Yeunun Choo
Professor Andras Zsom
DATA 1030, Project Final Report
December 7, 2021
https://github.com/yeununchoo/plasticbags

Predicting green customer behavior of avoiding single-use plastic bags

Introduction

The purpose of this project is to predict if a customer is going to purchase a new single-use plastic bag in a supermarket. While much remains to be discussed about the overall footprint of reusable bags (Cho, 2020; Thompson, 2017), it is still important to reduce plastic waste in our best ability (UNEP 2018), which does not exclude machine learning skills.

This project is not alone in trying to predict the plastic bag usage. A study by Lavelle-Hill (2020) has applied Logistic Regression to this problem with the additional purpose of identifying consumer-level characteristics that are associated with the bag usage. Their model's balanced accuracy was approximately 67%, which could be used as a benchmark for this project.

This project uses the data from a public policy study by Thomas et al. (2019), which examined the effect of charging a small fee of five pence in nudging the customers away from single-use plastic bags in England in 2015. This study compared the English customers' behavior between 2015 and 2016 to the Welsh who had already been under similar policy. The dataset (Poortinga, 2016) contains the observations from N = 3764 people in Cardiff and Bristol. The dataset has 31 variables in total, 18 of which describe the bag usage, 11 of which are about the customer and the supermarket branch, and the rest 2 are for the administrative use for the surveyors.

The target variable of this project was constructed from combining the 18 bag-related variables to classify if the customer has made the environmentally friendly choice of avoiding plastic bags, but exceptions were made for making one additional purchase when they had already brought own reusable bags ($y = 1$ for the green choice and $y = 0$ otherwise). For the feature variables, the 11 variables on the customer and supermarket branch and some selected interaction terms were used (see Figures 1, 2, and 3). A selection of several classification algorithms was then applied and evaluated, followed by an analysis on global and local feature importance.

Exploratory Data Analysis

Under the definition of the target variable, 3015 customers were labeled as "green" or y = 1 and the other 749 otherwise. The effect of the new English policy is clearly shown in the stacked bar plot, Figure 1. Customers at different supermarket brands reacted differently to the policy change (Figure 2) and the gender also mattered (Figure 3). Another interesting plot is the scatter plot (Figure 4) of the number of reusable bags versus the single-use plastic ones. The distributions of the bags are highly skewed and inversely correlated.

Methods

Preprocessing

First, the missing values hidden in various disguises were manually replaced with the Numpy *NaN* value – a floating-point representation to indicate the value is missing. In addition, the following five interaction terms were manually picked: year-country, year-country-time, year-country-supermarket, year-country-gender, and year-country-age. All the interactions include the year and country to capture the policy effect under different authorities.

Then, the dummy variables for the categorical features were created with Scikit Learn's OneHotEncoder. In this project, the year variable was interpreted as a categorical variable, as it is essentially a binary flag for before and after the policy. Note that the age variable was also treated as a categorical. The distances between the three age groups were not uniform, making OrdinalEncoder unapplicable. For the numerical columns *ObsSize*, the number of customers who were together, and *MaleN* and *FemaleN*, the numbers of male and female members in each group, StandardScaler was used. Collinearity was not a concern as *ObsSize* also included gender-non-binary people not counted in *MaleN* and *FemaleN*.

The initial dataset consisted of 1 custom-defined target variable, 11 initial feature variables, and 5 hand-picked interaction terms. The number of features has increased to 85 by the preprocessing step.

Data Split

Then 10% of the data were split into the test set, stratified on the target variable, and the Stratified 5-Fold split was employed for the training and cross-validation. Each observation in the data is independent and identically distributed as they are collected from spontaneously encountered groups of customers with no relation from the previous encounter. The data is not time-series, nor does it contain any group structure. As the data were collected in the two similar neighboring cities, there is not much reason to believe that English and the Welsh customers behaved differently other than the fact that policy was applied at different times. However, the distribution of the target variable is not balanced, as about four times more customers are labeled as "green", or y = 1, making it necessary to stratify on the target variable.

Models

In this project, 11 different classification algorithms of 5 different types were trained and evaluated. They are the Logistic Regression models with no, L1, L2, and Elastic Net regularizations, Support Vector Machine Classifiers with the Linear and Radial Basis Function kernels, Random Forest, K-Nearest Neighbor Classifier, and three Boost methods, which are Adaptive Boost, Gradient Boost, and XGBoost. The hyperparameters are tuned with a brute-force approach with the GridSearchCV function, and the searched values are summarized in Table 1. A particular attention was paid to make sure that the best fitted parameter values are well in the middle of the search space.

The metric to evaluate the performance was balanced accuracy, defined as the average of sensitivity and specificity:

$$\text{Accuracy}_{\text{balanced}} = \frac{1}{2}(\text{Sensitivity} + \text{Specificity}) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right)$$

The baseline score is 50%. If the model always predicts positive, it will achieve 100% sensitivity and 0% specificity, and the average is 50%. Conversely, if the model always predicts negative, it will achieve 0% sensitivity and 100% specificity, and the average is 50% again. This metric was chosen because the target variable was unbalanced, and it was equally important to predict actually positive cases to positive and the actually negative cases to negative. The regular accuracy measure would care about the actually positive cases four times more importantly, because there are four times more actually positive cases, but that is not the desired model behavior.

Overall Pipeline and Uncertainty

There is always a randomness in splitting the data into test and training sets. It is entirely possible that a particular split with a particular random seed would result in the test set full of a particularly untypical cases, just as it is entirely possible to see ten heads in a row if you throw a coin enough many times. Many algorithms are non-deterministic too, so it is entirely possible to have an unusually predictive model with a particular random seed. However, after repeating the whole process many times with a new seed each time, it is possible to find the average and the standard deviation of the model's performance, and that is exactly what this project does. In this project, the whole process is repeated with 10 different seeds, and the average and standard deviation is used to evaluate the models.

Results

The balanced accuracy scores for each model are plotted in Figure 5 and summarized in Table 2. Most of them are far above the baseline of 50%, about thirty to fifty standard deviations away, except for the KNN model which is still 10 standard deviations away. The most predictive model was the XGBoost. However, the mean balanced accuracy scores of all models are within the narrow range between 68% and 69.5%, other than the KNN model. This is on a par with the previous study by Lavelle-Hill (2020) with 67% balanced accuracy.

Global feature importance was measured three times: with the coefficients the Logistic Regression, a linear model, with L2 regularization (Figure 6), mean decrease in Gini impurity in AdaBoost, a tree-based ensemble model (Figure 7), and the Permutation Importance from any model of choice (Figure 8). There are some features that are identified as important by all three measures such as *YearCountry_Y2015England*, *Supermarket_Asda*, *Age_Age_g1*, and *Age_Age_g3*. It was interesting to see that main policy variable, *YearCountry_Y2015England*, repeatedly emerged as one of the most important. Some of the least important features are about whether or not the shopper was with a child. It was surprising to see that most people do not change their behavior in front of a child.

The Shapley Additive Explanations, or SHAP, was used to measure the local feature importance to any predicted cases. The first negative prediction and the first positive prediction are chosen as examples here. Figure 9 shows the negative example, and Figure 10 the positive example. We see the same features appearing again, such as being in England before the policy change, shopping at Asda, and being in a specific age group. From the public policy perspective, perhaps it is worth investigating how shopping at a specific supermarket brand or being in a specific age group motivates the customers to be greener.

Outlook

Many of the most important features are the interactions between Year, Country, Supermarket, Age, and Gender, so further improvements can be expected from exploring other possible interactions between the original features. The interpretability can be improved by calculating the SHAP local feature importance on a representative set of observations for each Country, Age, and Gender segments, instead of calculating it on the two randomly chosen observations. Other than the five interaction features, no further feature engineering was performed, making it the weakest spot of this project. Additional techniques to consider are the reduced-feature models and the imputation of missing values. Currently the missing values are left as missing, forcing the model to make predictions without them. Lastly, the data is highly local and rather outdated – it is from Bristol and Cardiff in 2015 and 2016 only. We need more recent and geographically diverse data.

References

Cho, Renee. “Plastic, Paper or Cotton: Which Shopping Bag is Best?” *Earth Institute, Columbia University*, 30 April 2020, URL: https://news.climate.columbia.edu/2020/04/30/plastic-paper-cotton-bags/. Accessed 9 October 2021.

Edgington, Tom. “Plastic or paper: Which bag is greener?” *BBC*, 28 January 2019. URL: https://www.bbc.com/news/business-47027792. Accessed 9 October 2021.

Lavelle-Hill, R., Goulding, J., Smith, G., Clarke, D.D. and Bibby, P.A., 2020. “Psychological and demographic predictors of plastic bag consumption in transaction data”. *Journal of Environmental Psychology*, 72, p.101473. doi: 10.1016/j.jenvp.2020.101473

Poortinga, Wouter, Sautkina, Elena, Thomas, Gregory O. and Wolstenholme, Emily 2016. “The English plastic bag charge: Changes in attitudes and behaviour”. [Project Report]. *Welsh School of Architecture, School of Psychology, Cardiff University*. URL: https://orca.cardiff.ac.uk/94652/

Poortinga, Wouter and Whitmarsh, Lorraine (2018). “The English plastic bag charge and behavioural spillover”. [Data Collection]. *Colchester, Essex: UK Data Archive*. 10.5255/UKDA-SN-852642

Thomas GO, Sautkina E, Poortinga W, Wolstenholme E and Whitmarsh L (2019) “The English Plastic Bag Charge Changed Behavior and Increased Support for Other Charges to Reduce Plastic Waste”. *Front. Psychol*. 10:266. doi: 10.3389/fpsyg.2019.00266

Thompson, Claire, “Paper, Plastic or Reusable?” *Stanford Magazine*, September 2017. URL: https://stanfordmag.org/contents/paper-plastic-or-reusable. Accessed 9 October 2021.

UNEP (2018). *SINGLE-USE PLASTICS: A Roadmap for Sustainability* (Rev. ed., pp. vi; 6) ISBN: 978-92-807-3705-9. URL: https://www.unep.org/resources/report/single-use-plastics-roadmap-sustainability

Appendix

.

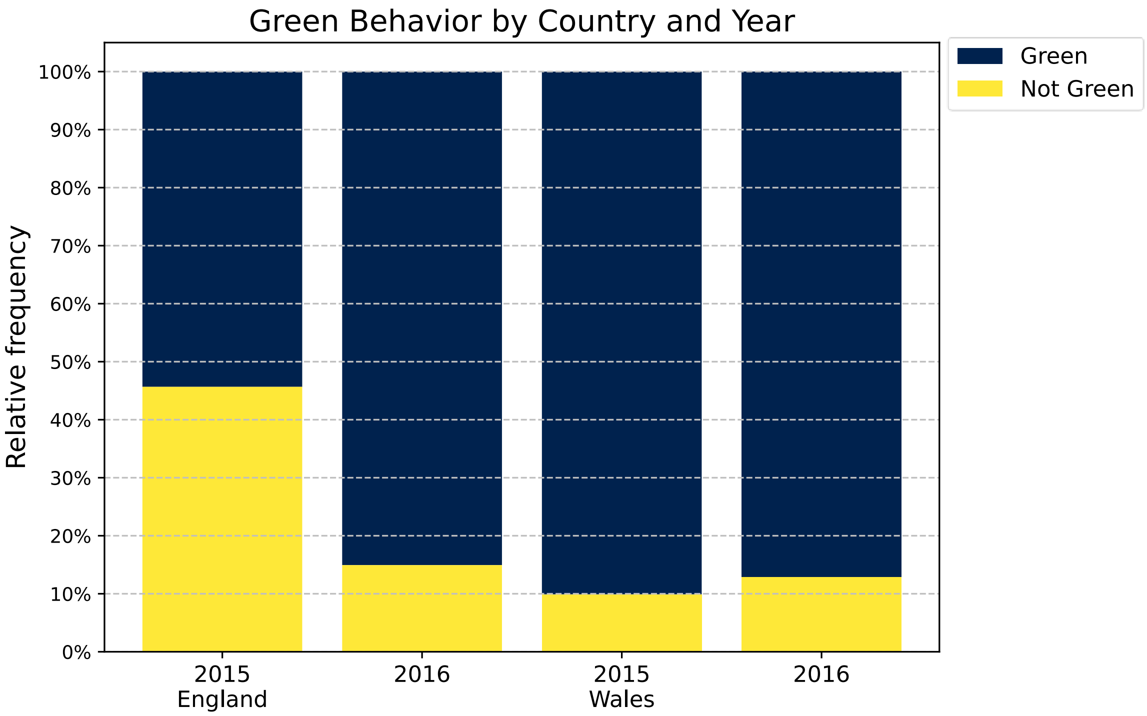

Figure 1 The effect of charging five pence for a single-use plastic bag in England. Similar policy was already in effect in Wales.

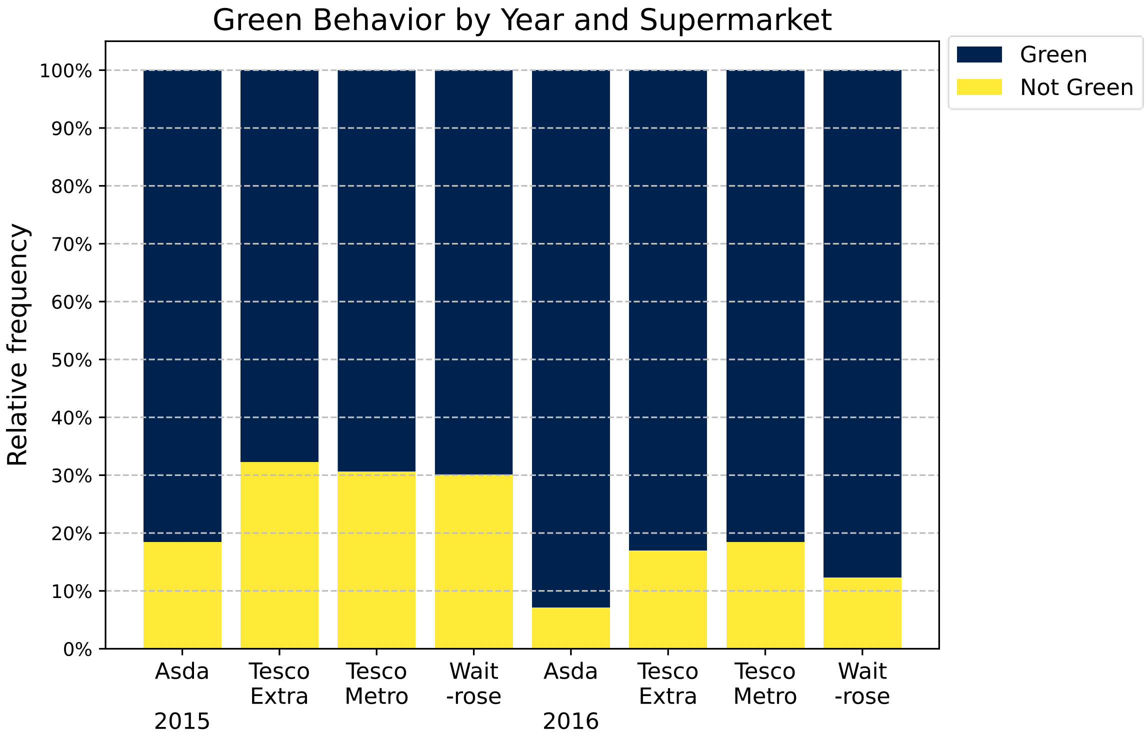

Figure 2. Customers at different supermarket brands reacted differently to the policy change.

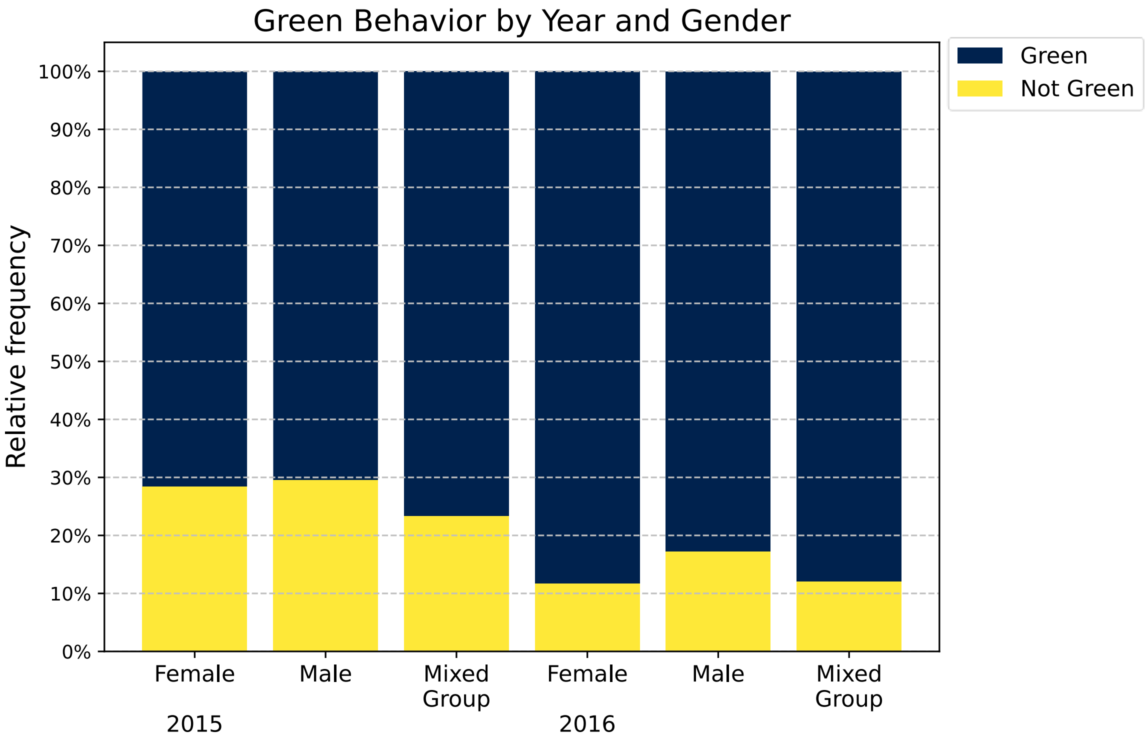

Figure 3 People of different gender reacted differently to the policy change.

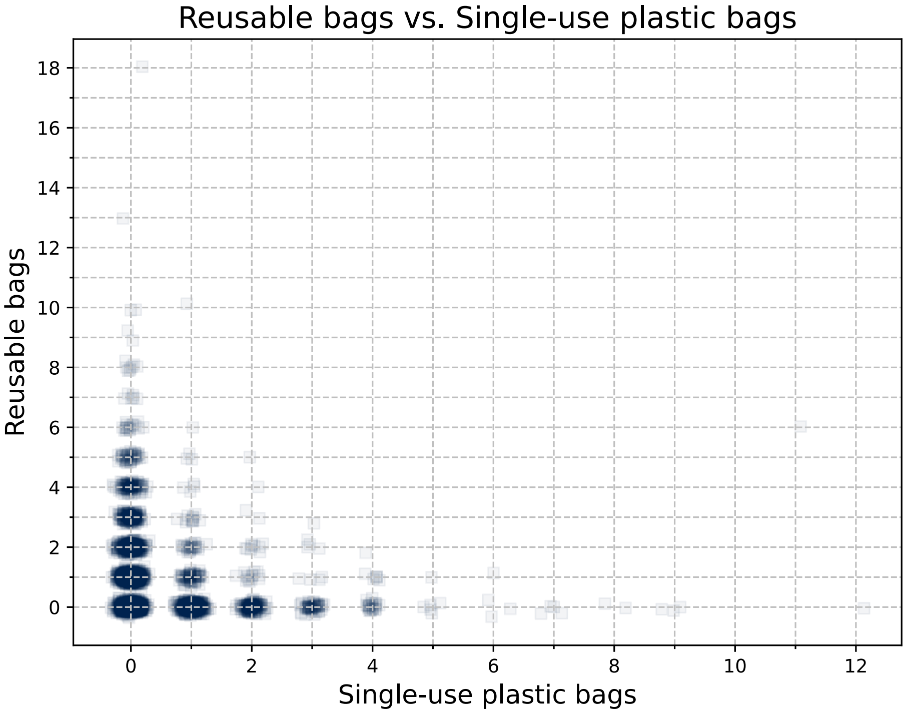

Figure 4 Beware that the number of plastic bags include both the newly purchased ones and the ones brought from home.

Table 1 Hyperparameters for each model

Type	Model	Parameter	Values
Logistic Regression	No Penalty	No Parameter	
	L1	C	np.logspace(-3, 2, 21)
	L2	C	np.logspace(-3, 2, 21)
	Elastic Net	C	np.logspace(-3, 2, 21)
		l1_ratio	[0.1, 0.2, .., 0.9]
SVM	Linear	C	np.logspace(-3, 2, 11)
	RBF	C	np.logspace(-3, 2, 6)
		gamma	np.logspace(-2, 2, 17)
Random Forest		max_features	[0.3, 0.35, .., 0.65, None]
		max_depth	[3, 4, 5, 6, 8, None]
		min_samples_split	[2, 3, .., 7]
K-Nearest Neighbors		n_neighbors	[2, 3, 5, 8, 9, 10, 11, 12, 15, 30]
		weights	['uniform', 'distance']
Boost Methods	AdaBoost	learning_rate	[0.3, 0.4, .., 1.1]
		n_estimators	[8, 12, .., 40]
		base_estimator__max_depth	[1, 2, 3, 4]
	Gradient Boost	learning_rate	[0.05, 0.1, .., 0.3]
		max_features	[0.1, 0.2, .., 1.0]
		max_depth	[2, 3, .., 6]
	XGB	learning_rate	[0.25, 0.3, .., 0.55]
		max_depth	[2, 3, .., 6]
		gamma	[0.5, 0.6, .., 1.2]
		n_estimators	[10, 12, 16, 20, 24, 30, 35, 40]

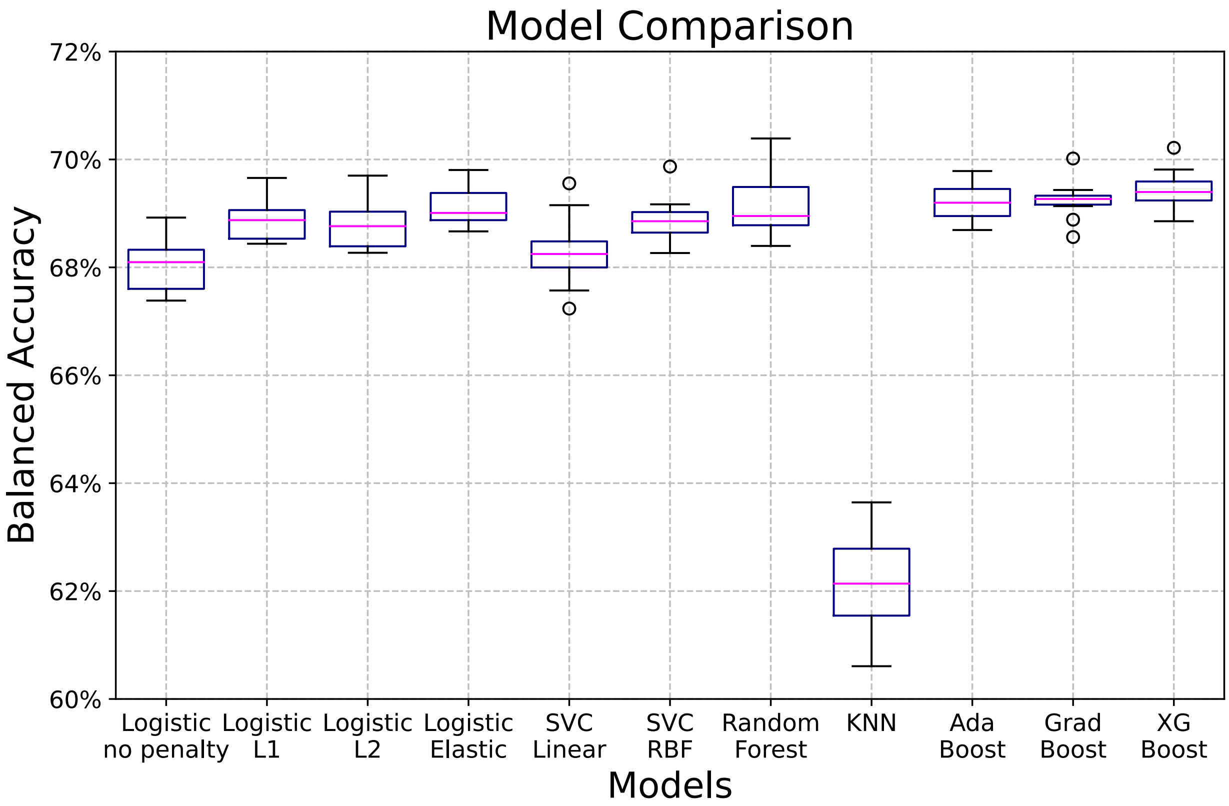

Figure 5 Model Balanced Accuracy Scores

Table 2 Balanced accuracy scores compared to the baseline

Type	Model	Mean	Std	Std above Baseline
Logistic Regression	No Penalty	68.11%	0.43%p	41.7
	L1	69.09%	0.45%p	42.5
	L2	68.91%	0.55%p	34.3
	Elastic Net	69.19%	0.49%p	39.3
SVM	Linear	68.2%	0.51%p	35.6
	RBF	69.07%	0.33%p	57.7
Random Forest		68.98%	0.29%p	66.5
K-Nearest Neighbors		61.51%	1.06%p	10.9
AdaBoost		69.28%	0.38%p	50.2
Gradient Boost		69.29%	0.36%p	53.8
XGBoost		69.44%	0.39%p	50.4

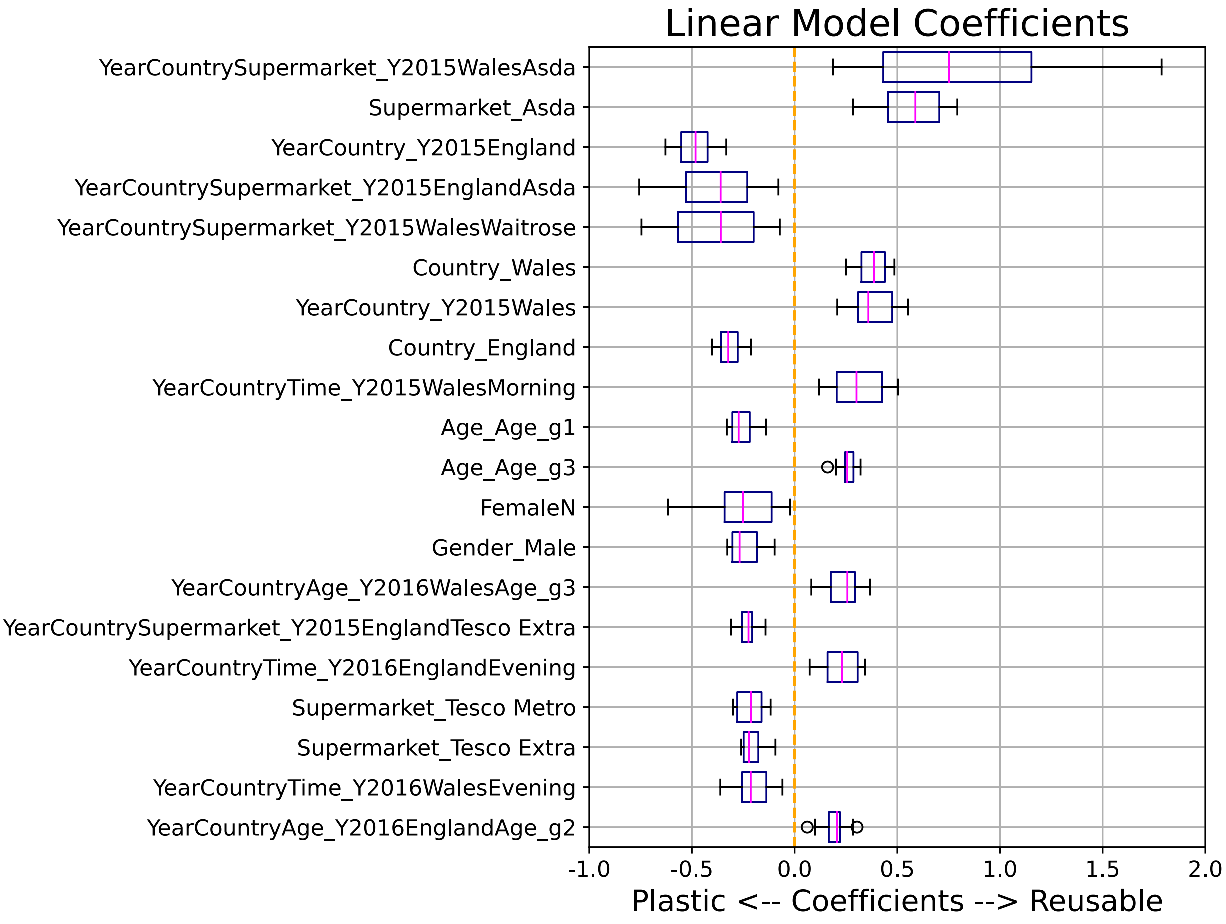

Figure 6 Global feature importance from the coefficients of the Logistic Regression with the L2 regularization

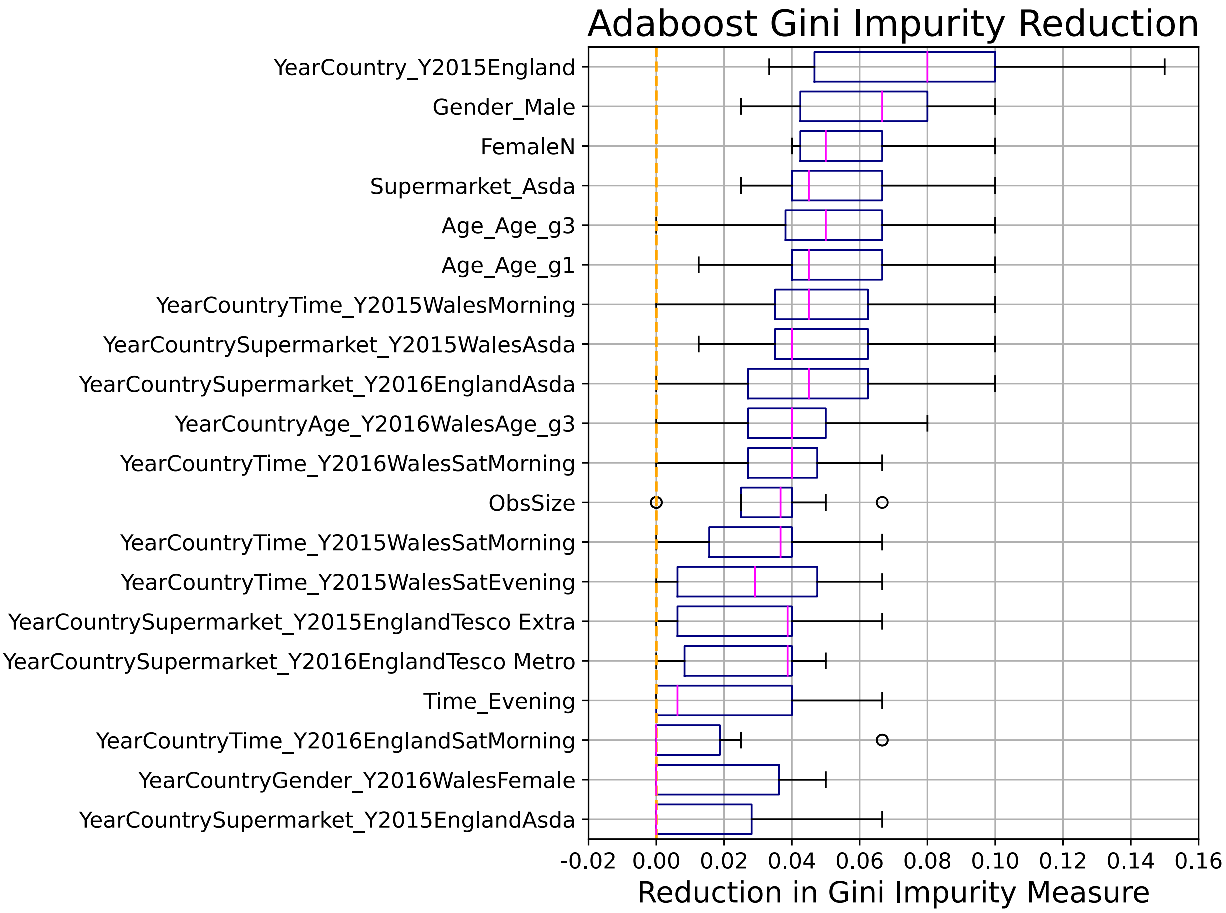

Figure 7 Global feature importance from the AdaBoost mean decrease in Gini impurity

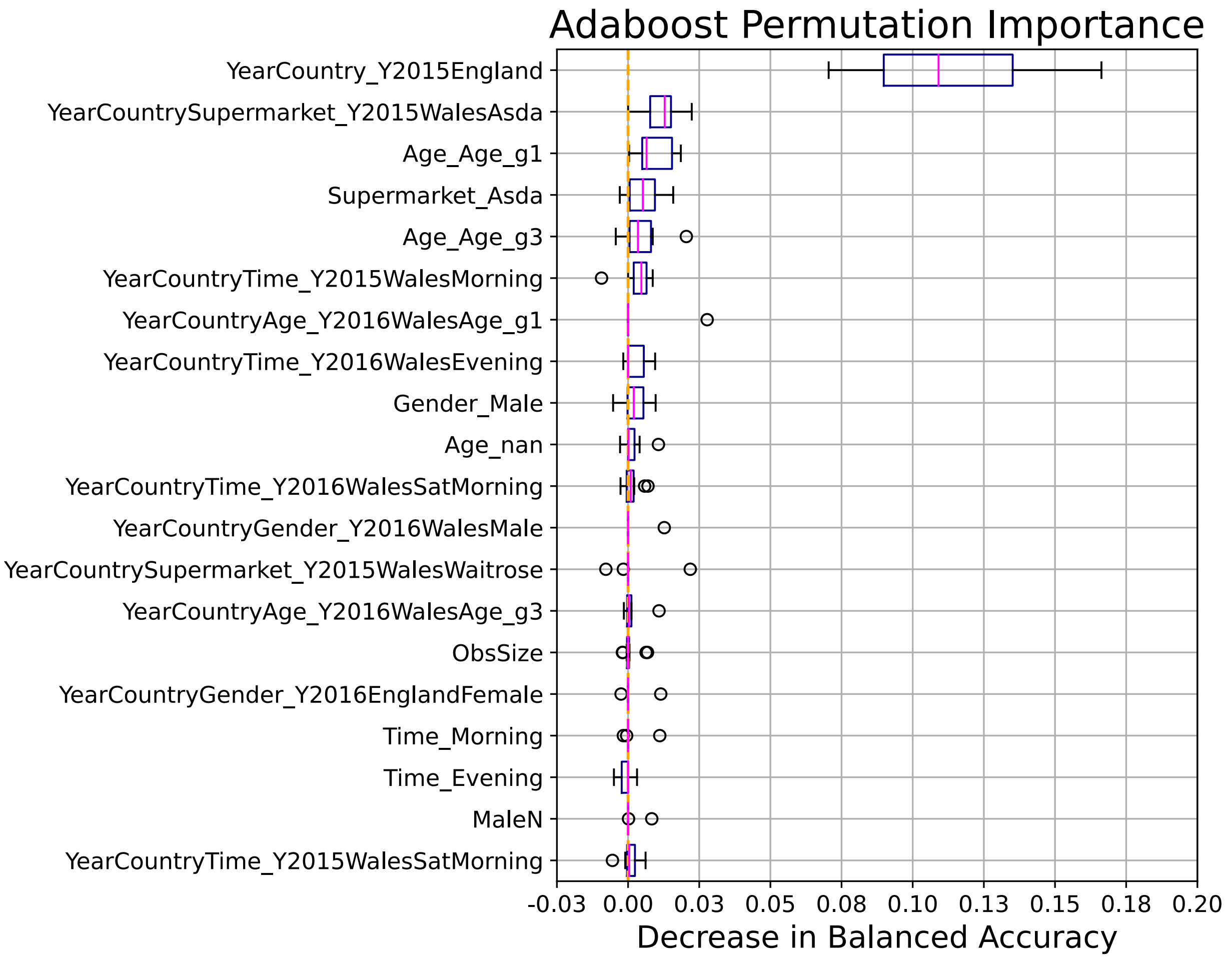

Figure 8 Global feature importance from Permutation Importance

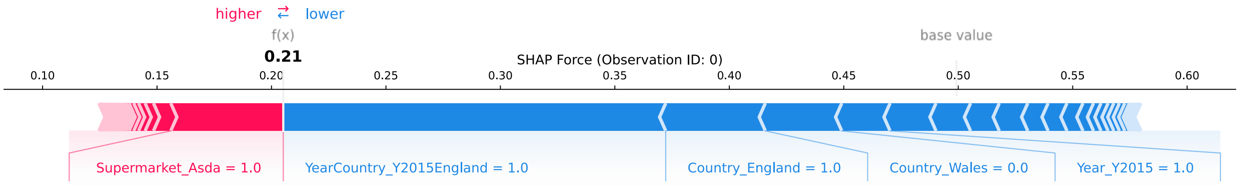

Figure 9 SHAP local feature importance for a negative prediction

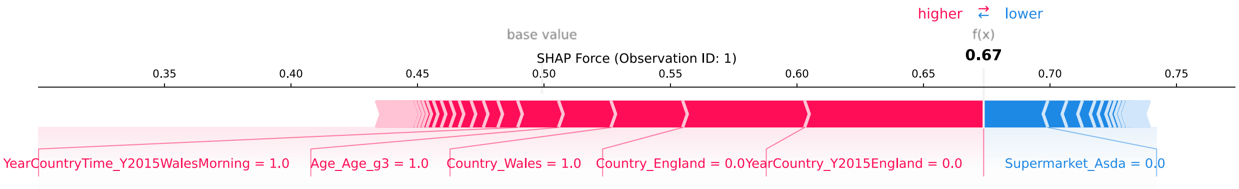

Figure 10 SHAP local feature importance for a positive prediction