Predicting food safety compliance for informed food outlet inspections: a machine learning approach

[1,3]Rachel A. Oldroyd, [1,2,4]Michelle A. Morris, [1,2,3]M. Birkin

[1] *Leeds Institute for Data Analytics, University of Leeds, LS2 9JT, United Kingdom.*
[2] *Alan Turing Institute, London, NW1 2DB, United Kingdom*
[3] *School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom.*
[4] *School of Medicine, University of Leeds, Leeds, LS2 9JT, United Kingdom.*

Corresponding Author:
Rachel A. Oldroyd, MSc
Leeds Institute for Data Analytics
University of Leeds
Worsley Building, Level 11
Clarendon Road
Leeds, LS2 9JT
United Kingdom
Phone: 44 113 34 ext 33309
Email: r.oldroyd@leeds.ac.uk

Keywords

Food safety, food environments, food hygiene, machine learning

Abstract

Consumer food environments have transformed dramatically in the last decade. The number of food outlets has increased, and a large proportion of the UK population now purchase food from a takeaway or restaurant at least once a week. Despite these developments, national spending on food control has reduced and many Local Authorities struggle to meet health inspection targets. This work presents a data driven approach to enhance current inspection processes with a view to reduce consumer risk of foodborne illness whilst eating outside the home. We explore the utility of three machine learning algorithms to predict non-compliant food outlets in England and Wales as defined by Food Hygiene Rating Scheme scores >= 2. Using socio-demographic, business type and urbanness features we experiment with under and over sampling strategies at five ratios to address problems of class imbalance in the dataset prior to analysis. We find that Synthetic Minority Over Sampling Technique alongside a Random Forest algorithm with a 1:1 sampling strategy provides the best predictive power. Our final model retrieves 84% of total non-compliant outlets in a test set of 92,595 (sensitivity=0.843, specificity=0.745, precision=0.274). We discuss the utility of machine learning algorithms to prioritise high risk establishments for inspection by Local Authority officials and make recommendations for weighting outcomes to improve their appropriateness in an applied setting.

1 Introduction

Patterns of national food consumption have changed dramatically in recent years. 43% of the UK population now purchase food from a take away or restaurant on a weekly basis (Food Standards Agency 2016). Although drivers of consumer behaviours are complex and multi-faceted, this change can be partially attributed to a recent proliferation of fast food outlets; equating to 34% between 2010 and 2018 (Office for National Statistics 2018b). With consumers eating fewer home cooked meals than ever before, the governance of food-serving businesses is increasingly important. Especially considering that an estimated 60% (Fleetwood et al. 2019) of 2.4 million cases of foodborne disease (Holland et al. 2020) are thought to be contracted whilst eating outside the home. Overseen by the Food Standards Agency (FSA), Local and Unitary Authorities (referred to as LA's for the remainder of this paper) are responsible for enforcing hygiene standards within food businesses in the UK.

Despite the changes in consumer food environments, since 2013 LA spending on food control has reduced from £125million to £101million (National Audit Office 2019). In 2019 the National Audit Authority (NAO) reported severe delays in routine food outlet inspections due to a decrease in food control staff (National Audit Office 2019). The NAO report that only 14% of LA's achieved their inspection targets in 2019. This lack of governance is problematic. Not only are food businesses not receiving the required support, but Critical Health Violations (CHVs), inappropriate procedures and structural problems can go unchecked, exposing consumers to unknown and potentially dangerous levels of risk. Aside from increasing resources, one solution is to adopt a data-driven approach to identify high risk food outlets. As under-reporting of foodborne illness is widespread at both the patient and GP level (Arendt et al. 2013), the use of national surveillance data for such purposes is limited. However, many studies have highlighted the potential of novel data and novel methodological approaches to identify hazardous outlets which show potential to enhance current inspection processes and reduce consumer risk whilst eating outside the home.

1.1 Machine Learning approaches for food Safety

Considered a subset of artificial intelligence, machine learning is increasingly being used across a variety of health-related studies for the purposes of surveillance, prediction, and hypothesis testing. It is concerned with the ability of a system to undertake a specific task without being explicitly programmed to do so. Rather, patterns and inference are used to automatically learn algorithms and improve upon them without a given set of instructions (Bishop 2006). In their work identifying novel and transferrable methods for monitoring foodborne illness, Oldroyd, Morris, and Birkin (2018) categorised approaches for public health and disease surveillance into four categories: basic approach, classification and regression, clustering approaches, and lexicon-based approaches. This research found that supervised classification algorithms, used in 42 of the 62 studies, prove to be the most frequently used. These approaches are often used to yield insight into big datasets, such as Twitter or Consumer Generated Data (Oldroyd, Morris, and Birkin 2018) with a view to rapidly detect disease or illness outbreaks, or supplement traditional surveillance by calculating incidence in a region or over a particular time frame.

Concerned with rapidly identifying reports of foodborne illness, Sadilek et al. (2013) used a popular machine learning model, Support Vector Machine (SVM) to classify 3.8 million tweets gathered from

restaurant visitors in New York over a four-month period. Sadilek et al. (2013) highlighted the power of a machine learning model utilising Consumer Generated Data by identifying 480 potential cases of food poisoning which had not been reported via the Department of Health (DoH). Similarly, Harrison et al. (2014) used a probabilistic classifier to analyse 294,000 restaurant reviews gathered from the online review platform Yelp. 16 cases of foodborne illness were identified which had not been reported via official channels, the results of which were validated via phone interviews. Although both studies identified unreported cases of foodborne illness, both also emphasise that the methods should be used to supplement traditional approaches, as opposed to replace them.

Alongside classifying online restaurant reviews to identify outbreaks of foodborne illness, Nsoesie, Kluberg and Brownstein (2014) also aimed to establish commonly reported pathogenic foods and food vehicles. A high level of agreement was found between the frequency and types of food vehicles reported by consumers in Yelp reviews and those present in the Centre for Disease Control's Foodborne Outbreak Online Database. This study highlights the utility of online reviews not only to identify locations of outbreaks but also implicated food vehicles. As discussed by Sadilek et al. (2013) and Harrison (2014), Nsoesie, Kluberg and Brownstein (2014), also stress the importance of a supplementary approach to enhance current methods of food safety surveillance.

Aside from analysing the content of online reviews, other studies have explored associations between Yelp review scores and CHVs as reported by public health authorities. Kang et al. (2013) reported negative correlations between review scores and frequency of CHVs (where a higher score indications more violations) and subsequently formed robust predictions of health inspection outcomes in Seattle. This suggests that consumer perceived hygiene standards can accurately reflect official DoH inspection scores. Further building on these findings, Schomberg (2016) utilised a logistic regression approach to predict CHV prevalence for food serving businesses in San Francisco and New York. Higher accuracy was reported in areas where Yelp review participation was highest and therefore the use of such data is reliant on a large user base. Although many studies have reported promising results from the use of machine learning approaches and Consumer Generated Data, few studies analysed the demographic representativeness of such data, which could also prove to be a methodological limitation (Oldroyd, Morris, and Birkin 2018).

1.2 Neighbourhood demography and food safety

Many studies have examined patterns of food safety, specifically CHVs at the neighbourhood level without using machine learning approaches (Roberts et al. 2011; Harris et al. 2015; Darcey and Quinlan 2011; Pothukuchi, Mohamed, and Gebben 2008). These studies primarily aim to investigate associations between demographic characteristics and public health inspection outcomes. For example, Darcey and Quinlan (2011) found that whilst deprived areas in Philadelphia had a higher frequency of food outlets, these food outlets had less CHVs than those in less-deprived areas. This study also found that establishments in predominantly Hispanic areas had an increased number of CHV. Although these findings could suggest Hispanic populations are at higher risk of foodborne illness than other ethnicities, the authors postulate that the analysis may also highlight inspection bias. Where underlying factors could influence the frequency of inspections and identification of CHVs.

Pothukuchi, Mohamed, and Gebben (2008) also discuss the associations between socio-demographic characteristics and CHVs in Detroit, Michigan. Specifically, this study found that food outlets in deprived areas and areas with primarily African American populations had an increased number of CHVs compared to other areas. The authors hypothesise that language problems, cultural differences and inexperience with food safety practices could act as potential explanations for these associations. However, they also discuss the likelihood of inspection bias and state that further fine-grained research is required to understand the complex interactions between compliance, neighbourhood ethnicity and the inspection process. In a review of the literature relating to food safety risks for populations of low socioeconomic status and minority ethnic groups, Quinlan (2013) further echoes this sentiment and states that further research is required to understand the socio-economic associations between foodborne illness incidence and both retail access and food handling behaviours.

In the UK, similar studies have undertaken analyses using the Food Hygiene Rating Scheme (FHRS), whereby food outlets are awarded a score on a scale of zero to five following inspection by a public health official. Scores of 3-5 represent compliant hygiene practices and 0-2 represent non-compliance. Oldroyd, Morris, and Birkin (2020) employed logistic regression to identify socio-demographic, urbanness and business type determinants of non-compliant food outlets in England and Wales. Specifically, this work reported that food outlets located in the most deprived quintile were 25% less likely to be compliant compared to those in the least deprived quintile. Although this work utilised small area measures, limitations include problems associated with the Modifiable Areal Unit Problem and ecological fallacy. Fleetwood et al. (2019) also utilised the FHRS data and found associations between low scoring food establishments and contaminated microbiological food samples. This suggest that the FHRS can be used to identify problematic food outlets and can be used as a proxy for foodborne illness risk.

Our review of the literature indicates that many studies have used machine learning approaches to explore patterns of food safety in the form of review scores and reports of foodborne illness via Consumer Generated Data. We also find that studies have explored associations between neighbourhood demography and both FHRS and CHVs, however to our knowledge, none have explored the utility of machine learning approaches to predict food establishment compliance in a UK setting using such neighbourhood data.

1.3 Compliance prediction and class imbalance

One of the main barriers to compliance prediction using FHRS data is a problem known as class imbalance. Supervised machine learning models organise unseen data into classes using an algorithm learnt from pre-labelled training data (Bennet and Campbell 2000). In binary classification problems, if classes are extremely imbalanced in the training phase, an algorithm will favour assignment of unlabelled records to the majority class in an attempt to improve its accuracy metric (Kotsiantis, Kanellopoulos, and Pintelas 2006). 7% of food establishments in England and Wales are non-compliant. Therefore, without sampling, a predictive algorithm could achieve 93% accuracy by labelling all unseen food outlets as compliant.

Many studies have documented ways in which class imbalance can be addressed. For example, in the aforementioned studies, Sadilek et al. (2013) used a method of human-guided machine learning,

whereby tweets belonging to the minority class were actively provided prior to model training. Schomberg (2016) found that Chinese restaurants in San Francisco had a higher number of CHV compared to other restaurants and therefore used only Yelp reviews from Chinese restaurants to train the model. Both studies reported improved outcomes when the models were trained under favourable conditions.

In this work we build upon the work undertaken by Oldroyd, Morris, and Birkin (2020) and utilise neighbourhood and business type predictors of non-compliant food establishments to employ supervised predictive machine learning methods. We analyse under and over sampling approaches to address class imbalance as reported by previous studies. We aim to answer the following research questions; can neighbourhood and business type features be used to predict non-compliant food establishments and if so, which are the most effective algorithms and sampling strategies?

2 Methodology

Three supervised machine learning approaches were trained and tested with a view to predict non-compliance in England and Wales: Linear SVM; Radial SVM; and Random Forest. In this case the algorithms organised unseen data into one of two classes: compliance and non-compliance. Under and over sampling techniques were adopted on the training set to address the problem of class imbalance discussed in section 1.3.

Input features were informed by Oldroyd, Morris, and Birkin (2020) and included business type, urbanness and socio-demographic variables at Output Area (OA) level: Rural and Urban Classification (RUC); region; Output Area Classification (OAC); age; ethnicity; car access; overcrowding; renting and unemployment (see table 1). OAs are chosen as they represent the smallest statistical geography and are designed to be internally homogenous. See Figure 1 for an overview of the analysis.

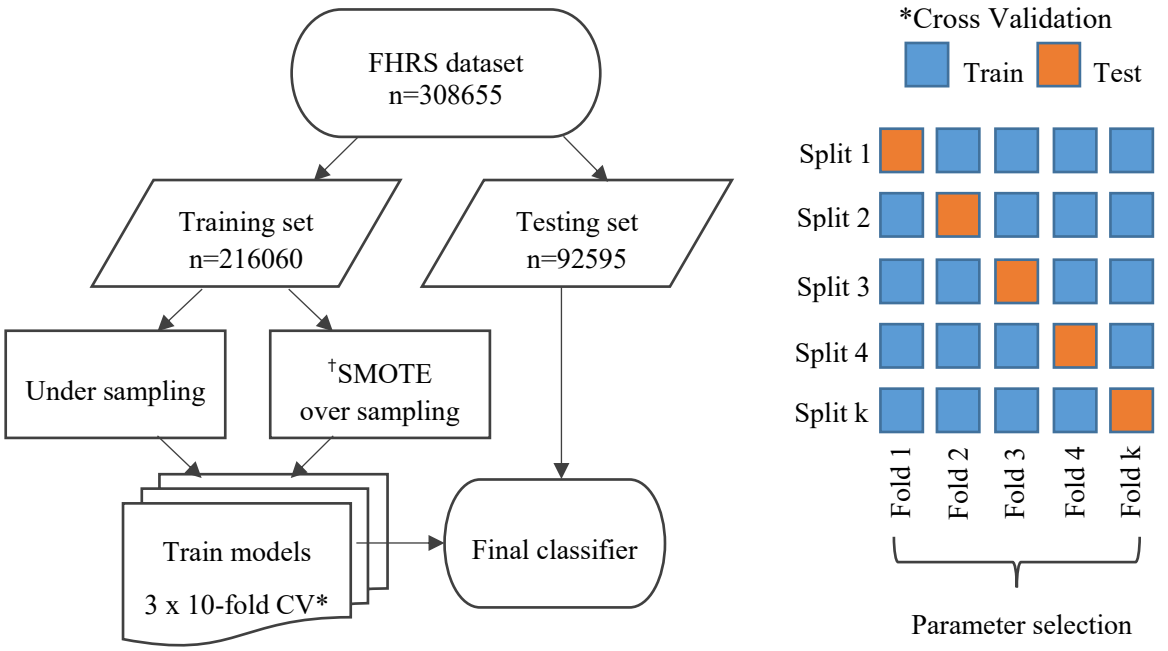

*Figure 1. An overview of the analysis process. The full FHRS dataset is split into training and testing phases prior to under and over sampling. †Synthetic Minority Over Sampling Technique. *Cross Validation.*

2.1 Data and data preparation

Each food establishment was matched to an OA via its postcode in the Office for National Statistics Postcode to OA lookup (Office for National Statistics 2018a). 99.7% of food outlet postcodes were matched to the FHRS dataset using this method. All variables were merged into one dataset using the R statistical programming language, v.5.3.1 (R Core Team 2008) and subsequent analysis was undertake in the Caret package (Kuhn 2008).

Data domain and source	Predictor variables
Food Hygiene Rating Scheme Scores (Food Standards Agency 2020)	Business Type: *restaurants, cafés and canteens; other retailers; super and hyper markets; other catering; pubs, bars and nightclubs; takeaways and sandwich shops; hotels, guesthouses, bed and breakfasts* Region: *East Midland, West Midlands, East of England, London, North East North West, South East, South West, Wales, Yorkshire*
Socio-demographic data (Office for National Statistics 2016)	Age (percentage of persons aged): *0-4; 5-14; 15-19; 20-24; 25-44; 45-64; 65+* Ethnicity (percentage of persons): *Asian, Black, Mixed, Other, White* Unemployment (percentage of persons) Overcrowding (percentage of households) No car access (percentage of households)
Rural Urban Classification (Office for National Statistics 2011b)	RUC Categories: *Urban cities and towns; Rural hamlets and isolated dwellings; Rural town and fringe; Rural village; and Urban conurbation*
Output Area Classification (Office for National Statistics 2011a)	OAC Supergroups: *(1) Rural residents; (2) Cosmopolitans; (3) Ethnicity central; (4) Multicultural metropolitans; (5) Urbanites; (6) Suburbanites; (7) Constrained city dwellers; (8) Hard-pressed living.*

Table 1, predictor variables.

Continuous variables were scaled and centred (percentage of persons in each age and ethnicity category; percentage of non-economically active persons; and percentage of households overcrowded; rented; and without access to a car or van). Categorical variables (RUC; OAC; Business Type; and Region) were included as binary dummies where 1 indicates presence and 0 indicates absence of the variable.

The outcome variable was converted to a categorical binary variable whereby establishments with a FHRS of 3 or above were labelled as ‘compliant’ and those with a score of 2 or below were labelled

as 'non-compliant'. The aligns with the FSA's definitions of 'broadly compliant' and 'not broadly compliant' respectively. Multivariate classification was not possible due to small numbers in individual classes.

2.2 Data Partition

The dataset was divided into training and testing sets with a 70:30 split respectively. Stratified sampling was used to maintain the ratio of compliant and non-compliant establishments in each set. The largest proportion of the dataset was used to train the algorithms (*n=216,060*); the remaining 30% was set aside for testing performance (*n=92,595*). Whilst the training set was under and over sampled, the testing set was not resampled to ensure performance metrics were a true reflection of the model's ability to predict the imbalanced dataset in an applied setting.

2.3 Sampling strategy

We utilised over and under sampling techniques prior to training the three models, to address the class imbalance problem outlined in section 1.3.

2.3.1 Under sampling

All non-compliant establishments in England and Wales with a FHRS score of 2 or less were selected from the training dataset (*n=14,226*). A subset of the majority class, comprised of compliant restaurants with a FHRS score of 3 or higher, were then selected using random sampling. The training set was resampled at ratios of 1:1, 3:2, 2:1, 2:3, 1:2 for non-compliant and compliant establishments. The named sets are 1, 2, 3, 4 and 5 respectively. Under sampling is the least complicated strategy, however it can lead to a weaker classifier due to the reduction of available training data.

2.3.2 Over sampling

Over sampling can be advantageous compared to under-sampling as it maximises the amount of data available during the training process. The minority class was over sampled by generating synthetic data points in the training dataset using the Synthetic Minority Over Sampling Technique (SMOTE). Chawla et al. (2002) suggest that SMOTE, whereby the minority class is over-sampled whilst the majority class is under-sampled, can achieve better classifier performance compared to under-sampling the majority class alone. The SMOTE function from the DMwR package in R (Torgo 2010) was used to generate class ratios as before. The SMOTE method utilises K-Nearest Neighbour (KNN) to generate new data points, see Altman (1992).

2.4 Training phase

We trained a total of 33 models using repeated cross validation (five under sampled, five SMOTE datasets and one unsampled dataset, across the three algorithms), whereby each dataset was split into 10 equal folds using stratified sampling. A subset of the parameters was used to train 9 folds of the data with the 10^{th} fold used to compute performance metrics for that parameter subset. For each fold, this process was repeated three times. Area Under the Receiver Operating Characteristic

Curve (AUC) was computed for each cross-validation iteration and used to select the optimum input and tuning parameters for the final model algorithms.

A ROC curve is a graph generated by plotting the proportion of correctly classified actual positives (in this case non-compliant outlets), known as Sensitivity or True Positive Rate (TPR), against the proportion of correctly classified actual negatives (compliant outlets), referred to as Specificity or the True Negative Rate (TNR), at various probability thresholds. There is often a trade-off between Sensitivity and Specificity whereby as one increases the other decreases. Therefore, a perfect ROC curve assumes the shape of a right angle, which passes through point (1,1) on the graph, indicating 100% specificity and sensitivity and maximising the AUC.

An AUC value of 1 indicates that 100% of the model predictions are correctly classified and a value of 0.5 indicates that only 50% of the classifications are correct, effectively allocating points at random. AUC accounts for both correctly and incorrectly classified data points, and is therefore considered superior to the accuracy metric when evaluating classifiers concerned with imbalanced classes (Ling, Huang, and Zhang 2003).

2.5 Model Specifications

A brief overview of the three model algorithms is provided in the following sections.

2.5.1 Linear Support Vector Machine

Support Vector Machine (SVM) is a non-probabilistic binary classifier which aims to find the optimum hyperplane between two classes in a 2D space. New data points are assigned to one of two classes depending on which side of the hyperplane they fall (Cristianini and Shawe-Taylor 2000). For further information see Vapnik (1998). In the Caret package in R, it is possible to impose a penalty for the misclassification of points during the training process, through the Cost parameter. The higher the Cost parameter the lower the probability of the model misclassifying a point. We vary the value of the Cost parameter throughout the training process using the tuneGrid function in Caret. The optimal and final value of which is reported in Table 2.

2.5.2 Radial Support Vector Machine

In addition to performing linear classification, SVM can also perform non-linear classification by applying a Kernel Trick (Aizerman, Braverman, and Rozonoer 1964). Whereby the model predictors are replaced with kernel functions. This enables the algorithm to operate in a high dimensional implicit feature space; for example, a 3D space. Coordinates of the data points in the newly transformed space are not explicitly calculated, which means this approach is more computationally efficient than others. Instead the relationship between pairs of data is calculated (Aizerman, Braverman, and Rozonoer 1964). Radial SVM in Caret automatically tunes the Cost parameter, the final values of which are reported in Table 2.

2.5.3 Random Forest

The third classification model we employ is Random Forest; an ensemble learning algorithm that averages the outcomes of several decision trees (Fawagreh, Gaber, and Elyan 2014). Combining multiple decision trees can address problems of overfitting, where individual classifiers often learn

highly irregular patterns based upon training data resulting in low bias and extremely high variance, limiting their application beyond the training set. For ensemble learners, the variance of the overall model is decreased without increasing the bias, usually resulting in better performance than individual learners.

Sampling Set / Ratio (non-comp: comp)	*Model Tuning Parameters*					
	Linear SVM (cost)		*Radial SVM (cost)*		*Random Forest (mtry)*	
	SMOTE	*US*	*SMOTE*	*US*	*SMOTE*	*US*
Set 1 (1:1)	1.895	0.842	32	32	5	3
Set 2 (2:1)	0.632	0.947	64	0.25	5	3
Set 3 (3:2)	0.316	0.105	64	2	6	3
Set 4 (2:3)	2	2	16	16	5	3
Set 5 (1:2)	1.368	2	16	0.25	5	3

Table 2, final model tuning parameters. For Linear and Radial SVM, the cost parameter represents the optimal penalty threshold for misclassifications. For the Random Forest models, mtry, the optimal number of randomly selected predictor variables is reported.

Random Forest differs from other ensemble learning methods as only a specific number of randomly sampled input features are available for each learner. The number of input features used in any one fold of the cross-validation process is represented by the $mtry$ argument; in Table 2 we report the final and optimal values for $mtry$. Commonly, learners opt for more predictive input features during training which can result in both overfitting and correlated outcomes between individual learners (Bernard, Heutte, and Adam 2010). The advantage of choosing only a set number of random predictive variables is that those which appear highly predictive in the training set, but which are not in the testing set, are not over sampled during the learning process. For further information see Breiman (2001).

2.6 Testing phase

Following cross-validation, model performance was assessed through the classification of unseen data points; the testing phase. Class probabilities were calculated for each unseen food establishment using the learnt algorithms. Class labels were then assigned based upon a probability threshold, extracted using the $coords$ function from the $pROC$ package (Robin et al. 2011). The optimal probability threshold is defined using Youden's J statistic (Youden 1950). This is the point on the ROC curve which maximises both the distance from the diagonal and therefore the AUC.

For the five top performing algorithms, we assess the impact of assigning a cost weighting to the probability threshold. This acts as a penalty for false negative outcomes, i.e. non-compliant outlets labelled as compliant by the classifier. We envisage a predictive system where highly probable non-compliant food outlets will be flagged for inspection by the LA. In other terms, a system that will supplement the current inspection process to prioritise resources, rather than replace it. Therefore, for this application, the cost of a false negative is far higher than a false positive and subsequently a relative cost of 30 was applied to reduce false negative labels.

For both weighted (applied cost penalty) and unweighted probability thresholds, class labels were assigned and model metrics were calculated to assess performance. Alongside AUC and Kappa, confusion matrices were also generated which include the number of True Positives (TP), non-compliant food outlets correctly classified; False Positives (FP), compliant food outlets incorrectly classified; False Negatives (FN), non-compliant food outlets incorrectly classified; and True Negatives (FN) compliant outlets correctly classified. We compare the predictive power of each algorithm across sampling strategies.

Finally, for the best performing model, we calculate predictor variable importance using the $varImp$ function from the Caret package in R. Importance scores for variables are generated by calculating the mean decrease in accuracy across trees when the variable is excluded. For more information see (Liaw and Wiener 2002). The variable importance scores are scaled between 1 and 100 to aid interpretation.

3 Results

In this section, we first present ROC curves and AUC values using unweighted probability thresholds for all models (figure 2) to give an overview of the predictive power of the algorithms across sampling strategies. For the top five performing models we then present model metrics for labels generated using weighted (applied cost penalty) and unweighted probability thresholds. We assess model performance across algorithms and sampling strategies (table 3). Finally, we present predictor variable importance scores for the top performing model (figure 3).

3.1 ROC curve analysis

ROC curves and AUC values for unweighted probability thresholds are displayed in figure 2. Figure 2 highlights that ROC curves for SMOTE RF models show the most tendency towards a right angle, indicating the best predictive performance. These models produce the highest AUC values (0.859-0.873) followed by SMOTE Radial SVM models (0.740-0.761) and US RF models (0.715-0.718). Linear SVM models for both SMOTE and US datasets reported the lowest AUC values at 0.608-0.698 and 0.660-0.696 respectively.

Of all models, SMOTE Linear 2 shows the shallowest ROC curve with an AUC of 0.608. This model is therefore the weakest classifier in terms of its ability to distinguish between compliant and non-compliant food outlets. We see a large range of AUC values for SMOTE datasets across different models (0.873-0.608), where RF models clearly outperform others, however there is smaller difference in performance between the AUC values for US datasets (0.718-0.660) indicating the sampling strategy has a greater influence over predictive performance than the model algorithm.

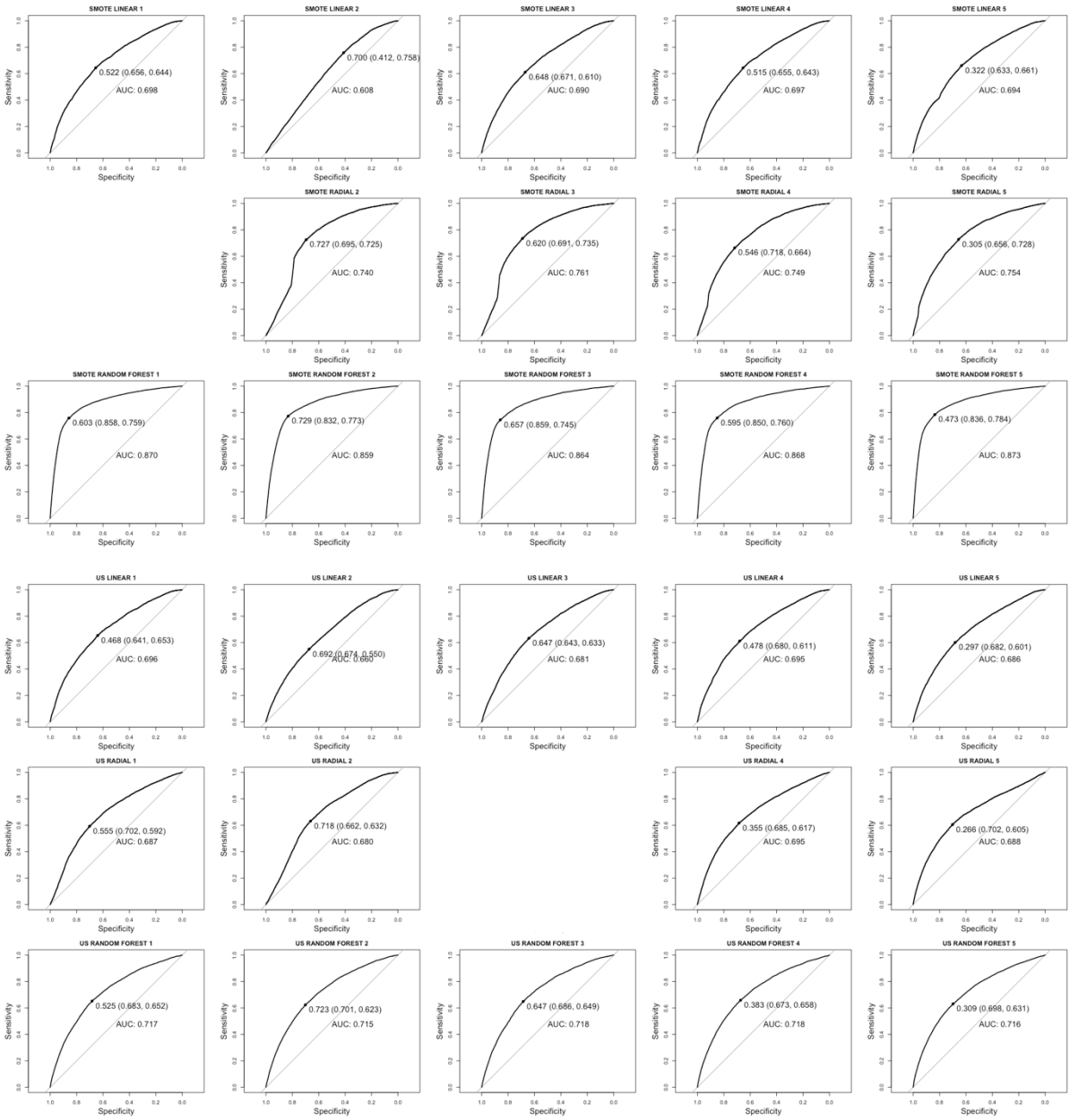

Figure 2, ROC curves and AUC values are generated for each model. Steep ROC curves and high AUC values indicate better performance than shallow ROC curves and low AUC values. SMOTE Radial 1 and US Radial 3 did not converge.

3.2 Application of cost penalty

Analysis of the ROC curves show that the SMOTE RF algorithms have the best predictive power based upon AUC, specificity and sensitivity values. In an attempt to further increase the sensitivity of these models, and therefore classification of the non-compliant class, we apply a cost penalty for false negative classifications as described in section 2.6. Weighted (applied cost penalty) and unweighted model metrics for the five SMOTE RF models are presented in table 3. Here we also include the results of the RF model trained on an unsampled dataset. The size of the testing set remains constant to allow comparisons.

Application of the cost penalty lowers probability thresholds compared to unweighted models, resulting in an increase in the number of records classified as non-compliant. Although this strategy increases the number of FP classifications, it results in a large reduction of FN classifications, equal to 35%, 26%, 36%, 39% and 30% for RF sets 1 to 5 respectively. Sensitivity measures for weighted models are much higher (0.833-0.853) than unweighted sensitivity measures (0.745-0.784); reflecting an increase in the predictive power of the non-compliant class.

Application of the cost penalty negatively effects the overall model metrics, with weighted models exhibiting lower AUC, Kappa and precision values compared to their weighted counterparts. Where precision represents the fraction of non-compliant outlets among those labelled as such. However, in an applied setting a reduction in FN classifications takes precedence over model metrics. Of all weighted classifiers, RF set 4 reports the lowest number of FN classifications (895) and the highest sensitivity (0.853), however this model also reports the highest number of FPs, indicating a move towards a 'catch-all' approach compared to other classifiers.

This 'catch-all' approach is clearly exemplified in the weighted model metrics for the RF model trained with an unsampled dataset. Here the rate of non-compliant records remains equivalent to the original data at 7%. The probability threshold for this model is extremely low at 0.021. Therefore, the model will label all data points above this probability threshold as non-compliant. Of 92,595 unseen records in the test set a total of 83,494 were labelled as non-compliant by the unsampled model and of these 77,591 were incorrect classifications with an overall precision of 0.071. Of all classifiers, on balance RF set 1 is adopted as the final classifier as it reports the highest kappa (0.230) and precision (0.192) values of the weighted models, indicating the lowest number of misclassifications of both classes whilst reporting low values of FN classifications (957).

	RF Set 1 n=92595		*RF Set 2 n=92595*		*RF Set 3 n=92595*		*RF Set 4 n=92595*		*RF Set 5 n=92595*		*RF unsampled n=92595*	
	unweighted	*weighted*	*unweighted*	*weighted*	*unweighted*	*weighted*	*unweighted*	*weighted*	*unweighted*	*weighted*	*unweighted*	*weighted*
Probability Threshold	*0.603*	*0.481*	*0.729*	*0.645*	*0.657*	*0.515*	*0.595*	*0.459*	*0.473*	*0.367*	*0.067*	*0.021*
AUC	*0.87*	*0.87*	*0.859*	*0.859*	*0.864*	*0.864*	*0.868*	*0.868*	*0.873*	*0.873*	*0.796*	*0.796*
Sensitivity	*0.759*	*0.843*	*0.773*	*0.833*	*0.745*	*0.838*	*0.76*	*0.853*	*0.784*	*0.849*	*0.661*	*0.859*
Specificity	*0.858*	*0.745*	*0.836*	*0.741*	*0.859*	*0.737*	*0.85*	*0.724*	*0.836*	*0.737*	*0.797*	*0.481*
True Positives	*4624*	*5139*	*4712*	*5076*	*4540*	*5107*	*4630*	*5201*	*4781*	*5175*	*4029*	*5903*
False Positives	*12264*	*21676*	*14572*	*22383*	*12180*	*22752*	*12976*	*23872*	*14210*	*22773*	*17571*	*77591*
TN	*74235*	*64823*	*71924*	*64116*	*74319*	*63747*	*73523*	*62627*	*72289*	*63726*	*68928*	*8908*
FN	*1472*	*957*	*1384*	*1020*	*1556*	*989*	*1466*	*895*	*1315*	*921*	*2067*	*193*
Kappa	*0.338*	*0.230*	*0.301*	*0.218*	*0.334*	*0.216*	*0.325*	*0.210*	*0.313*	*0.220*	*0.210*	*0.010*
Precision	*0.274*	*0.192*	*0.244*	*0.185*	*0.272*	*0.183*	*0.263*	*0.179*	*0.252*	*0.185*	*0.187*	*0.071*

Table 3 weighted and unweighted performance metrics for Random Forest models utilising SMOTE datasets across 5 sampling strategies. Where weighted observations have a cost penalty applied (30) when extracting the optimal probability threshold. Where precision is the proportion of correctly classified non-compliant outlets.

In figure 3 we present variable importance scores for SMOTE RF set 1 predictor variables. These are scaled between 1 and 100 to aid interpretation. Although we decide to calculate importance scores to further undertand the prediction outcomes, there are limitations associated with variable importance scores for entropy based classifiers which we discuss further in section 4.3. Furthermore, it is not possible to ascertain the directionality of the predictive strength, specifically whether a high or low value contributes towards a non-compliant class label.

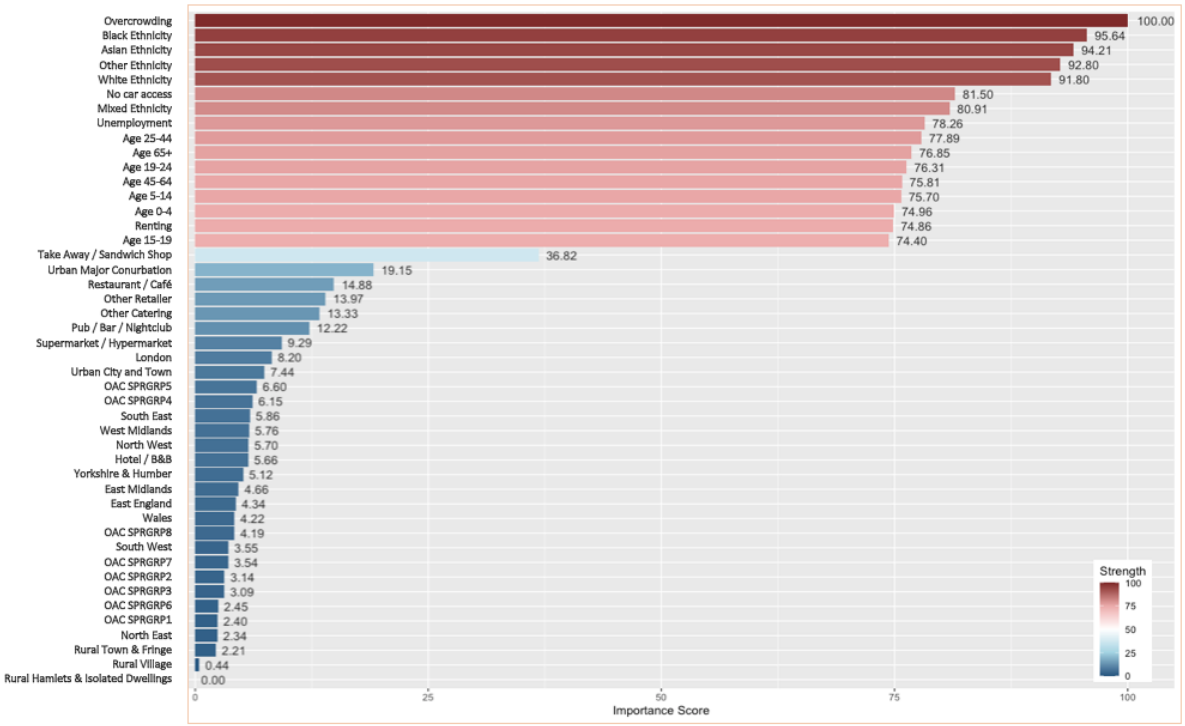

Figure 3 variable importance scores for SMOTE RF set 1. Where red variables have higher predictive strength and blue variables have lower predictive strength.

Overcrowding, with a score of 100, is reported as the most predictive variable for food establishment compliance in the SMOTE RF1 model. This is closely followed by Ethnicity categories, Black (95.64), Asian (94.21), Other (92.80) and White (91.80). Following the ethnicity categories, no car access scored 81.50, mixed ethnicity scored 80.9 and percentage of unemployed scored 78.26. Age categories were also reported to be highly predictive with scores between 74.40 and 77.89. For other variables we see a large drop in the reported predictive power. Takeaways and sandwich shops shows mild predictive strength with variable importance score of 36.82 and all other business types, RUC, region and OAC variables score below 20. Rural Hamlets and Isolated Dwellings are reported as the least predictive variable with a score of 0.

4 Discussion

Of the three models, we find that Random Forest algorithm produces the strongest predictive classifier. Of the adopted sampling strategies, models utilising SMOTE training data at a 1:1 ratio yield the best results and outperform models trained with under sampled data and SMOTE data at different ratios. Furthermore, we find that this sampling strategy greatly improves model metrics, and the frequency of FN predictions compared to unsampled data. In this section we discuss the

implications of our findings before discussing the model algorithms, sampling strategies and variable importance.

4.1 Policy Implications

Although the prediction of food outlet compliance does not help to meet the LA targets with regards to required inspections, it could reduce consumer risk by capturing high risk establishments earlier in the inspection cycle. It is clear how a classifier of this nature could advance current food inspection processes, however as we will discuss in section 4.2, we must ensure that predictions of this nature do not enforce biases which may or may not already be present in the system.

In an applied setting, if our test set were to represent foot outlets awaiting inspection, we show that approximately 26,815 (28%) of 92,595 outlets could be prioritised for inspection by the LA. Of these establishment inspections, we expect approximately 5139 to result in a FHRS score of 2 or below; capturing 84% of the total number of non-compliant outlets. 72% of outlets would be labelled as lower priority, easing the stain on Local Authority inspectors and reducing consumer risk.

It should be noted that the figures reported by our model represent averages across LA's in England and Wales and therefore the predictive strength of the model will vary geographically. In further work we will examine spatial variations in model predictions.

4.2 Appropriateness of algorithms and sampling strategies

Overwhelmingly, we find that the classification problem cannot be solved by Linear SVM. These models reported the lowest metrics across both SMOTE and US datasets, with some linear models reporting equivalent performance to a random classifier. Radial SVM, which transforms data into a non-linear space, performed slightly better than its linear equivalent, however the Random Forest algorithms reported the best predictive power for unseen data points. Therefore, future work will look at advanced learners such as gradient boosted decision trees which may further improve performance metrics.

Of sampling strategies, we found SMOTE to be the most effective for compliance prediction. Unlike under sampling, where a large proportion of the majority class is discarded resulting in the loss of important contextual information, SMOTE retains most of the data points in the majority class whilst adding synthetic points to the minority class. Subsequently, SMOTE training datasets are inherently larger than under sampled counterparts, which could somewhat explain the differences in predictive strength between models. We find that compliance prediction using unsampled data is not possible

With regards to sampling ratios, we generally see that models utilising training sets which best reflect real-world ratios, i.e. those with a higher frequency of compliant food outlets, report higher model metrics than those with a higher frequency of non-compliant outlets. Set 2, with the highest frequency of non-compliant outlets compared to compliant outlets, with a ratio of 2:1, reports the lowest AUC values for SMOTE and US datasets across all models. SMOTE Radial 3 and US RF 3, with a sampling ratio of 3:2 non-compliant to compliant outlets respectively, are the only sets with a higher frequency of non-compliant outlets which equal or outperform other sampling ratios for corresponding models. The difference in performance is marginal. Future work will look towards

supplementing the minority class with additional data such as historical records of non-compliant food outlets to attain more data whilst maintaining representative ratios.

4.3 Variable Importance

Although we calculate variable importance to further understand the outcomes of our model, the way in which scores are calculated for entropy based classifiers mean they should be approached with caution (Deng, Runger, and Tuv 2011; Breiman et al. 1984). Decision tree algorithms, such as random forest, attempt to reduce entropy at each division in the data. Variables with multiple levels or values inherently provide more flexibility for data partition compared to categorical variables and are subsequently afforded greater importance (Strobl, Boulesteix, and Augustin 2007).

We clearly see this effect in the variable importance scores for our predictive model. Continuous numeric variables representing overcrowding, ethnicity, age, unemployment, no car access and renting have high predictive importance compared to categorical variables: RUC, region, OAC and business type. These findings contrast those found previously by Oldroyd, Morris, and Birkin (2020) where large effect sizes were found for business type and RUC categories (takeaways, sandwich shops and major urban conurbations) alongside smaller but significant associations for Mixed, Asian, Black and Other ethnicities. This study did however find a clear gradient of association between increased deprivation and non-compliance, which could align with the predictive strength of unemployment, rented households, no car access and overcrowding in our model. All are variables which are frequently used in indices of deprivation (McLennan et al. 2019; Townsend, Phillimore, and Beattie 1988).

Many studies consider the relationship between ethnicity and food safety (Darcey and Quinlan 2011; Pothukuchi, Mohamed, and Gebben 2008; Byrd-Bredbenner et al. 2013). Many postulate that associations between perceived unsafe practices and food businesses serving ethnic cuisine may result from inspection biases rather than a causal relationship. Pham et al. (2010) state that many health inspectors report communication problems with business owners who do not speak English. Furthermore, Darcey and Quinlan (2011) discuss the influence of biases on health inspections, such as confirmation bias, which exists when inspection results align with a preconceived notion based on the type of neighbourhood within which the outlet is located.

Based on the limitations of variable importance score calculations in entropy based classifiers we do not hypothesise about the relationship between highly predictive variables and non-compliance in our model. As discussed, our findings maybe a result of inspection bias rather than a causal relationship. In future work we will explore methods such as partial permutations (Altman et al. 2010) or growing unbiased trees (Painsky and Rosset 2017) with the aim to provide a higher degree of confidence in the variable importance scores of predictive variables, such that we can postulate about associations. These algorithms are designed to reduce bias towards continuous variables during variable selection. We will also undertake further fine-grained analysis of the composite measures that make up the FHRS scores, such as confidence in management scores, to further unpick the relationships between highly predictive variables and model outcomes.

4.4 Methodological Limitations

As with all studies utilising area-based measures, there may be problems associated with the geographical scale of analysis. The Modifiable Areal Unit Problem arises when point data, such as individuals or households, are aggregated to polygons (Openshaw and Taylor 1979). The size and shape of the polygon will determine the aggregated units and therefore different results maybe reported when using an alternative geography. We use a small geography to minimise the effect, however this cannot be entirely mitigated without using individual level data.

Our analysis assumes that the FHRS data reflect current hygiene scores at food serving establishments across England and Wales. When calculating model metrics and the number of false and true positives and negatives we use the FHRS data as a gold-standard measure. However as only 14% of Local Authorities are up to date with their planned inspections, this measure may not entirely represent of current food hygiene practices and our calculated metrics may under or overestimate model performance as a result. Furthermore, False Positive predictions, i.e. food establishments predicted non-compliant but with a compliant FHRS score, may provide strong indication of where repeat inspections should be undertaken.

5 Conclusions

Using socio-demographic, business type and neighbourhood data we determine a Random Forest model that is able to predict non-compliant food outlets in England and Wales with 84% sensitivity. We show that the SMOTE technique with a 1:1 sampling strategy is effective at addressing problems associated with class imbalance. Namely model metrics are improved and False Negative classifications are reduced compared to a model utilising unsampled and highly imbalanced data. We discuss the problems associated with variable importance scores for entropy based classifiers and recommend that alternative methods be used to identify determinants of non-compliance. As Local Authorities find themselves under increasing pressure due to cuts in funding and demand from other services, this data driven approach to food safety control could prove useful in allocating scarce resources, reducing consumer exposure to unsafe food practices and by extension reduce the incidence of foodborne illness.

6 Funding Information

The authors wish to acknowledge funding from the Food Standards Agency and the Economic and Social Research Council (grant number ES/J500215/1).

7 Author Contributions

Formal analysis, Rachel A Oldroyd; Investigation, Rachel A Oldroyd; Methodology, Rachel A Oldroyd; Supervision, Michelle A Morris and Mark Birkin; Writing – original draft, Rachel A Oldroyd; Writing – review & editing, Rachel A Oldroyd, Michelle A Morris and Mark Birkin.

8 Conflicts of Interest

The authors have no conflicts of interest to declare.

References

Aizerman, M., E. Braverman, and L. Rozonoer. 1964. 'Theoretical foundations of the potential function method in pattern recognition learning.', *Automation and Remote Control* 25: 821-37.

Altman, A., L. Tolosi, O. Sander, and T. Lengauer. 2010. 'Permutation importance: a corrected feature importance measure', *Bioinformatics*, 26: 1340-47.

Altman, N.S. 1992. 'An introduction to kernal and nearest-neighbour nonparametric regression', *The American Statistician*, 46: 175-85.

Arendt, S., L. Rajagopal, C. Strohbehn, N. Stokes, J. Meyer, and S. Mandernach. 2013. 'Reporting of Foodborne Illness by U.S. Consumers and Healthcare Professionals ', *Int. J. Environ. Res. Public Health*, 10: 3684-714.

Bennet, K.P., and C. Campbell. 2000. 'Support Vector Machines: Hype or Hallelujah?', *SIG KDD Explorations*, 2: 1:13.

Bernard, S., L. Heutte, and S. Adam. 2010. *A Study of Strength and Correlation in Random Forests* (Springer Verlag: Heidelberg).

Bishop, C.M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag: Berlin, Heidelberg).

Breiman, L. 2001. 'Random Forests', *Machine Learning*, 45: 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and regression trees* (Chapman and Hall New York).

Byrd-Bredbenner, C., J. Berning, J. Martin-Biggers, and V. Quick. 2013. 'Food Safety in Home Kitchens: A Synthesis of the Literature', *Int J Environ. Res. Public Health* 10: 4060-85.

Chawla, N.V., K.W. Bowyer, L.O Hall, and W.P. Kegelmeyer. 2002. 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research* 16: 321-57.

Cristianini, N., and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernal-Based Learning Methods* (Cambridge University Press: Cambridge, UK).

Darcey, V.L., and J.J. Quinlan. 2011. 'Use of geographic information systems technology to track critical health code violations in retail facilities available to populations of different socioeconomic status and demographics', *J Food Protection*, 74: 1524-30.

Deng, H., G. Runger, and E. Tuv. 2011. 'Bias of importance measures for multi-valued attributes and solutions ', *Proceedings of the 21st International Conference on Artificial Neural Networks*: 293-300.

Fawagreh, K., M.M. Gaber, and E. Elyan. 2014. 'Random forests: from early developments to recent advancements', *Systems Science & Control Engineering*, 2: 602-09.

Fleetwood, J., S. Rahmanb, D. Holland, D. Millson, L. Thomson, and G. Poppy. 2019. 'As clean as they look? Food hygiene inspection scores, microbiological contamination, and foodborne illness', *Food Control*, 96: 76-86.

Food Standards Agency. 2016. 'The 'Food and You' survey ', Accessed 26/09/2018. http://www.food.gov.uk/science/research-reports/ssresearch/foodandyou.

Food Standards Agency. 2020. 'Food Hygiene Rating Schemes', Accessed 27/02/2020. https://www.food.gov.uk/safety-hygiene/food-hygiene-rating-scheme.

Harris, K.J., K.S. Murphy, R.B. DiPietro, and G.L. Rivera. 2015. 'Food safety inspections results: A comparison of ethnic-operated restaurants to non-ethnic-operated restaurants', *International Journal of Hospitality Management*, 46: 190-99.

Harrison, C., M. Jorder, H. Stern, F. Stavinsky, V. Reddy, H. Hanson, H. Waechter, L. Lowe, L. Gravano, and S. Balter. 2014. 'Using online reviews by restaurant patrons to identify unreported cases of foodborne illness - New York City, 2012-2013 ', *Morbidity and Mortality Weekly Report*, 63: 441-45.

Holland, D., L. Thomson, N. Mahmoudzadeh, and A. Khaled. 2020. 'Estimating deaths from foodborne disease in the UK for 11 key pathogens', *BMJ Open Gastroenterology*, 7.

Kang, J. S., P. Kuznetsova, Y. Choi, and M. Luca. 2013. "Using Text Analysis to Target Government Inspections: Evidence from Restaurant Hygiene Inspections and Online Reviews." In, 1-5. Cambridge, Massachusetts: Harvard Business School.

Kotsiantis, S., D. Kanellopoulos, and P. Pintelas. 2006. 'Handling imbalanced datasets: A review', *GESTS International Transactions on Computer Science and Engineering*, 30.

Kuhn, M. 2008. 'Building Predictive Models in R Using the caret Package', *Journal of Statistical Software*, 28: 1-26.

Liaw, A., and M. Wiener. 2002. 'Classification and Regression by randomForest', *R News*, 2: 18-22.

Ling, C.X., J. Huang, and H. Zhang. 2003. "Ling, C.X., Huang, J. and Zhang, H., 2003, August. AUC: a statistically consistent and more discriminating measure than accuracy." In *International Joint Conferences on Artificial Intelligence* Acapulco, Mexico.

McLennan, D., S. Noble, M. Noble, E. Plunkett, G. Wright, and N. Gutacker. 2019. "The English Indices of Deprivation 2019: technical report." In.: Ministry of housing, communities & local government

National Audit Office. 2019. "Ensuring food safety and standards." In, edited by National Audit Office.

Nsoesie, E.O., S.A. Kluberg, and J.S. Brownstein. 2014. 'Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports', *Preventative medicine*: pp.264-69.

Office for National Statistics. 2011a. "2011 OAC Clusters and Names." In.

Office for National Statistics. 2011b. "Rural and Urban Classification." In.

Office for National Statistics. 2016. "2011 Census Aggregate Data " In.: UK Data Service (Edition: June 2016).

Office for National Statistics. 2018a. "Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2018) Lookup in the UK." In, edited by ONS Geography.

Office for National Statistics. 2018b. "UK business: activity, size and location." In.

Oldroyd, R.A., M.A. Morris, and M. Birkin. 2018. 'Identifying Methods for Monitoring Foodborne Illness: Review of Existing Public Health Surveillance Techniques', *JMIR Public Health Surveill*, 4: e57.

Oldroyd, R.A., M.A. Morris, and M. Birkin. 2020. 'Food Safety Vulnerability: Neighbourhood determinants of non-compliant establishments in England and Wales', *Health and Place*, 63.

Openshaw, S., and P. Taylor. 1979. 'A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem', *Statistical Methods in the Spatial Sciences*.

Painsky, A., and S. Rosset. 2017. 'Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance', *EEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 2142-53.

Pham, M.T., A.Q. Jones, J.M Sargeant, B.J Marshall, and C.E. Dewey. 2010. 'A qualitative exploration of the perceptions and information needs of public health inspectors responsible for food safety', *BMC Public health* 10: 1-9.

Pothukuchi, K., R. Mohamed, and D. Gebben. 2008. 'Explaining disparities in food safety compliance by food stores: does community matter?', *Agriculture and Human Values*, 25: 319-32.

Quinlan, J.J. 2013. 'Foodborne Illness Incidence Rates and Food Safety Risks for Populations of Low Socioeconomic Status and Minority Race/Ethnicity: A Review of the Literature', *Int. J. Environ. Res. Public Health*, 10: 3634-52.

R Core Team. 2008. "R: A language and environment for statistical computing." In. Vienna, Austria: R Foundation for Statistical Computing.

Roberts, K., J. Kwon, C. Shanklin, P. Liu, and W.S. Yen. 2011. 'Food safety practices lacking in independent ethnic restaurants', *Journal of Culinary Science and Technology*, 9: 1-16.

Robin, X, N Turck, A Hainard, N Tiberti, F Lisacek, JC Sanchez, and M Müller. 2011. 'pROC: an open-source package for R and S+ to analyze and compare ROC curves', *BMC Bioinformatics*, 12: 2-8.

Sadilek, A., S. Brennan, H. Kautz, and V. Silenzio. 2013. "nEmesis: Which Restaurants Shold You Avoid Today?" In *First AAAI Conference on Human Computation and Crowdsourcing*, pp.138-46. Palm Springs, California: AAAI Press.

Schomberg, J.P., O.L. Haimson, G.R. Hayes, and H. Anton-Culver. 2016. 'Supplementing public health inspection via social media', *PLoS ONE*, 11 (3) (no pagination): 1-21.

Strobl, C., A. Boulesteix, and T. Augustin. 2007. 'Unbiased split selection for classification trees based on the Gini index', *Computational Statistics & Data Analysis*, 52: 483-501.

Torgo, L. 2010. *Data Mining using R: learning with case studies* (CRC Press).

Townsend, P., P. Phillimore, and A. Beattie. 1988. *Health and deprivation: inequalities and the north,* (Croom Helm).

Vapnik, V.N. 1998. *Statistical Learning Theory* (Wiley: New York).

Youden, J. 1950. 'Index for rating diagnostic tests', *Cancer*, 3: 32-25.