major work element in the order they were performed. Part A depicts data collection and preprocessing, Part B summarises the selection and feature engineering process, and Part C presents the iterative nature of the experimental design.

2. Materials and methods

This section details the data and methodology used to address the research questions (Refer to Fig. 1 for an overview of the methodology). We discuss the processes and rationale behind our data models, e.g., filtering, cleaning, and metric selection, as well as the development of the final methodology.

2.0.1. Retail data

Grocery shopping data was provided by Co-op Food, a large UK food retailer with over 2500 convenience stores and supermarkets. Their retail network coverage is the highest in the UK, with over 95% of the population living within 5 miles of a Co-op store (Rains and Longley, 2021).

The dataset observed grocery purchases made by Co-op members (equivalent to loyalty card holders in other retailers) between July 2019 and December 2021. This covered around 2.5 billion individual transactions for over 4 million members. Geo-location data anonymised to LSOA level was provided by Co-op to attribute purchases spatially while keeping member addresses anonymous (See Part A of Fig. 1). Aggregated data on dietary classification, nutrition and shopping behaviours were collated and extracted at the LSOA level for each quarter in the study period. Descriptions of specific data features are provided in Appendix B (Table B.1). More details on the methods used are described in Section 2.1.

2.0.2. Demographic data

The ground truth deprivation data paired with the retail data described was drawn from ONS Indices of Deprivation (IoD) datasets. Primarily, we included the IMD, which represents a weighted average score of the major subdomains of deprivation modelled by the IoD. The 2019 version of the IoD was used to match the time frame of the retail data. The IMD for an LSOA is typically represented by three values: the IMD score, its overall ranking, and its decile. For this study, we focused on the decile value of the IMD, allowing us to classify relative deprivation into 10 classes ranging from decile one (most deprived) to decile 10 (least deprived).

In particular, we were interested in the most deprived 20% of neighbourhoods in the UK. While there is no definitive threshold for what constitutes the most or least deprived, the top 10% and 20% of areas are considered eligible for 'remedial resources' through a wide range of intervention policies and funding to support struggling households (Bannister and Botta, 2021; Comber et al., 2022). Additionally, the most deprived 20% of the population is also used in NHS England's Core20PLUS5 programme (England, 2023), a health intervention policy operated in England by the NHS to reduce health inequalities. Accurately identifying such areas using only shopping data could be highly beneficial. It would provide a rapid, timely and context-sensitive methodological alternative to existing deprivation surveys, which are expensive and infrequent. Moreover, it could help socially-minded retailers like Co-op better understand the relative well-being of their members and support directed interventions for those most in need. The geospatial distribution of IMD deciles across England is shown in Appendix B B.1.

We sought to investigate the relationship between grocery purchases and deprivation, and between grocery purchases and health outcomes, especially as it affects children of school-going age in the UK. In England, the Office for Health Inequalities and Disparities (OHID) provides health data across a broad range of areas, including disease and poor health, lifestyle and mortality statistics as part of their local health portal (Office for Health Improvement and Disparities (OHID), 2024). This includes data on levels of childhood obesity recorded from NHS Digital's, National Child Measurement Programme (NCMP).

The NCMP data is available at the Middle Super Output Area (MSOA)[2] level of spatial aggregation and records the percentage of children measured as overweight or obese at school reception age (age 5) and school year 6 (age 11). The data used actual measurements of child weight and height measured by trained staff for over 95% of children between 2018 and 2019. Confidence in the accuracy of this data is extremely high since it is based on actual measurements in a controlled environment of an (almost) entire population rather than a sample or self-reported survey as is often the case with health data. This made the dataset of particular utility and value for our study, as we could be very confident that the data accurately reflected childhood obesity levels. Height and weight measurements are used to calculate the percentage of overweight and obese children based on their recorded Body Mass Index (BMI). For this work, we used the measure of relative obesity represented by the obesity quintile in the dataset. A mapping showing the distribution of childhood obesity at the MSOA level can be found in Appendix B Fig. B.2.

2.1. Methodology

We started by extending the work of Bannister and Botta (2021) to investigate whether secondary data from a UK-wide grocery retailer could provide timely and accurate deprivation estimates across neighbourhoods in England. To address RQ1, this study utilised grocery transaction data, aggregated at the neighbourhood (LSOA) level, to predict deprivation through machine learning (ML) methods. Key features included spending and purchase volumes across various food categories, their macronutrient composition, and derived nutritional indicators (Table 1), all available quarterly. The temporal framework was selected to align with the release timing of the latest deprivation data for England (last quarter of 2019).

[2] MSOAs are another geographic small area defined by the ONS. MSOAs are built from contiguous groups of LSOAs (usually four to five). Their resident populations are typically between 5000–15000.

Our initial analysis aimed to classify neighbourhoods into one of 10 relative deprivation deciles. While the models performed well in distinguishing the most and least deprived areas, they struggled to accurately classify neighbourhoods in the middle deciles (confusion matrix for the 10-classes classifier is shown in Appendix C, Fig. C.1), a challenge consistent with earlier findings by Bannister and Botta (2021) using Tesco data for the City of London. Since the focus of this study was on identifying highly deprived areas with significant levels of childhood obesity, we restructured the classification task into a binary model for greater interpretability. The binary model classified neighbourhoods into the most deprived quintile versus the rest (20/80, 2.1.2). For models involving childhood obesity data, which was available at the MSOA level, grocery data was aggregated accordingly to maintain consistency in granularity.

To address data imbalances inherent in each treatment of the binary classification task, we employed a stratified undersampling approach, ensuring that the proportions of deprivation levels were retained in the resampled dataset (Cateni et al., 2014). An alternative option might be application of oversampling techniques, such as SMOTE, that generate synthetic training examples to address imbalance. While such approaches can often improve model accuracy, especially in situations with limited data and well-defined feature distributions (Joloudari et al., 2023), use of synthetic samples does bring with it a risk of overfitting and making interpretation and explanation of model predictions more challenging (data being less representative of real-world observations — a factor potentially important in domains such as health and social care) (Wongvorachan et al., 2023). As such, and given the large number of datapoints available to our analysis, a stratified downsampling strategy was preferred on this occasion.

Feature selection played a critical role in improving model performance and interpretability. Starting with 92 features grouped into 10 domains (detailed in Table B.1), we used an ablation study to reduce redundancy and focus on the most predictive variables. Features were ranked using Shapley values (SHAP) to identify their contribution to model predictions, and less informative variables were systematically removed. This was followed by model class reliance (MCR, detailed in Section 2.1.6) resulting in a streamlined set of 39 features, achieving dimensionality reduction while maintaining classification accuracy.

This methodological framework demonstrated the potential of ML for predicting deprivation-related conditions, offering significant advantages over traditional econometric models (Amin et al., 2021; Villacis et al., 2023; Hossain et al., 2019; Bannister and Botta, 2021; McBride et al., 2022). Unlike conventional approaches, ML models are highly flexible and data-driven. These attributes make ML particularly well-suited for analysing complex relationships, such as those between deprivation and various predictors (input features), where the underlying patterns may be unknown or difficult to capture using parametric methods. By leveraging the high-dimensionality and richness of transactional data (2.5 billion grocery purchases for over 4 million members), this study contributes to the growing body of work showcasing how non-traditional data sources can provide actionable insights into socioeconomic conditions. The next sections detail the technical aspects of feature engineering, model selection, development, evaluation, and interpretation.

2.1.1. Feature engineering

Summary and descriptive statistics for the key features used in this work, and described in more detail below, can be found in Appendix B Table B.2 (for deprivation related features) and Table B.3 (for childhood obesity related features).

Dietary categorisation.

Products were classified using a taxonomy according to 17 broader categories, including grains, dairy, red meat, fish, soft drinks, etc. (a full list is given in Appendix B – Table B.1). An additional non-food category of cigarettes was also added to the data model, bringing the total number of categories to 18.

Behavioural shopping data.

Several features were identified and derived from the transactional data to represent behavioural patterns in the shopping records for each LSOA. These ranged from simple measures of the total number of customers, shopping baskets, stores and spending to more complex metrics, including average bundle entropy of customers (Mansilla et al., 2022), average products per shopping basket and average weekly spend per customer.

Additional measures relating to the number of shopping trips made at different times of the day (e.g. morning, afternoon, evening) were also included to examine whether the timing of shopping trips might have been in some way related to local health and deprivation.

Nutritional mapping exercise and metrics.

As one of the primary aims of this work was to examine potential links between nutrition and local deprivation and health outcomes, linking shopping purchases to their nutritional content was an essential element (See Part A of Fig. 1). However, this was challenging due to over 31,000 food products in the retail dataset. Each product had a text string called *item description*, which typically included the brand name, product name, and product size (in that order). However, not all item descriptions exactly followed this pattern. Products were also categorised into a three-level hierarchy based on their *department*, *section*, and *subsection*. This was the only information available to match the products to their nutritional content.

Three data sources were used to map grocery products to their nutritional content:

1. UK Composition of Foods Integrated Data set (CoFID) (GOV.UK, 2021), also referred to as 'McCance and Widdowson'.
2. Co-op's own brand nutritional data set provided by the retailer. This data did not have links to the product database, and nutritional mapping was still required.
3. The online Nutritics data set (Nutritics, 2024), an online food database containing nutritional data on over one million food products.

Where these data sources lacked nutritional information for high-sales products among our customers, the content was manually calculated from the ingredient list found via web search.

Retailer's own brand items were matched to their nutritional data source where possible as this represented the closest nutritional match. Aside from own-brand products, a three-stage method for matching nutritional data to product data was employed in this study. This process was performed for each of the three nutritional data sources in order of their primacy (i.e., CoFID, retailer nutrition, Nutritics). CoFID was selected as the primary nutritional data source as it contained a broader range of macro- and micro-nutrient data than the other nutritional data sources.

The three stages of nutritional mapping were:

1. Product and nutritional data sets were imported to a PostGreSQL database. The PostGreSQL *similarity* function, which performs fuzzy text matching using trigrams, was used to compare the item descriptions in both data sets and return those with a similarity above a specified confidence level. This level was initially set to 0.9 and then iteratively reduced to 0.5 to maximise the number of matches. Matches from each iteration of the confidence level were manually verified. Verified matches were added to a nutritional look-up table.
2. Unmatched products after step (1) were then matched using the *department*, *section*, and *subsection* fields in the product table. These fields were compared to the food description in the nutritional data source(s) using the PostGresSQL *similarity* function. Where verified matches were identified by this method, all products in the matched *department*, *section*, and *subsection* were added to the nutritional look-up table.

Table 1
Overall nutritional mapping results.

Data Source	Products matched	Matched Sales	Matched Spend	Percentage of total sales	Percentage of total spend
CoFID	4987	1,308,047,579	£1,977M	43.98	39.49
Retailer	5001	1,347,068,459	£2,374M	45.29	47.44
Nutritics[a]	214	44,405,507	£62M	1.49	1.25
TOTAL	**10,202**	**2,699,521,545**	**£4,413M**	**90.77**	**88.18**

[a] Also includes product matches based on manually calculated nutrition from product ingredients.

3. Food products still unmatched after steps (1) and (2) were then manually matched using their sales volume to prioritise matching, i.e., highest-selling products were matched first. This process was repeated until a specified sales volume threshold was reached.

Using the method described, just over a third of the total food products in the retailer's data set were matched to the nutritional data sources. This represented around 90% of total sales quantity and spend on food items in the retail data set (Table 1).

Nutritional content mapped to products (where available) included all macro-nutrient data (e.g. calories, protein, fat, sugar, fibre, etc.) and selected micro-nutrients, depending on the specific nutritional data source used. Nutritional features used in the analysis are described in Appendix B. As stated in the introduction, this work was focused on identifying and assessing nutritional metrics that could be useful in the assessment of local health and deprivation outcomes. To this end, three metrics derived from the nutritional data matched to grocery purchases were generated:

- ***Calorie-oriented purchasing*** (COP) was an extension of relative calorific pricing (RCP) or energy cost metrics as previously described in research into calorie-dense products and their links with health, diet and poverty in the literature (e.g., Headey and Alderman (2019), Drewnowski and Specter (2004)). Here, though, we extended the measure to the level of the individual shopper and/or local population by summing their total choices through time, so it thus became a general measure of the behavioural tendency to maximise calories for money spent, rather than a product-focused value. We calculated the measure where 'calories' is the total calories (kCals) in nutritionally matched food purchases and 'spend' is the sum of money spent on those food purchases (£) for a given geographic area and time period.

$$COP = \frac{\Sigma Calories}{\Sigma Spend} \tag{1}$$

  Variables were aggregated across quarterly time periods, and at LSOA (for the IMD model) and MSOA (for the childhood obesity model).
- ***Calorific Density*** (CD), or energy density (Drewnowski and Specter, 2004), is another measure of the calorific content of food normalised by weight. Examples of foods with high CD include fats, oils and red meat.

$$CD = \frac{\Sigma Calories}{\Sigma Weight} \tag{2}$$

  CD was assigned using the same method as COP but calculated by dividing the total weight (in kg) of matched food purchases by the total matched spend (in £)
- ***Obesogenic Potential*** (OP) is the proportion of calories obtained from fats and sugar. It shares similarities with CD but is a more nuanced measure and is more closely correlated with poor health outcomes where diets have higher levels of associated obesity. OP was assigned at the same level of spatiotemporal aggregation as COP and CD and is calculated using the following equation, where each nutrient is the sum (in grams) of that nutrient in matched

food purchases for the given area and period:

$$OP = \frac{(\Sigma fat \times 9) + (\Sigma sugar \times 4)}{((\Sigma fat \times 9) + (\Sigma carbohydrate \times 4) + (\Sigma protein \times 4))} \tag{3}$$

2.1.2. Outcome variables

We used food purchased by customers living in 32,844 LSOAs to predict the deprivation level of that area, as measured by the IMD. The modelling of deprivation included 10,547 LSOAs with at least 50 customers and at least 50% of food items matched against their nutritional composition (See Part B of Fig. 1). As previously detailed, the binary classifier (20/80 split) was designated as the main model. To test the deprivation predictor performance further and investigate the impact of attempting to classify different levels of deprivation, further ML models were developed. These included a model to identify the most deprived decile (a 10/90 split model) and a ternary-based model (using a 30/40/30 split) to classify deprivation into three categories. The ternary model was developed to see if the method was capable of discriminating into three groups representative of low, medium and high levels of relative deprivation.

We were also interested in using our engineered features to predict childhood obesity prevalence at school reception age and year 6, as recorded by the NHS Digital's NCMP. Data on levels of obese children in the 7201 MSOAs in England were collected and ranked to assign each area into quintiles, where the first quintile represent the lowest 20% of MSOAs with obese children and the fifth quintile represent the areas in the highest 20%.

2.1.3. Machine learning model selection

In binary classification, algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), XGBoost (XGB), Decision Tree (DT), and Random Forest (RF) are commonly utilised. However, given the large and high-dimensional dataset in this study, RF and XGB were particularly well-suited. These algorithms excel in handling complex, non-linear relationships and are resilient against outliers, noise, missing data, and feature correlation (Ali et al., 2012; Fernández-Delgado et al., 2014), which are typical in real-world datasets. While models like SVM and DT can struggle with high-dimensional data and overfitting issues (Singh et al., 2016; Guyon and Elisseeff, 2003), ensemble methods such as RF and XGB are designed to mitigate these challenges, improving accuracy and robustness. Nevertheless, simpler models like LR offer distinct advantages in terms of interpretability. Given these considerations, RF, XGB, and LR were chosen for this study, striking a balance between predictive performance and model transparency.

2.1.4. Training the models

Before training the models in an iterative process (See Part C of Fig. 1), we carried out a correlation analysis between the index of multiple deprivation (IMD) deciles, childhood obesity quintiles and food purchased in UK neighbourhoods. We assessed relationships between features using Pearson's correlation with Bonferroni's correction to limit the risk of false discoveries of significant results, given the large number of highly correlated features included in the hypothesis testing. Due to the large number of features tested, we limited the resulting correlation matrices to only include those features showing a significant correlation (correlation coefficient >+/- 0.2). The resulting matrices can be found in the appendices (see Figs. C.2 (IMD) and C.3 (childhood obesity at year 6).

We trained all classifiers in a similar way. We provided details for the RF model used to predict neighbourhood-level deprivation from the grocery features. We treated this as a classification task whereby the outcome variable denotes whether an area is more or less deprived along the deciles of the composite IMD. Data was split using stratified random sampling into training (80%) and test sets (20%), and each model's performance was tested using the held-out set. Grid search was also used to create an exhaustive search through the predefined hyperparameters of the RF (Dusmanu et al., 2017). Once the final model was selected, its performance was tested on the unseen (20%) test data. Testing the model on data previously unseen by the algorithm is a valuable indicator of how the model will generalise to areas previously not included in the modelling (Yarkoni and Westfall, 2017). The model evaluation is described further in Section 2.1.5.

The childhood obesity classification models used a subset of the grocery-related features employed in the deprivation classifier. Features related to total and proportional spending were excluded, as were any shopping categories inapplicable to children, i.e. alcohol and cigarettes.

2.1.5. Evaluation of model performance

The ML models described were evaluated against a range of standard criteria. These included overall accuracy, providing a broad sense of model effectiveness, precision, assessing the percentage of correctly classified results in the positive/relevant class (an important measure given the potential consequences of misclassification in policy and resource allocation), and F1-scores, combining both recall (the percentage of total relevant results correctly classified) and precision to indicate performance whilst balancing the false classification of more and less deprived areas (Pitsilis et al., 2018).

A dummy classifier, stratified on the dependent variable classes, was also generated to assess the performance of the ML models. Comparing the accuracy of the models against this baseline was crucial to determine whether the complex models were genuinely learning meaningful patterns in the data or merely performing at the level of random chance. Cohen's Kappa scores were also calculated and included in the results to provide an additional indication of the model's performance.

2.1.6. Model interpretation

We used a combination of ML techniques to examine which variables were the most important and to what extent they mattered. Although RF provides some insights into feature importance (i.e., can be used to rank the importance of features on the classification task), it only considers a single predictive model and not others that may perform equally or better (Altmann et al., 2010). Therefore, results and interpretation might be biased towards that specific model by just learning one of many equally well-performing relationships between the input features and the outcome. This means variables may be incorrectly considered unimportant, that model audits may not be robust to model retraining, and that the interpretation of potential causal features may be incomplete and misleading (Biecek et al., 2024). The same happens with other variable importance techniques, such as Unconditional Permutation Importance, Gain, Split Counts, and Minimal Depth.

Since we also sought to interpret the relationships that exist between identified levels of deprivation and our three derived nutritional metrics, as these would be of interest to policymakers, we derived SHapley Additive exPlanations (SHAP, (Futagami et al., 2021)). To further ensure that our feature importance results were consistent and not the cause of random fluctuations in a given model, Model Class Reliance (MCR) was used to fully explore the impact of our key features on model performance and accuracy (Fisher et al., 2019).

Although SHAP values are computationally expensive, they offer a valuable mechanism for understanding the relationship between predictors and the predicted output in ML models. SHAP values designate a weight to each predictor by evaluating all possible feature combinations, emphasising their relative relevance in determining the model's predictions. A positive SHAP value indicates an increase in the predicted response relative to the average prediction, whereas a negative value indicates a decrease. Predictors with SHAP values of zero do not deviate from the overall prediction. Hence, SHAP values provide a comprehensive and interpretable understanding of how each feature affects the model's output.

MCR is a measure that estimates the degree to which a given variable is relied upon by a set of models that achieve comparable predictive performance but employ distinct predictive factors. This set of models is known as a Rashomon set (Fisher et al., 2019). In contrast to other methods that only evaluate a variable's significance within a single predictive model, MCR provides a range of values that illustrate the variable's reliance across an entire set of models. The maximum model class reliance (MCR+) measures the maximum possible change in predictive performance that can be attributed to a specific variable within the Rashomon set. Conversely, the minimum model class reliance (MCR-) represents the minimum possible change in predictive performance attributable to the variable within the Rashomon set. Smith et al. (2020) extended MCR to non-linear algorithms, allowing MCR calculation for regression and classification Random Forests. By using both SHAP and MCR values, we gained a more comprehensive understanding of the significance of derived features in predicting childhood obesity and deprivation.

2.1.7. Implications of feature importance insights

Utilising SHAP and MCR features in machine learning models provides a thorough understanding of the factors affecting predictions and offers valuable opportunities for future research and targeted policy interventions.

In contrast to traditional feature importance methods, SHAP and MCR not only identify which features are influential but also elucidate the direction and magnitude of their effects. This facilitates a more nuanced comprehension of the complex relationships within the data. For researchers, these insights can inform the selection of variables for further investigation, particularly in longitudinal studies aimed at uncovering causal relationships between socioeconomic factors, dietary habits, and health outcomes.

For policymakers, feature importance analysis can reveal actionable areas for intervention. For instance, if variables related to the proportion of spend on nutrient-dense foods emerge as a significant predictor, it could guide the development of subsidy programs or public health campaigns that promote healthier choices. Furthermore, insights into behavioural patterns, such as shopping habits at different times of day or frequency of purchases, can help tailor interventions to specific community needs, thereby enhancing their effectiveness.

3. Results

Grocery purchases for over 4 million Co-op members covering around 2.5 billion transactions over a 30-month period were categorised, their nutritional value was calculated, and shopping behaviours were assigned to each LSOA and MSOA in England for each year and quarter in the sample. Data for the 4th quarter of 2019 was used in the ML results described in this section to closely match the time period of the deprivation and childhood obesity datasets that were the target variables. The results described relate solely to this time period, but we also tested the consistency of our results using data until the end of 2021. The consistency of model accuracy across the time frame of the data available is shown in the supplementary materials provided in Appendix C (Fig. C.4). As shown, the initial results were found to be consistent through time, including during and after the COVID period.

Only LSOAs/MSOAs with greater than 50 customers and with at least 50% of their food sales nutritionally matched were included in the sample set. This amounted to 10,547 LSOAs and 4667 MSOAs used in training and testing the ML classifiers.

3.1. Deprivation classifier results

Results of predicting deprivation using a binary classifier are shown in Table 2. All models perform significantly better compared to a dummy classifier (~50% accuracy). The highest performing models are the tree-based models for the 10/90 classifier (86%–88% accuracy) using all 92 features. However, the 20/80 classifier also achieves over 80% accuracy and remains accurate when using the reduced set of input features. Model accuracy across all three ML types (LR, XGB, and RF) is consistent in the majority of cases. Both binary classifier models show similar accuracy in identifying the most deprived 10%–20% of LSOAs and the less deprived 80%–90%.

Table C.1 assesses the performance of the ternary classifier using similar metrics. Overall accuracy is significantly reduced compared to our binary classifiers. However, the increase in performance over the dummy classifier is of similar magnitude to the binary classifiers. As shown in Table C.1, the ternary classifier's performance at identifying areas with extreme (low/high) deprivation areas is similar to the binary classifier models (~75%–80%). The ternary classifier's ability to discriminate areas of medium deprivation is where the drop in accuracy can be seen (e.g., LR having ~ 30% accuracy). The tree-based models are significantly better at ternary classification overall, with both RF and XGB showing similar levels of accuracy. Despite the drop in accuracy, these results demonstrate that a ternary model is capable of classifying deprivation into three categories — low, medium and high.

3.2. Childhood obesity classifier results

Results for predicting the areas with high levels of childhood obesity are shown in Table 3. In this instance, only a 20/80 classifier is shown. This was chosen as it provides a good comparison for our optimised IMD 20/80 classifier. Results for classifying childhood obesity at reception age and year 6 are shown next. The year 6 classifier is more accurate across all model types, with the tree-based models showing the highest overall accuracy (79%). The best-performing model at reception age, logistic regression, achieves an accuracy of 71%.

To ensure that the performance of the classifier was not significantly affected by seasonality, models were run for multiple quarters. In Appendix C (Fig. C.4), accuracy across all three classes of models was consistent throughout the time period considered with less than 10% difference in model accuracy across all quarters tested and an average overall accuracy of 80%. Notably, the models with the lowest accuracy are those for the second quarter of 2020. We hypothesise that this deviation can be accounted for by the Coronavirus pandemic and the first UK lockdown that occurred in this quarter. This would likely have led to significant changes in shopping behaviours and grocery purchases. Nevertheless, the model accuracy for this period was still high (76%–79%).

3.3. Feature importance

To address RQ2, it is necessary to assess the relative importance of each of the input features used in the ML classifiers developed. In addition to standard methods of ranking each feature, SHAP and MCR analysis was performed for each ML classifier as outlined in Section 2.1.6, to evaluate feature importance.

3.3.1. IMD classifier

SHAP summary plots for the RF classifiers predicting the most deprived 10% of LSOAs against the remaining 90% are shown in Fig. 2. A similar SHAP plot for the 20/80 split classifier is shown in Fig. 3. The top 20 features are ranked in terms of their importance, spending on foods in the "ready-made" category being the most important feature in the model using all 92 features. SHAP values are very useful in understanding the influence each feature has, and show how much each contributes to moving a prediction away from the average model prediction (i.e., the baseline prediction if we knew nothing about the feature values). The colour shows whether that feature was high (red) or low (blue) for that row of the dataset, and the SHAP value's sign indicates the *direction* it moves the prediction.

By way of example, in Fig. 2, observing high sales of ready-made food purchases (indicated in red) is shown to push the prediction

Table 2
Results of binary classifier from experimental models predicting extreme deprivation using food shopping data from October 2019 to December 2019.

Threshold	Inputs	Model	Results(ALL areas)				Results (Least deprived areas)				Results (Most deprived areas)			
			Acc	Prec	F1	Kappa	Acc	Prec	F1	Kappa	Acc	Prec	F1	Kappa
10–90 (deprived/not) percentile split	All	Dummy	0.51	0.51	0.50	0.02	0.51	0.50	0.52	0.02	0.50	0.51	0.49	0.0
		LR	0.83	0.83	0.83	0.64	0.84	0.84	0.84	0.65	0.82	0.83	0.84	0.63
		XGB	0.88	0.88	0.87	0.75	0.88	0.87	0.88	0.75	0.87	0.88	0.87	0.74
		RF	0.86	0.87	0.86	0.73	0.89	0.89	0.86	0.76	0.83	0.84	0.87	0.64
	Reduced	LR	0.78	0.78	0.78	0.59	0.78	0.78	0.78	0.59	0.78	0.78	0.78	0.59
		XGB	0.71	0.71	0.71	0.43	0.71	0.71	0.72	0.43	0.71	0.72	0.71	0.43
		RF	0.74	0.74	0.74	0.49	0.74	0.74	0.75	0.49	0.75	0.75	0.74	0.50
20–80 (deprived/not) percentile split	All	Dummy	0.47	0.47	0.47	−0.06	0.47	0.47	0.49	−0.06	0.47	0.47	0.46	−0.06
		LR	0.82	0.82	0.82	0.63	0.80	0.81	0.82	0.61	0.84	0.83	0.81	0.65
		XGB	0.81	0.81	0.81	0.62	0.81	0.82	0.81	0.62	0.80	0.80	0.81	0.61
		RF	0.81	0.81	0.81	0.62	0.82	0.82	0.81	0.63	0.80	0.80	0.81	0.61
	Reduced	LR	0.80	0.80	0.80	0.61	0.80	0.80	0.80	0.61	0.80	0.80	0.80	0.61
		XGB	0.82	0.82	0.82	0.64	0.82	0.83	0.82	0.64	0.82	0.81	0.82	0.64
		RF	0.82	0.82	0.82	0.64	0.82	0.85	0.82	0.64	0.80	0.81	0.83	0.61

Table 3
Results of binary classifier from experimental models predicting childhood obesity using food shopping data from October 2019 to December 2019.

Target	Model	Results(ALL areas)				Results (Least obese areas)				Results (Most obese areas)			
		Acc	Prec	F1	Kappa	Acc	Prec	F1	Kappa	Acc	Prec	F1	Kappa
Reception (Age 5)	Dummy	0.51	0.51	0.51	0.01	0.53	0.51	0.52	0.01	0.48	0.51	0.49	−0.01
	LR	0.71	0.71	0.71	0.43	0.74	0.70	0.72	0.45	0.68	0.73	0.70	0.40
	XGB	0.64	0.64	0.64	0.36	0.66	0.64	0.65	0.37	0.63	0.65	0.64	0.35
	RF	0.69	0.69	0.69	0.41	0.67	0.70	0.68	0.39	0.71	0.68	0.70	0.43
Year 6 (Age 11)	Dummy	0.50	0.50	0.50	0.0	0.48	0.50	0.49	−0.01	0.51	0.50	0.51	0.01
	LR	0.77	0.77	0.77	0.55	0.80	0.76	0.78	0.58	0.75	0.79	0.77	0.53
	XGB	0.79	0.79	0.79	0.57	0.78	0.78	0.79	0.56	0.79	0.79	0.79	0.57
	RF	0.79	0.79	0.79	0.57	0.75	0.80	0.78	0.53	0.82	0.77	0.79	0.60

towards lower deprivation (positive SHAP values), while low ready-made food purchases (indicated in blue) push the prediction towards higher deprivation (negative SHAP values). The magnitude of the SHAP value on the y-axis of the graph indicates the strength of this effect – larger absolute values indicate stronger influence on the prediction – and it is clear from the plot that ready-made meals is the feature most predictive of IMD levels.

Fewer purchases of ready-made foods, fish, wine, and fruits and vegetables are associated with higher deprivation in both models. In comparison, the SHAP plots show that areas with higher sales volumes of soft drinks and cigarettes (shown in red) are associated with higher levels of deprivation in both models (negative SHAP values), although their relative importance is higher in the reduced feature set model due to the removal of other features that shared information with these features, e.g. proportion/average spend of soft drinks/cigarettes. Our COP metric shows high importance in both models, with high values of COP associated with areas with the highest levels of deprivation. This indicates, as expected, that shoppers in deprived areas are getting more calories for every pound spent. The other two nutritional metrics developed, CD and OP, do not appear to be of high importance according to their SHAP values.

In terms of shopping behaviour features, a higher number of customers shopping later in the day (afternoon and night) is associated with areas of higher deprivation, potentially due to differences in employment in such areas (e.g., shift working and higher levels of unemployment). It also appears that customers in areas of extreme deprivation make more shopping trips (higher average basket count) and purchase more products. In the optimised, reduced feature set model, see Fig. 2 (right), features relating to macro-nutrient intake are shown to be of importance. Higher proportions of fibre and protein in food purchases are associated with areas of least deprivation. This result matches those identified in terms of the proportion of fibre in food purchases according to Bannister and Botta (2021) in their study of grocery purchases from Tesco in the city of London, UK.

MCR results for the reduced model predicting LSOAs in the most deprived 20% can be seen in Fig. 4. MCR values for each individual feature are shown, highlighting that several features are responsible for the model's accuracy.

The minimum importance of each feature (MCR-) is represented by the *lollypop* in the chart. Where this value is above zero, notably in the case of wine_spend_p, softdrinks_quantity_p and kcals_per_pound, this indicates that those features are required across all best-performing models. Features with MCR- values of zero are not essential in the set of best-performing models and could be replaced by another feature with which they could share information. For example, the proportion of fibre in grocery sales in an LSOA could be correlated with the proportion of fruit and vegetables sold in the same area since that food category is high in fibre.

3.3.2. *Childhood obesity classifier*

The SHAP summary plots for the classifiers used to predict areas in the lowest quintile for childhood obesity at reception age and year 6 (ages 5 and 11, respectively) are shown in Fig. 6. It is evident from the SHAP summary charts that feature importance for the two childhood obesity classifiers are similar, sharing five of the top seven features. The main differences between the two models appear to occur for the features with lower overall importance.

In both models, the proportion of soft drinks sold in each MSOA and the COP show the greatest importance. Areas with higher sales of soft drinks and grains and higher levels of COP (represented by red) lead to a lower SHAP value, indicating the 20% most obese areas. Conversely, higher sales of fruit and vegetables and fish lead to a higher SHAP value, which relates to the 80% of areas with lower levels of childhood obesity. Beyond the top-performing features, the best-performing model predicting obesity at year 6 (Fig. 5 [right]), reveals additional findings:

- Areas with higher sales of products that have higher levels of fat content (dairy products, saturated fats, fats and oils) are associated with lower levels of childhood obesity.

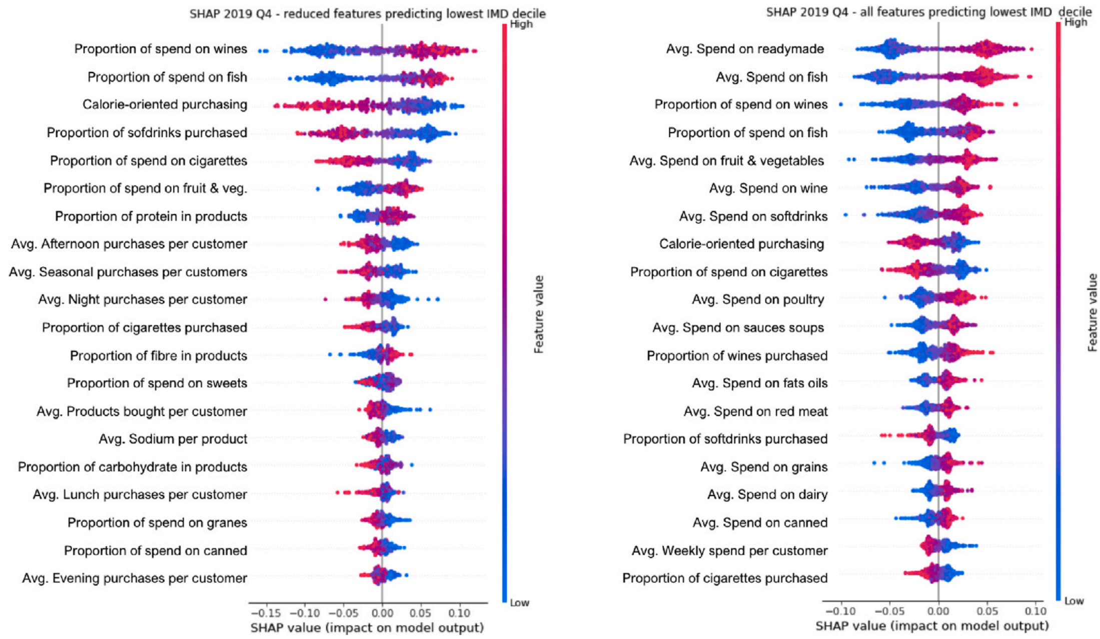

Fig. 2. SHAP summary plots for RF models classifying the most deprived decile. Right — Classifier using all 92 features. Left — Classifier using reduced feature set. Red values indicate the feature is a high value, Blue a low value. Positive SHAP values indicate the feature is shifting predictions towards a higher IMD estimate (Low Deprivation) and negative ones are shifting predictions towards a lower IMD estimate (High Deprivation). The magnitude of the SHAP value indicates the strength of that shift — for example the model has learnt that high calorie oriented purchasing (red) is a strong component in accurately predicting higher deprivation levels (negative SHAP). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

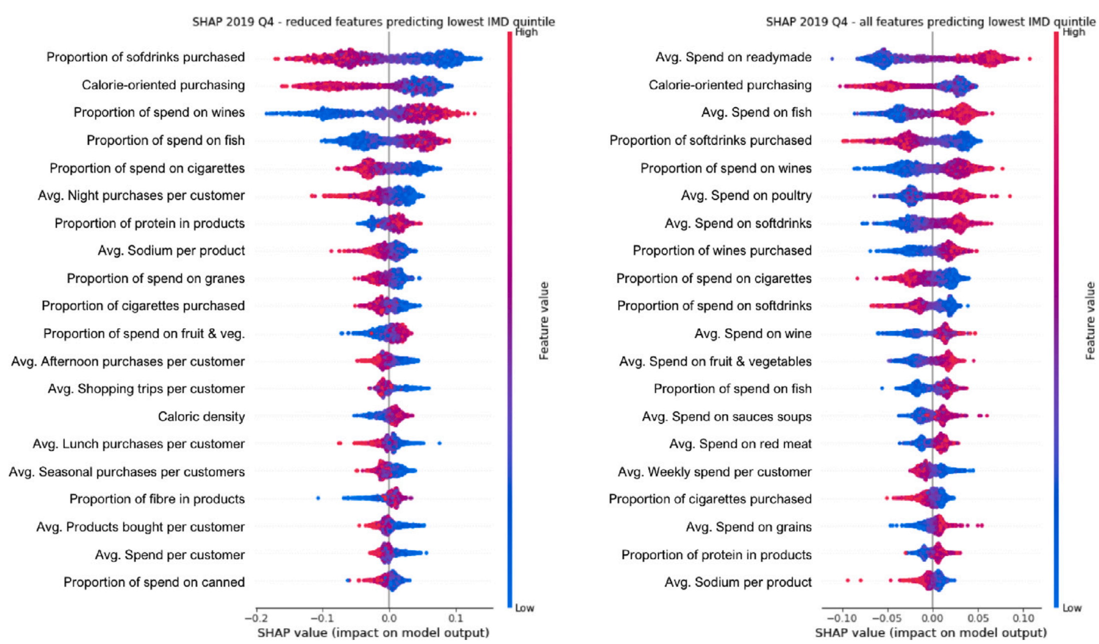

Fig. 3. SHAP summary plots for RF models classifying the most deprived quintile. Left — Classifier using reduced feature set. Right — Classifier using all 92 features. SHAP results broadly echo results for deciles (see Fig. 2), but high spend on soft drinks (red) and calorie-oriented purchasing (red) is even stronger in predicting the most deprived areas in quintiles (negative SHAP values) — perhaps due to the lowest quintile having slightly increased disposable income over the lowest decile, hence allowing such predictors to be more expressed in shopping behaviours. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- Similarly, areas with lower levels of CD are associated with lower levels of childhood obesity. Therefore, obesity policy should perhaps be more suitably focused on COP statistics than CD or OP when considering transactional data.

MCR was performed for both childhood obesity classifiers and the results are shown in Figs. 6 (reception age classifier) and 7 (year 6 classifier). MCR results for the two models show significant differences despite sharing similarities in the important features (e.g. soft drinks, COP, grains, fish and dairy).

The MCR results for the reception age classifier (Fig. 6) show a far greater number of features with similar levels of maximum importance (MCR+), represented by the size of the bars. Almost all features have a minimum importance (MCR-) greater than zero. This indicates that all the features are required to produce the set of best-performing ML models. The sales of soft drinks and COP have the highest values of both MCR+ and MCR- highlighting their criticality in predicting reception age obesity.

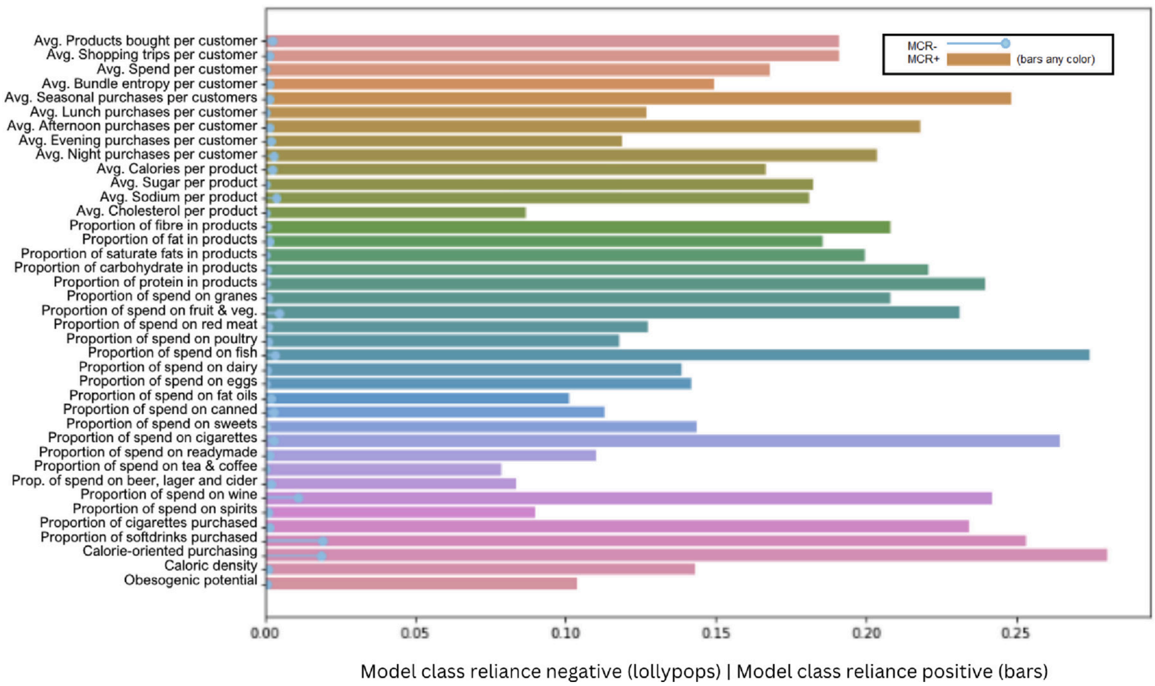

Fig. 4. Model Class Reliance chart showing feature importance across multiple RF models for the classifier using the reduced feature set to predict the most deprived 20% of LSOAs. MCR- (lollipops) and MCR+ (bars) represent the lower and upper bounds of each feature's importance, illustrating variability across models.

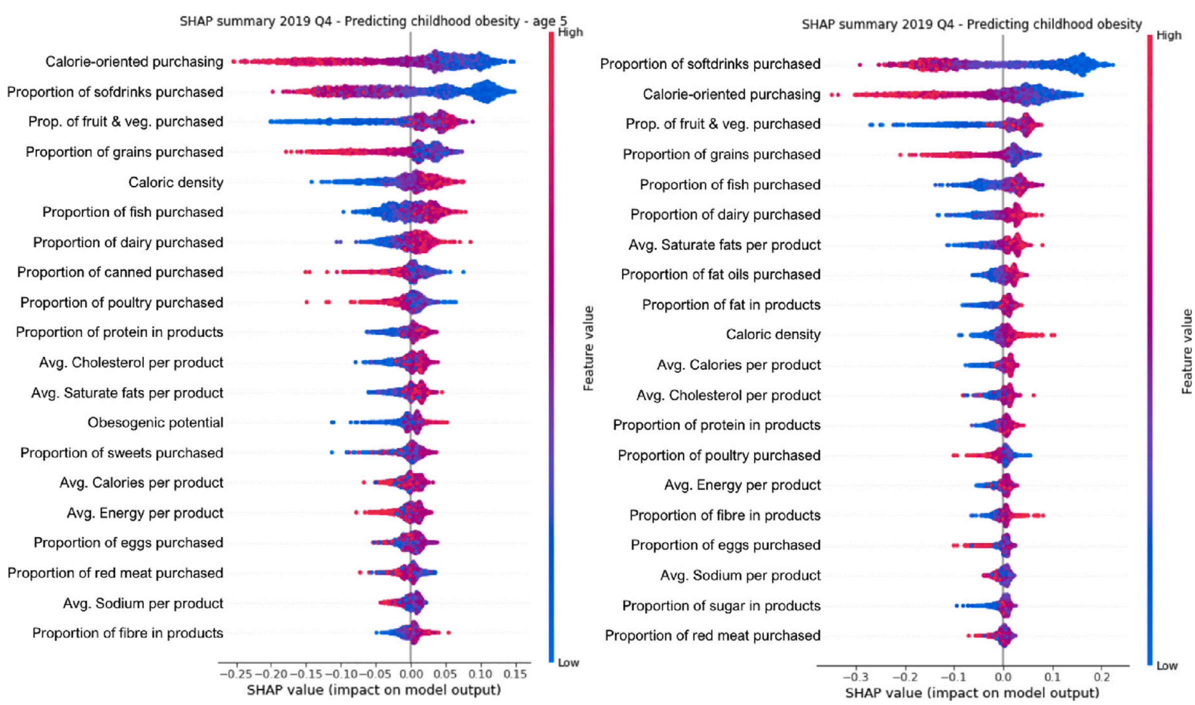

Fig. 5. SHAP summary plots for RF classifier predicting extreme quintiles of childhood obesity at [left] reception and [right] year 6. Here high calorie-oriented purchasing (left — red) is even more clearly indicative of predictions, intuitively shifting predictions towards predictions of higher childhood obesity risk (negative SHAP values) — and at a notably higher magnitude than in deprivation predictions, at > −0.25 in several cases.

Results for the year 6 classifier show a much greater divergence between the importance of features. The importance of soft drinks sales and COP is far greater than the other features in the model. Soft drinks sales show the highest potential importance, but COP has a higher minimal importance level. As shown in Fig. 7 the minimum importance of these two features exceeds the potential importance of the majority of the other features in the model. Only a small number of features have potential importance levels higher than the minimal level of soft drinks and COP (grains, proportion of fat and saturated fat, total carbohydrates and average number of baskets bought by customers). Although many of the remaining features show little potential importance, there are several features whose minimum importance level (MCR-) is above zero, indicating that they are required, even if their overall contribution is minimal. These include OP, total fibre, and the sales of fish, dairy and poultry products.

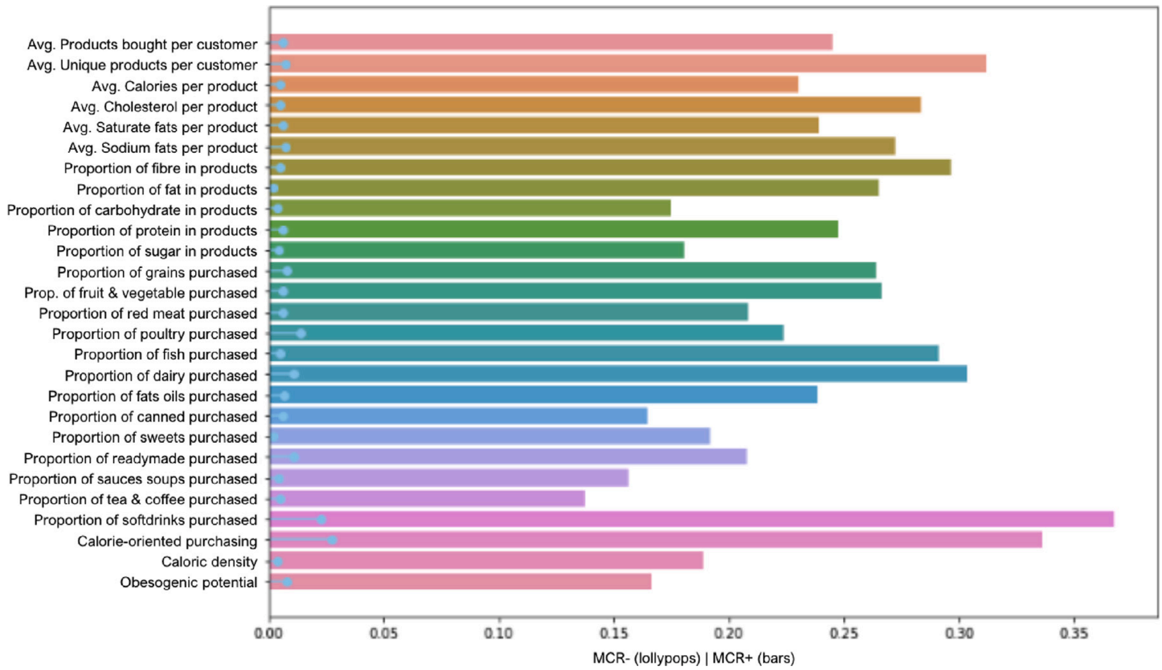

Fig. 6. Model Class Reliance chart showing feature importance across multiple RF models for the classifier predicting MSOAs with the highest 20% of obese children at reception age. MCR- (lollipops) and MCR+ (bars) represent the lower and upper bounds of each feature's importance, illustrating variability across models.

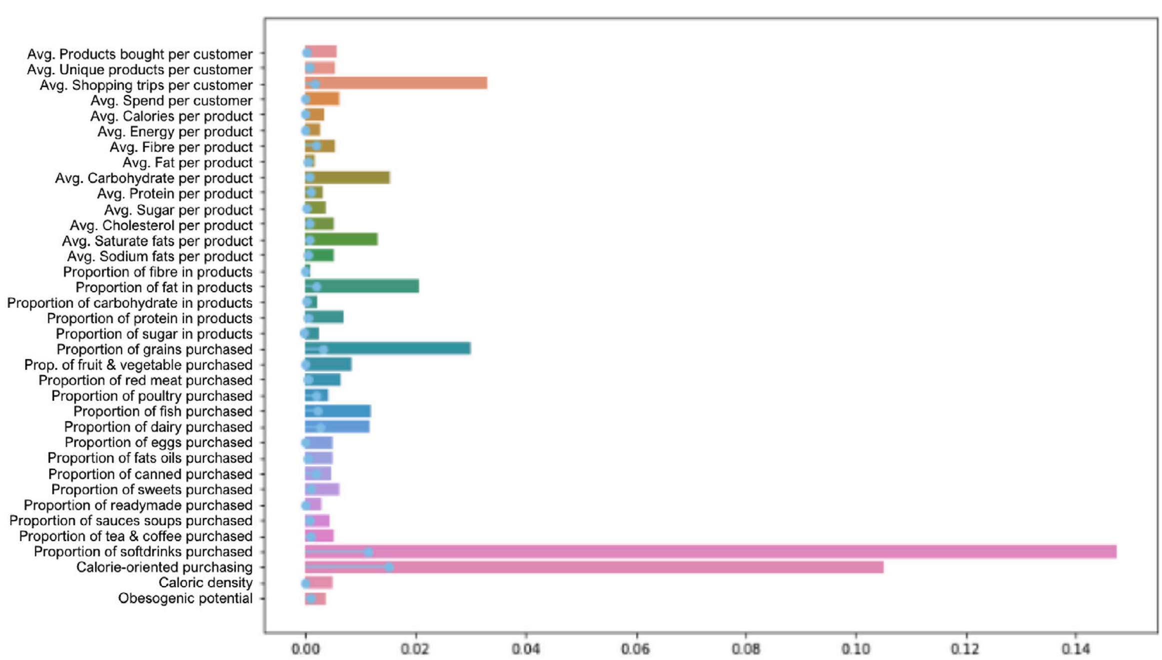

Fig. 7. Model Class Reliance chart showing feature importance across multiple RF models for the classifier predicting MSOAs with the highest 20% of obese children at year 6. MCR- (lollipops) and MCR+ (bars) represent the lower and upper bounds of each feature's importance, illustrating variability across models.

Overall, the MCR results for the year 6 classifier show greater robustness in the model than the reception classifier. A small number of features are required to predict areas with high levels of childhood obesity with high confidence. In contrast, the reception age classifier shows high levels of multicollinearity and the sharing of information across the features used.

4. Policy implications

The work described here offers new insights into neighbourhood-level deprivation and childhood obesity, as seen through the lens of grocery purchases. Current policy designed to alleviate extreme deprivation in English communities relies exclusively on the IoD and the IMD in particular. Only those areas in the lowest quintile of the IMD are eligible for remedial interventions and funding (Comber et al., 2022). Even assuming that the IoD accurately reflect the deprivation experienced in those areas, the time and cost of developing them means that they are only updated infrequently, every 3–5 years on average. Analysis of differences in the most recent IoD releases (2015 and 2019) shows that the deprivation deciles of around 40% of areas changed during that period. This means that a significant number of LSOAs will either not be eligible for the support they require or will be in receipt of support that is no longer warranted.

The work described here could be used as part of a national dietary monitoring model to highlight areas showing potential signs of high deprivation in near real-time, enabling national and local policymakers to deploy funding support as efficiently and effectively as possible. This could potentially be achieved by tracking the rates of change in key metrics such as Calorie-oriented purchasing or the sales of deprivation-related food categories such as cigarettes, soft drinks, wine and fish. Similarly, tracking changes in deprivation-related shopping metrics could offer a potential method to examine the impacts of local policy interventions. This would represent a significant step forward as methods for evaluating such interventions are limited and/or expensive (Nica-Avram et al., 2021).

Further research using shopping data from other grocery retailers would be warranted to investigate the generalisability of our findings, particularly the clear relationship identified between our COP metric and deprivation/obesity. Since COP does not rely on any subjective categorisation of grocery products and simply requires a good mapping of nutritional content, it is potentially a highly generalisable metric that could be applied across grocery retail markets. Another avenue of future research would be using shopping data covering the recent inflationary economic period, often referred to as the 'cost of living crisis' (Robinson, 2023; Keith Neal, 2022), and could provide clear evidence of the effects of increasing deprivation on grocery shopping behaviours and the associated nutritional impacts.

There are also policy implications for food retailers to help local communities suffering from deprivation. For example, in terms of the discounting of grocery staples or providing educational or informational content to help local communities. Co-op already engages with its members and customers to provide support for those in need. Co-op, like some other UK-based retailers, has already 'topped up' existing government voucher based schemes for the poorest in society (see for example the 'Healthy Start Scheme'). Previous research has demonstrated the value in dietary terms that such initiatives create, particularly regarding the breadth of fruit and vegetable diversity in consumption (Thomas et al., 2023).

Co-op has indicated an interest and commitment to the wellbeing of its members through discounting, particularly where discounting occurs in conjunction with government-supported initiatives to reach those most in need. Such community outreach currently uses traditional measures of deprivation, like the IMD (Hill-Dixon et al., 2018), but future support could integrate grocery sales as a metric for identifying local deprivation. For example, Co-op currently publish a public 'Wellbeing index' to help support their members that is constituted solely of open data but do not currently use the proprietary transactional data they have available.

From a childhood obesity perspective, the results confirm current thinking in terms of food policy, e.g. negative health impacts associated with excessive consumption of soft drinks, especially in children, and the positive health impact of increased consumption of fruit and vegetables. This finding backs up prior research on the impacts of the sales of sugary drinks on childhood obesity, e.g. Tedstone et al. (2015) and James and Kerr (2005), whilst highlighting that current policy interventions, and in particular the tax on sales of such products introduced in 2016, do not appear to have been sufficient to reduce levels of childhood obesity. This strengthens the policy arguments for greater intervention to encourage or discourage such behaviours, e.g. through increased support for 'Healthy Start Vouchers' for the poorest families or increases to the 'sugar' tax.

Additionally, the strong link between high levels of Calorie-oriented purchasing in a neighbourhood and the prevalence of childhood obesity in that area would appear to highlight the impact of poor diets on child health, This finding is also supported by the links to childhood obesity identified between low quantities of fruit and vegetables, and higher levels of grains purchased in areas with the highest quintile of obese children at ages 5 and 11. Such findings provide additional evidence linking poor dietary choices and childhood obesity and strengthen arguments that policy interventions need to go further than those currently implemented if childhood obesity is to be reduced.

5. Conclusions

In Section 1, we proposed two research questions. Responding to these two questions was the primary goal of this paper. For RQ1, we have clearly demonstrated that large-scale transaction data on grocery purchases can be useful in predicting deprivation and childhood obesity at the neighbourhood level. The ML-based classifiers produced predicted areas experiencing the highest levels of deprivation and childhood obesity with around 80% accuracy. In terms of the ML models tested, accuracy was similar for the three types of classifiers tested (LR, XGB, and RF).

As demonstrated by the results of the ternary classifier developed (Table C.1), the models are particularly of use when trying to predict extreme levels of deprivation/obesity and are less accurate at classifying neighbourhoods with average levels of deprivation/obesity. However, from a policy intervention context, it is these areas experiencing extreme deprivation/obesity levels that are of primary concern.

In terms of RQ2, several features derived from the transactional data were identified as being important predictors of deprivation and/or childhood obesity. These are highlighted in Section 3.3 for both the deprivation and obesity classifiers. One of our derived nutritional metrics, Calorie-oriented purchasing, was shown to be of high utility in predicting both extreme deprivation and childhood obesity. Both the SHAP and MCR analyses identify the metric as one of the most important features in the majority of the ML models produced during the study. Another metric that was found to be a very useful predictor of both deprivation and childhood obesity was the proportion of soft drinks purchased in a neighbourhood. Areas of high deprivation and childhood obesity are associated with higher levels of soft drinks purchases. While soft drinks strongly predict areas of high obesity in our models, this does not imply causality. The association may, for example, reflect a complex interplay of poverty and behaviour. Although the model points to areas where interventions are needed, it does not provide evidence for what those interventions should be.

For deprivation alone, it was found that lower levels of proportional spending on wine, fruit and vegetables, fish and ready-made meals were associated with areas with high deprivation. Conversely, higher levels of spending on cigarettes were associated with areas of high deprivation.

In the context of childhood obesity, lower levels of sales of fruit and vegetables, fish and dairy products appear to indicate higher levels of childhood obesity. In addition to Calorie-oriented purchasing and soft drinks, higher levels of purchasing of grains seem to lead to higher levels of childhood obesity.

Addressing the final point of RQ2, there are several ways in which these findings could inform national food policy, as described in the policy implications section of this paper (Section 4).

As with any research employing retail transactional data as its principal data source, there are inherent limitations to be considered. Firstly, the use of member data solely may be subject to gender and/or age biases (Nevalainen et al., 2018) compared to the general population. Secondly, membership is likely to be common for frequent and habitual shoppers (Rains and Longley, 2021) and may not accurately reflect the shopping behaviour of the wider population. These limitations are an unavoidable consequence of working with loyalty data and is necessary in order to be able to assign geographic locations to customers, a prerequisite to performing neighbourhood-scale analyses.

Another limitation is the reliance on a single retailer. This is not unusual in this type of study due to the complexities and commercial sensitivities involved in building long-term data-sharing agreements in competitive retail markets. Despite this, our findings appear in line with similar research (Bannister and Botta, 2021) using data from Tesco, the UK's largest supermarket chain (in terms of market share).

As stated, the results shown in Section 3 used data for 2019 Quarter 4. This was done to synchronise the data with the timing of the target datasets (IMD/childhood obesity) whilst excluding impacts of the COVID-19 pandemic and associated lockdowns. However, to ensure that the results were not a statistical anomaly of that specific time period, models were also generated for each quarter for which data was available. As shown in Fig. C.4, performance was consistent across the entire period with only minor fluctuations in accuracy identified.

Regarding our nutritionally derived metrics, we previously examined the effectiveness of the Calorie-oriented purchasing metric on the model's performance. The other two metrics developed, Calorific density and Obesogenic potential, did not share the relative importance of Calorie-oriented purchasing but did show some influence on model performance based on the SHAP and MCR analysis. Based on the SHAP analysis, it appears that higher levels of CD and OP are found in areas with lower levels of childhood obesity. This appears counter-intuitive as we would expect dietary consumption of calorifically-dense and obesogenic foods to lead to higher levels of obesity. However, it should also be noted that the MCR analysis showed CD and OP's variable importance to be relatively minor, see Fig. 7. The implication is that the simple prevalence of fat in the diet is not driving childhood obesity statistics and that COP is more likely reflecting the excessive consumption of very cheap, highly-processed and highly-palatable foods. What the feature importance analysis does not show is how features interact with one another, and their joint effect on the outcome; this could potentially be explored through partial dependence graphs to gauge how the probability of predicting deprivation/obesity levels changes when a factor is varied while all others are kept fixed.

Models predicting childhood obesity at year 6 performed significantly better than those predicting obesity at reception age. We hypothesise that the improved accuracy at year 6 may be related to both the higher overall levels of obesity at this age, and greater freedom for children to make their own dietary choices at this age. While testing this hypothesis is not within the scope of available data, further research is warranted to investigate this difference in model performance.

From a methodological context, we showed that the use of binary classification ML models to identify areas with high levels of deprivation/childhood obesity is an effective tool that could be easily extended to investigate other demographic and health outcomes. Results showed a significant level of performance with our models able to predict areas in the highest 10% and 20% of deprivation/childhood obesity with around 80% accuracy.

CRediT authorship contribution statement

Gavin Long: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Georgiana Nica-Avram:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **John Harvey:** Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Evgeniya Lukinova:** Writing – original draft, Methodology, Investigation. **Roberto Mansilla:** Writing – original draft, Visualization, Methodology, Investigation. **Simon Welham:** Writing – original draft, Methodology, Investigation, Conceptualization. **Gregor Engelmann:** Writing – original draft, Methodology, Formal analysis, Data curation. **Elizabeth Dolan:** Writing – original draft, Methodology, Investigation. **Kuzivakwashe Makokoro:** Writing – original draft, Methodology, Investigation. **Michelle Thomas:** Writing – original draft, Methodology, Investigation. **Edward Powell:** Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **James Goulding:** Writing – original draft, Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partly supported by the Future Food Beacon at University of Nottingham, the Horizon Centre for Doctoral Training at the University of Nottingham (grant EP/S023305/1), and UKRI CIVIC Project (grant EP/V053922/1).

Appendix A. Acronyms

BMI — Body Mass Index
CD — Caloric Density
CoFID — Composition of Foods Integrated Dataset
COP — Calorie-oriented Purchasing
DT — Decision Tree
IMD — Index of Multiple Deprivation
LR — Linear Regression
LSOA — Lower-layer Super Output Areas
MCR — Model Class Reliance
ML — Machine Learning
MSOA — Middle Super Output Area
NCMP — National Child Measurement Programme
NDNS — National Diet and Nutrition Survey
NHANES — National Health and Nutrition Examination Survey
OHID — Health Inequalities and Disparities
ONS — Office for National Statistics
OP — Obesogenic Potential
RCP — Relative Caloric Pricing
RF — Random Forest
SHAP — SHapley Additive exPlanations
SVM — Support Vector Machine
XGB — XGBoost

Appendix B. Set of predictor features and target variables

See Figs. B.1 and B.2.

Appendix C. Supplementary analysis

See Figs. C.1–C.4 and Table C.1.

Table B.1
Food purchased is organised into 17 categories including: grains; fruit and vegetables; red meat; poultry; fish; dairy; eggs; fat and oil; canned goods; sweets; cigarettes and tobacco; readymade; sauces and soups; softdrinks; beer, lager and cider; wine; spirits. For each LSOA, a total of 92 features were constructed to include the fraction of purchased products and money spent, and averages relative to the number of customers. Additional predictors are the fraction of macro-nutrients present in the purchased products – fibre, fat, carbohydrate, protein, sugar, sodium, cholesterol, saturated fats, iron, iodine, calcium, vitamin B12 content; and our derived nutritional metrics – COP, CD and Obesogenic Potential.

Feature domain	Feature name & shorthand	Feature description	Source/ Reference
Area-level deprivation	IMD decile (imd_decile)	Index of Multiple Deprivation for England (2019)	Ministry of Housing, Communities & Local Government (2019), Bannister and Botta (2021), Amin et al. (2021), Cetateanu and Jones (2014), Howard Wilsher et al. (2016)
Childhood Obesity	Childhood obesity quintile (obesity_quintile)	NCMP Childhood obesity data Prevalence of obesity (including severe obesity), 3-years data combined (2019 to 2020, to 2021 to 2022)	National Child Measurement Programme (NCMP), N.H.S. Digital (2023), Aiello et al. (2019), Cetateanu and Jones (2014), Howard Wilsher et al. (2016)
Population statistics	Mid-2019 median age (mid_2019_median_age) Population density (population_sq_km)	Median population age in mid-2019 Population density per square kilometer	Office of National Statistics (2020), Amin et al. (2021)
Food-access	Number of food stores per population (e.g., takeaway_sandwich_population)	Number of takeaways and sandwich shops _ Number of supermarkets _ Number of restaurants, cafes and canteens per population	Food Standards Agency (2021), Badruddoza et al. (2023), Cetateanu and Jones (2014), Amin et al. (2021)
Grocery shopping levels per customer	Number of customers (customers_population) Avg. number of products (avg_products_bought_customers) Avg. number of unique products (avg_unique_products_customers) Avg. number of shopping trips (avg_basket_count_customers) Avg. spend (£) (avg_spend_customers) Avg. number of unique stores (avg_unique_stores_customers) Avg. bundle entropy (avg_bundle_entropy_customers) Avg. weekly spend (£) (avg_weekly_spend_customers) Avg. spend (£) per shopping trip (avg_basket_spend_customers) Avg. number of products per shopping trip (avg_products_per_basket_customers)	Number of customers per LSOA population Avg. number of products per customer Avg. number of unique products per customer Avg. number of shopping trips per customer Avg. spend (£) per customer Avg. number of unique stores per customer Avg. bundle entropy per customer Avg. weekly spend (£) per customer Avg. spend (£) per shopping trip per customer Avg. number of products per shopping trip per customer	Mansilla et al. (2022)
Grocery shopping seasonality	Avg. seasonal purchases per number of customers (avg_seasonality_customers) Avg. number of shopping trips per customer (e.g., avg_morning_customers)	Avg. seasonal purchases per number of customers Avg. number of shopping trips during early morning, mid-morning, lunch-time, afternoon, evening and night per customer	
Grocery shopping spend	Avg. spend per number of products (e.g., avg_fruit_vegetables_spending_total) Proportion of spend on product categories (e.g., red_meat_spend_p)	Avg. spend per number of products across 17 categories Proportion of spend across 17 categories	
Grocery shopping volumes	Proportion of purchased items across product categories (e.g., red_meat_quantity_p)	Proportion of purchased items across 17 categories	Andreyeva and Tripp (2016), Schwartz et al. (2017), Berger et al. (2021), Brimblecombe et al. (2013), Howard Wilsher et al. (2016)
Obesogenicity	Caloric frugality (CF, kcals_per_pound) Caloric density (CD, kcals_per_kg) Obesogenic potential (OP, obesogenic)	Kilocalories per £spent Calorific content normalised by weight Proportion of calories obtained from fats and sugar	Headey and Alderman (2019), Drewnowski and Specter (2004)

(continued on next page)

Table B.1 (*continued*).

Feature domain	Feature name & shorthand	Feature description	Source/ Reference
Nutrient composition	Calories per number of products (calories_sum_products)	Calories per number of products	
	Energy content per number of products (energy_sum_products)	Energy content per number of products	
	Avg. macronutrients per number of products (e.g., fibre_sum_products)	Avg. macronutrients per number of products: fibre, fat, carbohydrate, protein, sugar, sodium, cholesterol, saturated fats, iron content	
	Proportion of macronutrient content (e.g., fibre_p)	Proportion of macronutrient content: fibre; fat, carbohydrates; protein; sugar; sodium; iron; iodine; calcium; vitamin B12	Brimblecombe et al. (2013), Andreyeva and Tripp (2016), Berger et al. (2021)

Table B.2
Summary statistics for reduced feature set and IMD target variable. Statistics shown for 10,547 LSOAs in England after exclusion criteria filtering was applied to the dataset.

Feature name	Min	Mean	SD	Median	Max
IMD Decile (Target variable)	1.00	6.11	2.66	6.00	10.00
Average products bought per customer	16.98	96.77	30.88	93.14	402.27
Average shopping baskets per customer	3.43	15.11	4.86	14.53	37.75
Average spend per customer (£)	34.58	177.49	57.89	171.11	739.96
Average bundle entropy per customer	0.25	0.38	0.04	0.38	0.56
Average basket spend per customer (£)	4.39	12.25	2.83	11.67	35.71
Average products per basket per customer	3.16	6.69	1.41	6.41	18.57
Average seasonal purchases per customer	1.25	24.15	13.05	21.75	114.62
Average lunchtime purchases per customer	0.53	2.68	0.81	2.61	7.26
Average afternoon purchases per customer	0.77	3.80	1.24	3.67	9.72
Average evening purchases per customer	0.24	2.92	1.35	2.76	11.87
Average night purchases per customer	0.00	0.86	0.61	0.73	5.49
Average calories per product (Kcals)	285.42	556.01	36.51	558.94	681.59
Average sugar per product (g)	16.48	30.22	2.64	30.27	43.85
Average sodium per product (mg)	170.90	390.85	71.36	385.90	1108.27
Average cholesterol per product (mg)	17.66	41.60	5.89	41.34	82.87
Average spending on grains (£)	0.66	1.01	0.07	1.00	1.40
Average spending on fruit and vegetables (£)	0.59	1.09	0.08	1.09	1.44
Average spending on red meat (£)	1.62	2.64	0.26	2.65	4.13
Average spending on poultry (£)	1.06	2.84	0.39	2.87	5.27
Average spending on fish (£)	1.04	2.79	0.46	2.86	4.07
Average spending on dairy (£)	0.81	1.36	0.09	1.36	2.04
Average spending on eggs (£)	0.66	1.47	0.15	1.47	2.16
Average spending on fats and oils (£)	1.04	1.90	0.18	1.91	3.07
Average spending on canned foods (£)	0.48	0.94	0.11	0.93	1.59
Average spending on sweets (£)	0.76	1.22	0.11	1.21	1.90
Average spending on readymade foods (£)	0.96	1.77	0.15	1.77	2.42
Average spending on sauces and soups (£)	0.53	1.20	0.15	1.18	2.02
Average spend on tea and coffee (£)	0.99	2.69	0.24	2.69	4.08
Proportion of fibre in products	0.03	0.05	0.004	0.05	0.08
Proportion of fat in products	0.14	0.21	0.01	0.21	0.30
Proportion of saturated fat in products	0.06	0.08	0.005	0.08	0.13
Proportion of carbohydrate in products	0.54	0.62	0.02	0.62	0.69
Proportion of protein in products	0.13	0.17	0.01	0.17	0.22
Proportion of spend on cigarettes	0.00	0.10	0.06	0.09	0.42
Proportion of spend on wine	0.00	0.12	0.05	0.12	0.38
Proportion of soft drinks purchased	0.02	0.10	0.03	0.09	0.25
Calorie-Oriented Purchasing (Kcals/£)	186.94	347.46	39.05	345.57	622.30
Calorific Density (Kcals/kg)	626.10	1096.31	87.24	1096.94	1470.43
Obesogenic Potential	0.50	0.60	0.02	0.60	0.68

Table B.3
Summary statistics for reduced feature set and Childhood obesity quintile target variable. Statistics shown for 4,677 MSOAs in England after exclusion criteria filtering was applied to the dataset.

Feature	Min	Mean	SD	Median	Max
Obesity quintile	1.00	3.20	1.36	3.00	5.00
Average products bought per customer	24.58	84.66	27.66	82.01	222.18
Average unique products bought per customer	13.78	43.78	12.76	42.43	113.89
Average number of shopping baskets per customer	4.09	13.71	4.25	13.45	29.56
Average spend per customer (£)	35.41	153.71	52.32	149.13	397.11
Average calories per product (Kcals)	330.33	539.11	40.81	545.16	650.84
Average energy per product (KJ)	1386.28	2262.73	171.54	2288.04	2732.81
Average fibre per product (g)	2.95	4.82	0.50	4.86	7.29
Average fat per product (g)	14.06	22.55	1.82	22.69	30.32
Average carbohydrate per product (g)	39.49	65.36	5.58	65.97	83.60
Average protein per product (g)	10.70	17.86	1.52	18.13	22.78
Average sugar per product (g)	18.31	29.45	2.53	29.61	39.41
Average cholesterol per product (mg)	17.10	39.54	6.11	39.98	74.94
Average saturated fat per product (g)	5.71	8.93	0.79	9.00	11.95
Average sodium per product (mg)	183.62	376.30	66.83	378.42	1205.48
Proportion of fibre in products	0.004	0.009	0.001	0.009	0.020
Proportion of fat in products	0.017	0.044	0.005	0.044	0.064
Proportion of carbohydrate in products	0.046	0.128	0.015	0.127	0.188
Proportion of protein in products	0.013	0.035	0.004	0.035	0.053
Proportion of sugar in products	0.020	0.058	0.007	0.057	0.092
Proportion of grains purchased	0.017	0.050	0.011	0.050	0.232
Proportion of fruit and vegetables purchased	0.030	0.161	0.044	0.156	0.406
Proportion of red meat purchased	0.006	0.029	0.007	0.030	0.069
Proportion of poultry purchased	0.001	0.018	0.004	0.018	0.048
Proportion of fish purchased	0.002	0.013	0.005	0.012	0.059
Proportion of dairy products purchased	0.077	0.151	0.019	0.152	0.337
Proportion of eggs purchased	0.001	0.012	0.004	0.012	0.032
Proportion of fats/oils purchased	0.002	0.015	0.004	0.014	0.047
Proportion of canned foods purchased	0.008	0.033	0.010	0.032	0.111
Proportion of confectionery purchased	0.049	0.133	0.025	0.132	0.289
Proportion of ready made foods purchased	0.092	0.170	0.027	0.165	0.384
Proportion of sauces/soups purchased	0.003	0.017	0.004	0.016	0.043
Proportion of tea/coffee purchased	0.004	0.017	0.005	0.016	0.061
Proportion of soft drinks purchased	0.030	0.091	0.027	0.086	0.245
Calorie Oriented Purchasing (Kcals/£)	200.10	341.85	34.86	340.71	541.43
Calorific Density (Kcals/kg)	556.12	1089.30	81.08	1092.44	1483.63
Obesogenic Potential	0.53	0.60	0.02	0.60	0.67

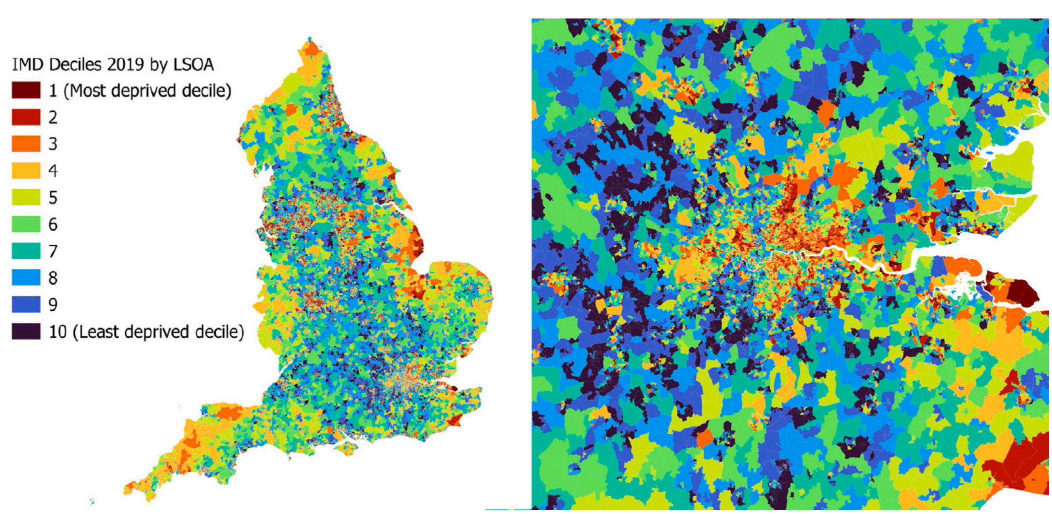

Fig. B.1. Map showing spatial distribution of IMD deciles across England [Left]. More detailed view of London and South East England is shown on the[right] highlighting smaller urban areas of higher deprivation that are not clearly visible at the national level. The geospatial accuracy of the 20/80 IMD classifier was highest in LSOAs with higher levels of co-op membership and sales. Misclassification errors were generally found in those LSOAs with lower sales and customer levels, e.g. West Midlands and the East of England.

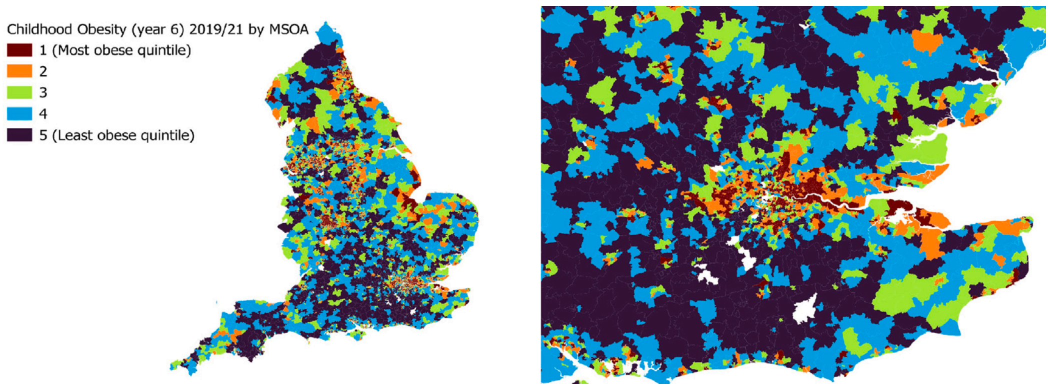

Fig. B.2. Map showing spatial distribution of Childhood Obesity (year 6) quintiles across England [Left]. More detailed view of London and South East England is shown on the[right] highlighting smaller urban areas with higher levels of child obesity that are not clearly visible at the national level. Similarly to the IMD Classifier, the Childhood Obesity results were spatially more accurate in those MSOAs with higher levels of sales. Misclassification errors were generally found in areas with borderline childhood obesity levels (e.g. 1st/2nd quintile) combined with low levels of co-op sales and membership were found such as the West Midlands and the East of England.

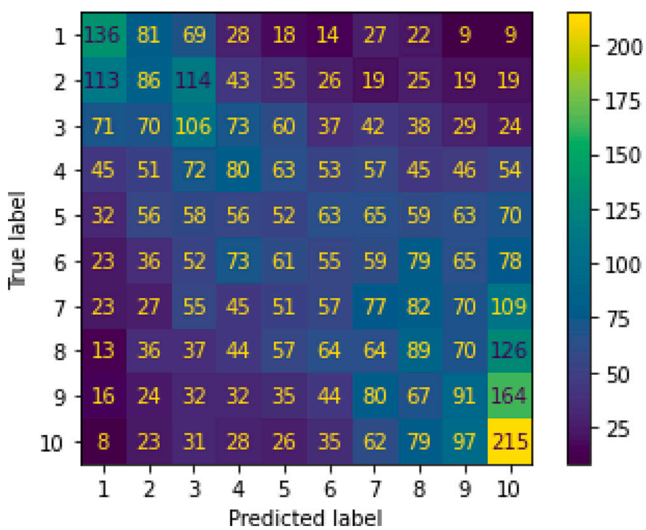

Fig. C.1. In the 10-class classifier, LSOAs in the middle range of deciles are predicted less accurately, as it is probably harder to differentiate among them. Only the deciles of pronounced deprivation (1–3, 8–10) appear to be discriminatory to the classifier.

Table C.1
Ternary classification results from experimental models predicting deprivation levels using food shopping data from October 2019 to December 2019.

Threshold	Inputs	Model	Results(ALL areas)				Results (Low deprivation areas)			
			Acc	Prec	F1	Kappa	Acc	Prec	F1	Kappa
30-40-30 (high/med/low) deprivation percentile split	All	Dummy	0.35	0.35	0.35	0.02	0.34	0.34	0.34	0.01
		LR	0.62	0.62	0.62	0.5	0.80	0.70	0.74	0.58
		XGB	0.75	0.75	0.75	0.63	0.76	0.75	0.79	0.63
		RF	0.74	0.74	0.74	0.62	0.77	0.76	0.79	0.64
	Reduced	LR	0.61	0.61	0.61	0.46	0.80	0.63	0.69	0.48
		XGB	0.69	0.69	0.69	0.53	0.74	0.73	0.73	0.57
		RF	0.70	0.70	0.70	0.55	0.73	0.74	0.74	0.59
			Results (Medium deprivation areas)				Results (High deprivation areas)			
			Acc	Prec	F1	Kappa	Acc	Prec	F1	Kappa
30-40-30 (high/med/low) deprivation percentile split	All	Dummy	0.35	0.35	0.36	0.02	0.34	0.34	0.34	0.01
		LR	0.32	0.53	0.40	0.41	0.79	0.64	0.71	0.52
		XGB	0.65	0.69	0.67	0.57	0.84	0.82	0.79	0.7
		RF	0.68	0.66	0.65	0.54	0.83	0.81	0.78	0.69
	Reduced	LR	0.33	0.52	0.40	0.37	0.76	0.68	0.74	0.53
		XGB	0.59	0.58	0.58	0.42	0.73	0.76	0.75	0.6
		RF	0.57	0.61	0.60	0.46	0.77	0.74	0.75	0.59

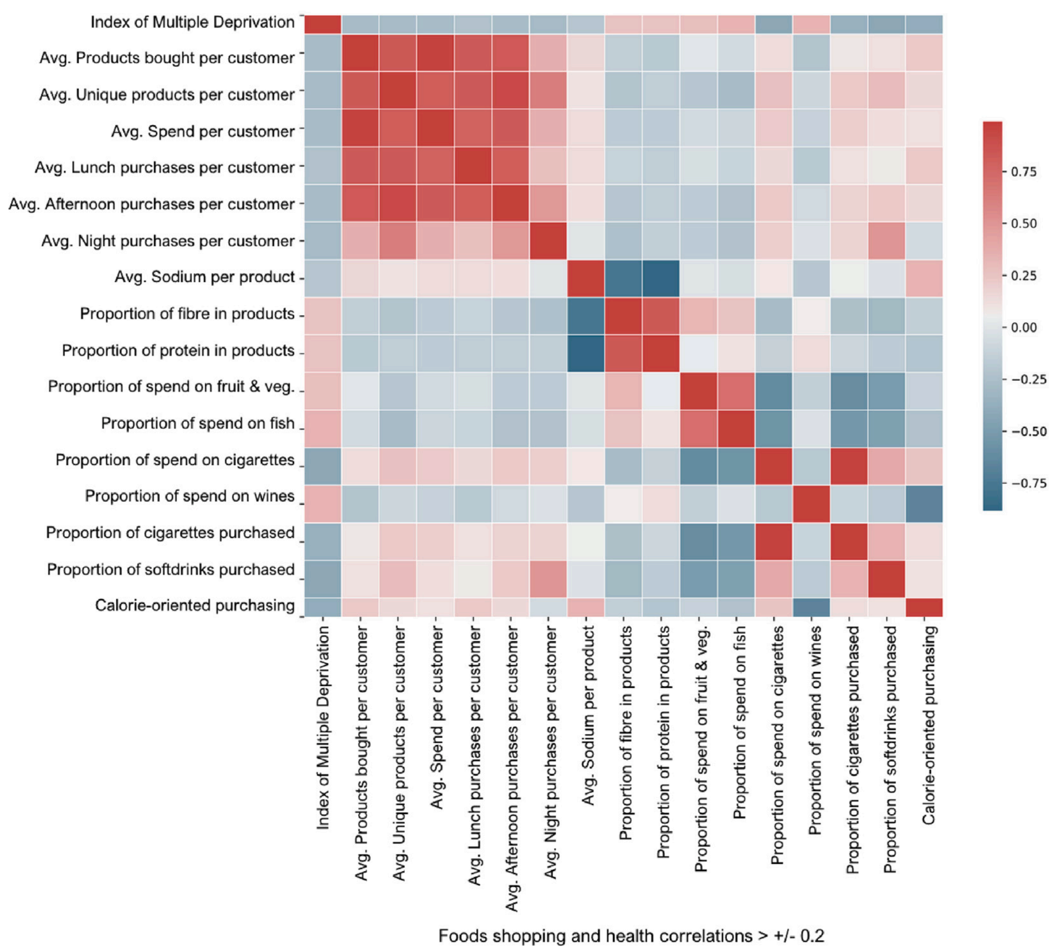

Fig. C.2. Correlation matrix showing relationships between the Index of Multiple Deprivation (IMD) and food purchased in UK neighbourhoods (LSOAs). We find that multiple features extracted from the grocery shopping data show a significant correlation with the IMD when performing Bonferroni corrected Pearson's correlation. To increase readability, only features showing a significant correlation coefficient (+/- 0.2) are depicted.

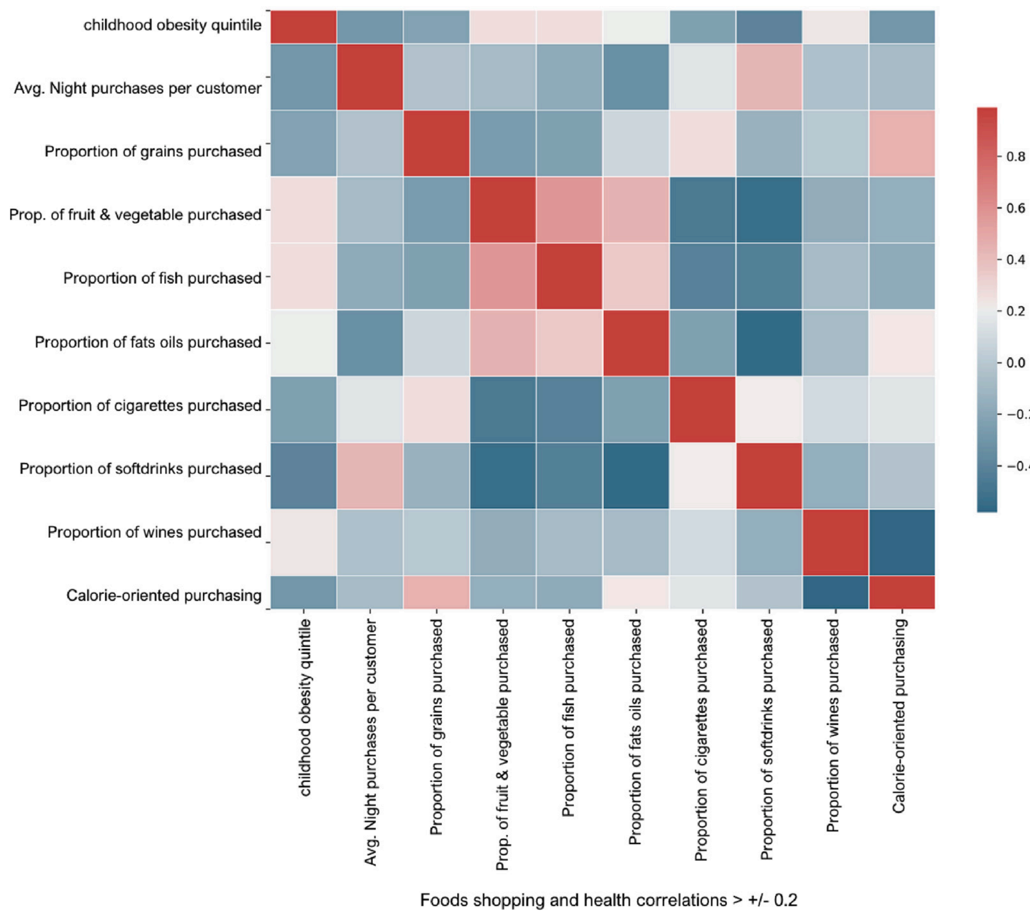

Fig. C.3. Correlation matrix showing relationships between the childhood obesity quintile (at year 6) and food purchased in UK neighbourhoods (MSOAs). We find that multiple features extracted from the grocery shopping data show a significant correlation with obesity when performing Bonferroni corrected Pearson's correlation. To increase readability, only features showing a significant correlation coefficient (+/- 0.2) are depicted.

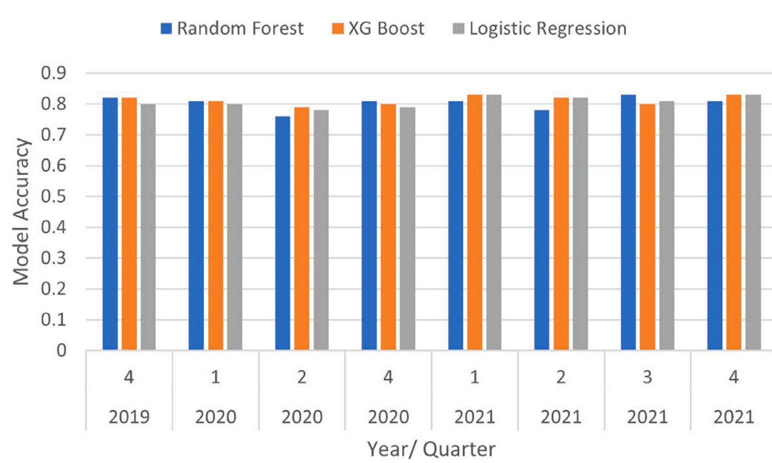

Fig. C.4. Temporal consistency of classifier accuracy by year and quarter. Results shown are for the 20/80 split IMD classifier using the reduced set of input features.

References

Aiello, L.M., Quercia, D., Schifanella, R., Del Prete, L., 2020. Tesco grocery 1.0, a large-scale dataset of grocery purchases in London. Sci. Data 7 (1), 57.

Aiello, L.M., Schifanella, R., Quercia, D., Del Prete, L., 2019. Large-scale and high-resolution analysis of food purchases and health outcomes. EPJ Data Sci. 8 (1), 1–22.

Ali, J., Khan, R., Ahmad, N., Maqsood, I., 2012. Random forests and decision trees. IJCSI Int. J. Comput. Sci. Issues 9 (5), 272–278.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347.

Amin, M.D., Badruddoza, S., McCluskey, J.J., 2021. Predicting access to healthful food retailers with machine learning. Food Policy 99, 101985.

Andreyeva, T., Tripp, A.S., 2016. The healthfulness of food and beverage purchases after the federal food package revisions: The case of two new England states. Prev. Med. 91, 204–210.

Badruddoza, S., Amin, M.D., McCluskey, J., 2023. Impact of food retailers' presence and composition on nutritional equity and health outcomes in the United States with machine learning. J. Nutr. Educ. Behav. 55 (7), 108–109.

Baker, C., 2022. Obesity statistics. URL https://commonslibrary.parliament.uk/research-briefings/sn03336/.

Bannister, A., Botta, F., 2021. Rapid indicators of deprivation using grocery shopping data. R. Soc. Open Sci. 8 (12), 211069.

Berger, N., Cummins, S., Smith, R.D., Cornelsen, L., 2021. Have socio-economic inequalities in sugar purchasing widened? A longitudinal analysis of food and beverage consumer data from british households, 2014–2017. Public Heal. Nutr. 24 (7), 1583–1594.

Biecek, P., Baniecki, H., Krzyziński, M., Cook, D., 2024. Performance is not enough: the story told by a rashomon quartet. J. Comput. Graph. Statist. (just-accepted), 1–6.

Brimblecombe, J., Liddle, R., O'Dea, K., 2013. Use of point-of-sale data to assess food and nutrient quality in remote stores. Public Heal. Nutr. 16 (7), 1159–1167.

Cateni, S., Colla, V., Vannucci, M., 2014. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing 135, 32–41. http://dx.doi.org/10.1016/j.neucom.2013.05.059.

Cetateanu, A., Jones, A., 2014. Understanding the relationship between food environments, deprivation and childhood overweight and obesity: evidence from a cross sectional England-wide study. Heal. & Place 27, 68–76.

Comber, S., Park, S., Arribas-Bel, D., 2022. Dynamic-IMD (d-IMD): Introducing activity spaces to deprivation measurement in London, birmingham and liverpool. Cities 127, 103733.

Davies, S.C., 2019. Time to solve childhood obesity. An Independent Report by the Chief Medical Officer. Annex A—Recommendations for Action.

Dolan, E., Goulding, J., Marshall, H., Smith, G., Long, G., Tata, L.J., 2023. Assessing the value of integrating national longitudinal shopping data into respiratory disease forecasting models. Nat. Commun. 14 (1), 7258.

Drewnowski, A., Specter, S., 2004. Poverty and obesity: the role of energy density and energy costs. Am. J. Clin. Nutr. 79 (1), 6–16.

Dusmanu, M., Cabrio, E., Villata, S., 2017. Argument mining on Twitter: Arguments, facts and sources. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2317–2322.

England, N., 2023. Core20PLUS5 (adults) – an approach to reducing healthcare inequalities. URL https://www.england.nhs.uk/about/equality/equality-hub/national-healthcare-inequalities-improvement-programme/core20plus5/.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15, 3133–3181.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20, 1–81.

Food Standards Agency, 2021. UK Food Hygiene Rating Data API. URL http://ratings.food.gov.uk/open-data/en-GB.

Futagami, K., Fukazawa, Y., Kapoor, N., Kito, T., 2021. Pairwise acquisition prediction with SHAP value interpretation. J. Financ. Data Sci..

GOV.UK, 2021. Composition of foods integrated dataset (CoFID). URL https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.

Harvey, J., Long, G., Mansilla, R., Welham, S., Rose, P., Thomas, M., Milligan, G., Dolan, E., Parkes, J., Makokoro, K., Goulding, J., 2023. Who consumes anthocyanins and anthocyanidins? Mining national retail data to reveal the influence of socioeconomic deprivation and seasonality on polyphenol dietary intake. In: 2023 IEEE International Conference on Big Data (BigData). IEEE, pp. 4530–4538.

Headey, D.D., Alderman, H.H., 2019. The relative caloric prices of healthy and unhealthy foods differ systematically across income levels and continents. J. Nutr. 149 (11), 2020–2033.

Hill-Dixon, A., Solley, S., Bynon, R., 2018. Being well together: the creation of the co-op community wellbeing index. URL https://communitywellbeing.coop.co.uk/media/1026/the_community_wellbeing_index_-_full_report.pdf.

Hossain, M., Mullally, C., Asadullah, M.N., 2019. Alternatives to calorie-based indicators of food security. Food Policy 84, 77–91. http://dx.doi.org/10.1016/j.foodpol.2019.03.001.

Howard Wilsher, S., Harrison, F., Yamoah, F., Fearne, A., Jones, A., 2016. The relationship between unhealthy food sales, socio-economic deprivation and childhood weight status: results of a cross-sectional study in England. Int. J. Behav. Nutr. Phys. Act. 13 (1), 1–8.

James, J., Kerr, D., 2005. Prevention of childhood obesity by reducing soft drinks. Int. J. Obes. 29 (2), S54–S57.

Jenneson, V.L., Pontin, F., Greenwood, D.C., Clarke, G.P., Morris, M.A., 2022. A systematic review of supermarket automated electronic sales data for population dietary surveillance. Nutrition Rev. 80 (6), 1711–1722.

Joloudari, J.H., Marefat, A., Nematollahi, M.A., Oyelere, S.S., Hussain, S., 2023. Effective class-imbalance learning based on SMOTE and convolutional neural networks. Appl. Sci. 13 (6), http://dx.doi.org/10.3390/app13064006, URL https://www.mdpi.com/2076-3417/13/6/4006.

Keith Neal, P.W., 2022. The 'cost of living crisis'. J. Public Heal. 44 (3), 475–476.

Kipnis, V., Midthune, D., Freedman, L., Bingham, S., Day, N.E., Riboli, E., Ferrari, P., Carroll, R.J., 2002. Bias in dietary-report instruments and its implications for nutritional epidemiology. Public Heal. Nutr. 5 (6a), 915–923.

Knai, C., Petticrew, M., Mays, N., 2016. The childhood obesity strategy. BMJ 354.

Mansilla, R., Long, G., Welham, S., Harvey, J., Lukinova, E., Nica-Avram, G., Smith, G., Salt, D., Smith, A., Goulding, J., 2024. Detecting iodine deficiency risks from dietary transitions using shopping data. Sci. Rep. 14 (1), 1017.

Mansilla, R., Smith, G., Smith, A., Goulding, J., 2022. Bundle entropy as an optimized measure of consumers' systematic product choice combinations in mass transactional data. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, pp. 1044–1053.

McBride, L., Barrett, C.B., Browne, C., Hu, L., Liu, Y., Matteson, D.S., Sun, Y., Wen, J., 2022. Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. Appl. Econ. Perspect. Policy 44 (2), 879–892.

Ministry of Housing, Communities & Local Government, 2019. English indices of deprivation 2019. URL https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.

Mooney, S.J., Pejaver, V., 2018. Big data in public health: terminology, machine learning, and privacy. Annu. Rev. Public. Health 39 (1), 95–112.

Morris, R., Carstairs, V., 1991. Which deprivation? A comparison of selected deprivation indexes. J. Public Heal. 13 (4), 318–326. http://dx.doi.org/10.1093/oxfordjournals.pubmed.a042650.

National Child Measurement Programme (NCMP), N.H.S. Digital, 2023. Prevalence of obesity (including severe obesity), 3-years data combined 2019 to 2020, to 2021 to 2022. URL https://fingertips.phe.org.uk/profile/national-child-measurement-programme/data#page/6/gid/1938133288/pat/159/par/K02000001/ati/15/are/E92000001/iid/93107/age/201/sex/4/cat/-1/ctp/-1/yrr/3/cid/4/tbm/1.

Nevalainen, J., Erkkola, M., Saarijärvi, H., Näppilä, T., Fogelholm, M., 2018. Large-scale loyalty card data in health research. Digit. Heal. 4, 1–10.

Nica-Avram, G., Harvey, J., Smith, G., Smith, A., Goulding, J., 2021. Identifying food insecurity in food sharing networks via machine learning. J. Bus. Res. 131, 469–484.

Nutritics, 2024. One platform for all food data management and digital publishing. URL https://www.nutritics.com/en/.

Office for Health Improvement and Disparities (OHID), 2024. Public health England. Local health [internet]. URL https://www.localhealth.org.uk/#c=home.

Office of National Statistics, 2020. Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019estimates.

PA Media, 2024. Experts lament 'appalling decline' in health of under-fives in UK. Guardian URL https://www.theguardian.com/society/2024/feb/05/experts-lament-appalling-decline-in-health-of-under-fives-in-uk.

Pitsilis, G.K., Ramampiaro, H., Langseth, H., 2018. Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433.

Rains, T., Longley, P., 2021. The provenance of loyalty card data for urban and retail analytics. J. Retail. Consum. Serv. 63, 102650. http://dx.doi.org/10.1016/j.jretconser.2021.102650.

Robinson, E., 2023. Obesity and the cost of living crisis. Int. J. Obes. 47 (2), 93–94.

Schwartz, M.B., Schneider, G.E., Choi, Y.-Y., Li, X., Harris, J., Andreyeva, T., Hyary, M., Vernick, N.H., Appel, L.J., 2017. Association of a community campaign for better beverage choices with beverage purchases from supermarkets. JAMA Intern. Med. 177 (5), 666–674.

Singh, A., Thakur, N., Sharma, A., 2016. A review of supervised machine learning algorithms. In: Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016. Bharati Vidyapeeth, New Delhi as the Organizer of INDIACom - 2016, pp. 1310–1315.

Smith, G., Mansilla, R., Goulding, J., 2020. Model class reliance for random forests. Adv. Neural Inf. Process. Syst. 33, 22305–22315.

Tarasuk, V., Mitchell, A., McLaren, L., McIntyre, L., 2013. Chronic physical and mental health conditions among adults may increase vulnerability to household food insecurity. J. Nutr. 143 (11), 1785–1793.

Tedstone, A., Targett, V., Allen, R., et al., 2015. Sugar reduction: the evidence for action. Public Health England, URL https://assets.publishing.service.gov.uk/media/5a7f928c40f0b623026904b7/Sugar_reduction_The_evidence_for_action.pdf.

The Food Foundation, 2023. The broken plate 2023: The state of the nation's food system. pp. 1–55, URL https://www.foodfoundation.org.uk/sites/default/files/2023-06/TFF_The%20Broken%20Plate%202023_DigitalFINAL_1.pdf.

Thomas, M., Moore, J.B., Onuselogu, D.A., Dalton, A., Rains, T., Lowry, E., Sritharan, N., Morris, M.A., 2023. Supermarket top-up of healthy start vouchers increases fruit and vegetable purchases in low-income households. Nutr. Bull. 48 (3), 353–364.

Timmins, K.A., Green, M.A., Radley, D., Morris, M.A., Pearce, J., 2018. How has big data contributed to obesity research? A review of the literature. Int. J. Obes. 42 (12), 1951–1962.

Townsend, P., Phillimore, P., Beattie, A., 2023. Health and deprivation: inequality and the North, vol. 8, Taylor & Francis.

Villacis, A.H., Badruddoza, S., Mishra, A.K., Mayorga, J., 2023. The role of recall periods when predicting food insecurity: A machine learning application in Nigeria. Glob. Food Secur. 36, 100671.

Wongvorachan, T., He, S., Bulut, O., 2023. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information 14 (1), http://dx.doi.org/10.3390/info14010054, URL https://www.mdpi.com/2078-2489/14/1/54.

World Health Organization, 2020. Children: new threats to health. URL https://www.who.int/news-room/fact-sheets/detail/children-new-threats-to-health.

World Health Organization and others, 2016. Report of the commission on ending childhood obesity. World Health Organization, URL https://www.who.int/publications/i/item/9789241510066.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. Perspect. Psychol. Sci. 12 (6), 1100–1122.