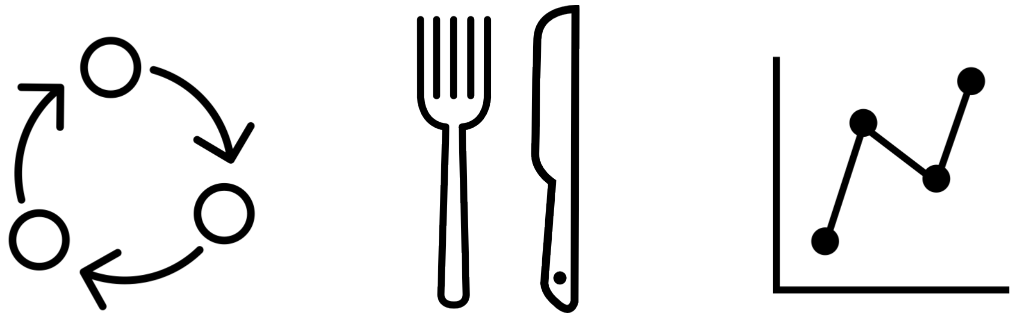
Predicting non-compliant food outlets in England and Wales using neighbourhood characteristics: a machine learning approach

Rachel Oldroyd, Dr Michelle Morris, Prof Mark Birkin

GEOG5927 Predictive Analytics| 19 July 2021

Contents

- Background & Rationale
- Data & Methods
- Results
- Discussion
- Real world application
- Future work



Background

Local Authorities (LAs) enforce food standards

- Overseen by Food Standards Agency (FSA)

Every food serving business is inspected by a Food Hygiene Officer & awarded a Food Hygiene Rating Scheme (FHRS) score*

Routine inspections occur every 6 months – 5 years

New businesses should be inspected within 4 months of opening.

*Scotland operates a pass / fail system

Consumer Data Research Centre

An ESRC Data Investment

Rationale

LAs are struggling to meet their inspection targets:

- Only 2% of LA's in the UK have no overdue inspections*
- 18% of LA's have over 20% of businesses overdue an inspection*
- Recent work suggests this has worsened during 2020

* National Audit Office 2019

Rationale

Business owners not receiving support

Consumers are exposed to unknown levels of risk

Extremely problematic -> 60% of foodborne illness is contracted outside the home

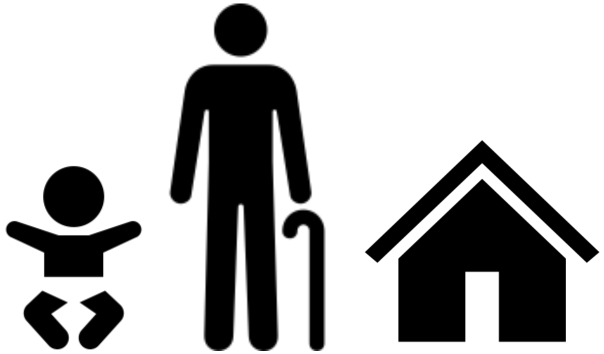
Foodborne illness affects ~2 million people annually

- At a cost of 1.6 billion GBP

Context

Previous studies have reported significant associations between food outlet compliance and neighbourhood characteristics:

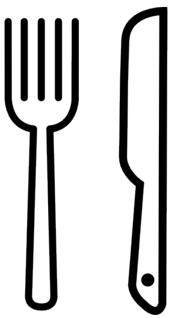
- Urbanness
- Demographics
  - Age
  - Ethnicity
  - Deprivation



Context

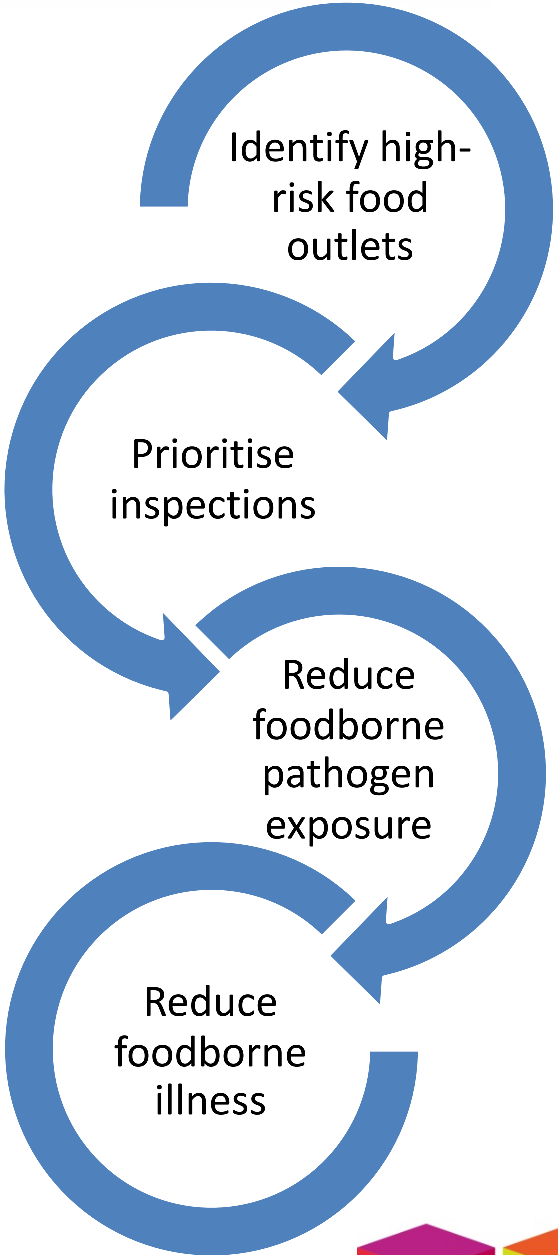
Oldroyd, Morris, Birkin (2020)*:

- Food outlets in the most deprived areas and large urban areas are less likely to comply with hygiene standards (25% & 32% respectively)
- Takeaways, sandwich shops are 50% less likely to be compliant than restaurants
- Small but significant associations were also found between some age categories, all non-white ethnicities and non-compliance
  - Some populations at higher-risk



*https://pubmed.ncbi.nlm.nih.gov/32217280/

Aim



Explore the utility of machine learning to predict high-risk (non-compliant) food outlets in England and Wales

- Using neighbourhood characteristics

Consumer Data Research Centre

An ESRC Data Investment

Data

Food Hygiene Rating Scheme (FHRS): All food businesses rated from 0-5 to reflect hygiene standards at time of inspection.

- Confidence in management, hygiene, structural integrity

Data

5	Hygiene standards are very good
4	Hygiene standards are good
3	Hygiene standards are generally satisfactory
2	Some improvement is necessary
1	Major improvement is necessary
0	Urgent improvement is required

Not broadly compliant = FHRS ≤ 2

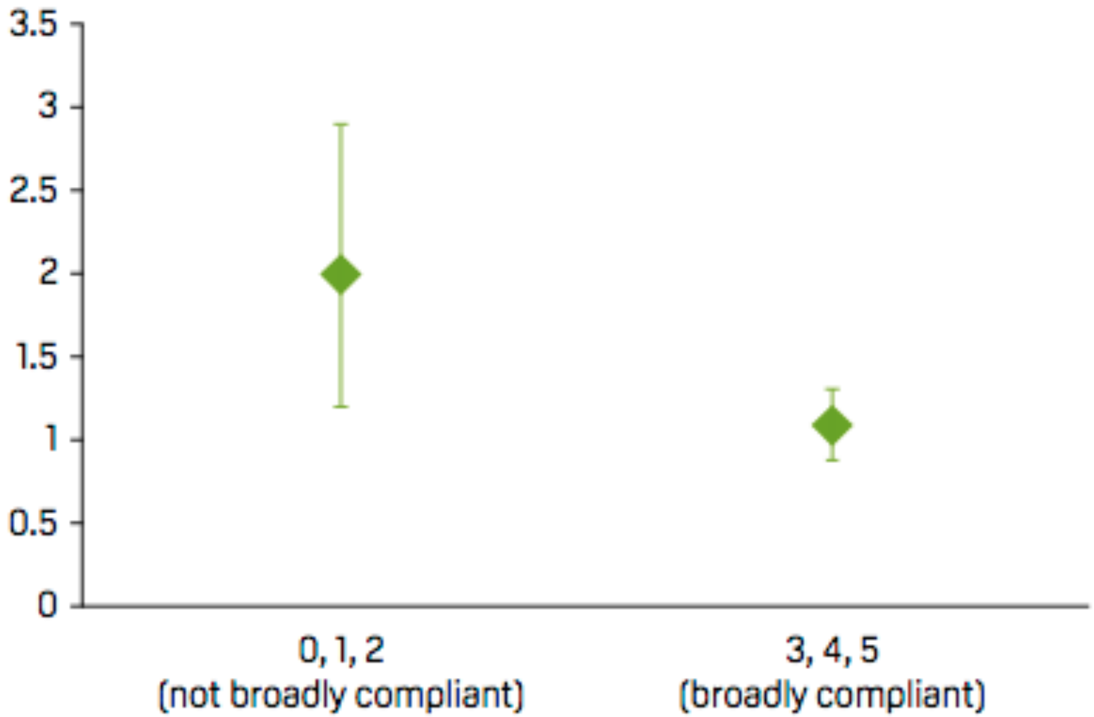
Broadly compliant = FHRS ≥ 3

Consumer Data Research Centre

An ESRC Data Investment

Data

Outbreaks per 10,000 restaurants per year



Note: The graph includes error bars. Error bar are a graphical representation of the variability of data used to show the error, or uncertainty in a reported measurement. Errors bars illustrated here show the 95% confidence interval.

G Poppy. 2017. Chief Scientific Advisor Report

2x

Outbreaks are twice as likely at non-compliant establishments compared to compliant ones.

Data

Outcome variable

- FHRS score converted to binary variable (0,1)
- 1 = non-compliant outlets (FHRS ≤ 2)
- 0 = compliant outlets (FHRS ≥ 3)

Consumer Data Research Centre

An ESRC Data Investment

Data

Predictor variables:

- Business type
- Region

- Age (% individuals)
  - 0-4, 5-24, 25-44, 45-64, 65+
- Ethnicity (% individuals)
  - Asian, Black, Mixed, Other, White
- No car access (% households)
- Renting (% households)
- Overcrowding (% households)
- Unemployment (% individuals)
- Rural Urban Classification (RUC)
- Output Area Classification (OAC)

Output Areas

2011 Census

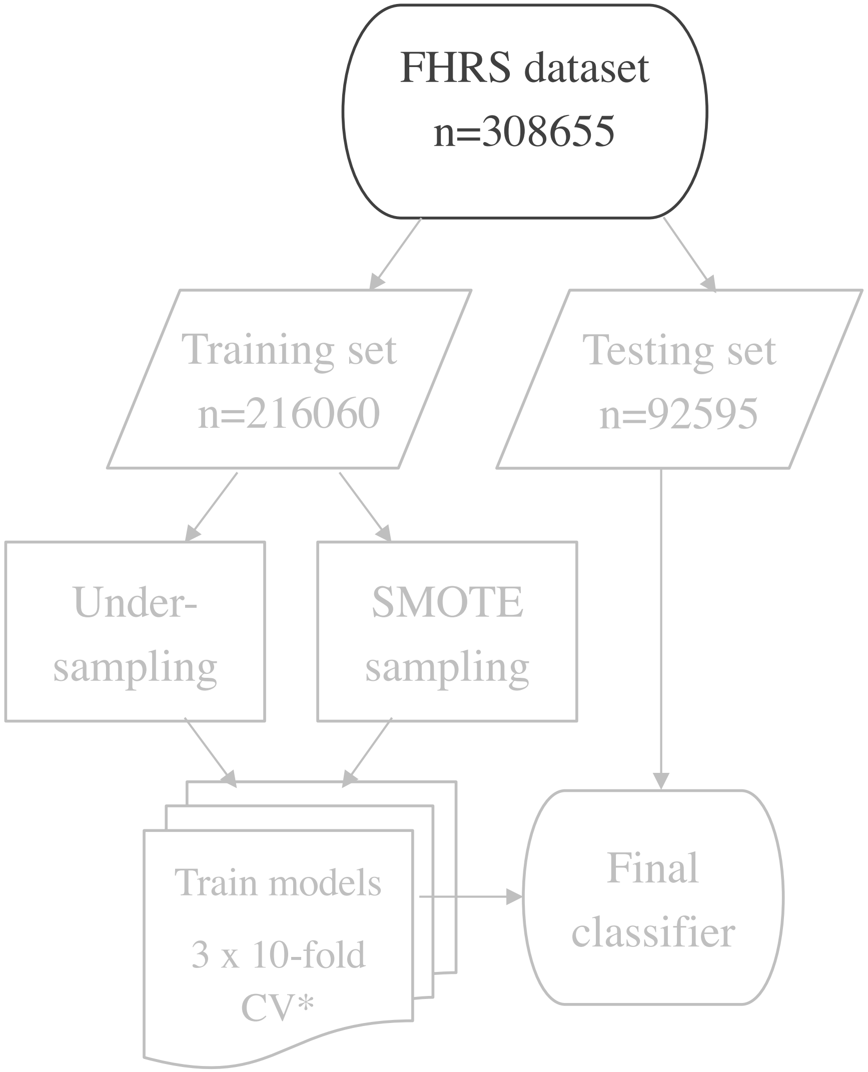
Consumer Data Research Centre

An ESRC Data Investment

Data

Data domain / source	Variable	Categories /levels
Food Hygiene Rating Scheme Scores (Food Standards Agency 2020)	*FHRS score (ordinal)*	*0 (Improvement necessary), 1, 2, 3, 4, 5 (Very good)*
	Business Type (categorical)	*Restaurants, cafés, & canteens; other retailers; super & hyper markets; other catering; pubs, bars & nightclubs; takeaways & sandwich shops; hotels, guesthouses, bed & breakfasts*
	Region (categorical)	*East Midland, West Midlands, East of England, London, North East, North West, South East, South West, Wales, Yorkshire*
Socio-demographic 2011 census data (Office for National Statistics 2016)	*Age (% of persons)*	*0-4; 5-14; 15-19; 20-24; 25-44; 45-64; 65+*
	Ethnicity (% of persons)	*Asian, Black, Mixed, Other, White*
	Unemployment (% of persons)	
	Overcrowding (% of households)	
	No car access (% of households)	
	Renting (% of households)	
Rural Urban Classification (Office for National Statistics 2011b)	*RUC (categorical):*	*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 (categorical):*	*(1) Rural residents; (2) Cosmopolitans; (3) Ethnicity central; (4) Multicultural metropolitans; (5) Urbanites; (6) Suburbanites; (7) Constrained city dwellers; (8) Hard-pressed living.*

Method Overview



- Office for National Statistics (ONS) postcode to OA lookup -> attach neighbourhood characteristics to food outlets (99.7% match)

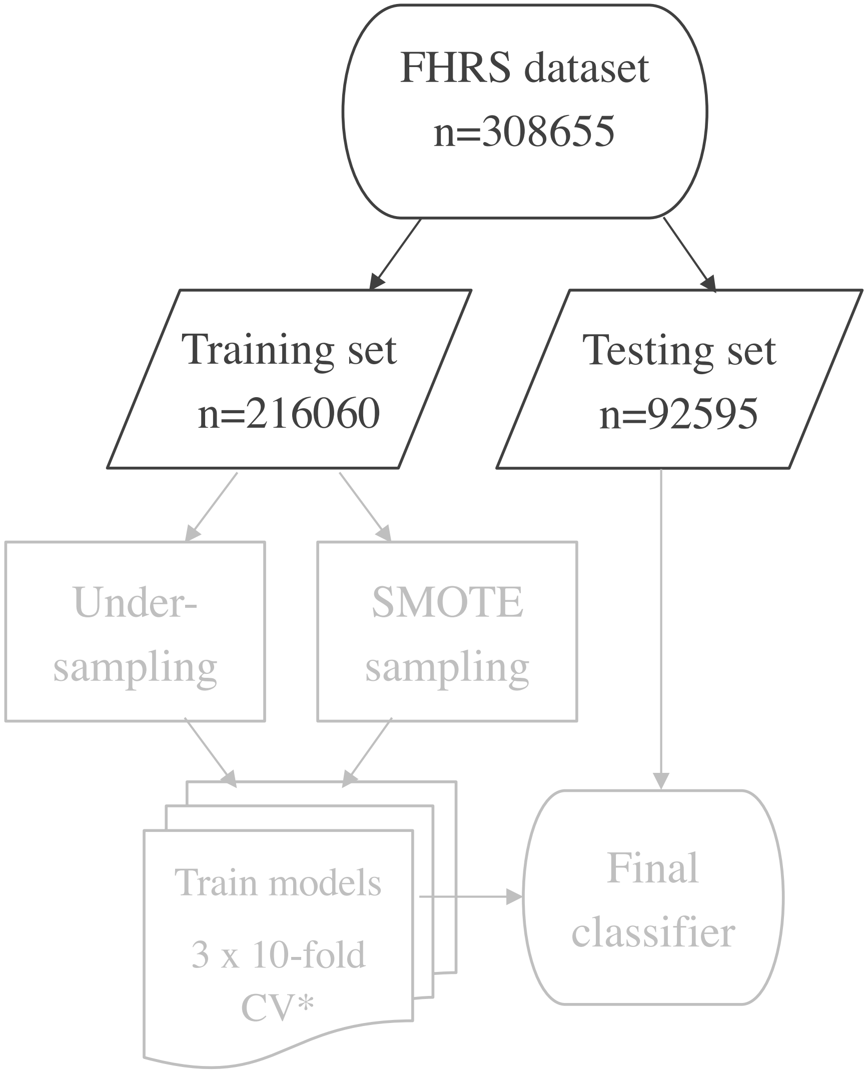
- FHRS scores converted to binary variable (0,1)

- Categorical variables converted to dummy variables (0,1)

Consumer Data Research Centre

An ESRC Data Investment

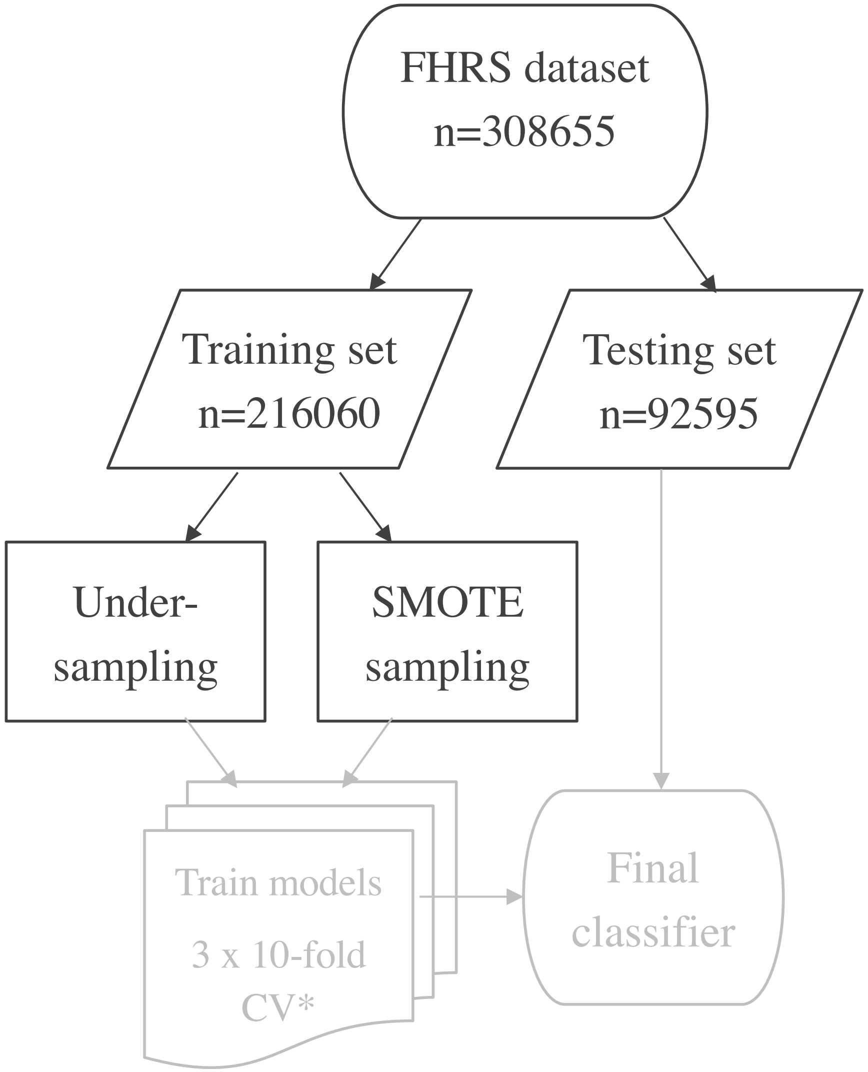
Method Overview



Split data using stratified sampling:

- Training set (70%)
- Test set (30%)

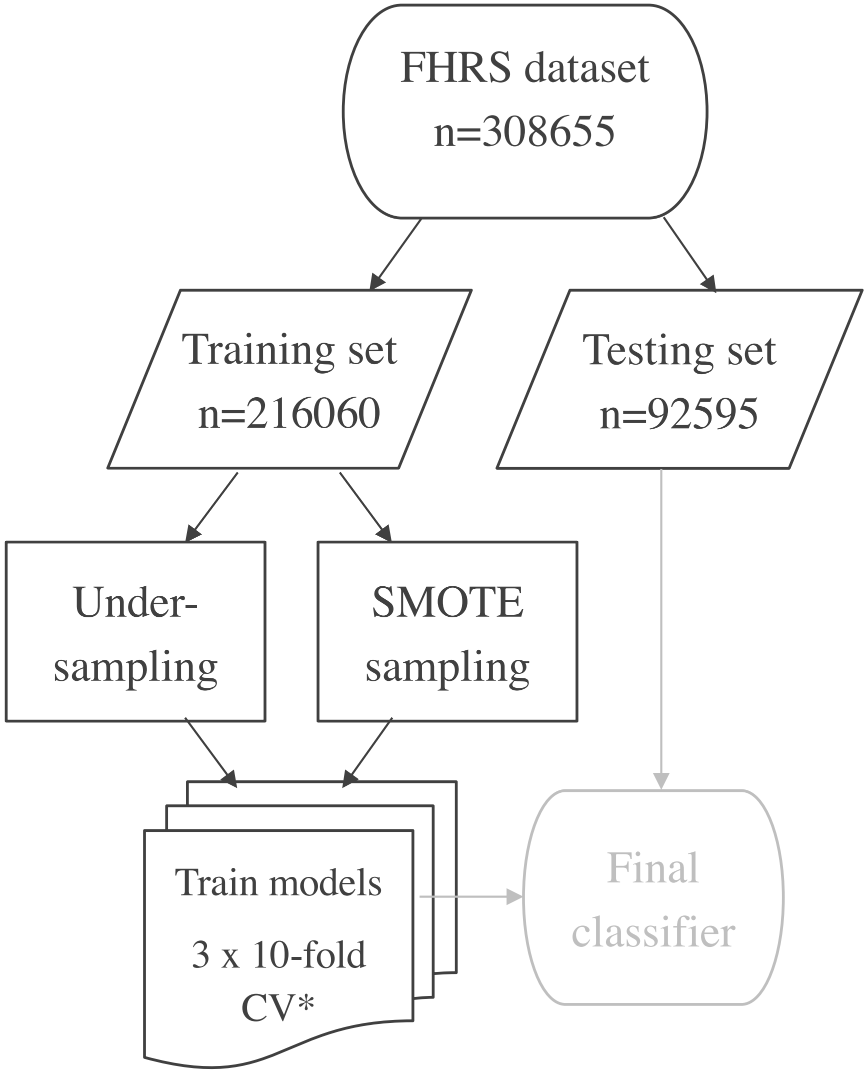
Method Overview



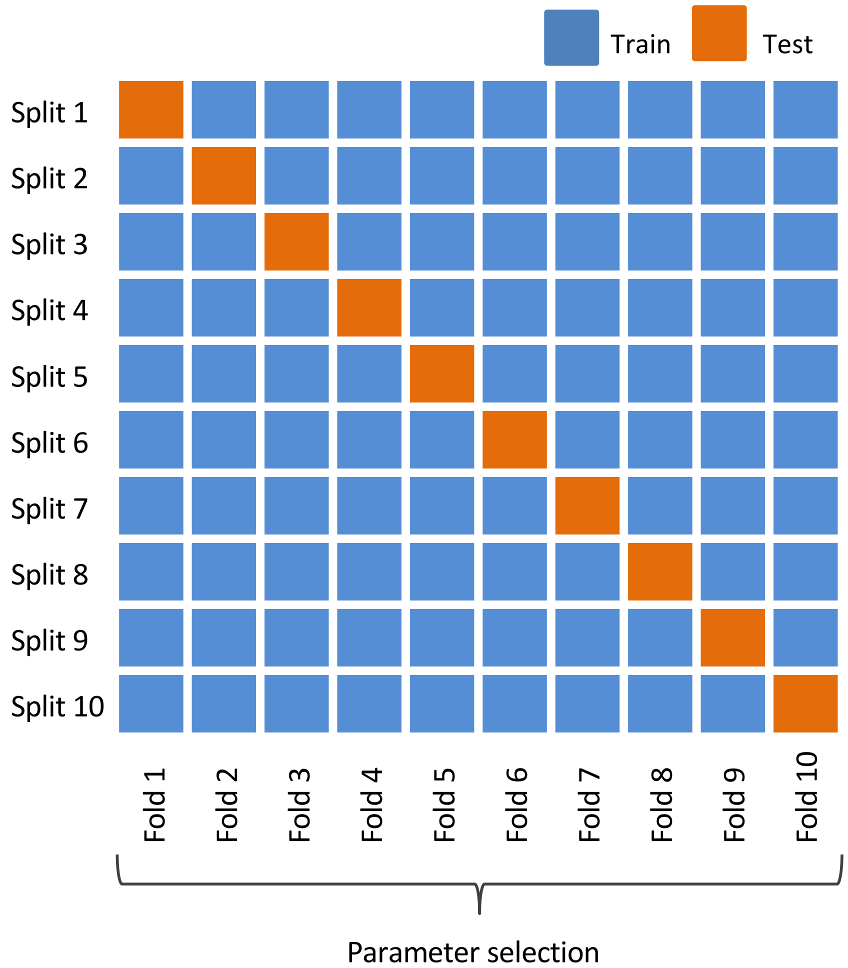
Different sampling strategies & ratios to address class imbalance (7% non-compliant outlets)

- Under-sampling
  - Straight forward
  - Reduces size of training set
- Synthetic Minority Over Sampling Technique (SMOTE)
  - Add synthetic data points to minority class –KNN
  - Under sample majority class
  - Maximises data for training
  - Time intensive
- Both methods repeated
  - 5 ratios (non-comp:comp)
  - 1:1, 2:1, 1:2, 3:2, 2:3
  - Resulting in 11 training sets (+ unsampled dataset)

Method Overview



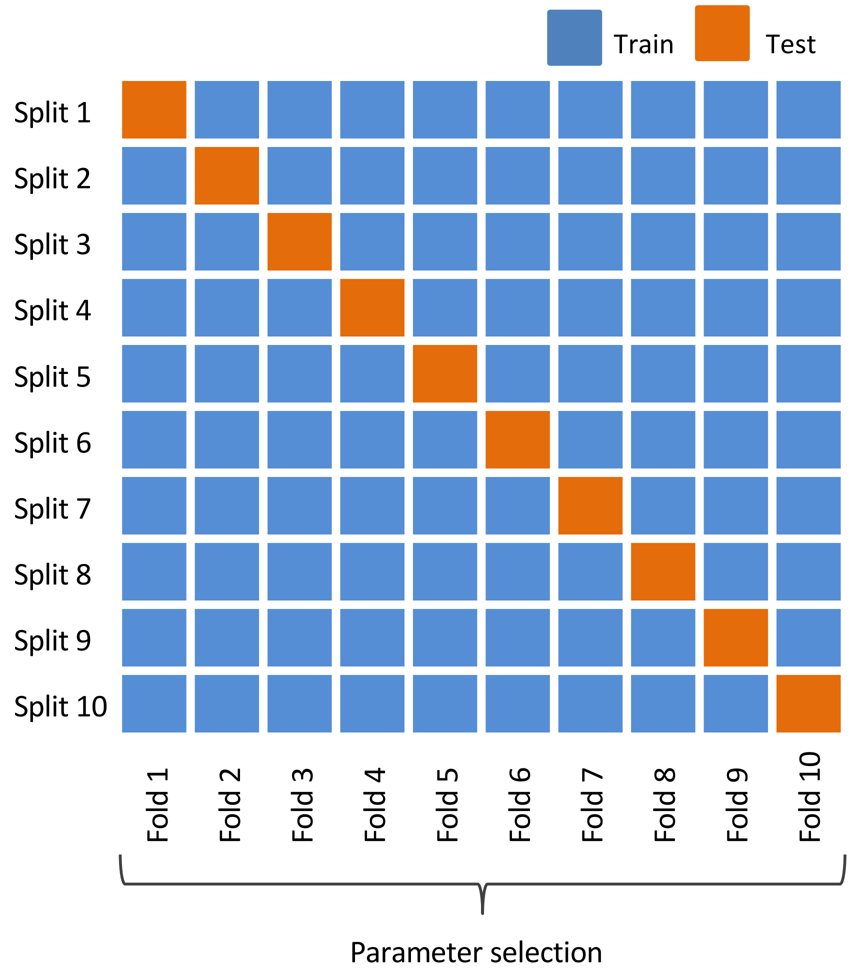
10 fold Cross Validation to train algorithms & select parameters:



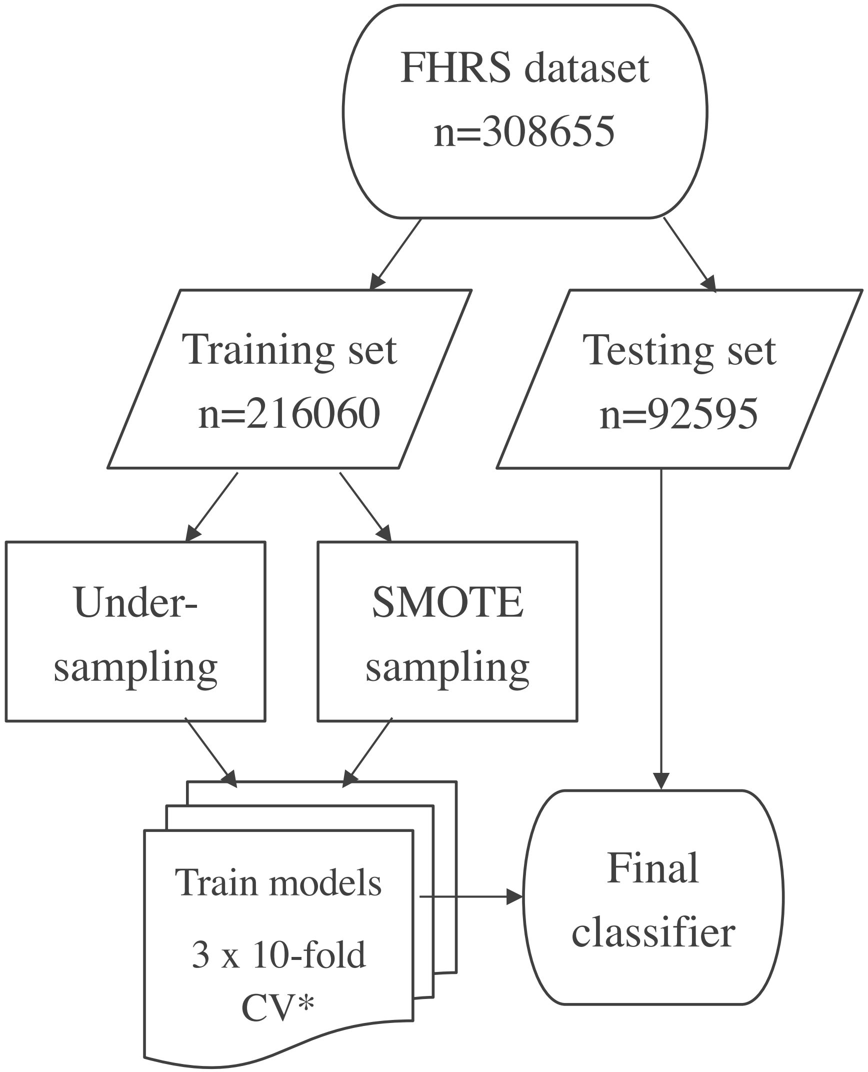
Method Overview

- Divide observation into k (10) folds of equal size
- Use 1st fold as validation set and fit model on remaining k-1 folds
- Repeat for each fold
- Use optimal parameters for final algorithm measured using:
  - Sensitivity (true positive rate)
  - Specificity (true negative rate)
  - Kappa (an accuracy measure which accounts for class size)

10 fold Cross Validation to train algorithms & select parameters:



Method Overview



Repeat model training & testing across sampling strategies, ratios and three algorithms:

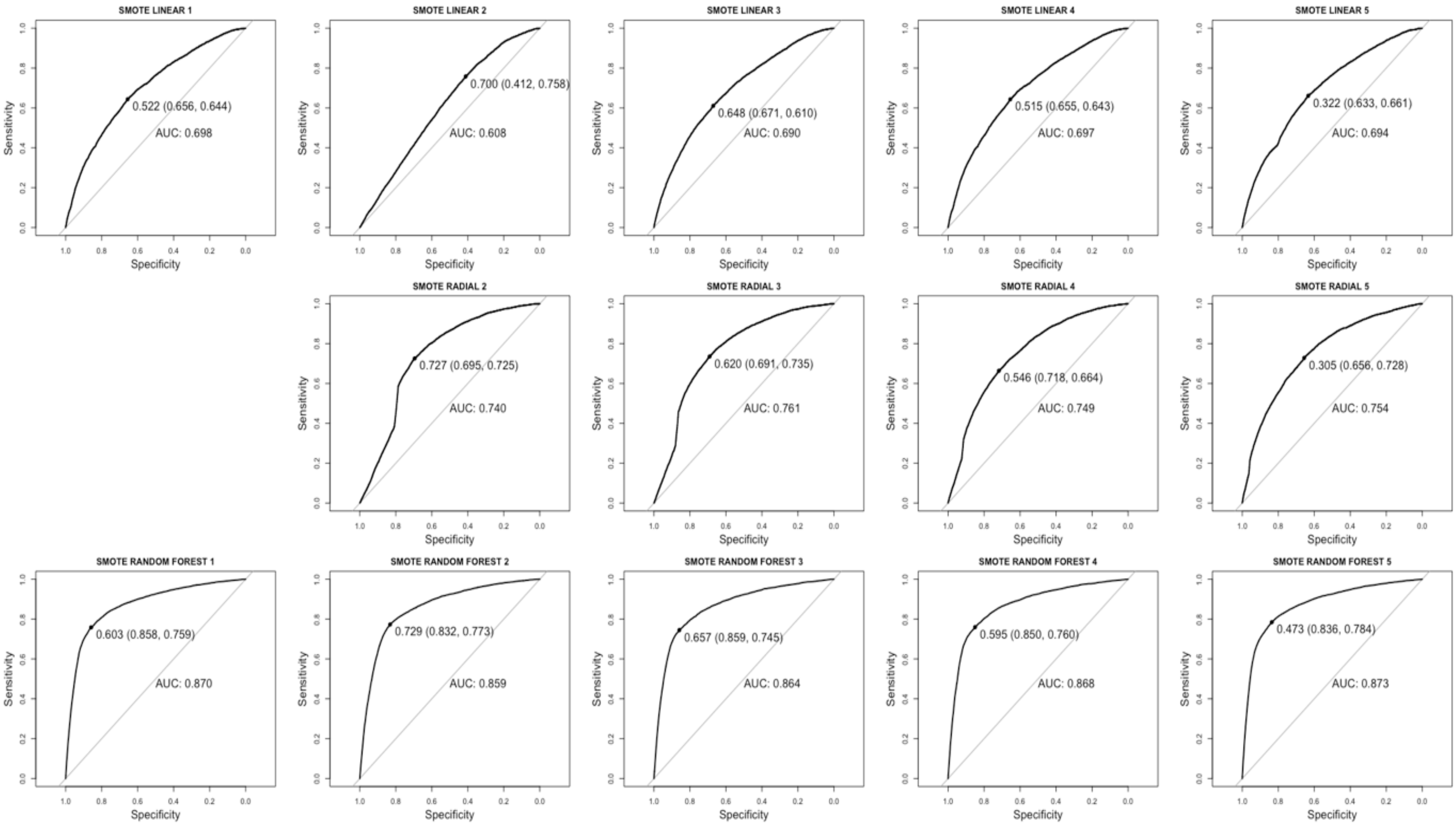
- Linear SVM
- Radial SVM
- Random Forest

=33 models in total run on HPC

Class probabilities calculated for each record -> compare metrics:

Results

SMOTE models reported best predictive power:



Consumer Data Research Centre

An ESRC Data Investment

Results

SMOTE, Random Forest, 1:1 adopted as final model:

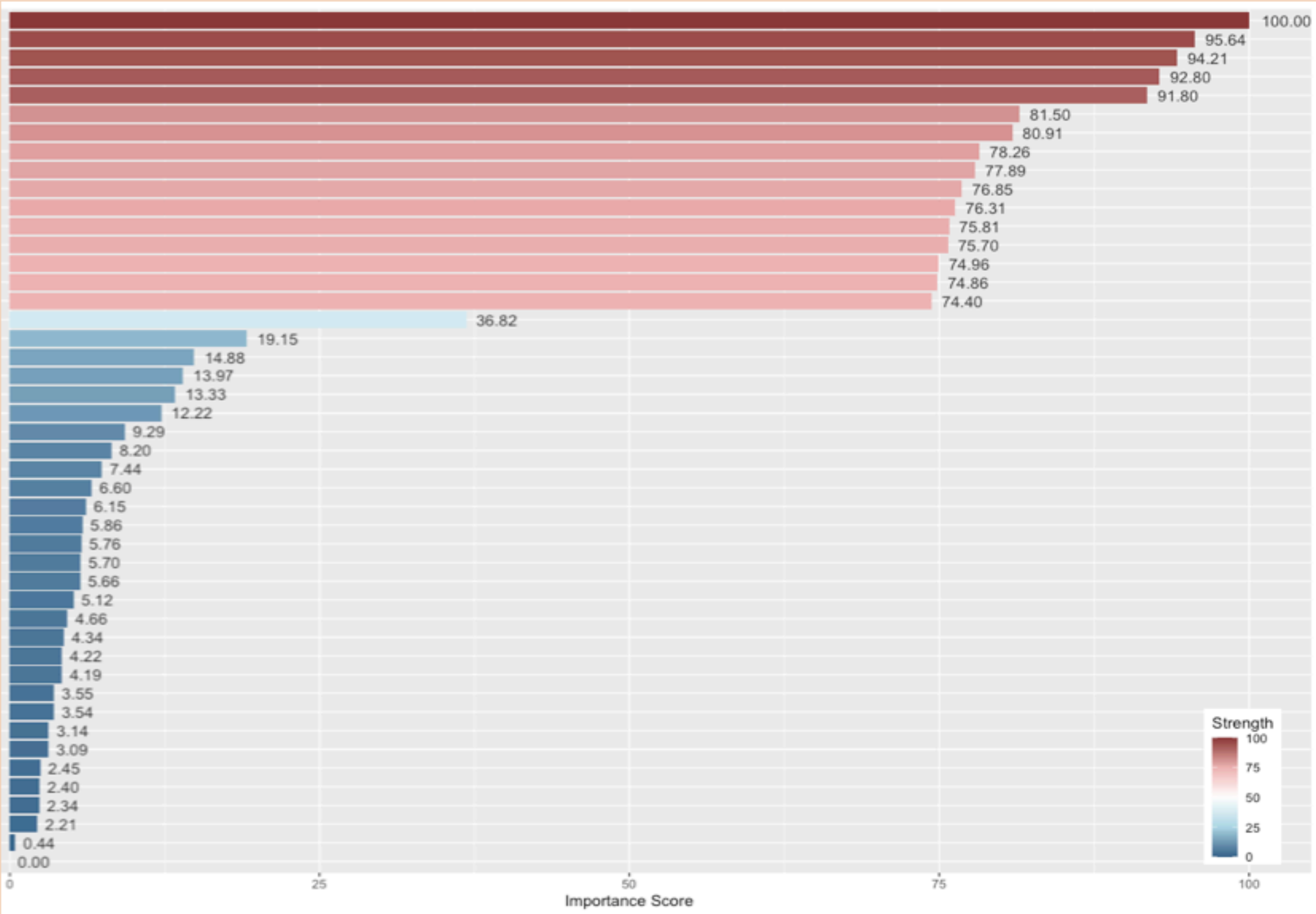
	RF Set 1 n=92595		RF unsampled n=92595	
	unweighted	weighted	unweighted	weighted
Probability Threshold	0.603	0.481	0.067	0.021
AUC	0.87	0.87	0.796	0.796
Sensitivity	0.759	0.843	0.661	0.859
Specificity	0.858	0.745	0.797	0.481
True Positives	4624	5139	4029	5903
False Positives	12264	21676	17571	77591
True Negatives	74235	64823	68928	8908
False Negatives	1472	957	2067	193
Kappa	0.338	0.230	0.210	0.010
Precision	0.274	0.192	0.187	0.071

Apply a weighting to penalise False Negatives

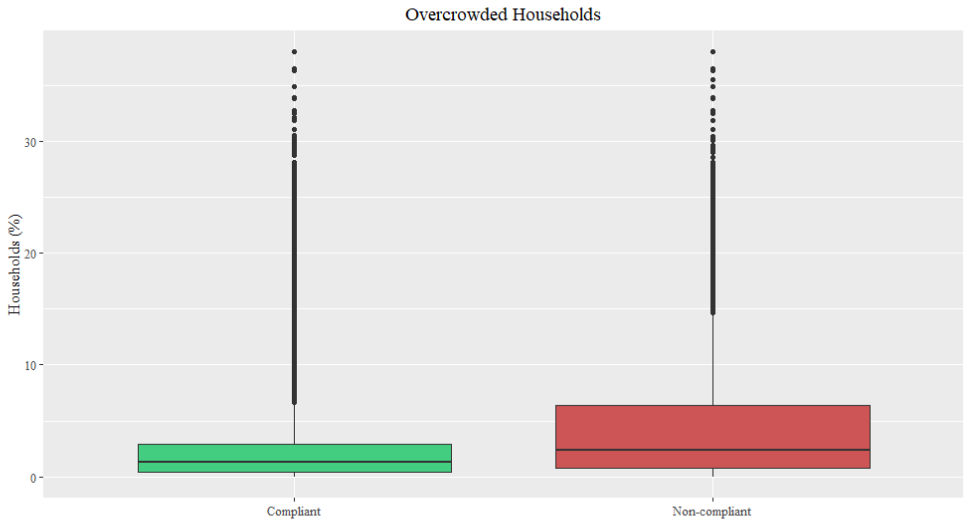
- Decreases the prob. threshold
- Increases no. of outlets classed as non-comp
- Decreases some model metrics
- Important to consider the context of the work
- 84% non-comp outlets correctly identified

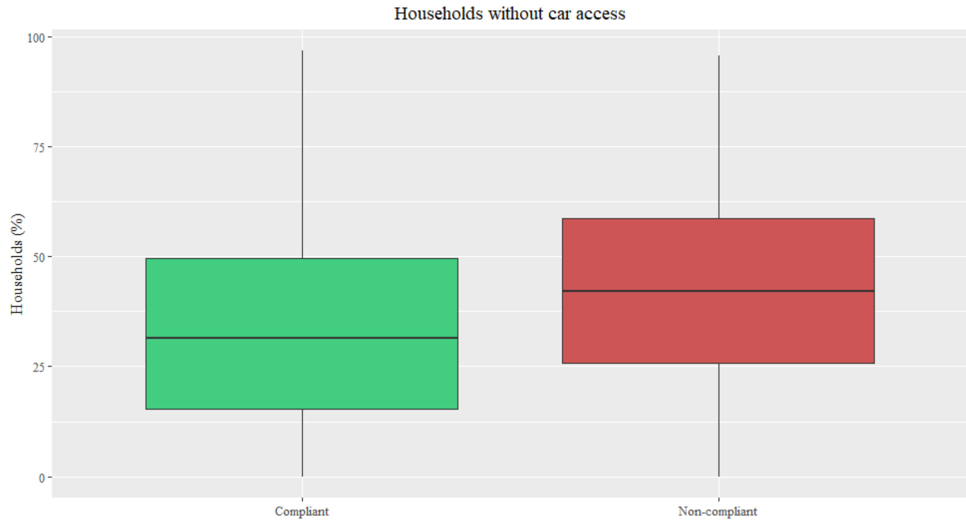
Variable Importance

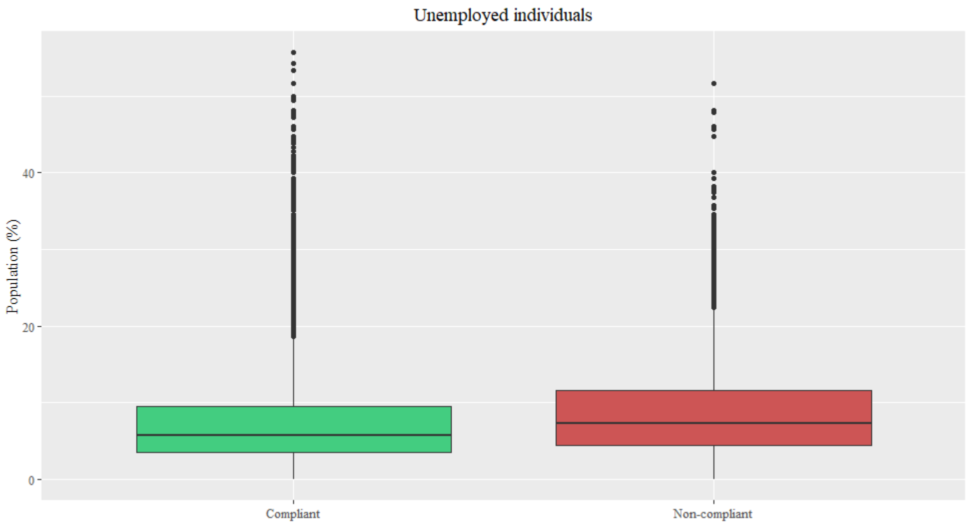
Scores calculated using Caret in R:

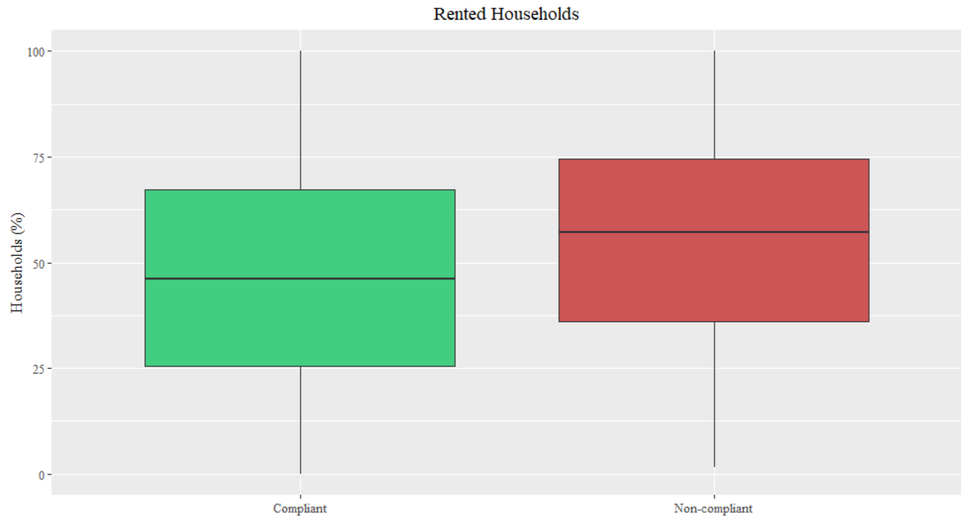


Variable Importance









Variable Importance

Problems with entropy based classifiers:

Numeric variables provide more flexibility for data partition → Minimize entropy in resulting class → Afforded more importance than categorical variables

Variable importance scores should be interpreted with caution

Consumer Data Research Centre

An ESRC Data Investment

Discussion

Highly predictive variables:

- Characteristics of deprived neighbourhoods
- Non-white ethnicities
- Some age variables
- Large urban areas
- Takeaways / sandwich shops

Further research required to unpick these relationships

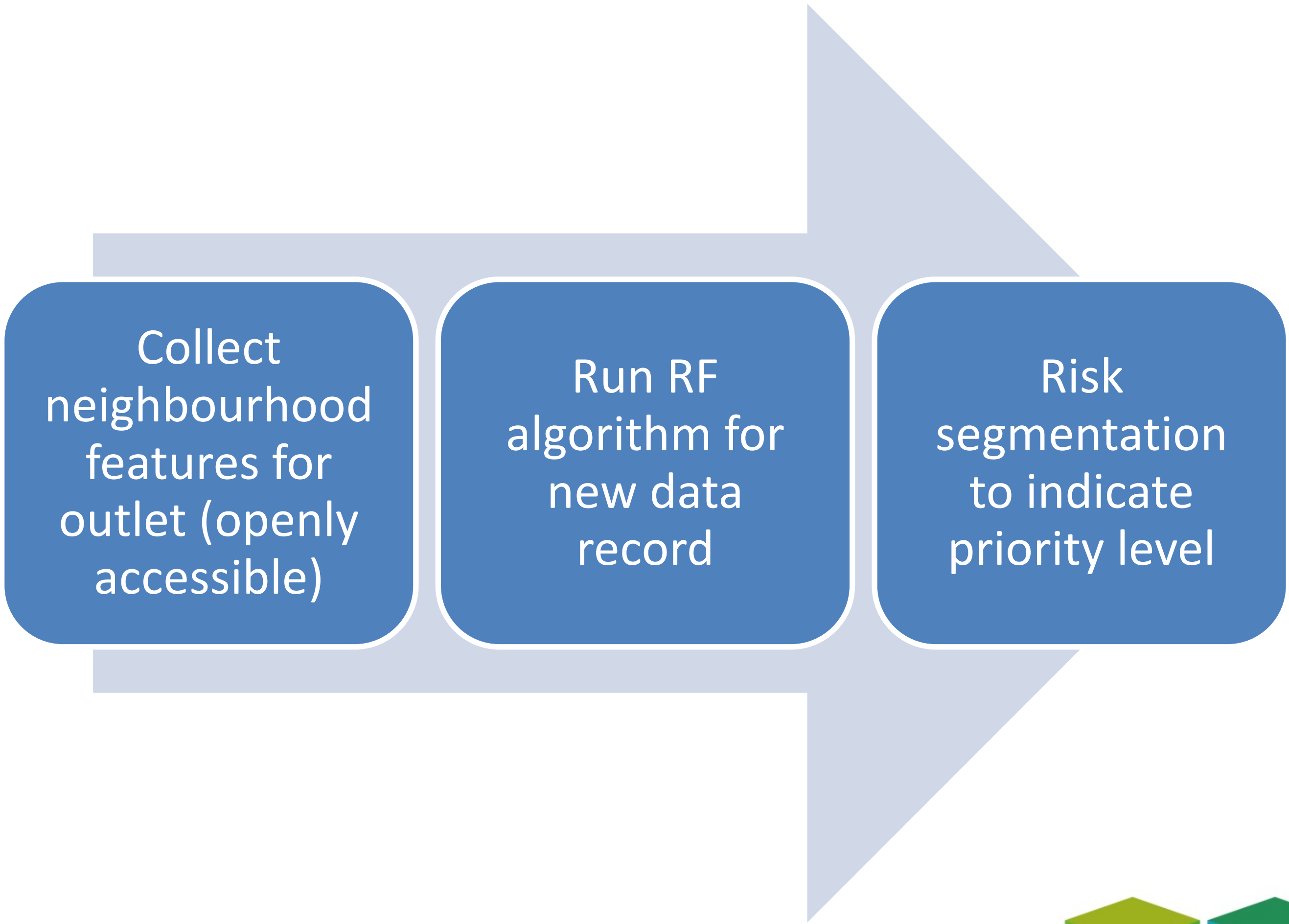
- High population turnover -> high staff turnover
- FHRS score display -> incentive to improve
- The role of cuisine type

Consumer Data Research Centre

An ESRC Data Investment

Real world application

For a newly opened outlet or routine inspection:



Limitations

- Data
  - FHRS -> Snapshot in time (some inspection data > 5 years old)
  - Inspection bias (deprivation, ethnicity)
  - Census 2011 outdated (esp. in large urban areas)

- Model doesn’t take behaviours into account
  - Food hygiene in the home
  - Habits of eating outside the home

- Problems with entropy based classification
  - Future work with look at alternative algorithms
  - Partial permutations (Altman et al. 2010), unbiased trees (Painsky and Rosset 2017)

Consumer Data Research Centre

An ESRC Data Investment

Publication

Oldroyd RA., Morris MA., Birkin M. Predicting high risk food outlets in England and Wales using neighbourhood characteristics: a machine learning approach. 2021. International Journal for Environmental Research and Public Health (in review).