From Rags to Riches:
Using web-scraped data to derive a clothing price index

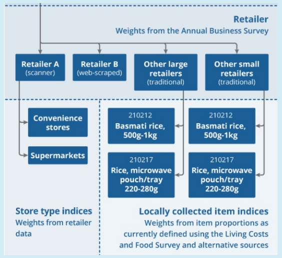
Laura Christen

Data Scientist | Prices Division

Office for National Statistics | UK

5th June 2023

The clothing project is part of a wider programme of transformation



Alternative data sources (ADS)

Incorporate scanner and web-scraped data into the production of major consumer price statistics

Clothing

5% of the CPI basket in the UK

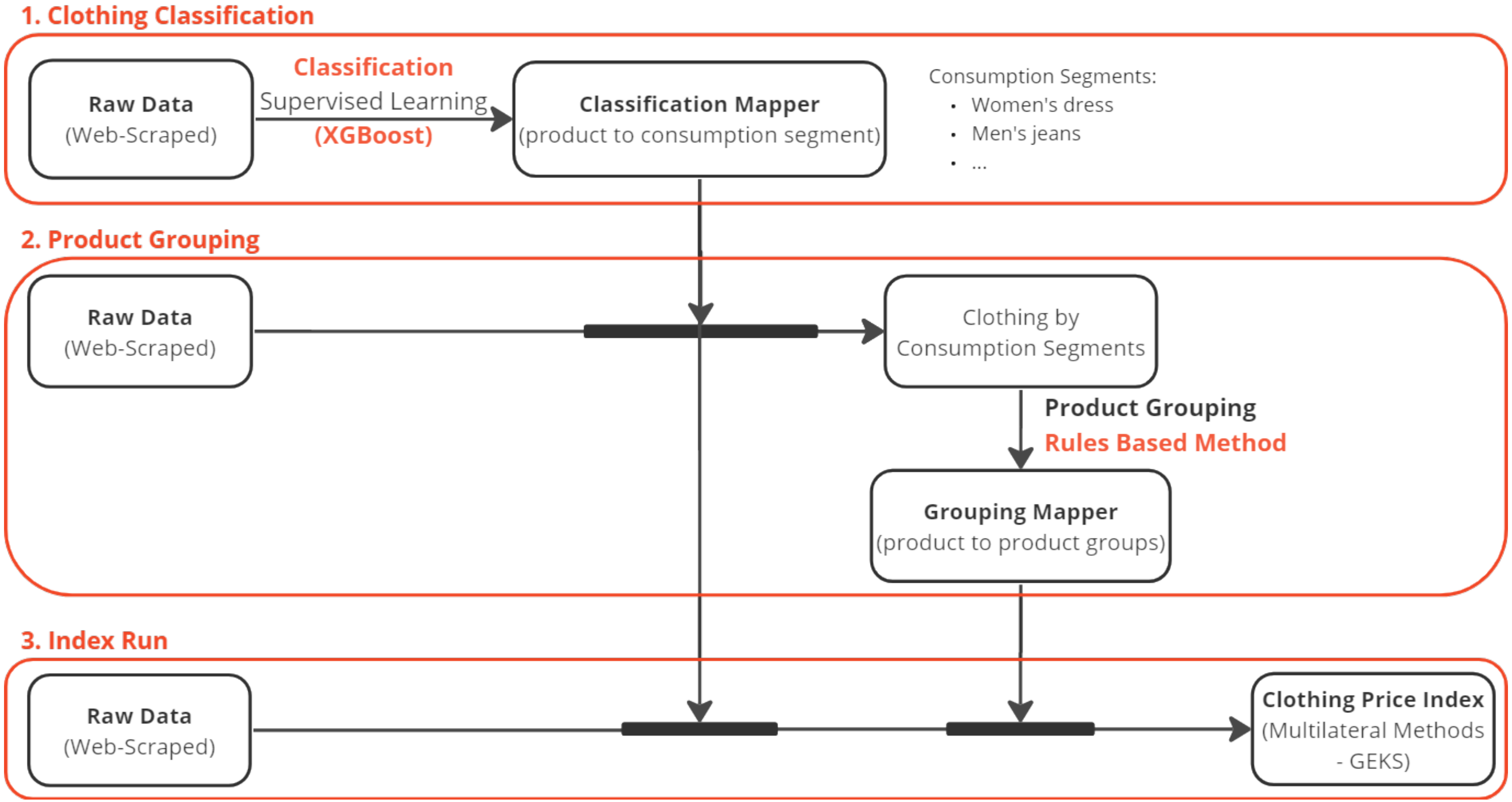
Webscraping

17 retailers

1000 brands

Dataset from June 2020

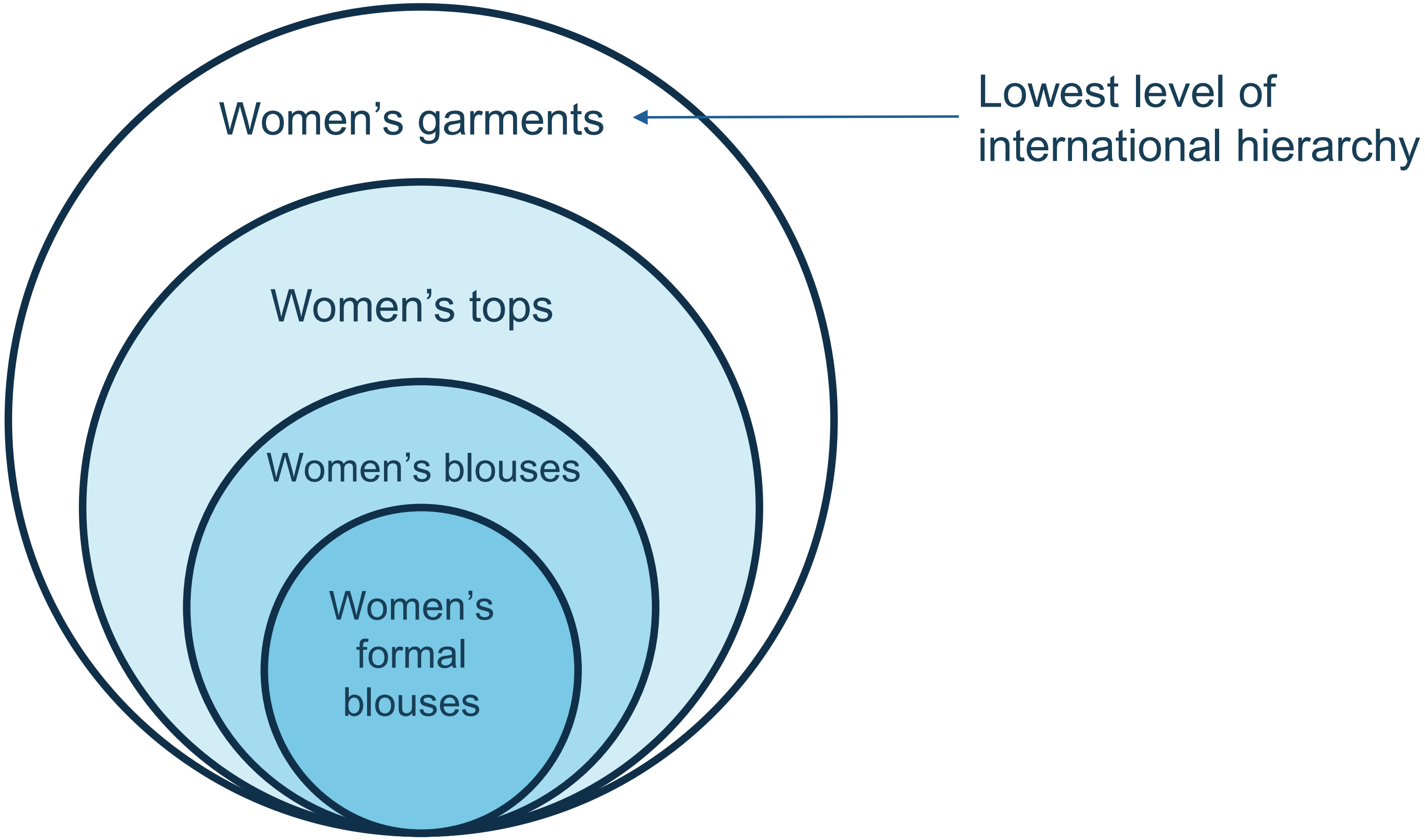
There are 3 key pipelines in the clothing project



We classify our data into UK-specific "consumption segments"

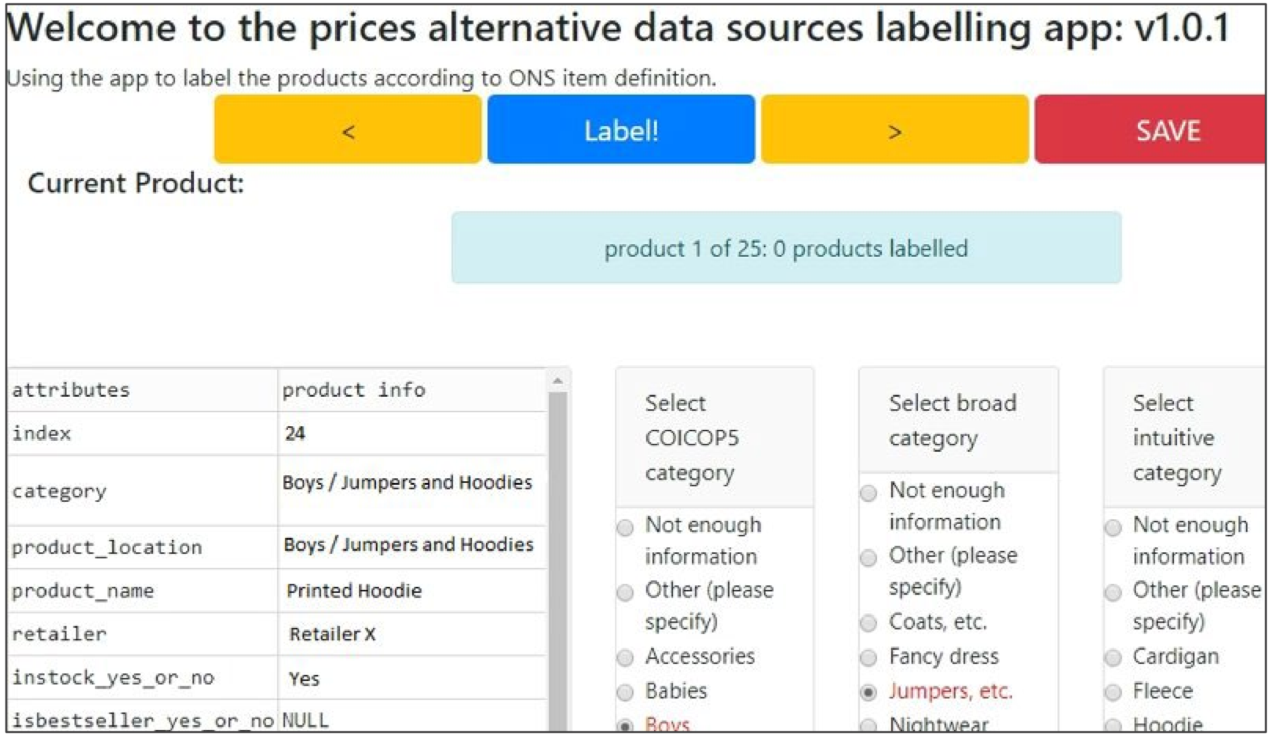
They must be:

- ✓ Relatively homogeneous
- ✓ Simple to classify
- ✓ Right size to produce reliable statistics

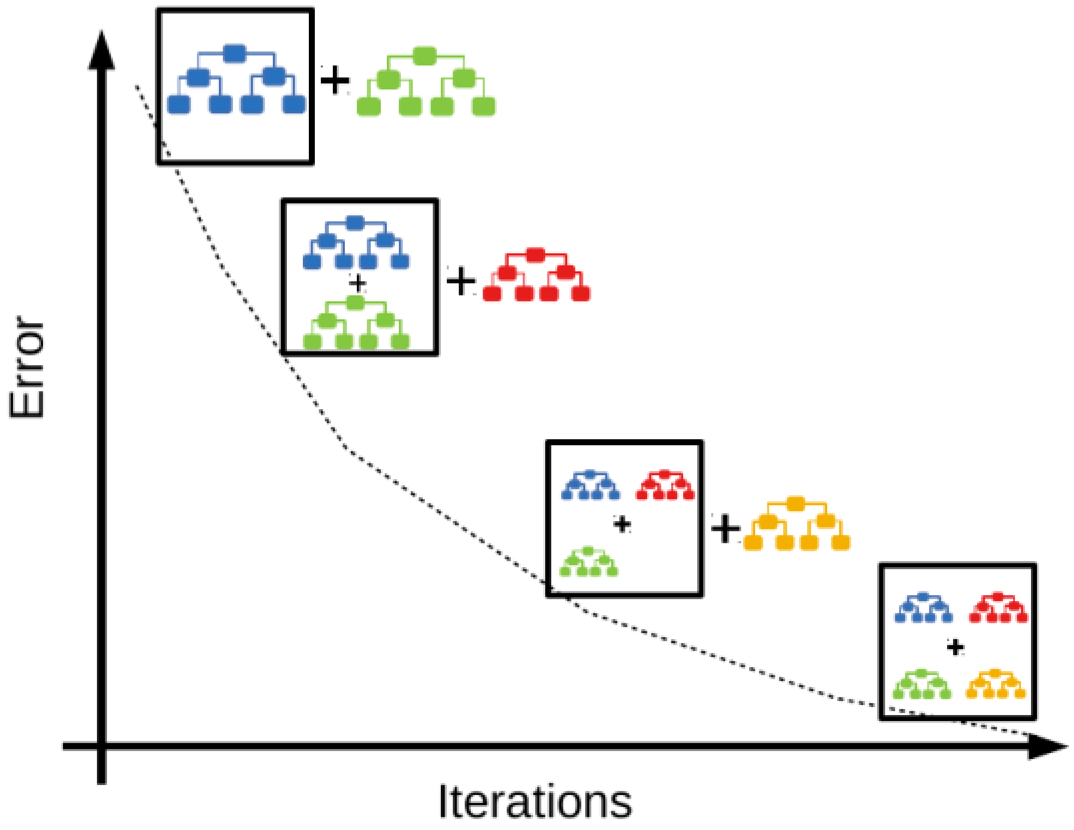


We used an in-house app to produce a labelled dataset

- Supervised machine learning requires a large, labelled dataset (162,700 products labelled)
- We achieved 89% consistency across categories; this varied by class
- Testing and training datasets



After testing multiple classification models, XGBoost best fit our needs



XGBoost:

- High performance metrics
- Acceptable training time (with GPU support)
- Confidence scores

Image source: https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

We’ve recently investigated two methods of model improvement

1) Confidence Threshold

- Defines a prediction probability that is the “threshold” for allocating a product to a class
- Increases precision at the expense of recall – this could be preferable **for our task**
- Could result in use of an fbeta score

Threshold	Precision	Recall	F1 Score	F0.33 Score
None	0.86	0.84	0.85	0.86
0.70	0.91	0.69	0.77	0.88
0.75	0.92	0.66	0.75	0.89
0.80	0.92	0.61	0.72	0.88

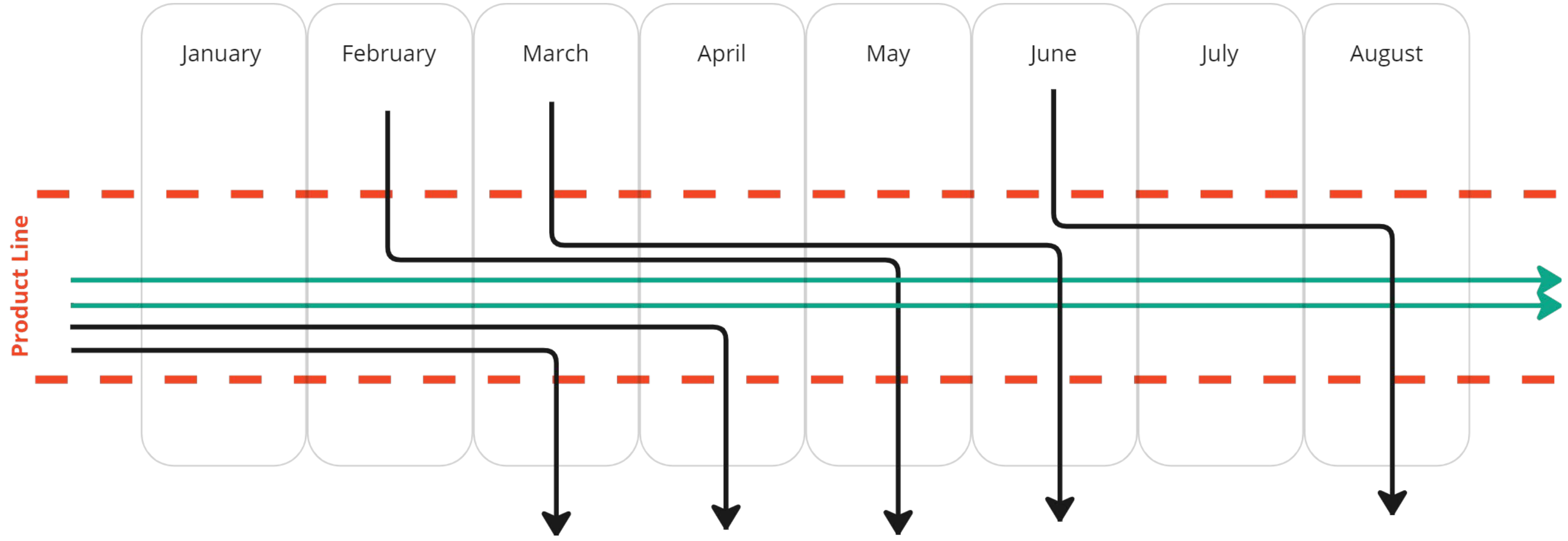
We’ve recently investigated two methods of model improvement

2) Confusion matrix

- Homogeneity vs. simplicity vs. size
- Compares predicted value to actual value for each class, providing us with points of contention
- Use weight, F1-score, and change in F1 score to decide which classes to combine

Class	Point of Contention
Girls’ sports top	Girls’ top/t-shirt/crop-top
Boys’ outfit set	Boys’ full tracksuit
Men’s sports top	Men’s t-shirt
Women’s sports top	Women’s top/t-shirt/crop-top
Boys’ vest	Boys’ t-shirt

Our index tracks prices over time, but this is hindered by product churn in clothing



- Rapid product entry and exit → Product churn
- Group similar products together to follow through time
- Reducing the impact of churn on the index

We form our product groups using “rules”, or keywords, from each column

- Retailer, Brand, Product Name, Description, Style, Material
- N most common words from each attribute column

	Rules Dictionary	
Attributes:	Product Name	Material
	v-neck	polyester
	maxi	cotton

	Product Name	Material	Group Identifier
Product 1	v-neck dress	polyester	v-neck_polyester
Product 2	floral maxi dress	100% cotton	maxi_cotton
Product 3	white maxi dress	cotton elastic	maxi_cotton

The quality of our groups are measured by the MARS Score

- Ideally, group items a consumer would consider to be similar

  → Homogeneous

- Increase product match by having large enough groups to survive

  → Match Rate

- Homogeneity vs. Match Rate
- $\mathrm{MARS}_t = \mathrm{R}_t \, \mu_t$

A “quality adjustment” of rules helps to improve the MARS score

- Basic approach

  N most common words from each column

- **Quality adjustment**

  Hedonic regression

  - Quantify the impact of key words on price
  - Keep words with significant impact
  - Re-rank rules dictionary according to their contribution to the price

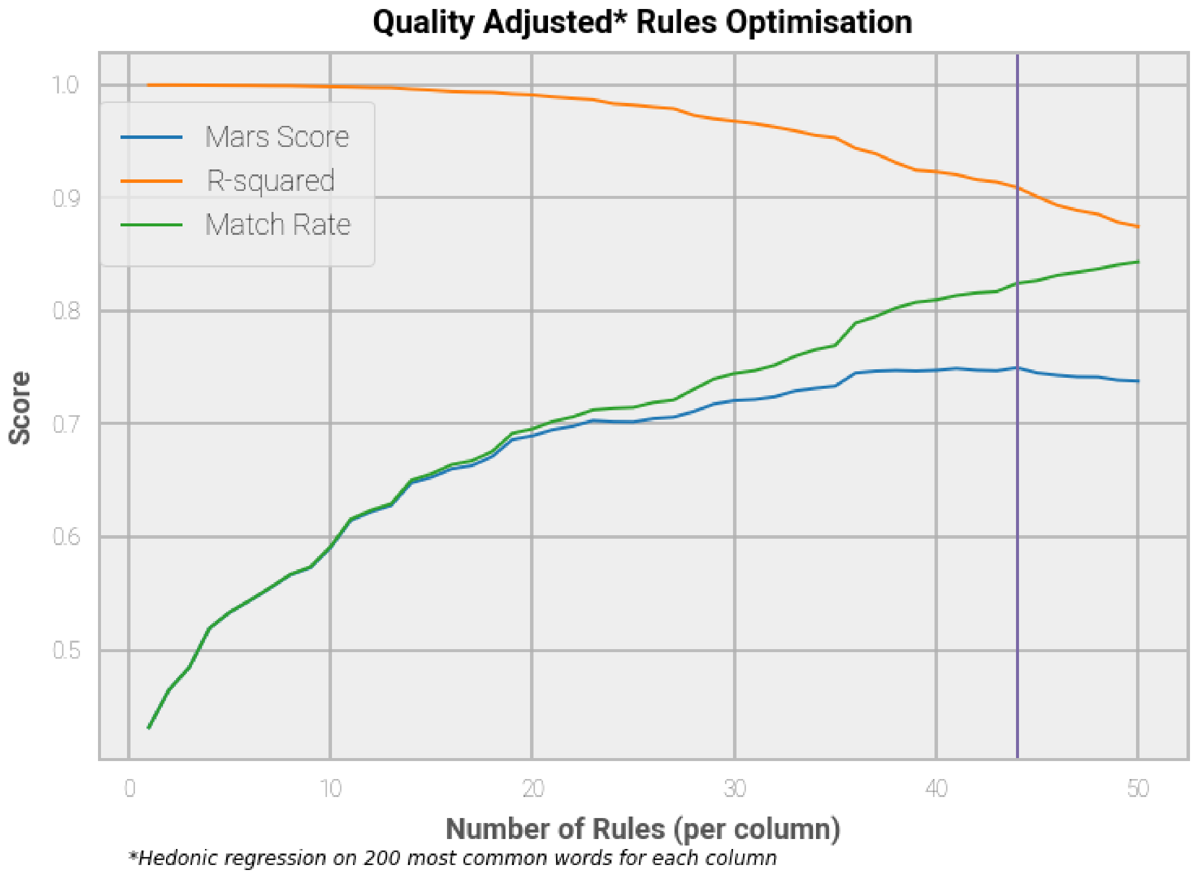
MARS scores for grouping with 20 rules from each column

We can also optimise the number of rules

- Basic approach

  Fixed number of rules from each column

- **Optimisation**
  - Find number of rules which maximises MARS score
  - Start with single group for each retailer and add one rule from each column in each step



MARS scores in the latest month for grouping with 1 to 50 rules from each column

The final output is a clothing price index

- We calculate price indices for each consumption segment and retailer
- These are aggregated up to get an online clothing market consumer price index
- Web-scraped data require more advanced index number and weighting methods

- Can read more about our ongoing research into index methods here:

New index number methods in consumer price statistics - Office for National Statistics

Thank you!

For more information please contact:

Laura.christen@ons.gov.uk
(Classification)

Ahmet.aydin@ons.gov.uk
(Product Grouping)