Predicting green consumer behavior of avoiding single-use plastic bags

Yeunun Choo

DATA 1030 Project Final Presentation

Brown University

December 8, 2021

GitHub Repository

Recap: Goal & Data

Predict Green Customer Behavior

- **avoid** purchasing new single-use plastic bags

(Poortinga, 2016, “Bags for life“)

Public policy study Thomas et al.

- Observational data (N = 3764) with CC-By-SA 4.0
- Bag Usage → **Target Variable**
- 11 on Customer & Supermarket → **Feature Variables**

(Poortinga, 2016, "Single-use plastic bags“)

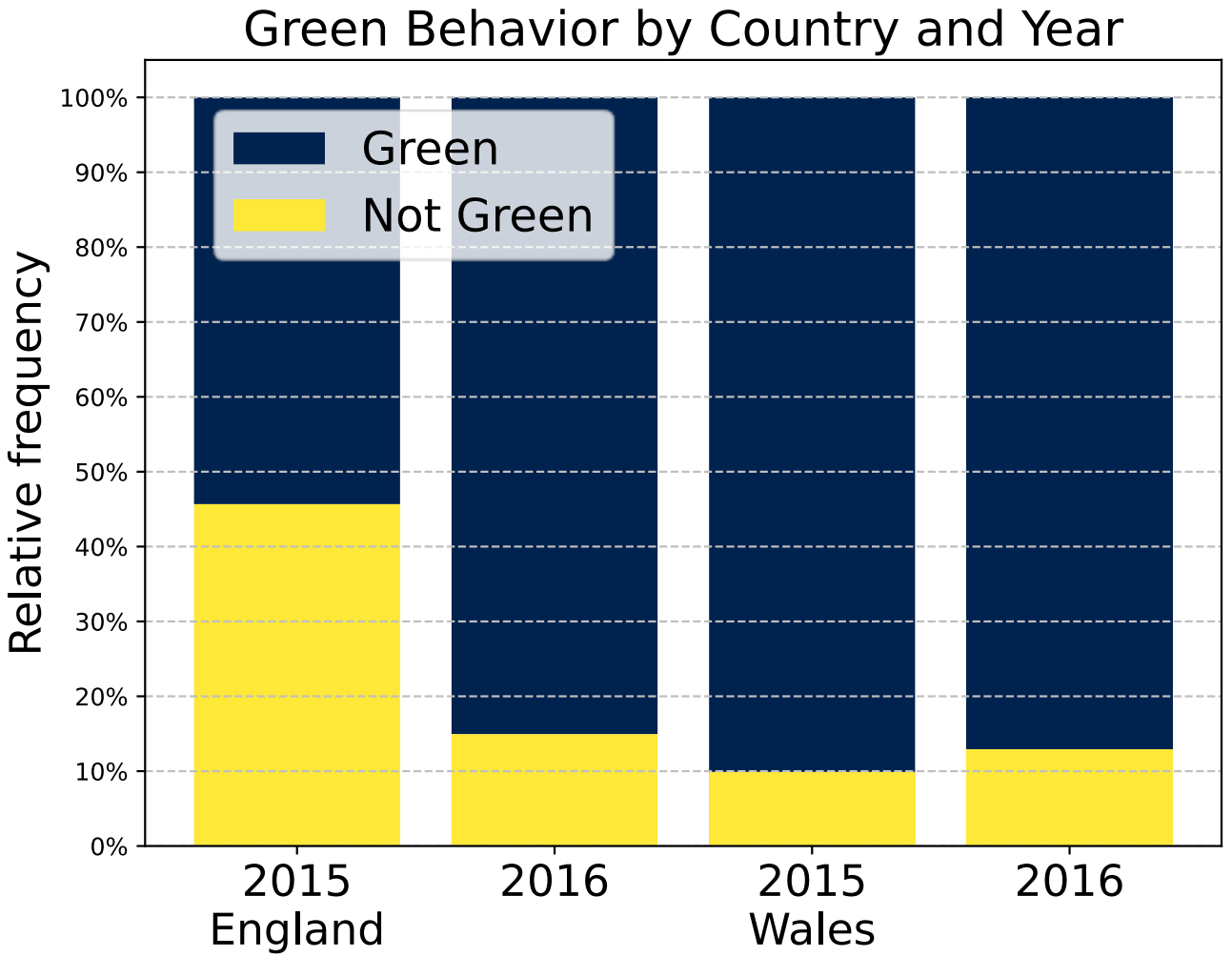
Recap: Model

Classification Model

- Y = 1 for “Green”
- No new purchase of SUPB
- **Re-using** a SUPB from home
- One new SUPB, **if reusable bags**

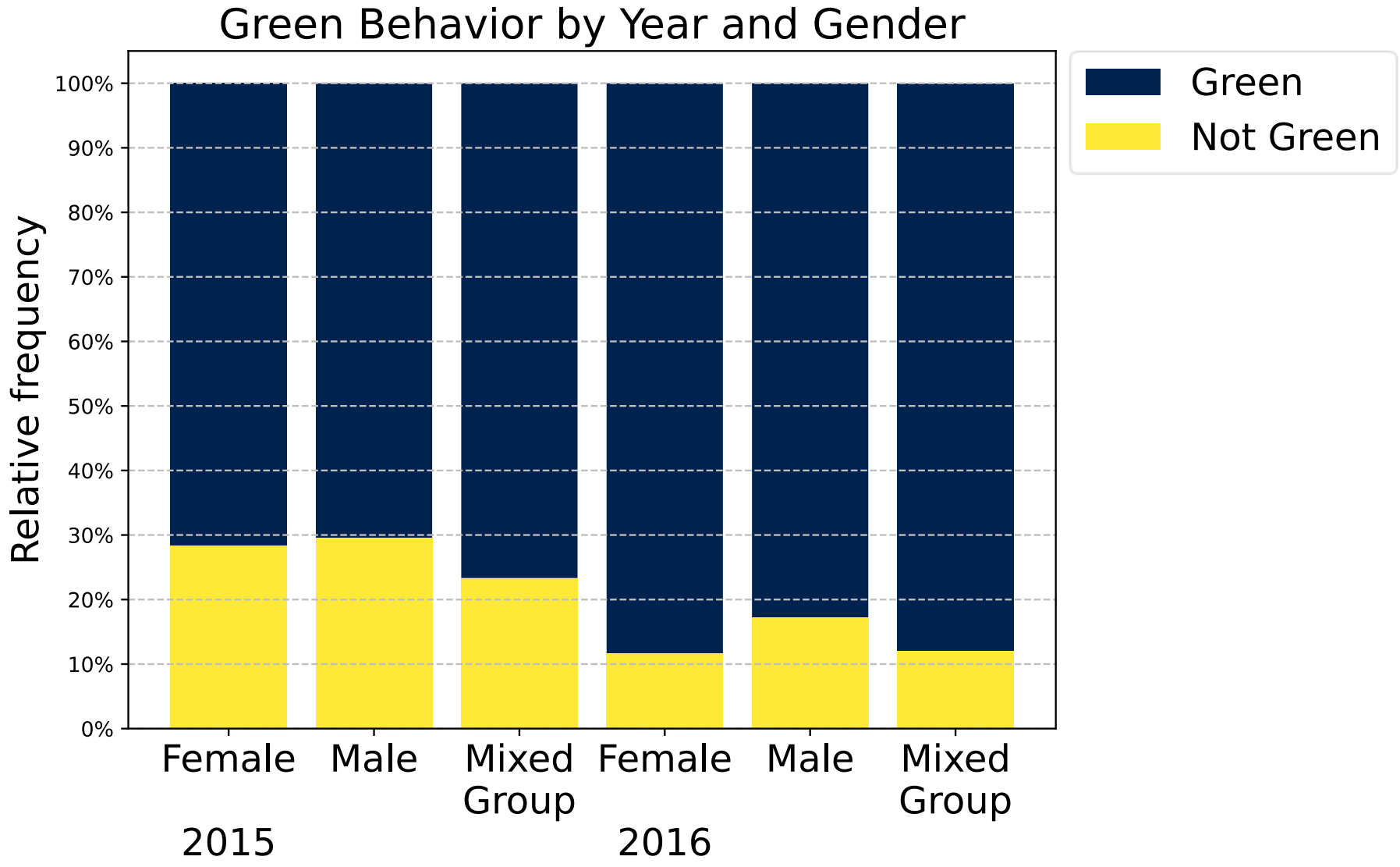
Feature Variables

- `Year`, `Country`, `Gender`, etc.
- + **Interaction Terms**



Recap: EDA

GENDER



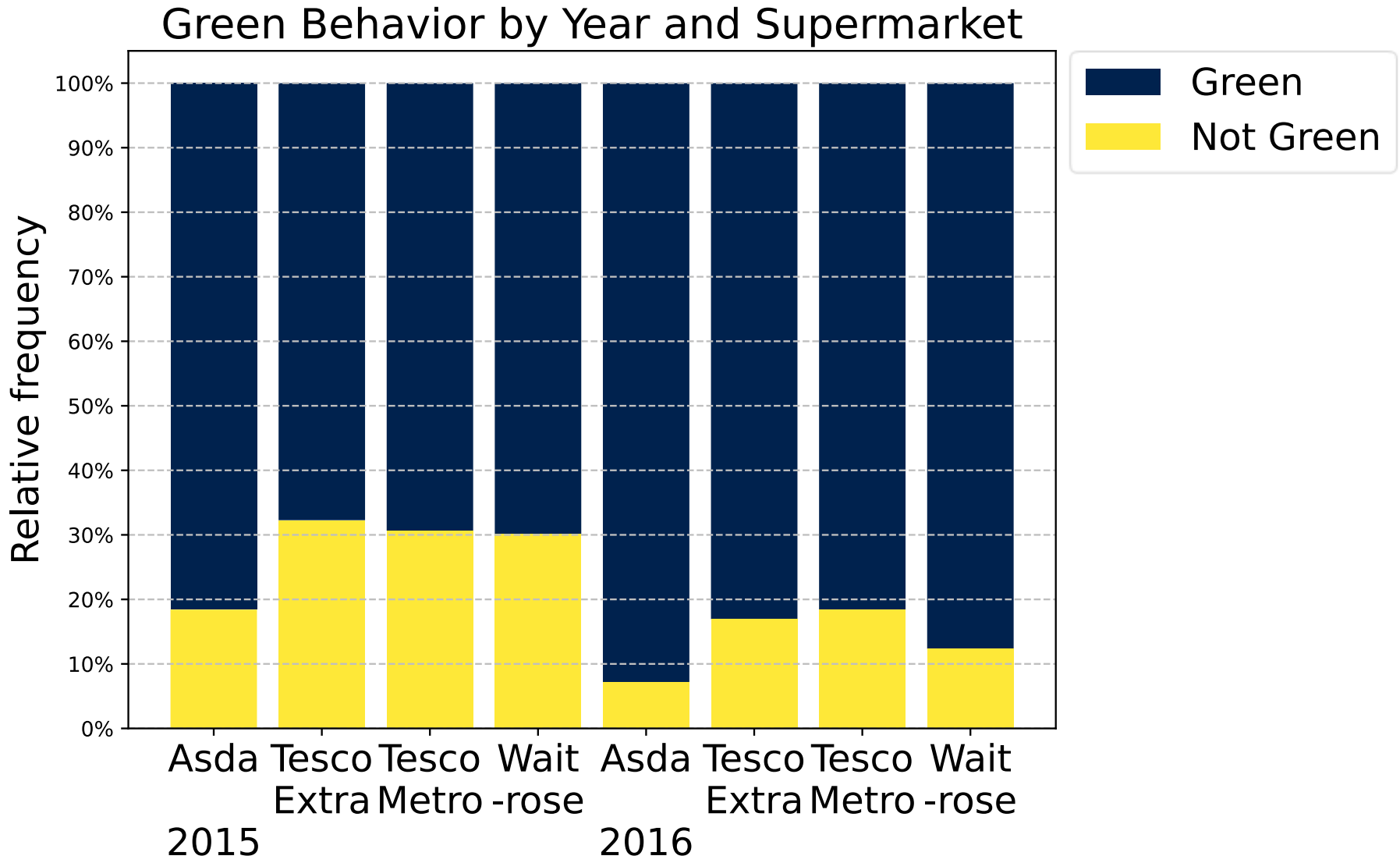
Recap: EDA

BUDGET

MID-CLASS

CITY

PREMIUM



Cross-Validation

Preprocessing
- Missing: np.nan
- OneHotEncoder
- StandardScaler
- Interaction Terms

```
for rs in random_states:
  Preprocessing
  other_set, test_set = strfied_split(10%)
  gridsearchCV(ML_algo,
               cv = 5)
  gridsearchCV.fit(other_set)
```

I.I.D. & Target **unbalanced**
- Test: **stratified** split 10%
- CV: **stratified** K-Fold (K = 5)
- 10 random states

Balanced Accuracy

Accuracy score, modified

- Sensitivity = **P(**predict pos **|** actually pos**)**, true positive rate
- Specificity = **P(**predict neg **|** actually neg**)**, true negative rate
- $\text{Acc}_{\text{balanced}} = \frac{1}{2}(\text{Sen} + \text{Spe}) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)$

Baseline

- Always predict pos → 100% Sen, 0% Spe → 50% Balanced Accuracy
- Always predict neg → 0% Sen, 100% Spe → 50% Balanced Accuracy

https://scikit-learn.org/stable/modules/model_evaluation.html#balanced-accuracy-score

Logistic Regression

No Penalty

L1 Penalty

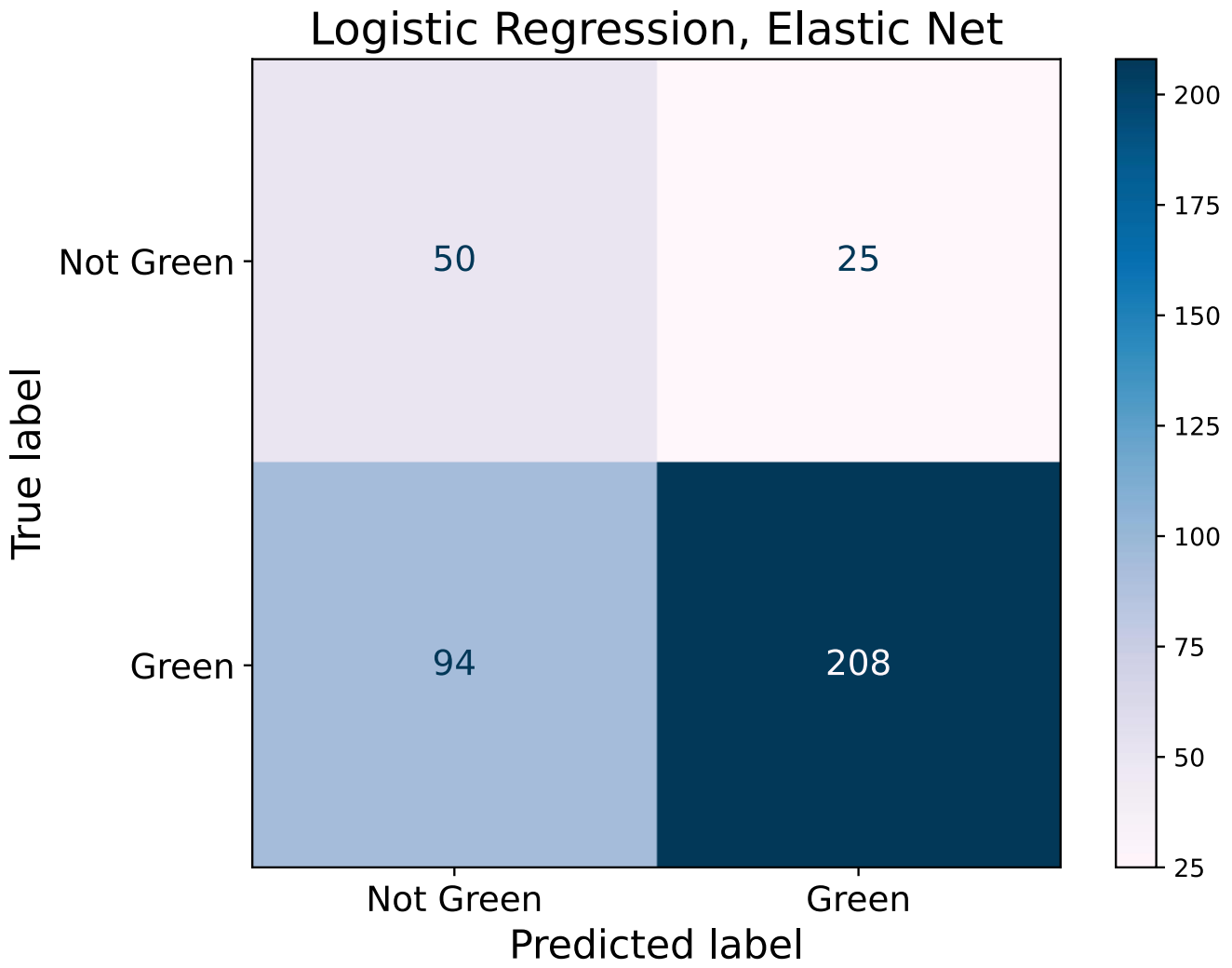
- `C: np.logspace(-3, 2, 21)`

L2 Penalty

- `C: np.logspace(-3, 2, 21)`

Elastic Net

- `C: np.logspace(-3, 2, 21)`
- `l1_ratio: [0.1, 0.2, .., 0.9]`



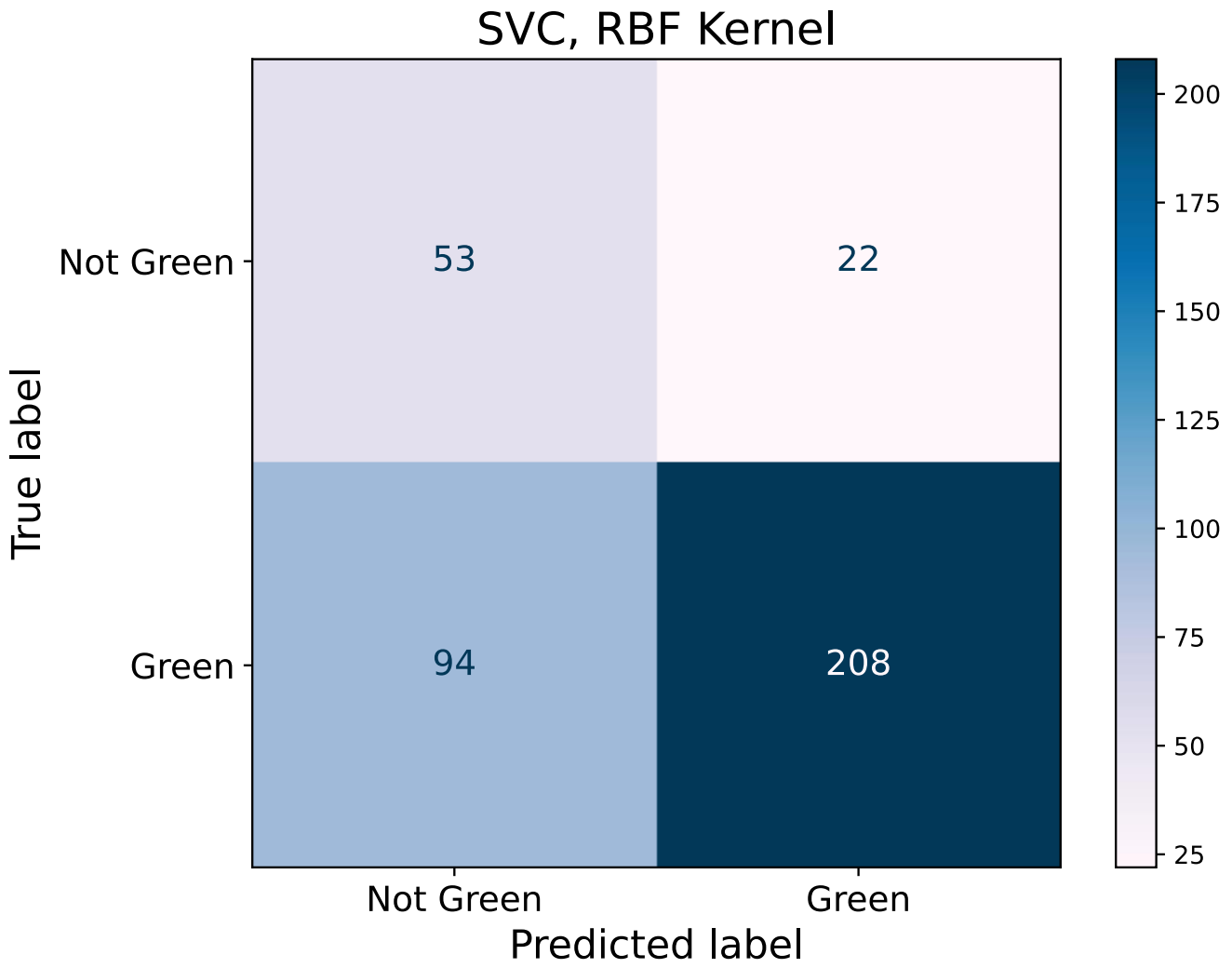
Support Vector Machine

Linear Kernel

- `C: np.logspace(-3, 2, 11)`

Radial Basis Function Kernel

- `C: np.logspace(-3, 2, 6)`
- `gamma: np.logspace(-2, 2, 17)`



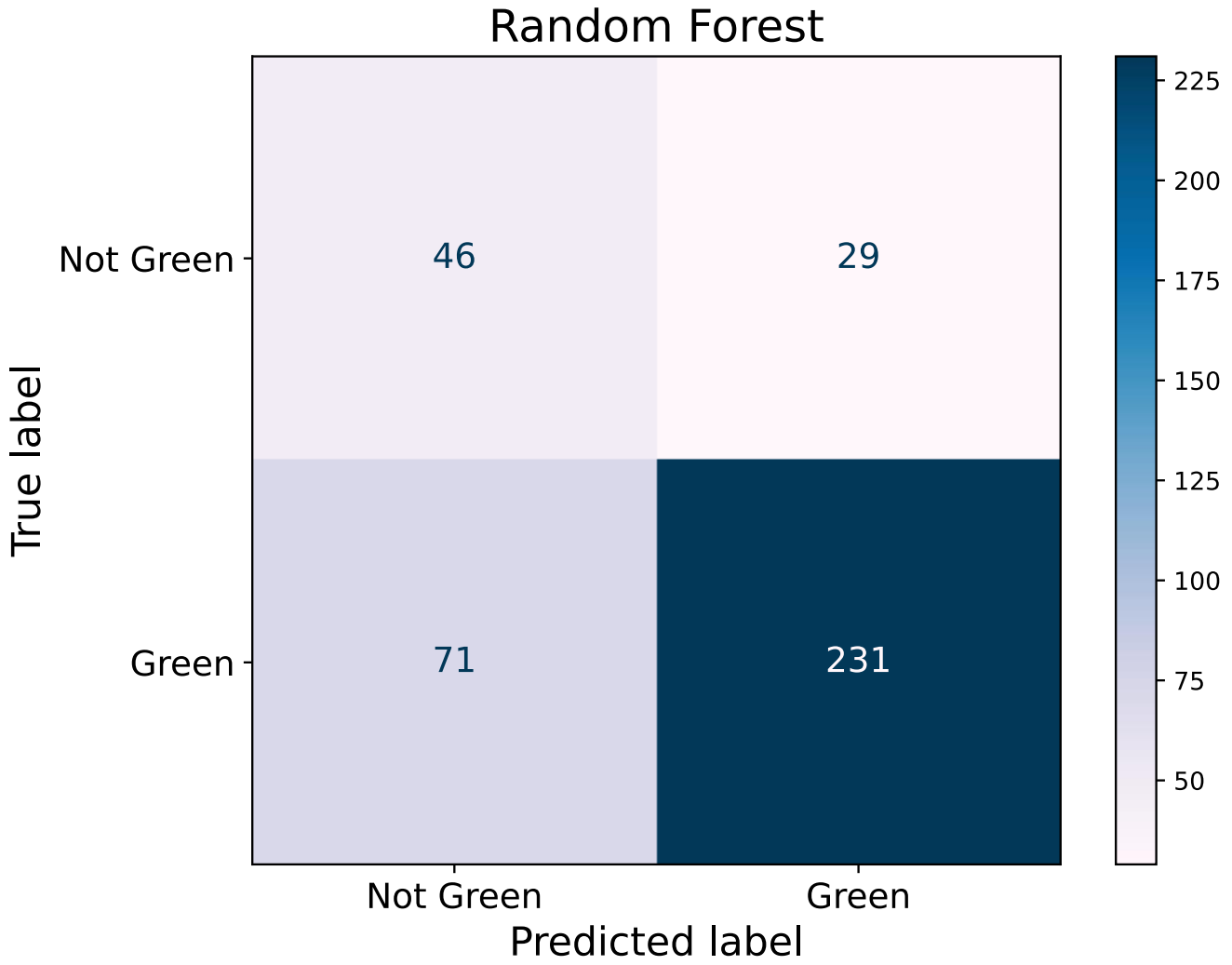
Random Forest

max_features:
- [0.3, 0.35, .., 0.65, None]

max_depth:
- [3, 4, 5, 6, 8, None]

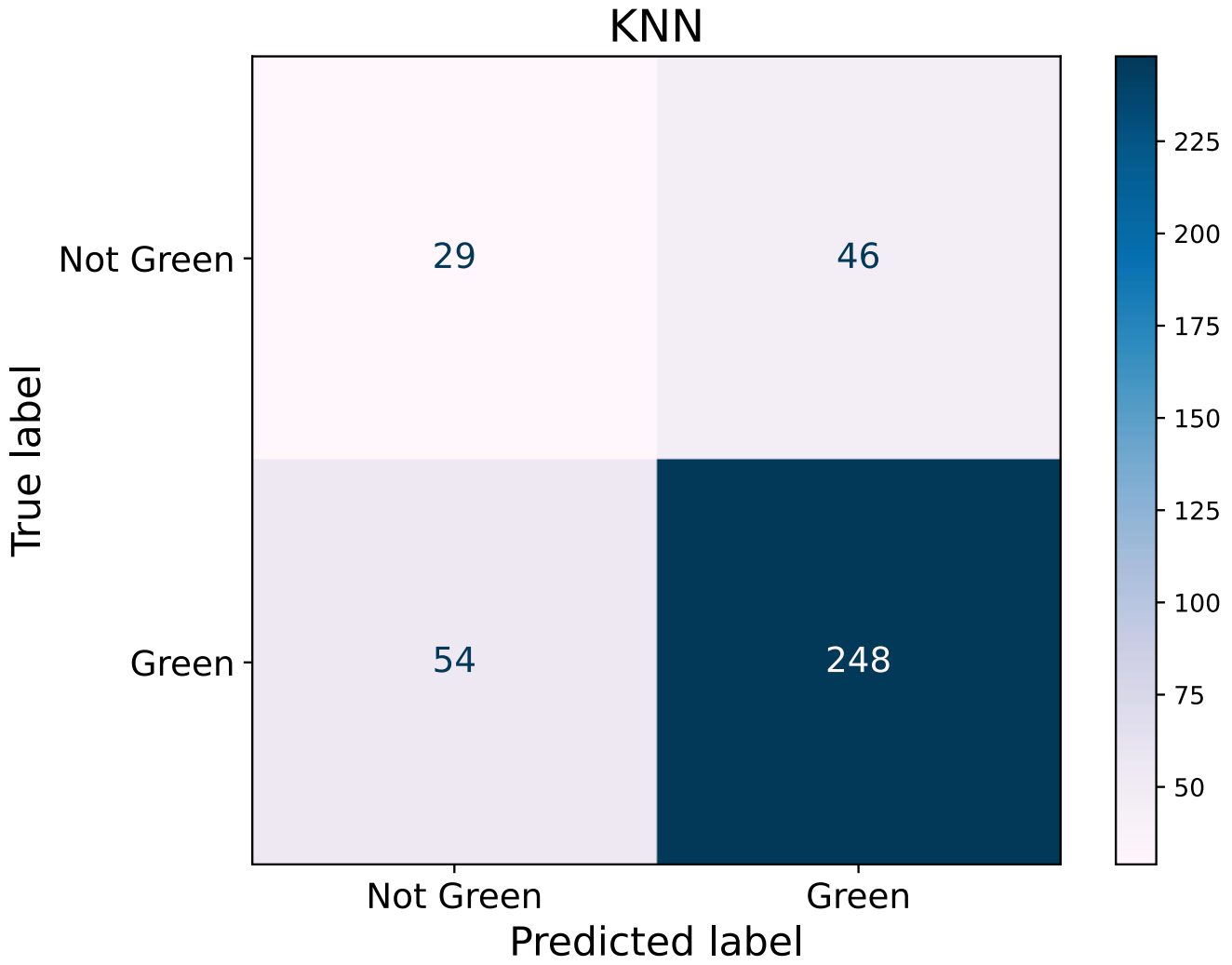
min_samples_split:
- [2, 3, .., 7]

n_estimators = 100



K Nearest Neighbors

- n_neighbors:
  - [2,3,5,8,9,10,11,12,15,30]
- weights:
  - ['uniform','distance']
  - *'class_weight' not balanced*



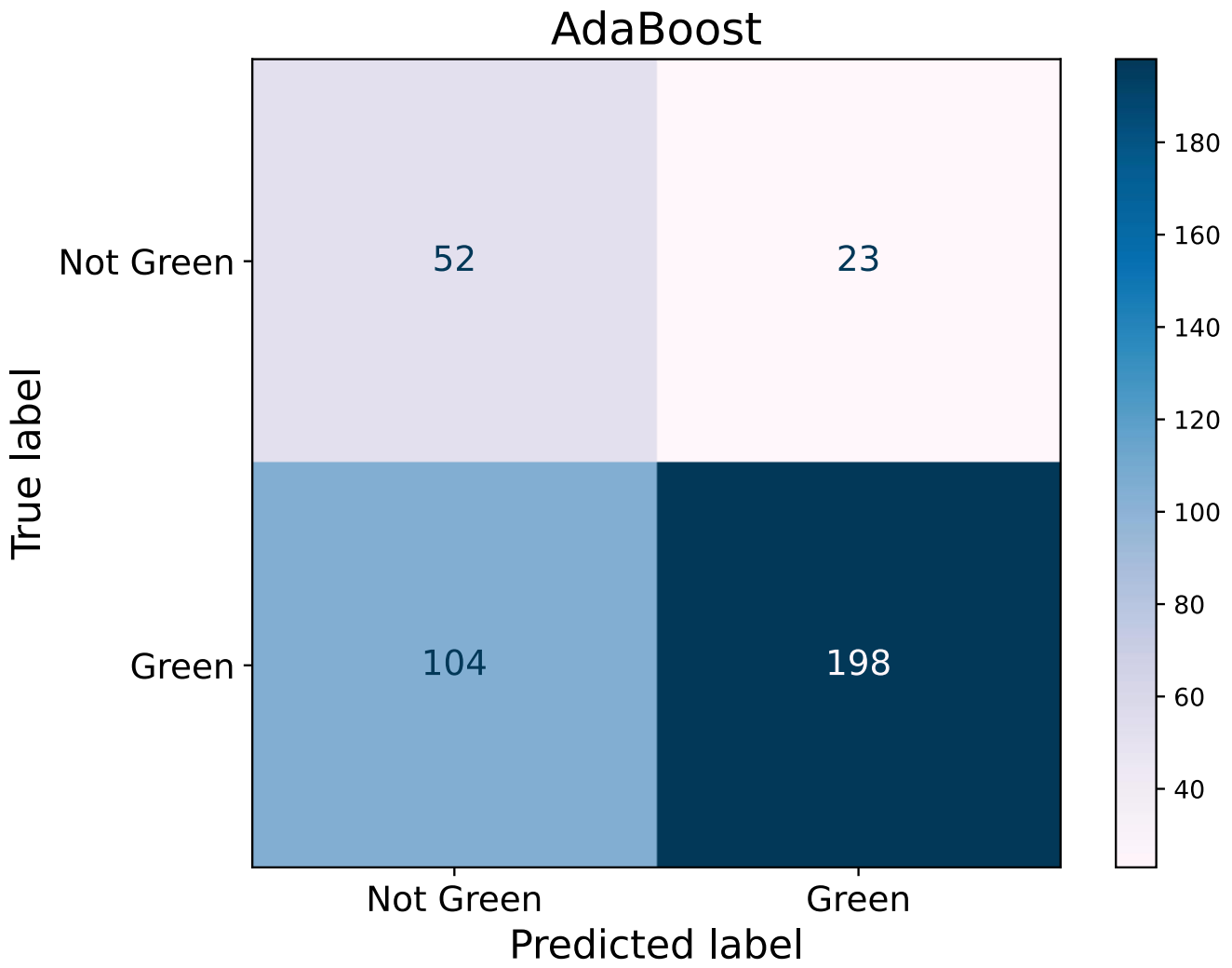
Boost Methods

AdaBoost

- learning_rate: [0.3, 0.4, .., 1.1]
- n_estimators: [8, 12, .., 40]
- base__max_depth: [1, 2, 3, 4]

Gradient Boost

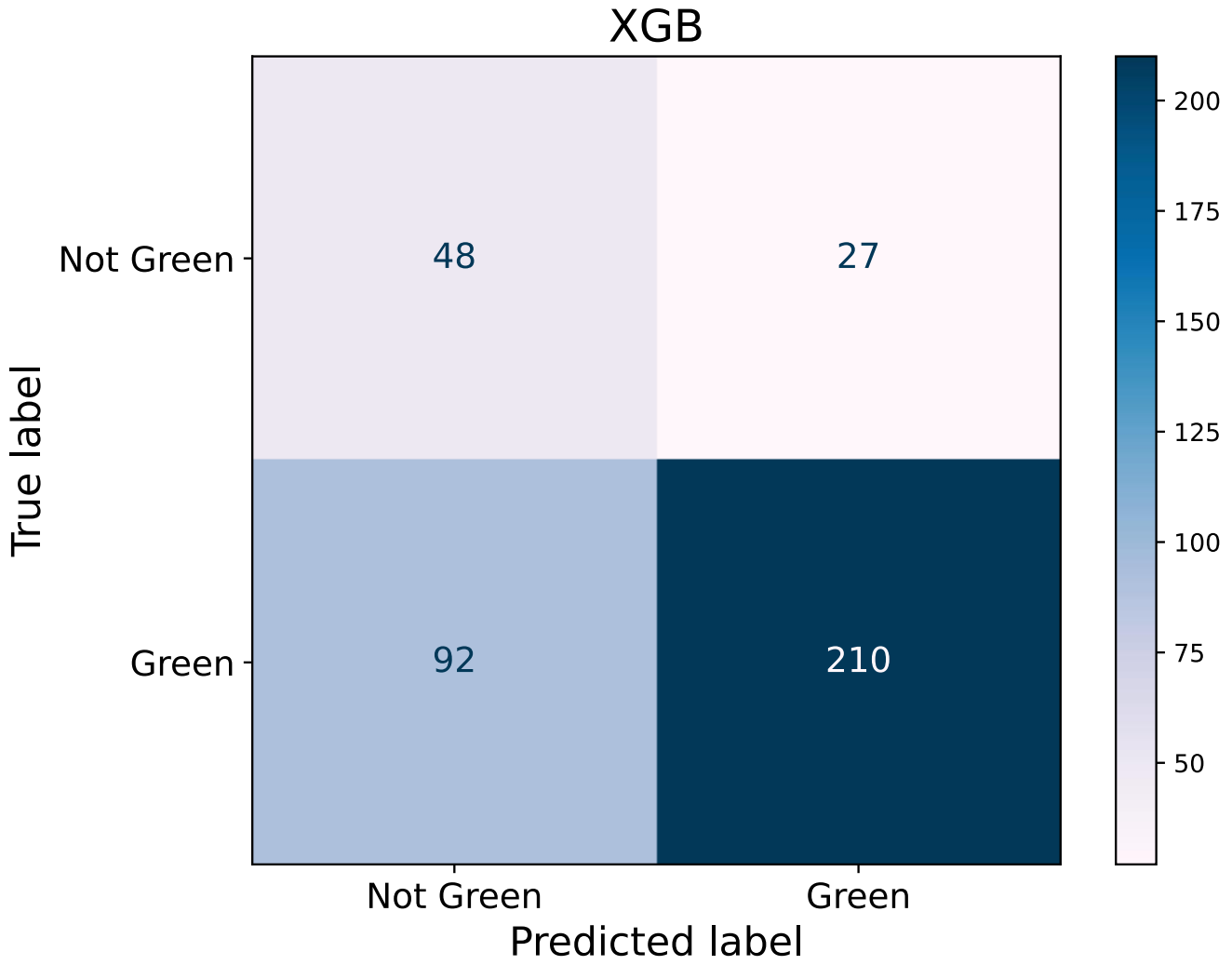
- learning_rate: [0.05, 0.1, .., 0.3]
- max_features: [0.1, 0.2, .., 1.0]
- max_depth: [2, 3, .., 6]
- n_estimators = 50



Boost Methods

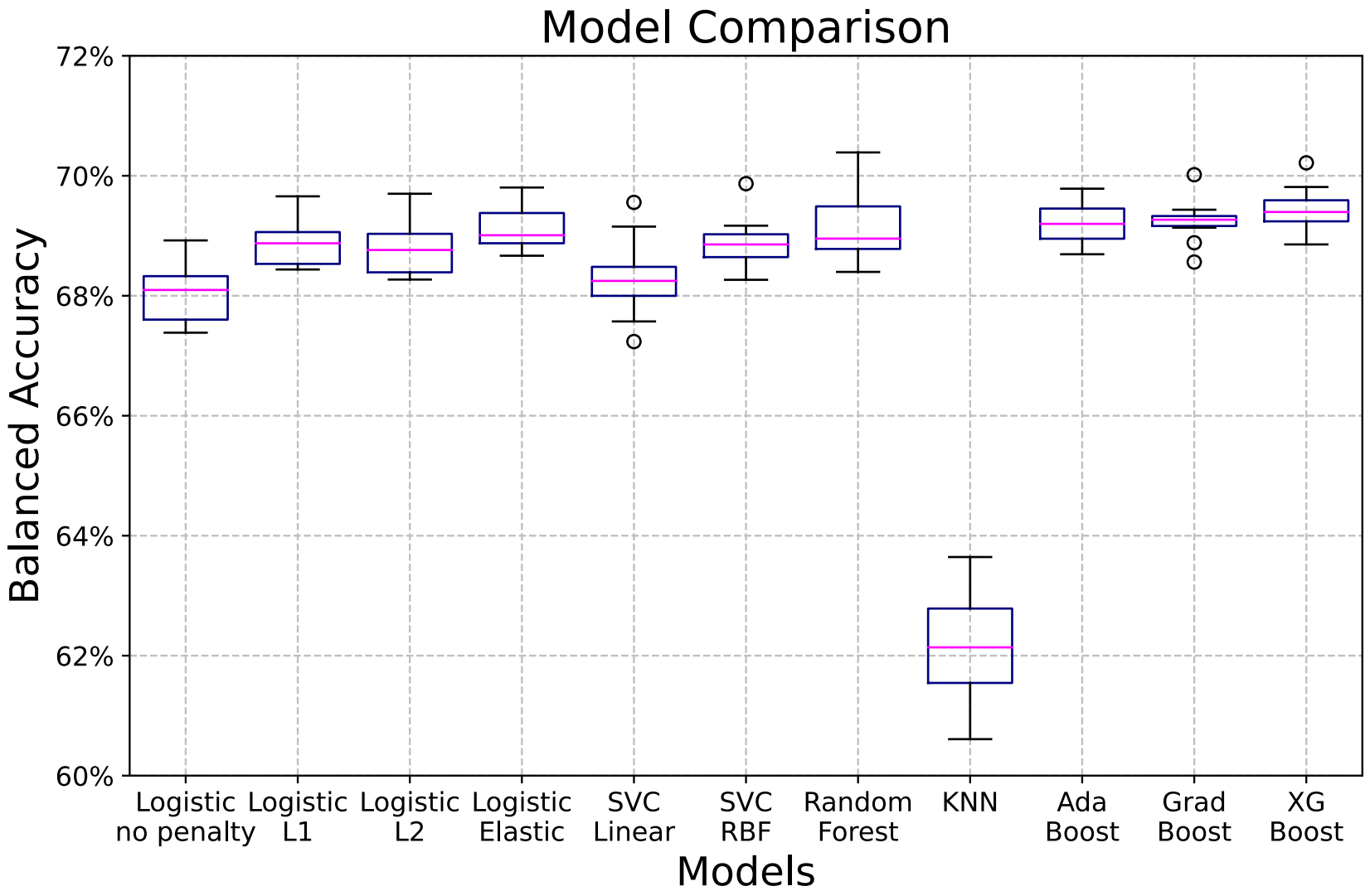
XGBoost

- 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]



Results

KNN Class Weight

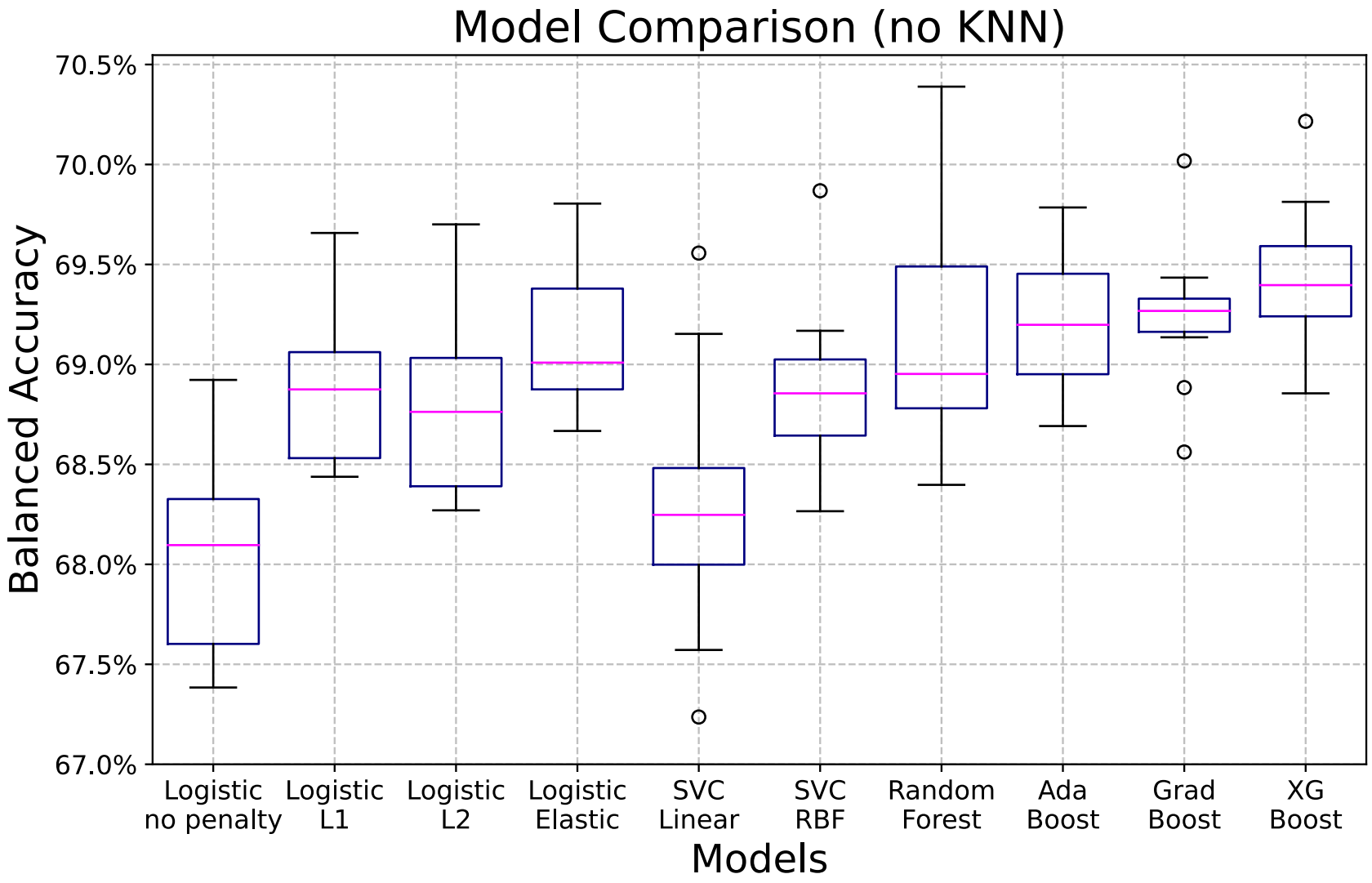


Results

XGB best

All within

68% - 69.5%

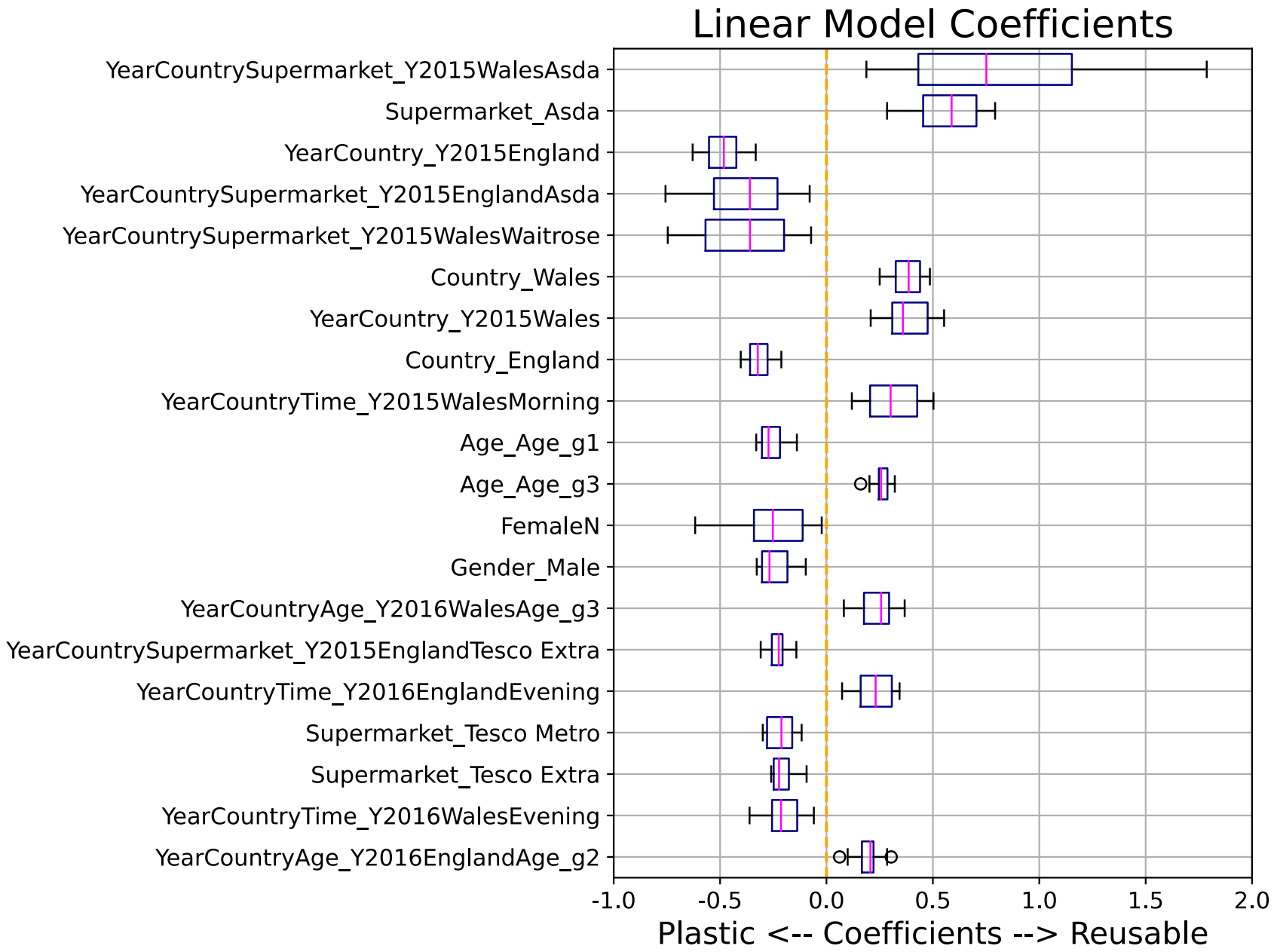


Feature Importance

Global Importance

Logistic Regression

L2 penalty

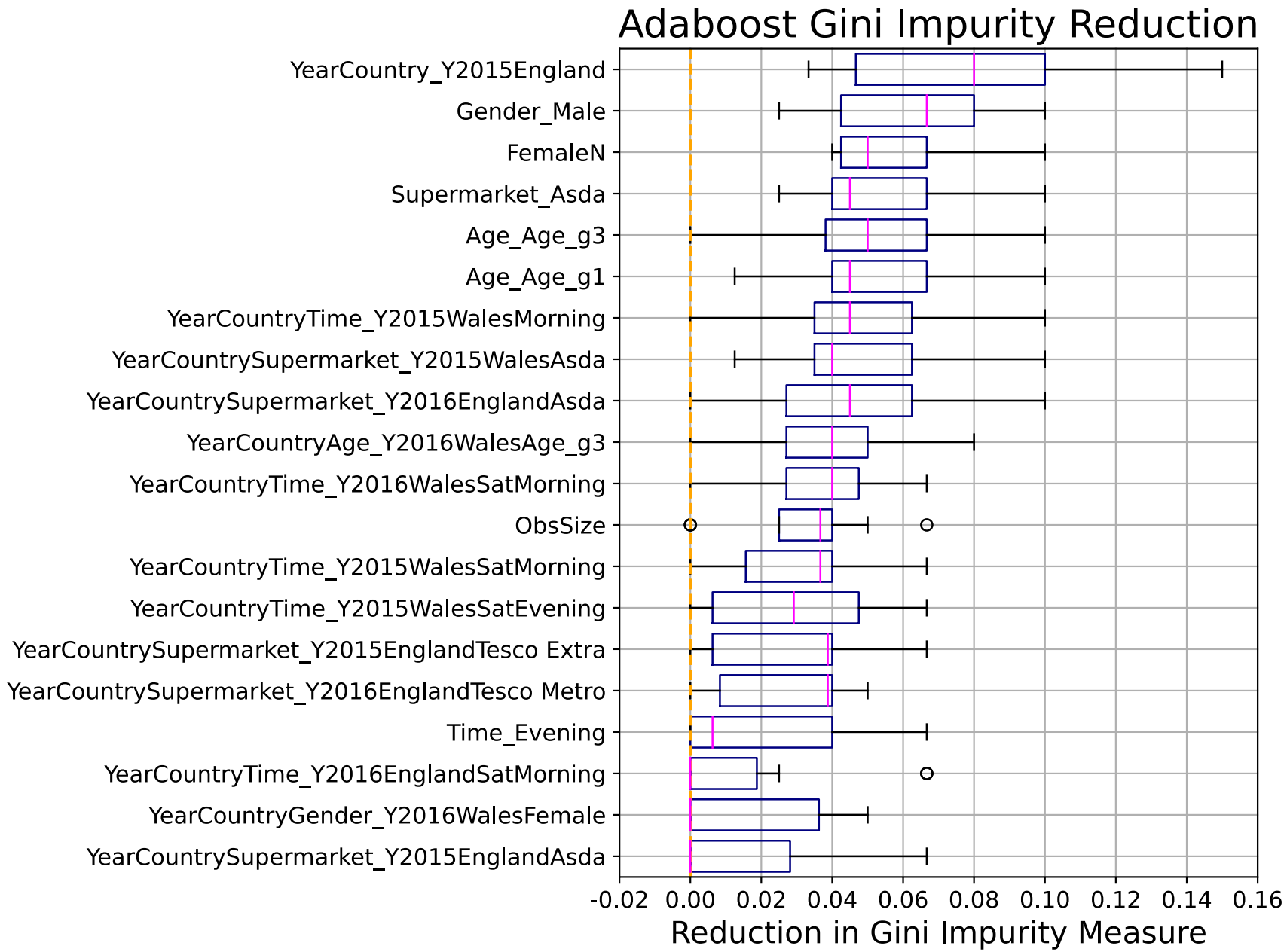


Feature Importance

Global Importance

AdaBoost

Mean Decrease in Gini Impurity



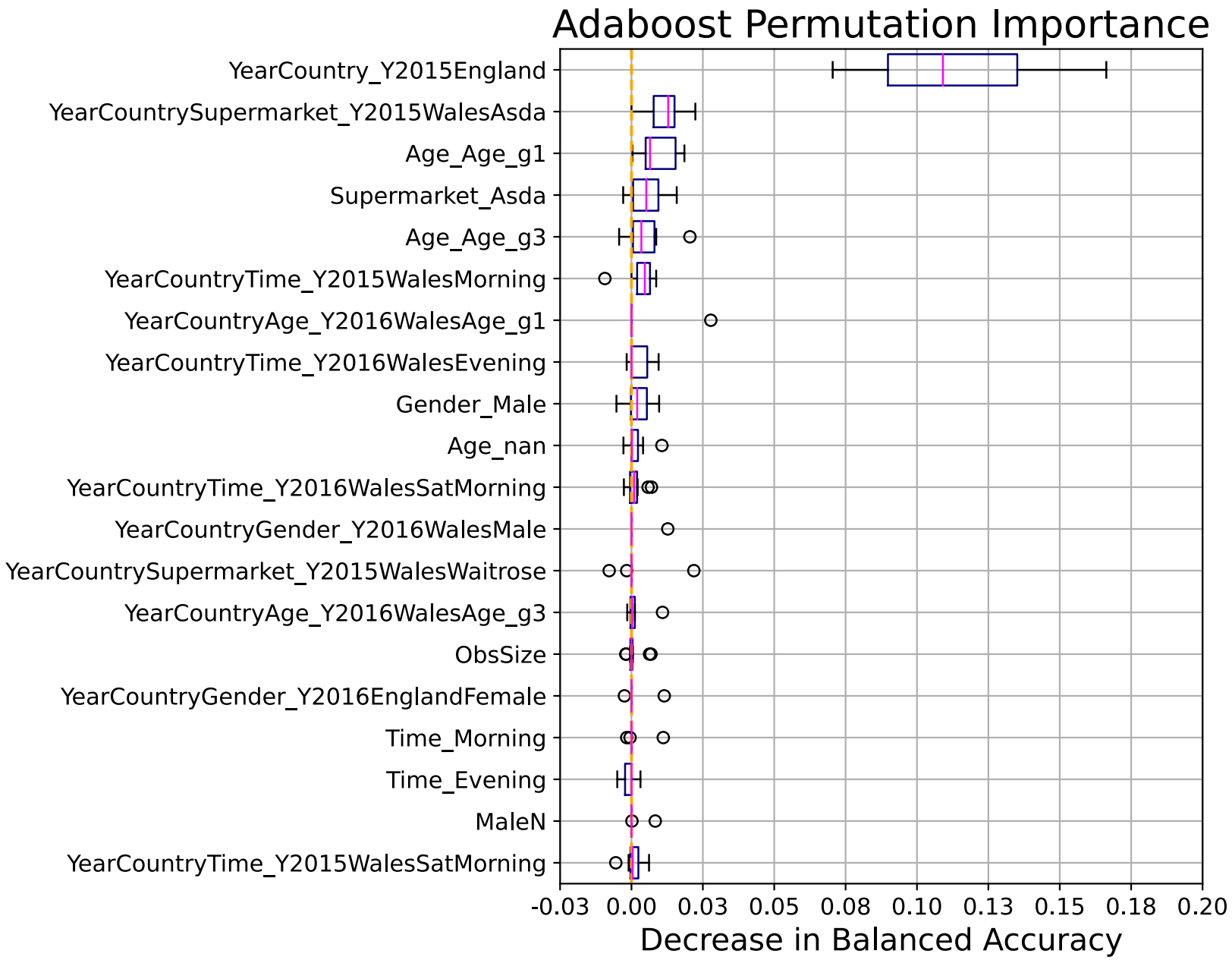
Feature Importance

Global Importance

AdaBoost

Permutation Score

Balanced Accuracy

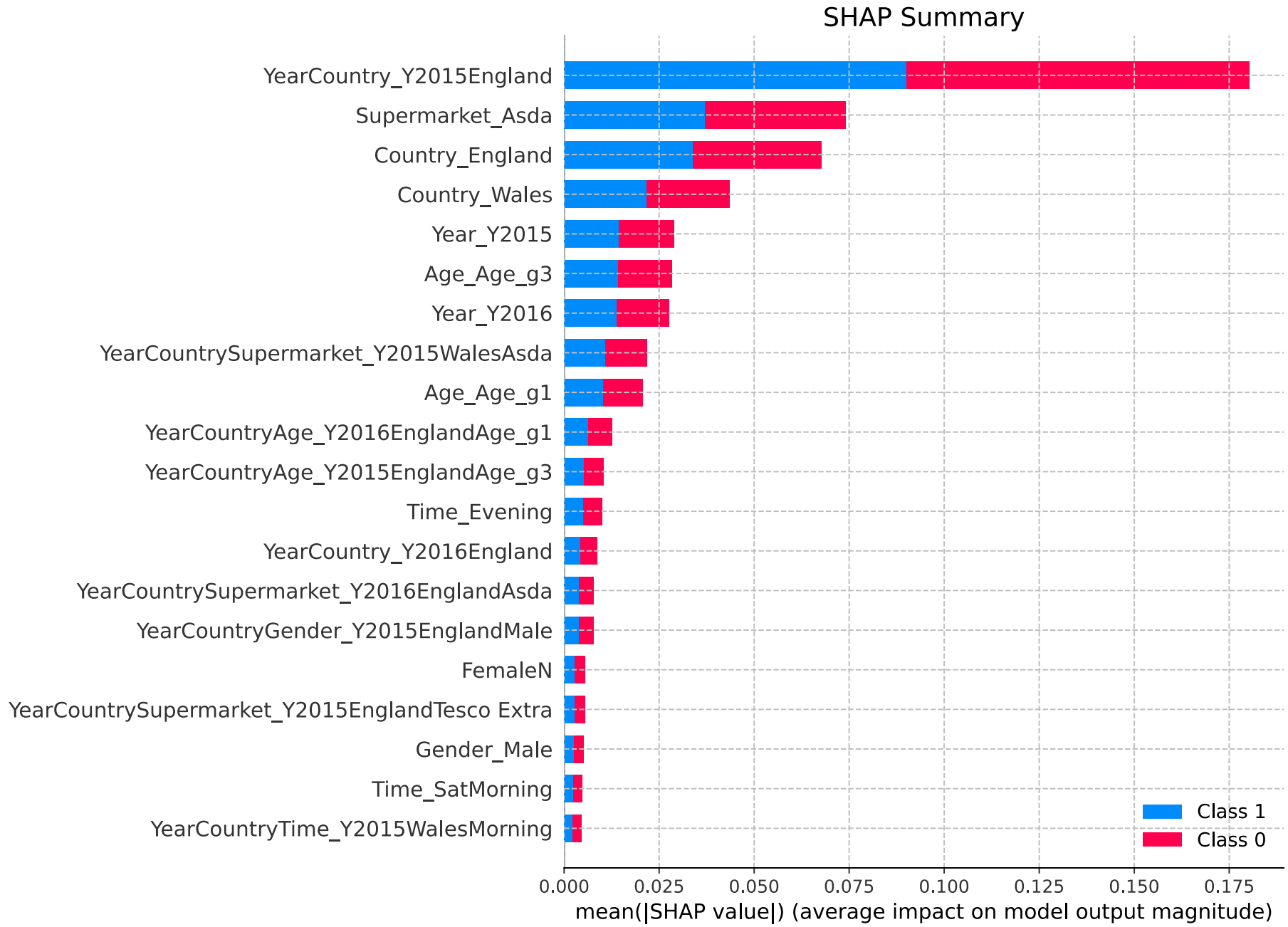


Feature Importance

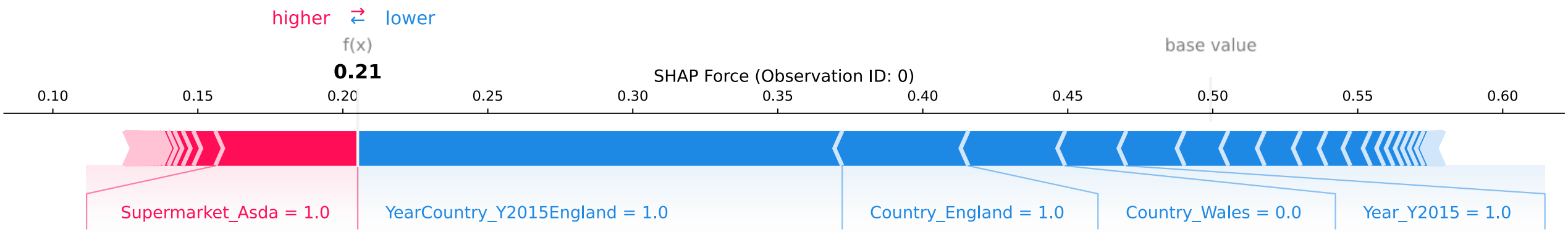
Global Importance

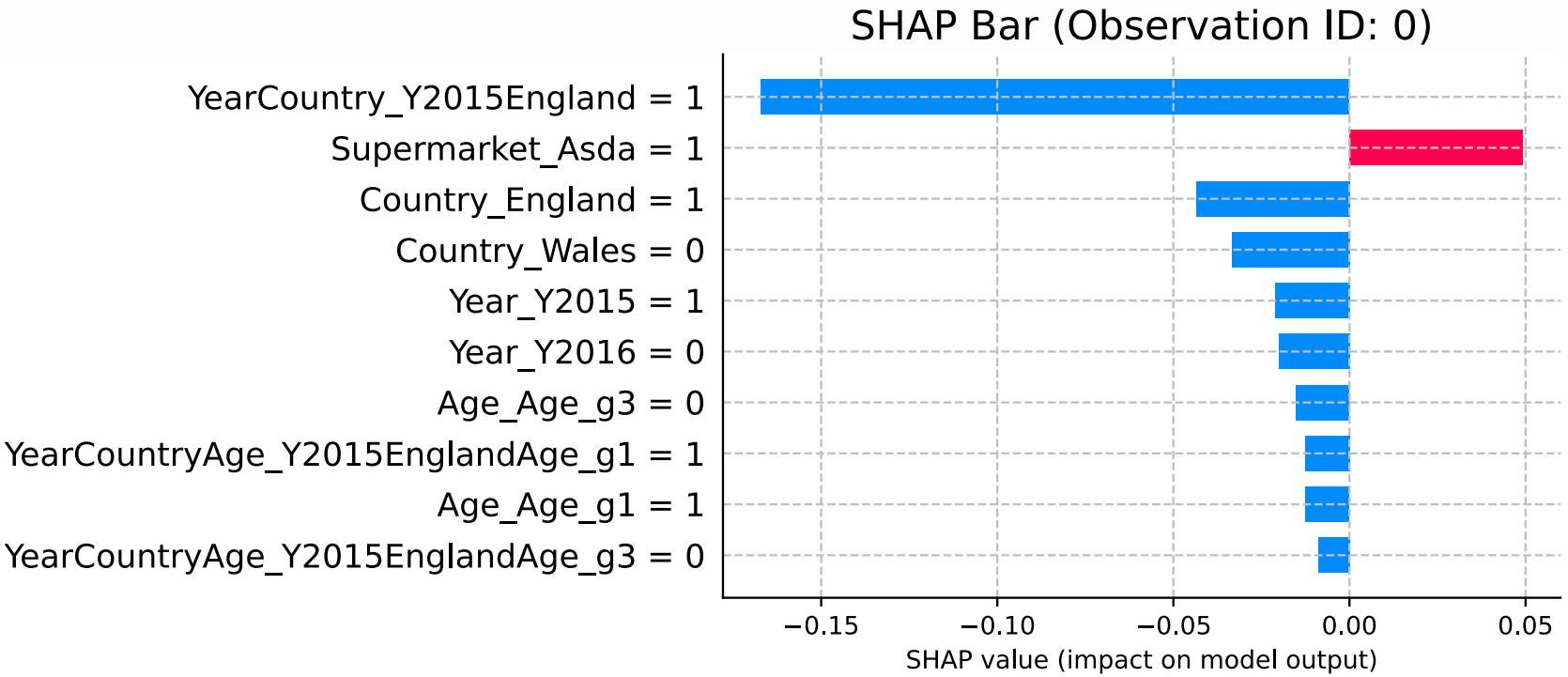
Random Forest

Average SHAP score

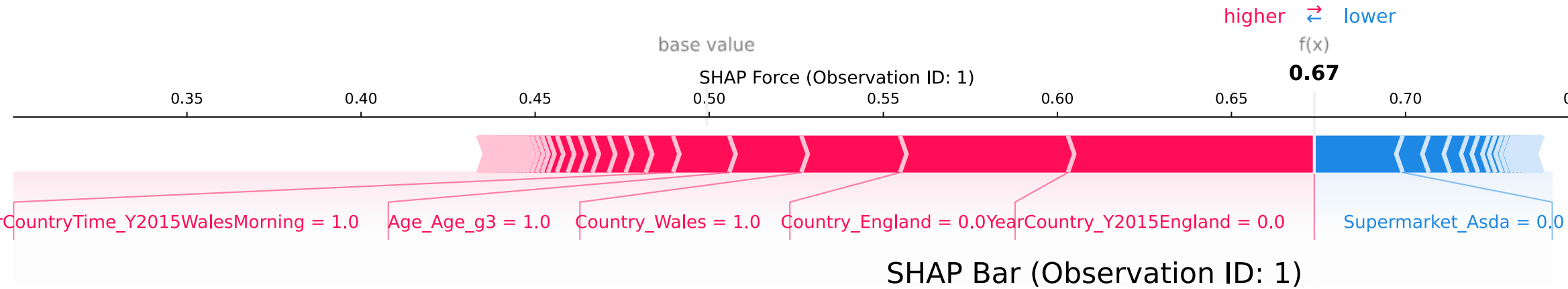


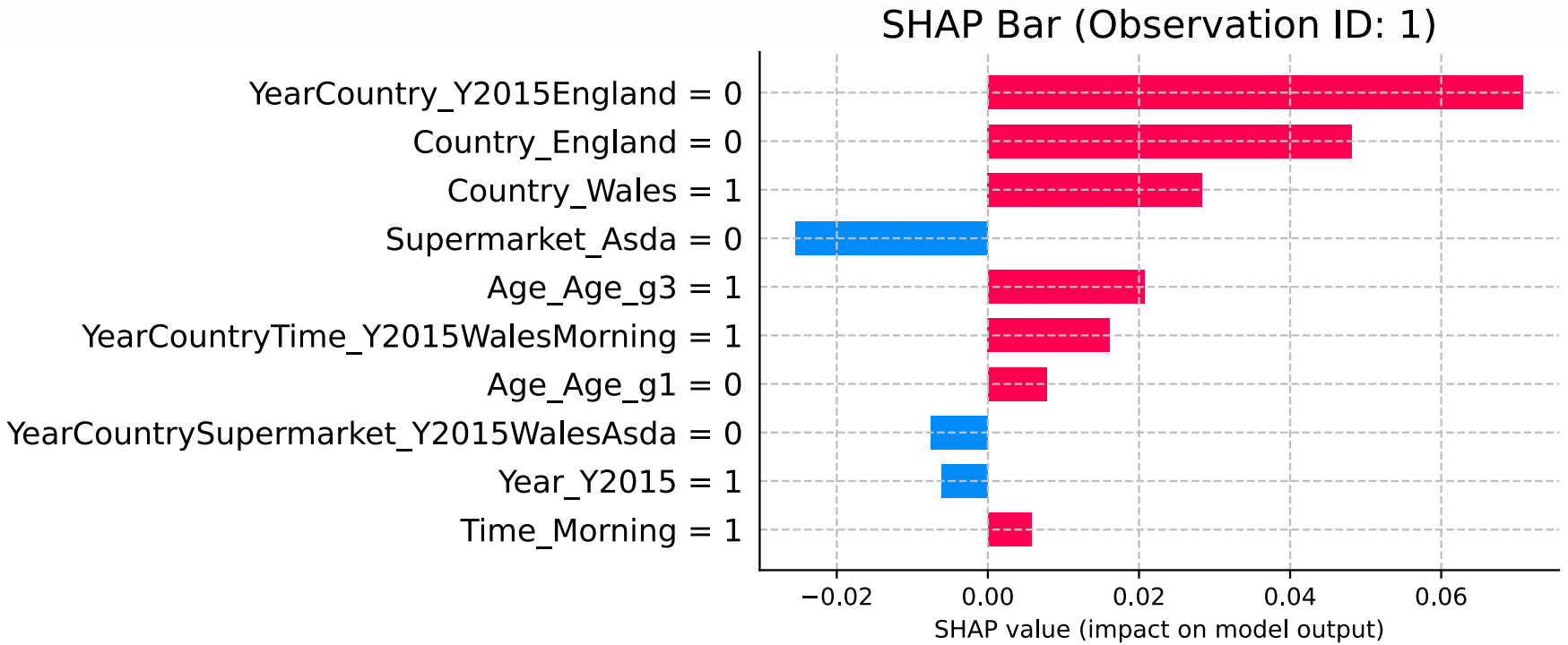
SHAP Local Feature Importance





SHAP Local Feature Importance





Outlook

Feature Engineering

More Recent Data

Minimize the SUBP usage

(Poortinga, 2016, "Single-use plastic bags“)

References

1. 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.

2. 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.

3. 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

4. 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/

5. 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

6. 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

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

8. 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

Any questions?

Thank you very much!!
More in the Repo!

Appendix

Logistic Regression

No Penalty

L1 Penalty

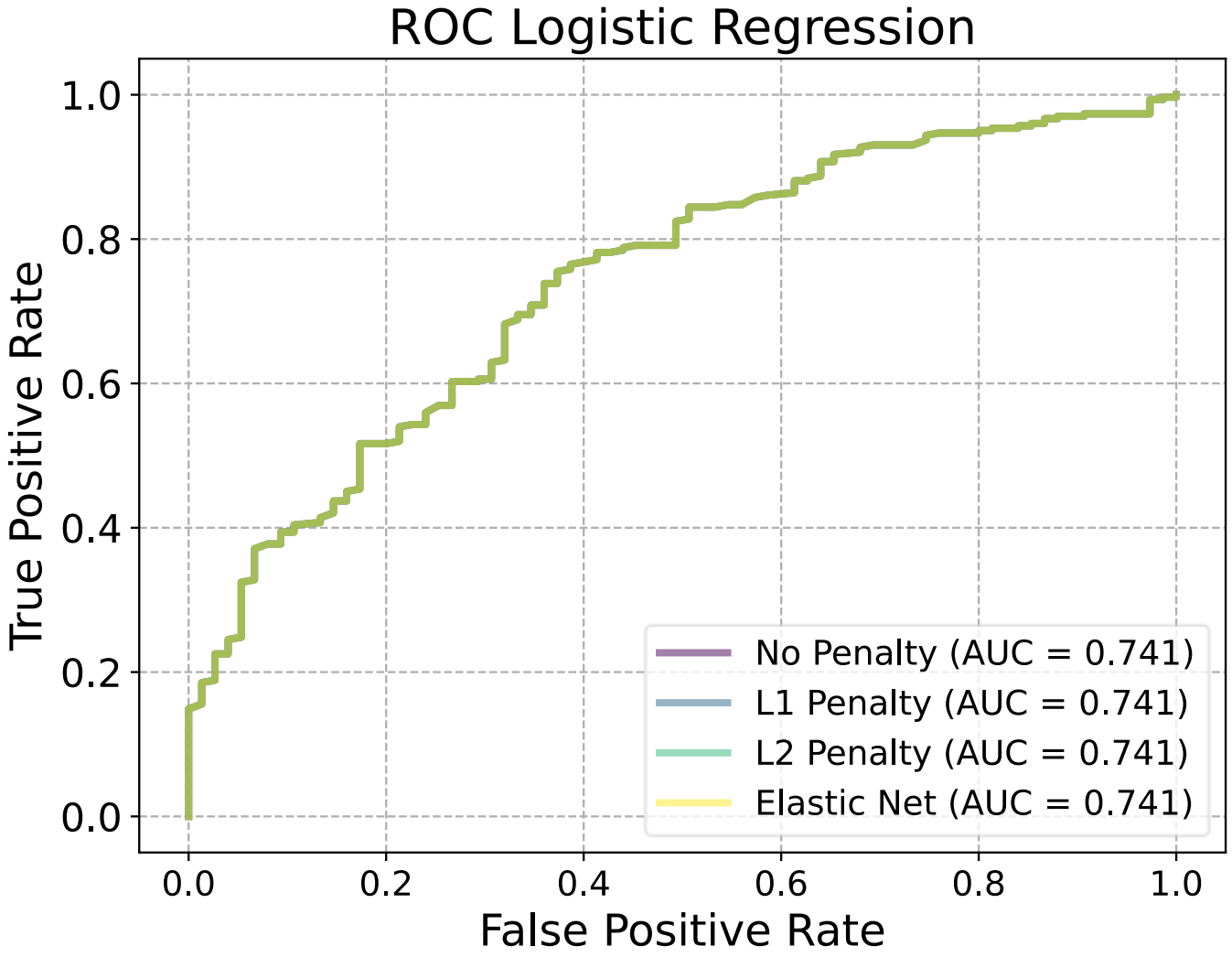
- `C: np.logspace(-3, 2, 21)`

L2 Penalty

- `C: np.logspace(-3, 2, 21)`

Elastic Net

- `C: np.logspace(-3, 2, 21)`
- `l1_ratio: [0.1, 0.2, .., 0.9]`



SHAP Matplotlib Color

Cannot change the colormap when matplotlib == True

https://github.com/slundberg/shap/issues/62