Predicting dietary choices with machine learning

UK attitudes towards vegetarianism and reduced meat consumption

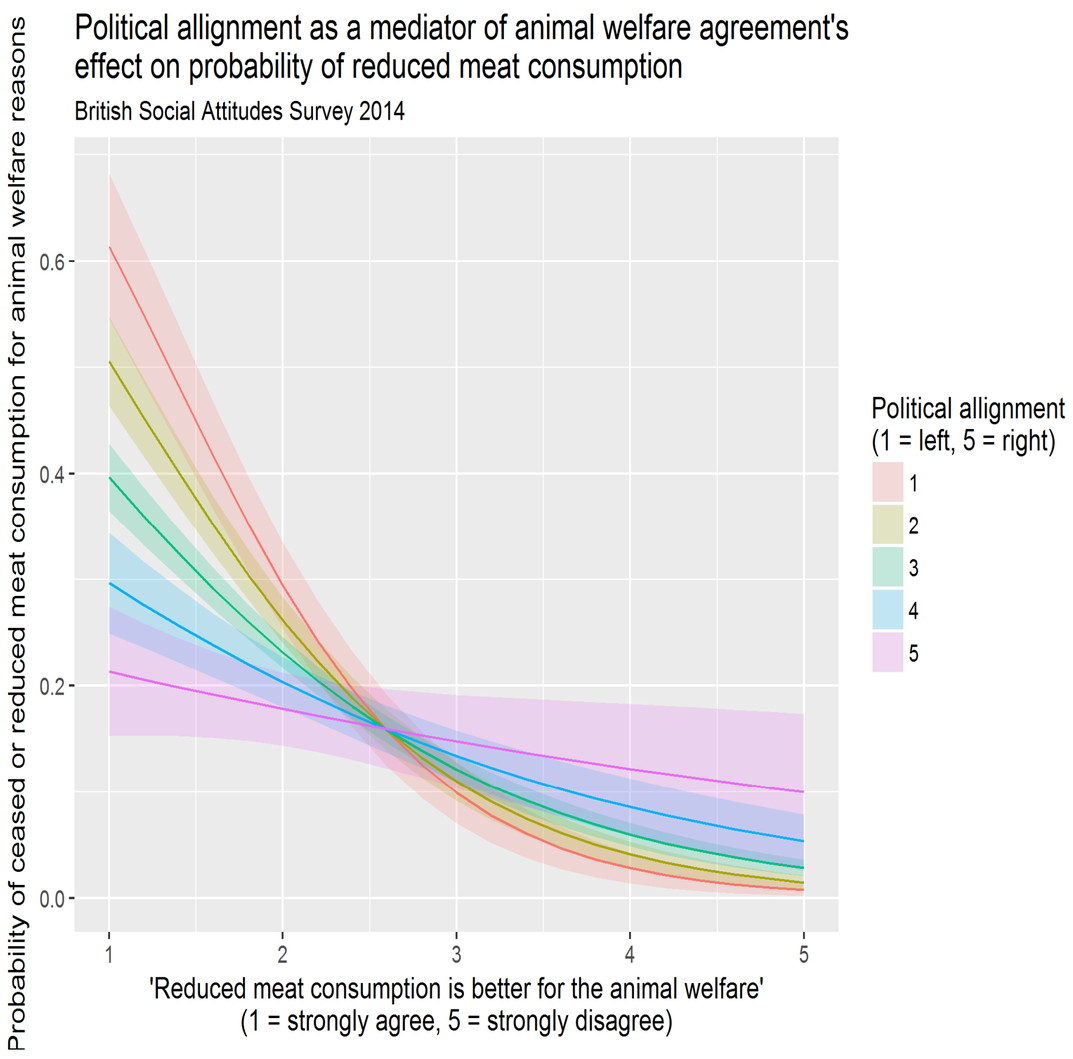
Chris Newton

Globally, meat production is a major driver of habitat loss and environmental pollution, and has high associated water resource use and greenhouse gas emissions (Macdiarmid et al. 2016; Poore and Nemecek 2018). In developed countries such as the UK, average meat consumption is on average nearly 50% higher than deemed healthy by healthcare experts (Wellesley et al. 2015).

This project attempted to develop machine learning models to classify and predict self-reported meat consumption behaviours based on demographic criteria and the answers to 3 survey questions on attitudes towards meat consumption.

Using data from the 2014 British Social Attitudes Survey, a multinomial logistic regression model and a random forest classifier were applied and compared as classification and prediction methods.

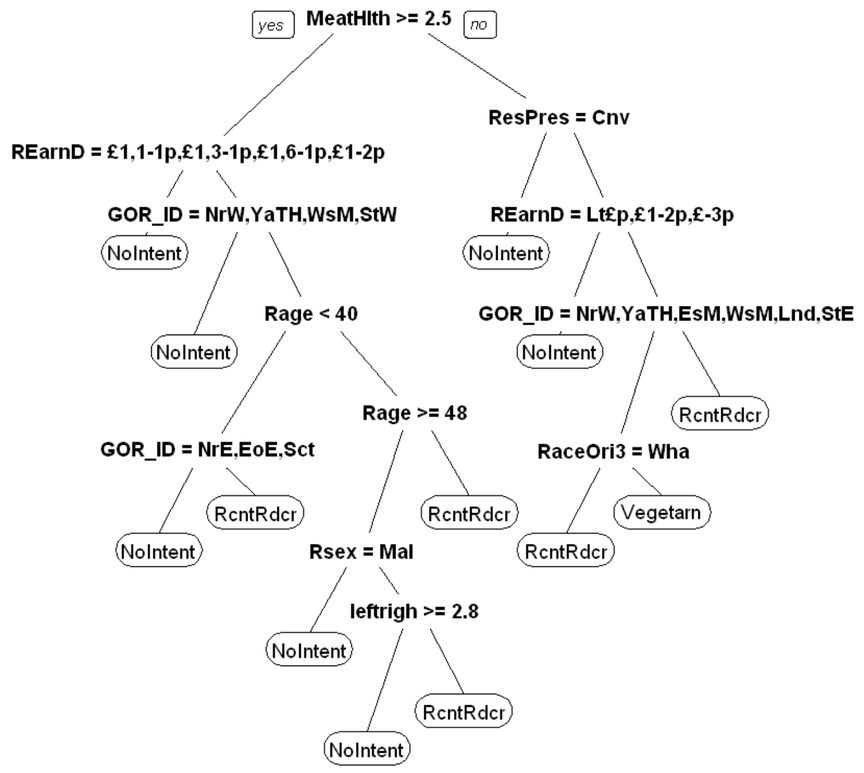
Logistic regression



Binomial logistic regression is a method of calculating the probability of a binary outcome based on one or more predictor variables, which may interact. The figure above shows how respondent's estimated measure of political alignment interacts with animal welfare beliefs to determine the probability of having reduced or stopped meat consumption for animal welfare reasons.

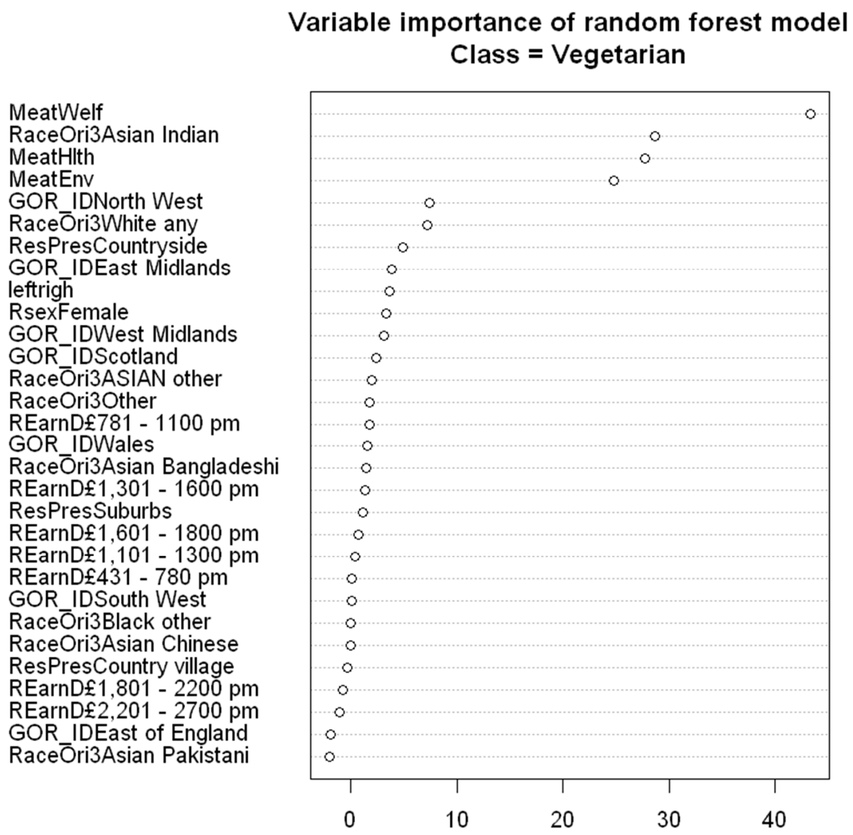
Multinomial logistic regression is similar in principle to binomial methods except rather than predicting a binary outcome, separate probabilities are calculated for multiple class outcomes, such as 'already vegetarian', 'recent reducer', or 'not interested', based on the predictor variables selected.

Random forest classifiers and decision trees



Random forests are an ensemble decision tree-based machine learning method which are adept at capturing non-linear relationships and interaction effects (Strobl et al. 2009). They are particularly suited to datasets with unbalanced class distribution, and have been shown to be robust to noisy data (Breiman 2001).

Variable importance for classification outcomes



Interpreting an ensemble of potentially thousands of decision trees would be difficult, so random forests provide a ranked measure of each predictor variable's importance in determining each outcome class by measuring changes in the out-of-bag error rate when each predictor variable is omitted (Strobl et al. 2009).

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), pp.5–32.

Macdiarmid, J.I., Douglas, F. and Campbell, J. (2016). Eating like there's no tomorrow: Public awareness of the environmental impact of food and reluctance to eat less meat as part of a sustainable diet. *Appetite*, 96, pp.487–493.

NatCen Social Research. (2016). British Social Attitudes Survey, 2014. [data collection]. 2nd Edition. UK Data Service. SN: 7809, http://doi.org/10.5255/UKDA-SN-7809-2

Poore, J. and Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. *Science*, 360(June), pp.987–992.

Strobl, C., Malley, J. and Gerhard Tutz. (2009). Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods*, 14(4), pp.323–348.

Wellesley, L., Happer, C. and Froggatt, A. (2015). *Changing Climate, Changing Diets Pathways to Lower Meat Consumption*. [online]. Available from: https://www.chathamhouse.org/publication/changing-climate-changing-diets.

Chris Newton

chris.newton@postgrad.manchester.ac.uk

ORCID: 0000-0002-5662-7481

Data Analytics & Society

The University of Manchester

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL