



Enhancing Size Ratios through Forecasting, Clustering, and Classification

Nicole Brown, Marks & Spencer - Enterprise Analytics

Table of contents

Table of contents	2
Chapter 1: Introduction	4
1.1 Outline the Business Problem	4
Chapter 2: Methodology and Justification	6
Chapter 3: Project Scope	8
3.1 Project Management	8
3.2 Scope for POC	9
3.3 Trade-offs and Future Enhancements	10
Chapter 4: Data Selection, Creation & Pre-processing	11
4.1 Data Architecture	11
4.2 Data Protection & Policies	12
4.3 Datasets	12
4.3.1 Sales, Returns, Stock & Allocations	12
4.3.2 Product Attributes	13
4.4 Feature store	13
4.4.1 Custom Features	13
4.4.2 Feature Selection	14
4.5 Initial Exploration	16
Chapter 5: Model Selection	17
5.1 Forecasting Unconstrained Sales	17
5.2 Size Ratio Baseline	18
5.3 Clustering Existing Products	18
5.3.1 Colour Clustering with LAB	18
5.3.2 Product Clustering	20
5.4 Classification of New Products	22
5.4.1 Random Forest Model	22
5.4.2 Logistic Regression Model	23
5.4.3 Gradient Boosting Machine (LightGBM)	23
5.4.4 Model Evaluation	23
Chapter 6: Results	24

6.1 Recommendations	24
6.2 Risks, Limitations, and Implications	25
Chapter 7: Conclusion	26
Chapter 8: Appendices	27
8.1 Declaration	27
8.2 Project Checklist Table	29
8.2.1 Pass Criteria	29
8.2.2 Distinction Criteria	32
8.3 Definitions & Abbreviations	34
8.4 Code, Development & Documentation	35
8.4.1 Data Processing & Pipeline	35
8.4.2 Experimentation	38
8.5 References	39
8.6 List of Tables & Figures	39

Chapter 1: Introduction

Marks & Spencer is a renowned British retailer specialising clothing, homeware, beauty, and food products. Operating in over 250 clothing stores, alongside an online business that processes over 500,000 orders weekly, we are committed to delivering quality and value to our customers.

As an analyst within the Data Enterprise team at Marks & Spencer, my role centres on identifying new problem spaces and developing data-driven solutions that support informed decision-making. I collaborate with stakeholders from various departments, including Clothing & Home, Foods, Online, and Retail. Working closely with Data Scientists, Data Engineers, ML Ops, and Product Managers. I focus on the creation and enhancement of data products, benchmarking performance, and effectively communicating insights to ensure they are understandable and actionable for all business users.

1.1 Outline the Business Problem

This project focuses on enhancing the accuracy of size ratios of clothing - an integral part of the buying process that determines the allocation of each size to stores and online. This problem, initially identified by buying teams, is critical for both existing and new products.



Figure 1 Product Lifecycle highlighting the buying process and Size Ratio Estimates.

As shown in Figure 1, the M&S buying process begins with planning at a range level using an internal tool known as Range Planner. Selected products are catalogued in Product Lifecycle Management (PLM) system, where articles (defined by product and colour) and eventually article variants (defined by product, colour, and size) are

specified, along with initial product attributes. Sales estimates are generated using another tool, SSI, followed by the Size Ratio Tool (SRT), which estimates the quantities of each size to be purchased for both stores and online.

The current Size Ratio Tool has several inefficiencies:

- Reliance on Product Hierarchy and Historical Sales Data: This data is often sparse and an inaccurate representation of demand.
- Manual Approach: Requires significant analysis from the buying teams, often manually selecting similar products to use.
- Performance for New Products: The high error of size ratio predictions for new products with a MAPE of 21.75% (see Section 5.2, Size Ratio Baseline).

Due to these inefficiencies, we face 9% in missed potential sales due to shortages in certain sizes and 10% unsold stock into markdown, leading to a max potential opportunity of £40M a year in additional profit.

The problem statement is: “How can we improve size ratio estimates to recover lost sales and reduce inefficiencies?”.

I initiated this project as part of my role within a squad dedicated to exploring new problem areas. This opportunity in improving the accuracy of size ratios was first identified by our Clothing & Home Development Team in 2019. Despite its significance, the project had not been pursued further due to resource constraints and conflicts with other ongoing data initiatives.

Recognising the potential impact to the business, I took the lead on this project, primarily working independently while collaborating with key stakeholders when required. My interactions spanned from Buying Systems Specialists to Data Scientists and ML Ops teams.

Chapter 2: Methodology and Justification

To address the challenges, I applied advanced data analysis and machine learning techniques to:

- Forecast actual demand: create a reliable view of demand (Unconstrained Sales), combining actual sales and time-series forecasted sales for article variants out of stock.
- Automate clustering and improve accuracy for existing products: Surpass the current accuracy of size ratio estimates through clustering based on size demand and attributes like colour and price range.
- Improve accuracy for new products: Classify new products into clusters, to increase accuracy of size ratio estimates, efficiency and reduce dependence on manual inputs.

Enhancing size ratio estimates will not only improve full-price sell-through but also enhance customer satisfaction by better fulfilling demand for various sizes. The application of machine learning techniques, including forecasting, clustering, and classification, have been instrumental in these improvements.

The analytical and ML methods used, and rationale includes:

- Unconstrained Sales Forecasting:
 - A straightforward time-series model was used due to the requirements of the dataset to have demand units per day.
 - As time-series are quick to build and implement, this also provided a swift solution for advancing to size ratio estimation.
 - Developed through collaboration with analysts across other data products. It's also scalable, ensuring uniform forecasting logic within the data team and aligning with broader objectives.
- Clustering Similar Products:
 - I chose to use an ML clustering technique over a rules-based solution due to the superior handling of complex patterns, product attributes, the ease to adapt to future changes and enhanced reliability.
 - Machine learning algorithms, especially clustering techniques, can efficiently handle vast amounts of data and a high number of product features. This scalability is essential as number of products and features expand over time.
- Classifying New Products:
 - Rather than have buyers select where new products should sit a ML classification model has better consistency, accuracy, and objectivity.
 - ML models can make classification decisions much faster than humans, enabling real-time decision-making.

When applying these methods, I assumed that historical sales patterns and specific product attributes are reliable predictors of future size distributions. For new

products or those with limited historical data, I presumed leveraging data from similar products would yield accurate predictions. Additionally, I assumed the historical sales data, stock information, and product attributes were both accurate and robust, providing a solid foundation.

Chapter 3: Project Scope

Given the complexity and extended lead times (up to 18 months) in the buying process, this report centres on developing an initial proof of concept (POC), validated through back testing rather than a live trial.

Due to constraints in computational resources, the scope of this POC was confined to data from 2022 onwards and focused solely on the Menswear BU. This demonstrated the solution's potential value, laying the groundwork for possible future expansion and ensuring optimal use of resources. The initial success of the POC paves the way for scaling the solution in 2025 across additional BUs.

Although no specific deadline was set by stakeholders for delivering the solution, I targeted completion by Peak 2023, concluding the final development sprint in January. Figure 2 provides a detailed outline of the project plan. I strategically scheduled project milestones to avoid busy phases. For example, I deliberately avoided model training and testing during markdown periods, which typically overlap with numerous Data Science trials.

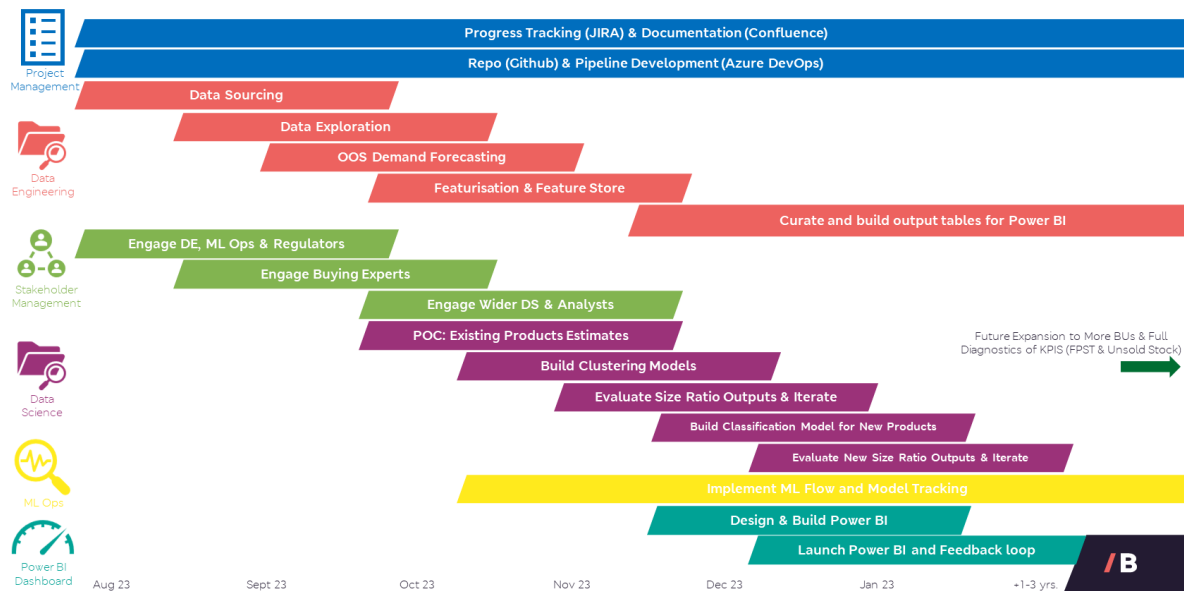


Figure 2 Project Plan showing POC, Model development and stakeholder management.

3.1 Project Management

I adopted an agile project management approach, prioritising the end-user needs, specifically the buying team. This approach was pivotal in defining the project's scope, addressing the inefficiencies in the current process, and allowed for iterative developments. This facilitated gradual scaling and refinement of models.

Work was organised into bi-weekly sprints, aligning with other data projects I worked on, using JIRA for task management and prioritisation based on shifting project demands. This flexibility was crucial given the variable time allocation week-to-week,

though it presented challenges in long-term planning, contrasting with the more structured timelines of a waterfall approach. To improve prioritisation, I held weekly review meetings with my manager.

My initial discovery involved value and feasibility analysis, determining the potential impact of the proposed ML solutions within existing systems. I collaborated with Laura, a non-technical buying process expert, who provided invaluable insights by explaining and identifying inefficiencies in the current system. Additionally, by quantifying unsold stock by size and lost sales due to out-of-stock sizes, I calculated the business impact of current inefficiencies. This ensured that solution was strategically directed towards enhancing business outcomes.

Advised by Data Scientists, I followed best practices in model development by initiating with Databricks development notebooks, following coding standards, and enforcing version control through GitHub. Each project stage, from data sourcing to performance evaluation, was documented on Confluence, linking to relevant JIRA tickets and GitHub repositories for traceability. Additionally, Luis, our ML Ops product owner, guided the adoption of innovative tools like ML Flow, enhancing model deployment strategies.

An agile and collaborative project management approach was well-suited to the dynamic environment of our team. It also ensured that all tasks were strategically targeted, adaptable, and transparent, leading to informed decision-making and improved project outcomes. However, the iterative nature of the sprints inadvertently led to scope creep, particularly as new features such as colour clustering were integrated. Furthermore, unforeseen priority tickets from other data products occasionally diverted resources, limiting the availability of key stakeholders like ML Ops and Data Engineers. This required me to take on additional responsibilities, complicating project management and extending project timelines beyond initial estimates.

3.2 Scope for POC

The business scope:

- Create a dataset, from 2022 onwards, calculating unconstrained sales, optimal size ratios and products attributes for all Menswear products launched in this period up to the first product markdown or sell through.
- Utilise clustering techniques to estimate size ratios for existing Menswear products.
- Develop classification methods to estimate size ratios for new Menswear products.
- Present these estimates in a user-friendly Power BI dashboard.

The technical scope:

- Use Pyspark in Databricks notebooks for data joining and transformation, monitoring includes checking article counts and null value checks.
- Establish and maintain a feature store within ML Flow.
- Train and monitor clustering and classification models in ML Flow.
- Maintain comprehensive documentation of models.

The measures of success for the project were:

- The technical and business scope achieved.
- Estimated outputs outperforming existing SRT estimates measured by MAPE (see Section 5.2) and time efficiency.

3.3 Trade-offs and Future Enhancements

Considering my project timeframe, the complexities in buying processes and the need to minimise computational and resource cost, the following trade-offs were made:

- The POC is limited to Menswear articles with data availability to ensure robust model training and outcomes, while keeping computational costs low. In future, this solution could be scaled include additional BUs.
- I opted not to include store-level size ratios in the initial scope, focusing instead on aggregated online and store level estimates. While store-specific estimates could provide deeper insights, integration complexities and time constraints led to this decision. However, incorporating store-level ratios could be considered as a future enhancement.
- Integration with existing tools like SSI was not considered. Instead, model outputs have been made accessible through a Power BI dashboard, offering buyers the flexibility to use the size ratio estimates as needed, alongside current tools.

Chapter 4: Data Selection, Creation & Pre-processing

4.1 Data Architecture

M&S operates a complex data ecosystem, with various data types including transactional, stock, customer information, product attributes and planning data from sources, including M&S App, transactional software, and planning systems like SSI. This data is stored in a cloud-based data warehouse, Microsoft Azure Data Lake Storage (ADLS), as shown in figure 3. ADLS is accessible via Pyspark on the Databricks platform and can be presented in a user-friendly manner through Power BI to stakeholders.

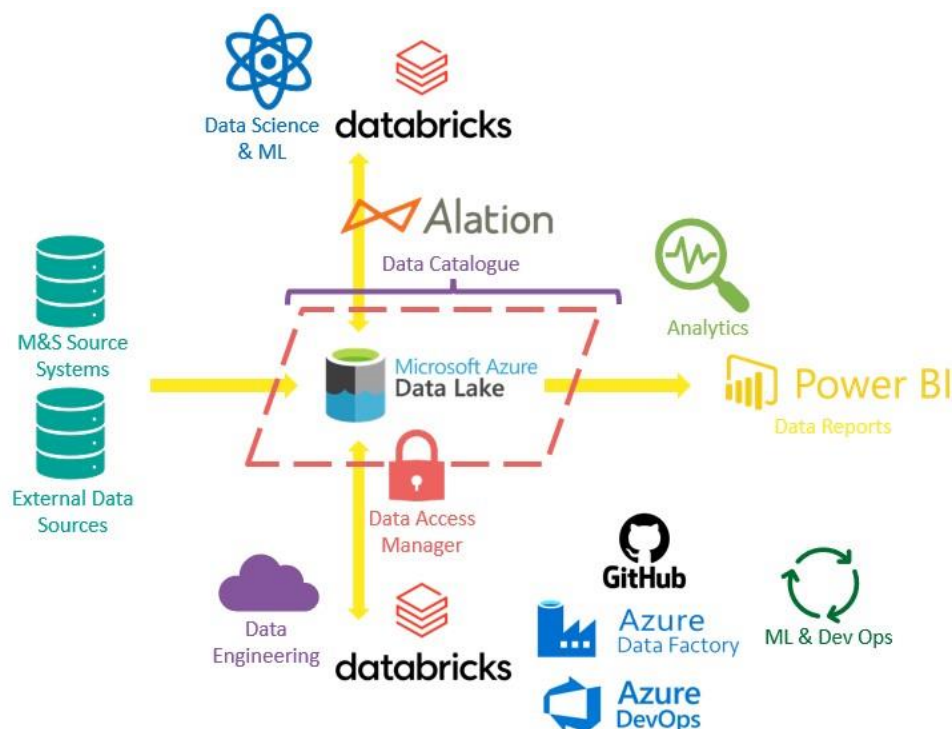


Figure 3 M&S Architecture: Flow of M&S source systems to ADLS that can be access in Databricks and presented in Power BI

Within this architecture, I have opted to use:

- Python: Chosen for its widespread use in data science, provides robust support through a rich library environment and an active community both at M&S and online, enhancing collaboration and troubleshooting.
- Databricks: Supports Python and interactive notebooks, integrates with GitHub, and offers an Apache Spark cluster for distributed computing. Selected for high-performance capabilities and its seamless integration with existing M&S systems, facilitating efficient data processing and analysis workflows.

- Power BI: Selected for its compatibility with M&S's data architecture, enabling the development of accessible reports & dashboards for end-users.

4.2 Data Protection & Policies

In my data selection process, I complied strictly to GDPR and M&S's Data Protection and Privacy Policy. This policy mandates compliance with all relevant data protection laws, ensuring personal data is secured, handled transparently, and anonymised where possible.

I specifically worked with non-sensitive, anonymous sales and returns data at the article variant and store level, avoiding any customer-specific data to mitigate re-identification risks. All tables and feature stores were designated as non-sensitive and named following recommended conventions. I documented the justification for each table's use, focusing on data minimisation.

All data processing was performed securely on M&S's internal platforms and via the company VPN. Data was stored in Azure Data Lake Storage (ADLS) and managed using Unity Catalogue, ensuring high standards of data security and integrity. This approach aligns with regulatory requirements and M&S's internal guidelines, reinforcing our commitment to responsible data management.

4.3 Datasets

For this project, I chose data based on recommendations from analysts experienced in this field and buying specialists familiar with the current process. The datasets include daily sales, returns, stock levels, and allocations for both retail and online channels, as well as product attributes.

My approach to data engineering was guided by best practices for code refactoring. This practice significantly improved code efficiency and facilitated the development of modular code structures, which are crucial for maintaining clarity and simplifying updates. To ensure the comprehensibility and maintainability of my work, I focused on thorough documentation. This involved writing clear comments and providing detailed descriptions within the code to clarify its purpose and functionality for future users. All created tables were carefully optimised with keys, such as the article identifier, to enhance performance and retrieval speed.

4.3.1 Sales, Returns, Stock & Allocations

Initially, I created a historical view of sales, returns, stock and allocations from 2022 onward. This historical data was crucial for forecasting unconstrained sales for out-of-stock article variants.

To compile this data, I used custom Pyspark functions (see Appendix 8.4.1) to extract and pre-process this data from strategic tables. I selected only essential columns

such as article variant, store key, date, allocation quantity, sales volume, returns volume, total stock, and markdown stock. The processing of this data involved restricting the date range from the first allocation (article launch) to the initial markdown or sell-through. Additionally, filtering to only include articles catalogued in at least 30 stores, achieved at least 100 lifetime sales, and had a sales history of at least three weeks. This ensured that the dataset used was robust and representative, improving the reliability and accuracy of forecasts.

4.3.2 Product Attributes

In the initial phase of the project, I conducted a review of the existing system (SRT) alongside an expert. This system was primarily based on hierarchical features to recommend product size ratios. Recognising the potential for improvement, I engaged with analysts and data scientists experienced in handling product attribute data. Their insights proved invaluable, not only in selecting appropriate features but also in shaping the overall strategy of the project.

I collated product attributes, again using Pyspark functions (see Appendix 8.4.1), selecting key columns from the product attribute table, such as:

- Product Hierarchical columns: BU, dept, l2, l4, l5, range, l8
- Season of launch (e.g., AU22)
- Price band
- Product lifecycle (year-round or seasonal)
- Colour name

Additionally, I incorporated advanced features from the Product Feature Repository (PFR), including fabric types and LAB colour features. During this process, I checked for NULL or missing values and decided to exclude any articles where these attributes were missing to ensure the integrity and accuracy of our dataset.

4.4 Feature store

4.4.1 Custom Features

I created several custom features to analyse sizes and unconstrained sales:

- size_vector: vector representing sizes offered by article.
- online_sum_total_sales: total unconstrained sales for online.
- retail_sum_total_sales: total unconstrained sales for retail.
- online_total_sales_vector: vector of unconstrained sales by size for online.
- retail_total_sales_vector: vector of unconstrained sales by size for retail.
- online_size_ratio_sales: vector of optimal size ratios for online.
- retail_size_ratio_sales: vector of optimal size ratios for retail.

(2) Spark Jobs

New result table: ON

Table	+								
article	size_vector	online_total_alloc_vector	online_total_sales_vector	1.2 online_sum_total_alloc	1.2 online_sum_total_sales	retail_total_alloc_vector	retail_total_sales_vector		
1	0000000000060357...	[157,154,35,86,70]	[122,85,8,26,21]	482	262	[125,106,43,83,58]	[85,97,53,57,43]		
2	0000000000060567...	[956,825,260,602,53,203,239,77]	[1127,1016,289,584,30,273,175,...	3095	3509	[3351,3309,1104,1041,0,656,114,0]	[3483,3241,994,2004,799,165,11		
3	0000000000060590...	[88,58,14,27,14,11,21,2]	[157,165,55,97,6,96,7,14]	216	597	[331,326,65,209,0,125,36,17]	[386,494,119,256,146,70,1]		
4	00000000000605763...	[53,47,21,50,34,15]	[54,45,14,54,32,10]	220	209	[534,412,120,338,154,31]	[621,503,154,438,228,57]		
5	00000000000605679...	[267,233,59,107]	[191,175,35,57]	666	458	[892,713,167,218]	[475,414,76,178]		
6	0000000000060558...	[70,70,15,65,30]	[39,44,6,47,18]	250	154	[381,362,122,214,131]	[266,248,95,168,103]		
7	00000000000605549...	[216,171,44,183,90,43]	[327,267,85,223,168,107]	747	1177	[2351,1459,368,1774,1027,496]	[2639,1647,458,2292,1355,789]		
8	00000000000606085...	[90,81,36,79,55]	[66,64,23,37,33]	341	223	[200,196,91,267,99]	[190,196,86,106,93]		
9	00000000000606366...	[264,336,96,120,24,24]	[156,141,54,73,43,26]	864	493	[717,757,372,425,257,202]	[415,480,160,242,89,42]		
10	00000000000606289...	[90,121,42,63,50]	[164,239,26,88,48]	366	565	[1514,1309,392,773,363]	[1463,1141,380,890,473]		
11	00000000000605502...	[240,187,61,183,21,64,72,12]	[252,255,69,161,2,23,29,1]	840	792	[1710,1473,470,1074,0,580,120,0]	[1801,1650,642,1177,614,82]		
12	00000000000605482...	[2177,1559,379,1445,60,803,453,...	[2599,1777,410,1682,5,752,223,...	7126	7473	[12885,9029,1750,8280,3,3507,626,8]	[13163,9218,1503,8735,11,3919,		
13	00000000000605291...	[84,70,17,55,22,12,0]	[83,46,8,63,20,19,14]	260	263	[147,107,15,99,62,38,2]	[106,88,14,78,63,34,0]		
14	00000000000605909...	[375,406,147,456,275,203]	[278,159,56,187,87,38]	2062	805	[444,450,164,248,132,77]	[504,475,213,352,215,177]		

Figure 4 Example of curated features.

Most of these features are stored as vectors, enhancing both efficiency and scalability. Additionally, I applied feature scaling techniques, such as log transforming sales data, to mitigate any overshadowing caused by scale disparities. This ensured a more balanced and fair comparison across different metrics.

During initial evaluation of features, I recognised the high number of vague subjective colour names that could introduce measurement bias. To mitigate this, I clustered colours based on LAB values (see section 5.3.1) consolidating these names into broader categories, reducing bias and generalisability for modelling.

4.4.2 Feature Selection

I evaluated possible features prior to any modelling activities. I included data testing with checks for nulls and errors. Novel attribute values were flagged, and products with minimal representation were excluded to maintain data consistency when new products are added.

I identified high error rates in some hierarchical attributes that proved to be redundant across 90% of products. Based on the risk of introducing reporting biases into the model I eliminated the redundant hierarchical features, to reduce potential biases and improve the accuracy of the model's predictions.

In my initial feature analysis, I utilised heatmaps to evaluate the correlation between categorical features and size distribution. This helped identify which features would be useful for clustering to estimate size ratios. For instance, the analysis revealed clear correlations within departments and fabric types, shown in figures 5 and 6. These correlations are likely a reflection of varying customer preferences based on size. Fabrics, for example, show distinct preferences in size distributions; certain materials like synthetics might be favoured in specific sizes due to attributes such as stretchability.

Additionally, colour trends indicated that lighter and brighter colours tend to be more popular in larger sizes. This could be linked to consumer perceptions influencing colour choices across different body types. Price band trends also provided insightful correlations; lower-priced items generally had higher sales in smaller sizes, possibly reflecting economic factors.

However, not all features demonstrated useful patterns. The analysis of seasonal trends and product lifecycle, for instance, did not show any significant correlation with size distribution and were excluded. I then processed the remaining categorical features using StringIndexer and OneHotEncoder, and manually mapped ordinal categories such as price band. Finally, I used MinMaxScaler instead of the StandardScaler to maintain the size proportionate boundaries (0 to 1).

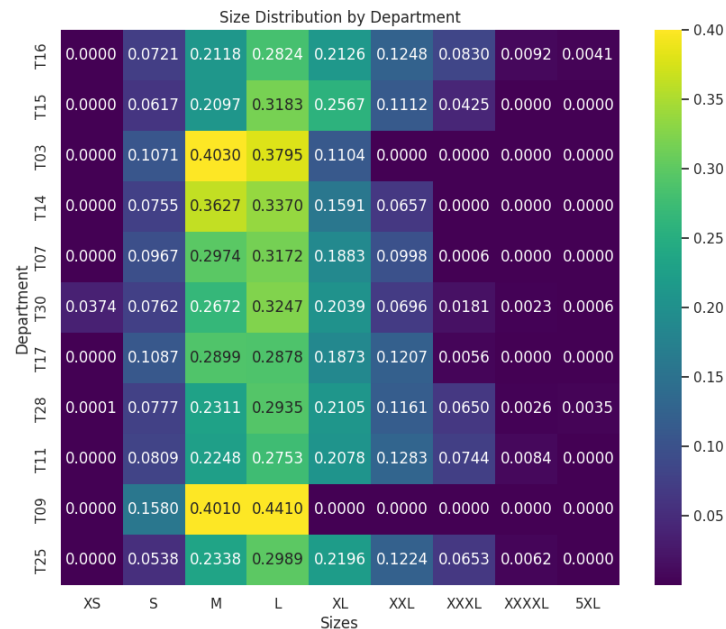


Figure 5 Heatmap of Department and Size Distribution.

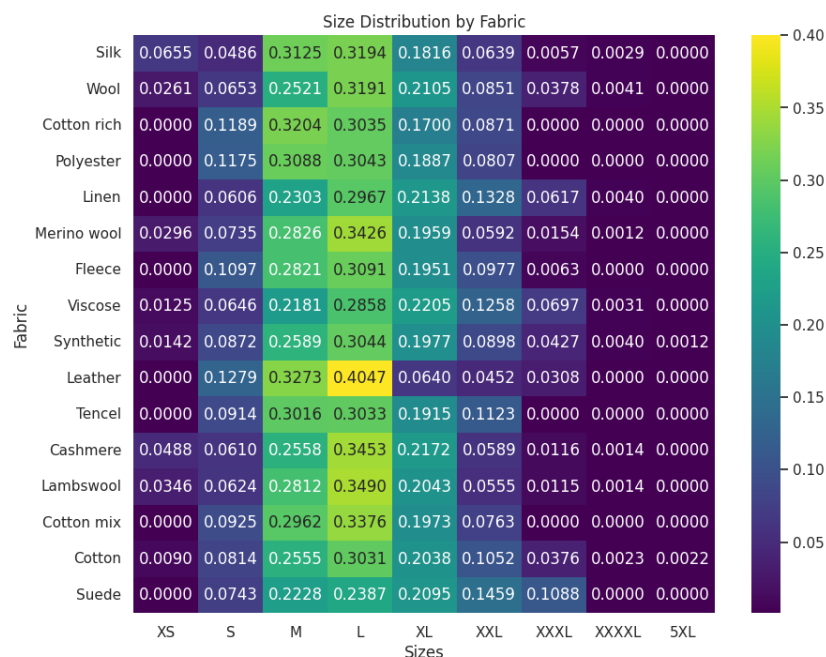


Figure 6 Heatmap of Fabric and Size Distribution.

4.5 Initial Exploration

Further analysis of menswear data revealed a distinct pattern of unsold stock accumulating in certain sizes, while other sizes sold out, leading to lost sales. For the example article in Figure 7, there is significant surplus in XXXXL, with 40% of stock remaining unsold, while other sizes like S experienced higher lost sales.

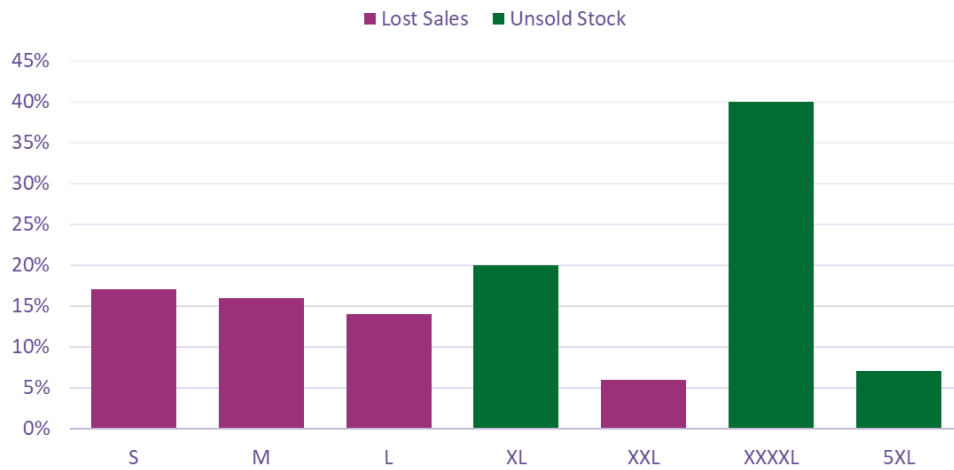


Figure 7 Unsold Stock and Lost Sales by Size for an example article.

To quantify the inefficiencies in size ratios for menswear, I calculated the current FPST % and Lost sales %, defined as:

$$FPST \% = \frac{SUM(Sales)}{SUM(Allocated\ Stock + Returns)}$$

$$Lost\ Sales\ \% = \frac{SUM(Lost\ Sales)}{SUM(Sales)}$$

Size Ratio	Lost Sales	FPST
Online	8%	70%
Retail	13%	81%
Overall	11%	79%

Figure 8 Lost Sales and FPST of Menswear.

By optimising size ratios in menswear, we could potentially recover 9% of lost sales. This recovery estimate is based on matching the unsold stock of sizes to the lost sales occurring in other sizes of each article. With optimal size ratios, I project an additional annual profit of £11 million.

Chapter 5: Model Selection

5.1 Forecasting Unconstrained Sales

To address the issue of uncaptured demand while OOS, I used a forecasting approach based on historic sales. Figure 9 illustrates this concept, showing how, over the lifecycle of an article variant, sales diminish due to stock shortages, resulting in lost sales. These combined represent the 'unconstrained sales'—an estimate of what could have been sold had stock been available throughout the lifecycle.

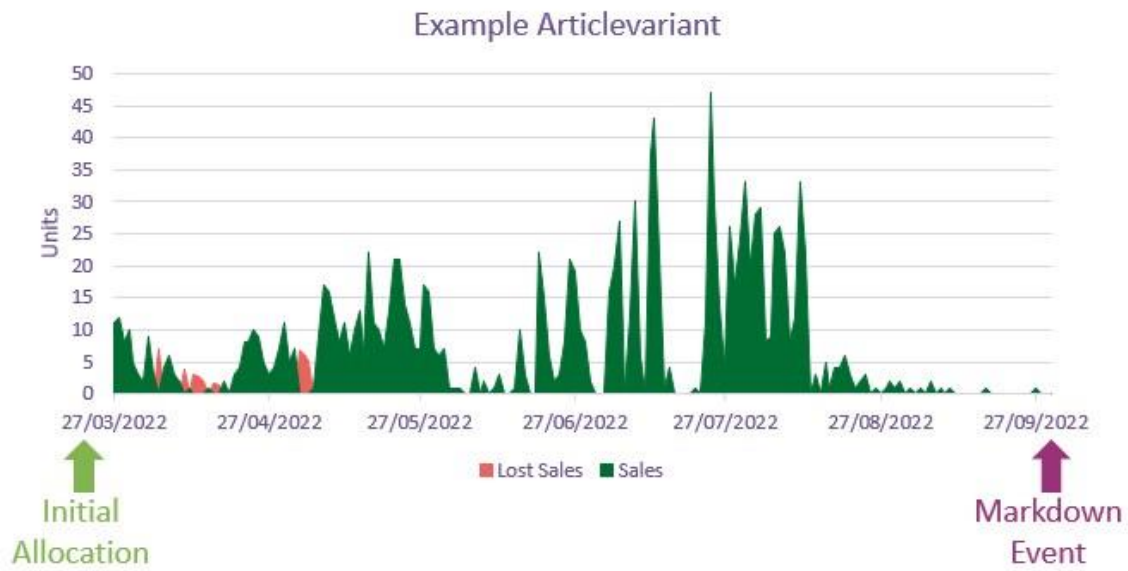


Figure 9 Unconstrained sales for an example variant.

I chose a moving Average Rate of Sale (ARS) model for its simplicity and ease of understanding for stakeholders. This choice also aligns with methodologies use by other data products, ensuring a consistent approach across our team. To assess each Simple Moving Average (SMA) model, I used key performance metrics such as Mean Absolute Percentage Error (MAPE) and Accuracy, applying a 0.1 threshold as detailed below:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Actual_i - Forecast_i}{Actual_i} \right|$$

$$Accuracy = \frac{\sum_{i=1}^n I((1 - threshold) \cdot a_i \leq p_i \leq (1 + threshold) \cdot a_i))}{n}$$

Here, I denotes the indicator function that is 1 if the condition is true and 0 otherwise, and n is the total number of observations in your dataset.

Forecasting Model	Accuracy	MAPE
SMA	34%	87%
SMA (stock)	54%	88%
SMA (10-day stock)	72%	77%
SMA (30-day stock)	52%	83%

Figure 10 Performance of unconstrained sales forecasts.

After evaluating each model variant, I decided to adopt the 10-day in stock moving average rate of sale, as the method for estimating demand in out-of-stock scenarios. This decision was based on the model's high accuracy, 72%, and its relatively lower MAPE of 77% compared to the alternatives. While more advanced time series method and further tuning might have given even better results, the high accuracy of this model suggests that it is sufficiently robust for our current purposes.

5.2 Size Ratio Baseline

Before selecting the optimal clustering and classification models, I conducted a review of the current size ratio estimates. This involved assessing the MAPE of size ratio estimates compared to demand, represented by unconstrained sales. This evaluation served as the baseline against which the performance of my models would be measured. Below is a summary of these findings:

Size Ratio	MAPE
Online	55.09%
Retail	17.87%
Overall	18.22%
Online New	68.38%
Retail New	21.08%
Overall New	21.75%

Figure 11 Baseline performance of current size estimates.

5.3 Clustering Existing Products

5.3.1 Colour Clustering with LAB

In exploring the use of colour names as a feature, I identified over 100 different colour names selected by individuals, which introduced a level of label bias. To address this, I undertook the additional task of creating new colour names by clustering products based on their LAB colour scale attributes. This approach was

inspired by Mouw (2018). LAB colour attributes for each product were already present in a well-maintained product feature table.

I began by visualising the colours in a 3D plot, which helped evaluate potential clusters. This visualisation was facilitated by converting LAB attributes to RGB using the `mpl_toolkits.mplot3d` and `skimage.color` packages.

I chose a straightforward k-means clustering model due to the task simplicity. I then employed the elbow method to determine the optimal number of clusters (K), selecting K=6 with squared Euclidean distance of 0.6. However, the 3D plots with displayed colours were the most significant indicators of successful clustering. As illustrated in Figure 13, clear clusters were established. Based on the centroid colours, I created the following colour name mapping: 0: "Lights", 1: "Darks", 2: "Green/Yellow", 3: "Red/Orange", 4: "Greys" and 5: "Blues". The new colour names were more accurate and unbiased.

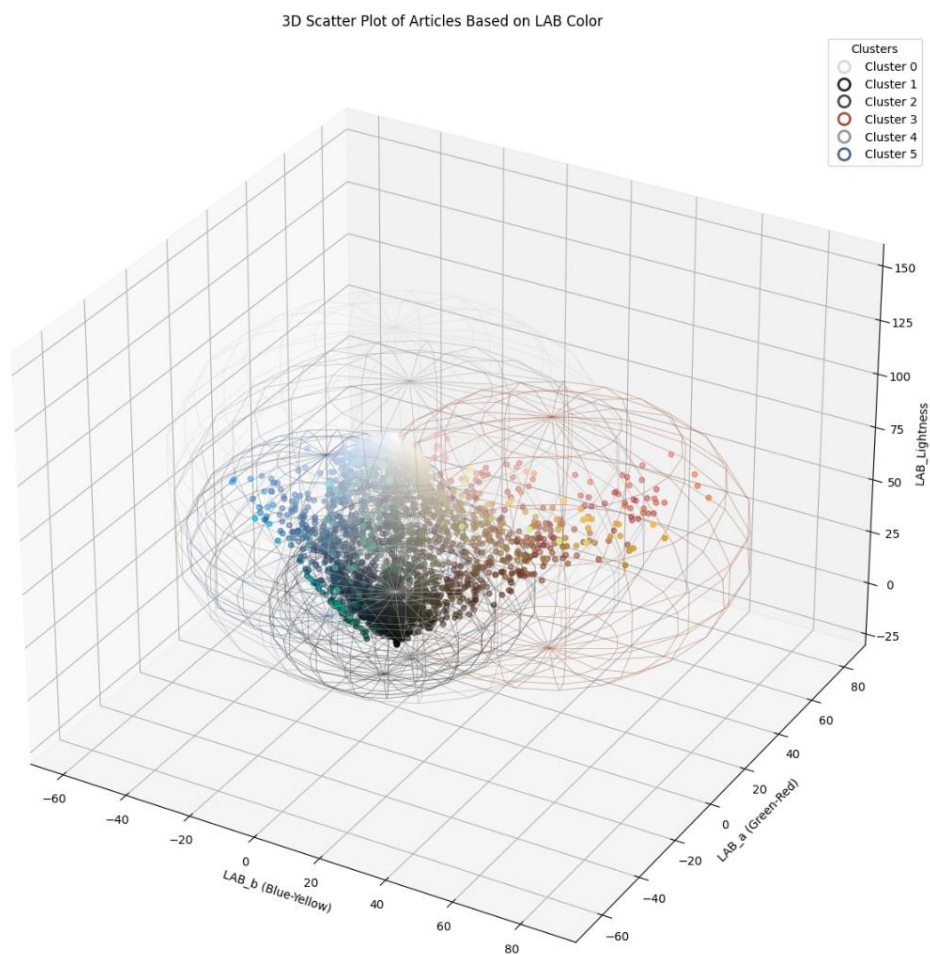


Figure 12 3D plot of clustered colours based on LAB colour values

5.3.2 Product Clustering

When selecting an appropriate clustering model, both transparency and specific data requirements were considered. The chosen technique needed to be straightforward to explain and visualise, computationally efficient, and scalable to accommodate any future expansions. Additionally, it had to effectively handle the high dimensionality and potential skewness in the data, possible from variables such as price band, sales category, and size ratios.

While Gaussian Mixture Models (GMM) offer more complex modelling of normally distributed data and adaptability to diverse cluster structures, they struggle with high-dimensional, skewed data and are inherently complex to explain. Similarly, DBSCAN excels in dealing with outliers without a predefined number of clusters. However, its performance diminishes with high-dimensional data, and it risks excluding articles by classifying them as outliers.

Given these considerations, I chose to explore K-Means and Bisecting K-Means. Both methods are not only suited for handling high-dimensional data but also accomplish simplicity, execution speed, and scalability. Throughout experimentation, I used ML flow to track runs, as shown in Appendix 8.4.1.

To optimise the number of clusters in both the K-means model and Bisecting K-Means, I employed the elbow method, supplemented by domain knowledge to ensure alignment with the Menswear BU. Considering the existing product groupings and the extensive range of menswear articles, which exceeds 1700, it was determined that the number of clusters should be no fewer than 10 and no more than 30. For Bisecting K-Means, the optimal number of clusters was identified as 12, shown in Figures 12. For K-Means, the optimal number of clusters was determined as 22.

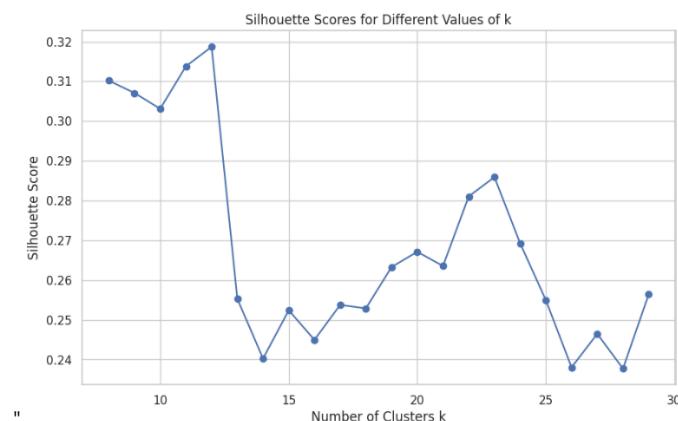


Figure 12 Elbow method for Bisecting K-Means.

The silhouette score confirmed both models did a relatively good job in creating clusters that are compact and well-separated from each other. K-means, 0.369, while not close to 1 is reasonably good and Bisecting K-means 0.319 is slightly lower.

Model	Size Ratio	MAPE
K-means	Online	18.68%
K-means	Retail	16.61%
K-means	Overall	13.08%
Bisecting K-Means.	Online	19.53%
Bisecting K-Means.	Retail	17.58%
Bisecting K-Means.	Overall	15.29%

Figure 13 Performance of clustered estimates by model.

Using the current size ratio estimates as a baseline, both models showed improvement, as detailed in Figure 13. However, the K-means model performed best, achieving a MAPE of 13.08%. Based on this result, I have chosen to proceed with the K-means model and will continue to monitor its performance using ML Flow.

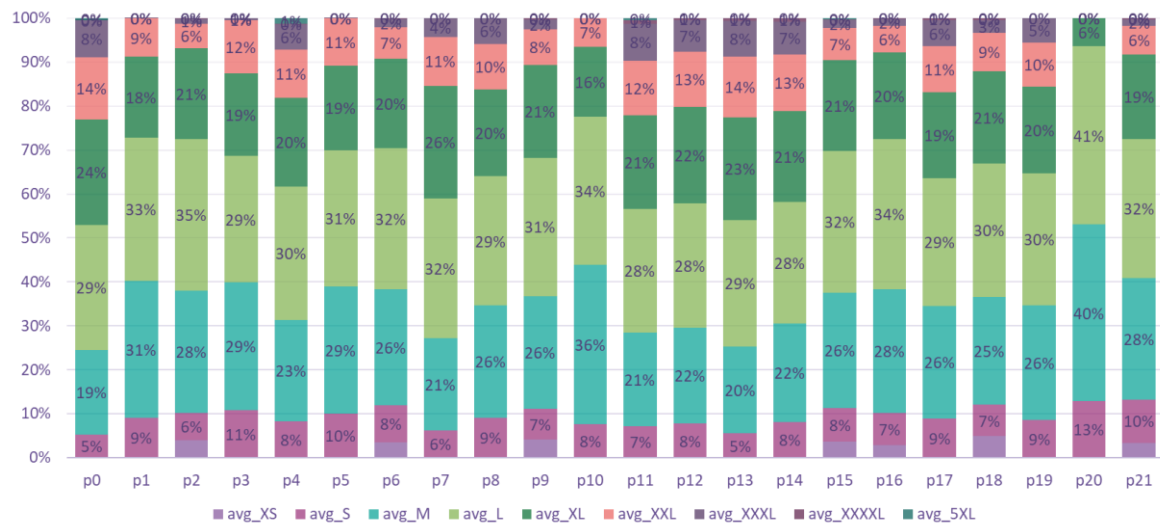


Figure 14 Average Size Ratios by K means clusters.

To maintain transparency in the clustering process, I consciously chose not to use dimensionality reduction techniques like PCA. Although these methods can effectively simplify complex datasets, they often obscure the direct relationship between the original features and the model's outputs. This decision ensured that the model remained easy to understand, facilitating easy analysis and clear visualisations in the final Power BI reports.

5.4 Classification of New Products

In classifying new products, I prioritised speed, high accuracy and low costs due to existing dependencies on cluster performance and need to prove value. Throughout the experiments, I used accuracy metrics from `pyspark.mlevaluation` to assess the effectiveness and reliability of various techniques. Additionally, I employed MLflow for streamlined experiment tracking and management, ensuring a robust, scalable, and efficient workflow (refer to Appendix 8.4).

5.4.1 Random Forest Model

Figure 15 shows accuracy of Random Forest Models against a simple Decision Tree:

Model	Accuracy
Decision Tree	0.397895
Random Forest	0.791579
Random Forest with 5-fold CrossValidator & ParamGrid	0.989474

Figure 15 Decision Tree and Random Forest Model Accuracy.

Advantages:

- Significant improved accuracy from 0.791579 to 0.989474 with tuning.
- Random Forest is robust to overfitting, using averaging.
- Handles categorical and ordinal data well which is effective with features like `price_band` and `sales_category`.
- Generally, requires less fine-tuning than models like GBM, easier to achieve good results.

Considerations:

- Slower performance with complex hyperparameter tuning such as 5-fold cross-validation and grid search.
- Poor for scaling due to possible large, cumbersome models.
- Can be outperformed by other models: Logistic Regression or when GBM is well-tuned.
- High computational cost, especially with large datasets and many trees.

5.4.2 Logistic Regression Model

Advantages:

- Runs quickly and efficiently, making it ideal for real-time predictions.
- Achieved a high initial accuracy of 0.993684, without the need for additional tuning.
- Validated generalisability with Cross-Validation.

Considerations:

- While the model's simplicity aids in transparency and ease of use, it may not capture complex patterns as effectively as more sophisticated models. This could limit performance with future scaling to more BUs.

5.4.3 Gradient Boosting Machine (LightGBM)

Advantages:

- Initial LightGBM models developed through experimentation in AutoML achieved high accuracy of 0.99, demonstrating high performance and speed.
- More accurate than Random Forest, especially when fine-tuned.
- Optimised for performance and ease for scaling, thanks to development using AutoML.

Considerations:

- The model is complex to understand and may incur higher computational costs in training.
- Susceptible to overfitting if the number of trees is excessive or the learning rate is not appropriately adjusted.
- Requires meticulous tuning of parameters, including the number of boosting stages, tree depth, and learning rate, which can be challenging.

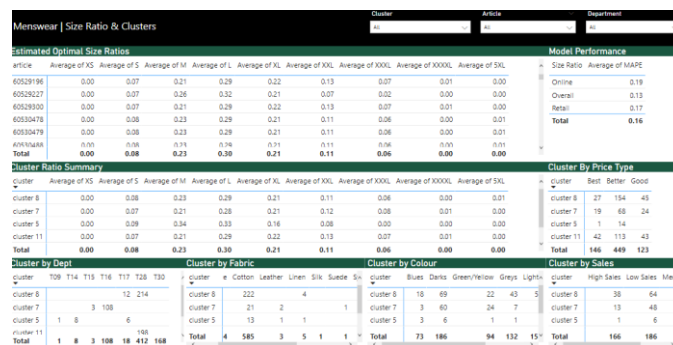
5.4.4 Model Evaluation

For my current project, where speed and efficiency are paramount, especially where making real-time decision-making, Logistic Regression is highly advantageous due to its rapid processing and impressive initial accuracy. Looking ahead, as the project scales to include more business units and integrates more complex features, a transition to a more advanced model such as LightGBM might become necessary. However, for this proof of concept, Logistic Regression fulfils the requirements. Regardless of the model chosen, ongoing monitoring will be essential to ensure its reliability and effectiveness in real-world scenarios.

Chapter 6: Results

The success of the POC was measured by its technical and business scope achievement and by outperforming existing benchmarks in terms of production speed and accuracy for size estimates of menswear. Notably, the MAPE improved from 18.22% to 13.08% (see Section 5.3.2), and the project introduced an automated, size estimation process without the need for manual inputs or grouping, indicating a moderate success of the POC.

The output, a Power BI dashboard (Figure 16) updated weekly via Databricks (Appendix 8.4), can provide the buying teams with actionable insights by presenting size estimates based on clustering similar products and demand instead of solely historical sales data. This new approach enhances accuracy and is expected to improve full-price sell-through rates and reduce excess stock.



article	Average of XS	Average of S	Average of M	Average of L	Average of XL	Average of XXL	Average of XXXL	Average of XXXXL	Average of SKL
60529196	0.00	0.07	0.21	0.29	0.22	0.13	0.07	0.01	0.00
60529227	0.00	0.07	0.26	0.32	0.21	0.07	0.02	0.00	0.00
60529300	0.00	0.07	0.21	0.29	0.22	0.13	0.07	0.01	0.00
60530478	0.00	0.08	0.23	0.29	0.21	0.11	0.06	0.00	0.01
60530479	0.00	0.08	0.23	0.29	0.21	0.11	0.06	0.00	0.01
60530484	0.00	0.08	0.23	0.29	0.21	0.11	0.06	0.00	0.01
Total	0.00	0.08	0.23	0.30	0.21	0.11	0.06	0.00	0.00

cluster	Average of XS	Average of S	Average of M	Average of L	Average of XL	Average of XXL	Average of XXXL	Average of XXXXL	Average of SKL
cluster 8	0.00	0.08	0.23	0.29	0.21	0.11	0.06	0.00	0.01
cluster 7	0.00	0.07	0.21	0.28	0.21	0.12	0.08	0.01	0.00
cluster 5	0.00	0.09	0.34	0.33	0.16	0.08	0.00	0.00	0.00
cluster 11	0.00	0.07	0.21	0.29	0.22	0.13	0.07	0.01	0.00
Total	0.00	0.08	0.23	0.30	0.21	0.11	0.06	0.00	0.00

cluster	Best	Better	Good
cluster 8	27	154	45
cluster 7	19	68	24
cluster 5	1	14	
cluster 11	42	113	43
Total	89	249	112

cluster	T08	T14	T15	T16	T17	T28	T30
cluster 8							
cluster 7							
cluster 5							
cluster 11							
Total							

cluster	Cotton	Leather	Linen	Silk	Suede	Synthetic
cluster 8						
cluster 7						
cluster 5						
cluster 11						
Total						

cluster	Blues	Dark	Green/Yellow	Greys	Light
cluster 8					
cluster 7					
cluster 5					
cluster 11					
Total					

cluster	High Sales	Low Sales	Medium Sales
cluster 8			
cluster 7			
cluster 5			
cluster 11			
Total			

Figure 16 Power BI presenting recommended Size Ratios.

6.1 Recommendations

To convey the potential POC value of recovered lost sales (see Section 4.3.1) to non-technical stakeholders, I used PowerPoint presentations, simplifying complex information and carefully explaining the logic behind it. I also took the time to address any concerns, ensuring clarity and confidence in the data presented.

For end-users, I present results through Power BI, selected for its ability to effectively communicate complex data, such as recommended size ratios and cluster details, in a user-friendly format. The report utilises tables and simple filters, enhancing ease of understanding and usability.

I recommended the adoption of this predictive model and dashboard as a strategic tool for informed decision-making. It's important to note that the dashboard includes most recent accuracy measures, providing transparency and ensuring that the buying teams can make decisions with a clear understanding of the underlying assumptions.

6.2 Risks, Limitations, and Implications

Risks & Limitations:

- The effectiveness of the solution may vary depending on how buyers use and interpret the Power BI recommendations. This emphasises the need to gather user feedback once a trial is live. The system is not fully automated, maintaining a necessary human element in decision-making to complement, not replace, the expertise of buying and merchandising teams.
- Significant changes in product attributes or hierarchy could impact the model's performance. While data testing for input tables helps monitor these changes to an extent, a fully automated monitoring and alert system through Azure pipeline could be beneficial.
- Current models may not scale effectively, requiring more advanced techniques. Additionally, estimations for fringe sizes (e.g., XS at 40.75% MAPE) were relatively poor, suggesting a need for future refinement or elimination strategies.
- Evaluating new product size ratio recommendations was challenging without a live trial to validate the solution.
- Enhancing size ratios is part of a complex ecosystem involving buying, forecasting, and allocation decisions. Effective communication of this interdependency is vital to set realistic expectations about the project's impact.

Implications:

- A significant success of the project was developing a reliable view of unconstrained sales, providing a forecasted demand for periods when products are OOS. This view is now used and continuously improved by other data teams, aligning key measures to present a unified version of the truth to the business.

Chapter 7: Conclusion


The POC has demonstrated significant opportunities, paving the way for extended applications in allocation strategies and size adjustments. The next step involves conducting a live trial to rigorously monitor improvements in FPST and reductions in lost sales, which will further validate the project's impact. Beyond mere cost savings and recovered sales, this initiative promises to significantly boost customer satisfaction by aligning product availability more precisely with demand.

As we move forward, iterative improvements and adaptability to evolving business strategies will be crucial for leveraging the full potential of this solution, particularly as we introduce new products.

Chapter 8: Appendices


8.1 Declaration

Apprentice Declaration

I declare this work is my own work and is completed as specified.	
Apprentice Title and Name: Nicole Brown, Data Analyst ULN:	Signature  Date: 30/05/2024
Project Start Date	
Week One (Time spent on Project)	Date: 25/04/2024 Number of hours spent on Project Report and Presentation: 16
Week Two (Time spent on Project)	Date: 02/05/2024 Number of hours spent on Project Report and Presentation: 16
Week Three (Time spent on Project)	Date: 09/05/2024 Number of hours spent on Project Report and Presentation: 16
Week Four (Time spent on Project)	Date: 16/05/2024 Number of hours spent on Project Report and Presentation: 16
Week Five (Time spent on Project)	Date: 23/05/2024 Number of hours spent on Project Report and Presentation: 16
Week Six (Time spent on Project)	Date: 30/05/2024 Number of hours spent on Project Report and Presentation: 16
Project completion date	30/05/2024

(This should be six weeks from start date of Project)	
Project and Presentation Submission Date	13/06/2024
Project Word Count (This must be 5000 words +/- 10%)	5485
Project Report Assessment Dates Please detail any times or dates over the next few weeks from submission of Project Report and Presentation that you would NOT be able to attend an assessment. This will help to avoid having to reschedule your assessment when its booked.	17 th -21 st June, 28 th June, 4 th -5 th July 9 th – 11 th July 1 st – 13 th August 29 th – 2 nd September

Observer(s) Employer Declaration

I declare that the above-named apprentice has been given sufficient time and resources to undertake the work-based project report and presentation.	
Company Name: Marks & Spencer Name of Observer: Kristian Buckstone Role: Lead Analyst	Signature  Date: 30/05/2024

Training Provider Declaration

I declare that following the completion of the project report and presentation by the above-named apprentice all required KSBs for this assessment method have been met. (eg ensuring the below table has been fully completed).	
Training Provider: Click here to enter text. Name: Click here to enter text. Role Click here to enter text.	Training Provider Signature: Date:

8.2 Project Checklist Table

8.2.1 Pass Criteria

Awareness of the opportunities of AI and data science to create business value and growth. (K13,K14)		
	Project Mapping	
AI and data science solution developed within the project addresses a business need in line with quality standards and timescales. The business value of a data product / solution and any constraints making trade-offs accordingly have been considered.		Chapter 1.1 Outline the Business Problem, pg4 Chapter 3.3 Trade-offs and Future Enhancements, pg10 (K13) (K14)
Critically evaluate the effectiveness and performance of proposed AI and data science solutions (K23, S3, S17)		
	Project Mapping	
Critically evaluates the performance of developed AI and machine models and the steps taken to mitigate sources of error and bias.		Chapter 5 Model Selection & Evaluation, pg17-23 Chapter 5.2 Size Ratio Baseline, pg18 Chapter 4.4 Feature Store, pg13-14 Chapter 5.3.1 Colour Clustering with LAB, pg18 Chapter 5.4.1 Random Forest Model, pg22 (S17) (K23) (S3)
Considers and selects from a range of appropriate principles, techniques and solutions to enhance the robustness of decisions at all stages.		Chapter 4.3 Datasets, pg12 Chapter 5 Model Selection & Evaluation, pg17-23 (S3)
Critically evaluates the arguments, assumptions, abstract concepts and data to make business focused recommendations.		Chapter 2 Methodology and Justification, pg6 Chapter 6 Results, pg24-25 (S3)

Demonstrates how, from the range of possible solutions presented, they contributed to identifying the optimal solution.	Chapter 3 Project Scope, pg8 – 9 (S3)
Explains how they implement data curation and data quality controls in line with organisational and regulatory requirements.	Chapter 4.2 Data Protection & Policies, pg12 Chapter 4.3 Datasets, pg12 (S17)
Apply systematic methodology and project management principles in the delivery of innovative, stable and robust solutions (S2, S9, S10, S22, S25)	
	Project Mapping
Selects and uses datasets, programming languages, tools and scientific methodologies to research business problems, providing a clear justification for their selection.	Chapter 4 Data Selection, Creation & Pre-processing, pg11- 16 Chapter 5 Model Selection & Evaluation, pg17- 23 (S10) (S22) (S2) (S9) (S25)
Analyses and critically evaluates test data and proposed solutions, considering current and future business requirements.	Chapter 3.3 Trade-offs and Future Enhancements, pg10 Chapter 4.4 Feature Store, pg13 - 14 Chapter 6.2 Risks, Limitations, and Implications, pg25 (S2)
Manipulates and analyses complex datasets and critically evaluates arguments, assumptions, abstract concepts and data (that may be incomplete) to make recommendations and to enable a business solution or range of solutions to be achieved.	Chapter 4 Data Selection, Creation & Pre-processing pg11- 16 (S2) (S9) (S10) (S22) (S25)

AI Project and Development Management (K6, S24)	
	Project Mapping
Correctly selects and applies development, research methodology and project	Chapter 3.1 Project Management, pg8 (K6) (S24)

management techniques to engage with customers and solve the business problem being addressed.	
Use of communication and influencing skills across teams (K28, S4, S5, S7, S27, B2, B6)	
	Project Mapping
Describes how they have worked with a range of technical and non-technical stakeholders adapting their approach successfully to meet their diverse needs.	Chapter 1.1 Outline the Business Problem, pg4 Chapter 6 Results, pg24-25 Chapter 4 Data Selection, Creation & Pre-processing, pg11 – 13 (S5) (B2) (K28) (S4) (S7) (B6)
Explains how to work autonomously and collaboratively with multidisciplinary teams indicating when each would be appropriate.	Chapter 1.1 Outline the Business Problem, pg4 Chapter 6 Results, pg24-25 Chapter 4 Data Selection, Creation & Pre-processing, pg11 – 13 (S5) (S7) (B2)
Describes how they have analysed information and data, using questioning and discussions with subject matter experts to scope new AI and data science requirements.	Chapter 4 Data Selection, Creation & Pre-processing pg11 – 16 (K28) (S4) (S5)
Written and verbal communication is clear, structured and appropriate for the audience.	Chapter 6 Results, pg24-25 Chapter 3.1 Project Management, pg8 (B6)
Explains how to work with software engineers to ensure suitable testing and documentation processes are implemented.	Chapter 1 Introduction, pg4 Chapter 4 Data Selection, Creation & Pre-processing, pg11 – 16 Chapter 3.1 Project Management, pg8 (S7) (B6) (S4)
Application of technical knowledge (K1, K3, K5, K26, S11, S15, S18)	
	Project Mapping
Describes how they applied appropriate scientific and	Chapter 5 Model Selection & Evaluation, pg17-23

technological methods for machine learning, AI and data science solutions, services and platforms to deliver business outcomes outlining successes and challenges.	Chapter 4 Data Selection, Creation & Pre-processing, pg11 – 16 Chapter 6 Results, pg24-25 (K5) (K1) (K3) (S11) (K26) (S15) (S18)
--	--

8.2.2 Distinction Criteria

Awareness of the opportunities of AI and data science to create business value and growth. (K13,K14)	
	Project Mapping
Articulates a commercial awareness of organisational priorities. Explains how the practical trade-offs in implementing an AI or data science solution for the particular business context have been addressed and shape the solution accordingly to optimise outcomes.	Chapter 1.1 Outline the Business Problem, pg4 Chapter 3 Project Scope p8 – 10 Chapter 5 Model selection & Evaluation p17 – 23 (K13) (K14)

Critically evaluate the effectiveness and performance of proposed AI and data science solutions (K23, S3, S17)	
	Project Mapping
Critically evaluates and adapts practice making recommendations for communicating technical methodology.	Chapter 6 Results, pg24-25 (K23) (S3)
Explains when they have effectively communicated technical information in a team context which has influenced others and impacted positively on decisions or working practices.	Chapter 6 Results, pg24-25 (S3) (S17)
AI Project and Development Management (K6, S24)	
	Project Mapping

Can evidence suitable methodology and tools have been selected with understanding of the impact of this choice on working practice, along with the risks to continuity of working practice that may arise if such solutions are not utilised.	Chapter 3.1 Project Management, pg8 Chapter 6 Results, pg24-25 (K6) (S24)
---	---

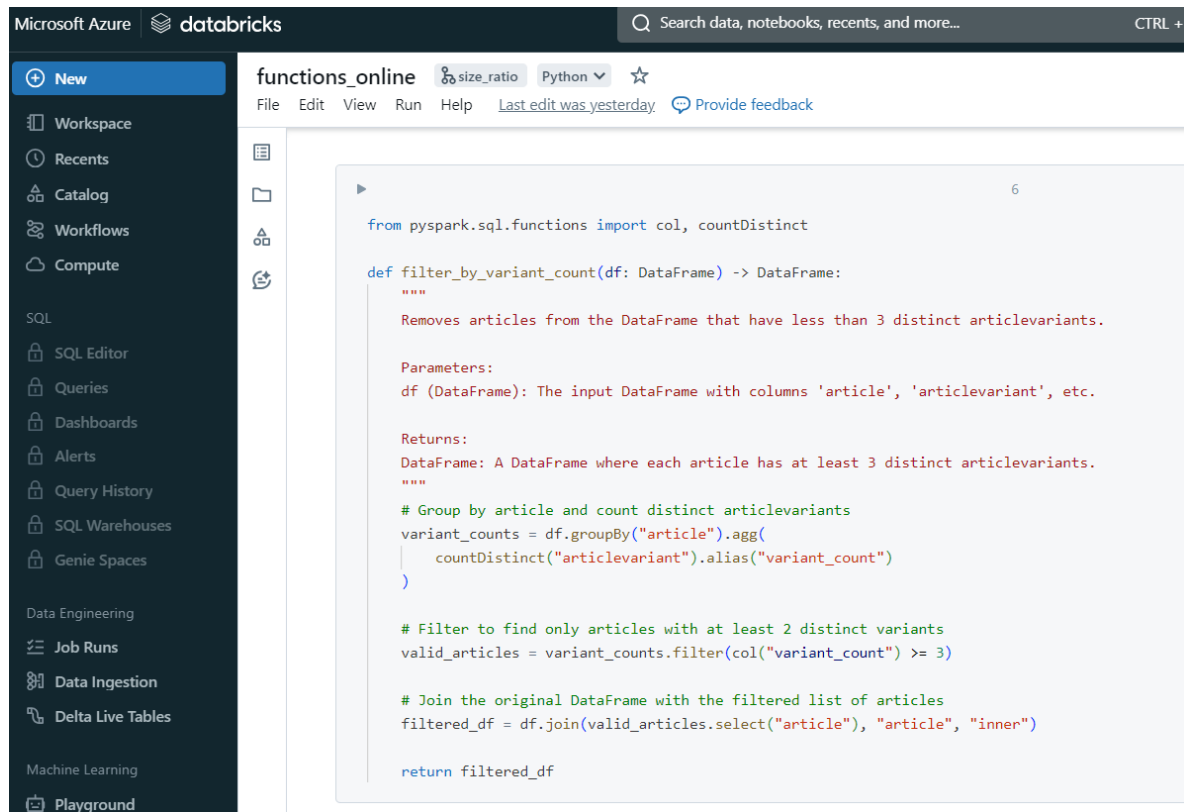
Use of communication and influencing skills across teams (K28, S4, S5, S7, S27, B2, B6)	
	Project Mapping
Explains how they adapted their approach with a range of technical and non-technical stakeholders and in different situations in order to achieve the best outcome for the business.	Chapter 3.1 Project Management, pg8 Chapter 6.1 Recommendations, pg24 (S5) (S7) (S27) (B6) (B2)
Evaluates solutions and explains the risks and implications of the AI data science requirements and alternative approaches and ways to address them.	Chapter 5 Model selection & Evaluation p17-25 Chapter 6 Results, pg24-25 Chapter 7 Conclusion, pg26 (B2) (S5)
Application of technical knowledge (K1, K3, K5, K26, S11, S15, S18)	
	Project Mapping
Explains the rationale for selecting particular technical solutions, including the relevant consideration of scientific benefit and suitability for working practices.	Chapter 5 Model selection & Evaluation p17 -23 Chapter 2 Methodology and Justification, pg6 (K1) (K3) (K26) (S11)
Appraises AI and/or Data solutions and explains the risks and implications of the process, alternative approaches and ways to address them.	Chapter 6 Results, pg24-25 (K5) (S15) (S18)

8.3 Definitions & Abbreviations

Term	Abbreviation	Definition
Clothing & Home	C&H	The name for our non-food business. Was previously known as GM or General Merchandise
Article		SAP terminology for a stroke / colour. Each colour within a stroke has a separate Generic Article in SAP. It is an 8-digit number
Article Variant		SAP terminology for below article which is stroke/ colour/ size level
Range		Level of hierarchy that sits directly below the sub dept - Allocation & Replenishment parameters are set at this level
Department	Dept	Level of hierarchy that sits directly below the BU
Business Unit	BU	AKA CBU and formerly called Business Group; Mens, Womens, Lingerie, Beauty, Kids, Home, Food
Range Planner		Tool for range plan analysis, ranking, catalogue planning and attribution
Product lifecycle management	PLM	Product lifecycle management software enables multidisciplinary teams to strategically collaborate with partners and customers using trusted, up-to-date product information
System Application and Product in Processing	SAP	Central system on which Master data for Article, Site and Vendor is held (it may originate from other sources though). Some data elements of those master files are maintained directly in this system
Sales, Stock, and Intake Tool	SSI	Detailed line level sales, stock and intake planning and forecasting tool. Increased visibility allows more effective trading of stock leading to reduced commitment and markdown. Issues POs to SAP
Purchase Order	PO	An M&S (SAP) document that authorises a purchase transaction
Size Ratio Tool	SRT/ESRT	Current tool used to help Merchandisers to create size ratios to estimate distribution of buy to stores and online
Full Price Sell Through	FPST	% of stock sold before markdown event (1 - unsold stock%)
Out of Stock	OOS	When products have sold through
Lost Sales		Demand for periods of OOS, what we could have sold if we had stock
Unconstrained Sales		Actual Sales and Lost Sales combined
Product Feature Repository	PFR	A maintained database of product features
LAB colour space	LAB	A three-axis colour system where L stands for lightness, and A and B are colour-opponent dimensions
RGB (red, green and blue)	RGB	A system representing the colours used on a digital display screen

8.4 Code, Development & Documentation

8.4.1 Data Processing & Pipeline



The screenshot displays the Databricks interface. At the top, the Microsoft Azure logo and 'databricks' branding are visible, along with a search bar and a 'CTRL +' shortcut. The left sidebar contains a navigation menu with options: New, Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Genie Spaces, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, and Playground. The main area shows a notebook titled 'functions_online' with a 'size_ratio' icon, a 'Python' language selector, and a star icon. Below the title are tabs for File, Edit, View, Run, and Help, along with a 'Last edit was yesterday' message and a 'Provide feedback' link. The notebook content is a Python function named 'filter_by_variant_count' that takes a DataFrame 'df' and returns a filtered DataFrame. The function uses PySpark SQL functions to group by 'article', count distinct 'articlevariant' values, filter for counts greater than or equal to 3, and then join the results back to the original DataFrame.

```
from pyspark.sql.functions import col, countDistinct

def filter_by_variant_count(df: DataFrame) -> DataFrame:
    """
    Removes articles from the DataFrame that have less than 3 distinct articlevariants.

    Parameters:
    df (DataFrame): The input DataFrame with columns 'article', 'articlevariant', etc.

    Returns:
    DataFrame: A DataFrame where each article has at least 3 distinct articlevariants.
    """
    # Group by article and count distinct articlevariants
    variant_counts = df.groupBy("article").agg(
        countDistinct("articlevariant").alias("variant_count")
    )

    # Filter to find only articles with at least 2 distinct variants
    valid_articles = variant_counts.filter(col("variant_count") >= 3)

    # Join the original DataFrame with the filtered list of articles
    filtered_df = df.join(valid_articles.select("article"), "article", "inner")

    return filtered_df
```

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Genie Spaces

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

online_table_data_processing

size_ratio

Python

File Edit View Run Help

Last edit was yesterday

Provide feedback

Yesterday (<1s)

1

%run `"./config"`

Yesterday (2s)

2

%run `"./functions_online"`

Yesterday (<1s)

3

cnh_online = aggregate_online_data(online_table)

cnh_online: pyspark.sql.dataframe.DataFrame = [date: date, articlevariant: string ... 3 more fields]

Yesterday (<1s)

4

product_df = join_data(cnh_online, product_table)

product_df: pyspark.sql.dataframe.DataFrame = [date: date, articlevariant: string ... 5 more fields]

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Genie Spaces

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Playground

Experiments

Features

Models

Serving

product_features_functions

size_ratio

Python

File Edit View Run Help

Last edit was 17 hours ago

Provide feedback

Workspace

notebooks

Sort: Name

bootstrap

classification

colour_clustering

colour_functions

config

feature_store

feature_store_functions

functions_online

functions_retail

online_table_data_processing

product_features

product_features_functions

retail_table_data_processing

unconstrained_sales

unconstrained_sales_functions

version

21 hours ago (<1s)

return joined_df

```

from pyspark.sql import DataFrame
from pyspark.sql.functions import col, array_sort

def drop_mismatched_size_vectors(df: DataFrame, col1: str, col2: str) -> DataFrame:
    """
    Returns a DataFrame with rows dropped where the specified array columns do not match.

    Parameters:
        df (DataFrame): The Input DataFrame.
        col1 (str): The name of the first array column to compare.
        col2 (str): The name of the second array column to compare.

    Returns:
        DataFrame: A DataFrame with rows where the array columns match.
    """
    # Add a column for checking if arrays are equal (sorted for order-independence)
    df = df.withColumn(
        "size_vector_equal",
        array_sort(col(col1)) == array_sort(col(col2))
    )

    # Filter the DataFrame to keep only rows where the size vectors are equal
    filtered_df = df.filter(col("size_vector_equal"))

    # Optionally drop the temporary column if you don't need it anymore
    filtered_df = filtered_df.drop("size_vector_equal")

    return filtered_df

```

```

from pyspark.sql import DataFrame
from pyspark.sql.functions import col, expr

```

Microsoft Azure databricks

Search data, notebooks, recents, and

New

- Workspace
- Recents
- Catalog
- Workflows
- Compute

SQL

- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses
- Genie Spaces

Data Engineering

- Job Runs
- Data Ingestion
- Delta Live Tables

Machine Learning

- Playground
- Experiments
- Features
- Models
- Serving

product_features size_ratio Python

File Edit View Run Help Last edit was 17 hours ago Provide feedback

Workspace

- notebooks
- Sort: Name
- bootstrap
- classification
- colour_clustering
- colour_functions
- config
- feature_store
- feature_store_functions
- functions_online
- functions_retail
- online_table_data_processing
- product_features
- product_features_functions
- retail_table_data_processing
- unconstrained_sales
- unconstrained_sales_functions
- version

17 hours ago (1)

```
%run "/config"
```

17 hours ago (<1)

```
%run "/product_features_functions"
```

18 hours ago (1)

```
joined_df = join_online_retail(size_ratio_online, size_ratio_retail, ["article", "bu", "dept", "corenewness"])
joined_df: pyspark.sql.dataframe.DataFrame = [article: string, bu: string ... 30 more fields]
```

18 hours ago (<1)

```
cleaned_df = drop_mismatched_size_vectors(joined_df, "online_size_vector", "retail_size_vector")
cleaned_df: pyspark.sql.dataframe.DataFrame = [article: string, bu: string ... 30 more fields]
```

18 hours ago (1)

```
product_attributes = process_products(product_table, cleaned_df)
product_attributes: pyspark.sql.dataframe.DataFrame = [article: string, bu: string ... 38 more fields]
```

17 hours ago (<1)

```
colour_attributes = join_colour_attributes(product_attributes, colour_predictions)
colour_attributes: pyspark.sql.dataframe.DataFrame = [article: string, bu: string ... 39 more fields]
```

17 hours ago (<1)

```
fabric_attributes = join_fabric_attributes(colour_attributes, fabric_table)
fabric_attributes: pyspark.sql.dataframe.DataFrame = [article: string, bu: string ... 42 more fields]
```

Workflows

Jobs Job runs Delta Live Tables

Filter jobs		All	Owined by me	Accessible by me	Favorites
Name	Tags	Created by		Trigger	
feature_run		Brown, Nicole		Scheduled	
feature_store		Brown, Nicole		Scheduled	
online_table_data_processing		Brown, Nicole		Scheduled	
product_features		Brown, Nicole		Scheduled	

Home

My workspace

Create

Browse

OneLake data hub

Apps

+ New

Upload

Databricks

{ "host": "northeurope.azuredatabricks.net", "httpP...

8.4.2 Experimentation

Microsoft Azure databricks

Experiments

Filter experiments

Name	Created by
clustering (2) /	nicole.brown@mnsCorp.net
clustering (1) /	nicole.brown@mnsCorp.net
clustering /	nicole.brown@mnsCorp.net
classification /	nicole.brown@mnsCorp.net
colour_clustering /	nicole.brown@mnsCorp.net
prediction_original_product_feature_predictions_ns_tbl-2023_12_19-15_10	nicole.brown@mnsCorp.net
Clustering K means - FINAL /	nicole.brown@mnsCorp.net
prediction_product_feature_predictions_ns_tbl-2023_12_19-14_28	nicole.brown@mnsCorp.net
prediction_product_feature_predictions_ns_tbl-2023_12_19-14_02	nicole.brown@mnsCorp.net
cluster_product_predictions_ns_tbl-2023_12_18-15_46	nicole.brown@mnsCorp.net
Classification /	nicole.brown@mnsCorp.net
Clustering K means /	nicole.brown@mnsCorp.net
OOS Sales /	nicole.brown@mnsCorp.net
Clustering /	nicole.brown@mnsCorp.net

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P dta-eun-lab-dbk-entr

AutoML Evaluation

Overview Warnings (5)

Training dataset: beam_prod.dataanalytics_azlab_prod.new_feature_store_class_ns_tbl
Target column: prediction
Evaluation metric: val_accuracy_score
Timeout: 120 minutes

AutoML Evaluation
All runs have completed, and have been added to the table below. Click a specific run to view details.

Model with best val_accuracy_score
The model is ready to be registered and deployed. Or, access the source code for the model training to make modifications by clicking a notebook under the [Source](#) column in the table below.

AutoML generated notebooks are now saved as MLflow artifacts. Click [here](#) to learn more.

Runs Evaluation **Preview** Traces **Preview**

Group by

	Run Name	Created	Dataset	Duration	Source	Models	Metrics	Tags
<input type="checkbox"/>	bedecked-eel-328	27 minutes ago	dataset (19a190a1) Train	33.1s	-	sklearn	val_accuracy_score: 0.993865030...	model_type: lightgbm_d...
<input type="checkbox"/>	suave-steed-810	29 minutes ago	dataset (19a190a1) Train	1.3min	-	sklearn	0.993865030...	xgboost_cla...
<input type="checkbox"/>	rogue-lynx-943	31 minutes ago	dataset (19a190a1) Train	31.4s	-	sklearn	0.993865030...	lightgbm_d...
<input type="checkbox"/>	luminous-hare-577	31 minutes ago	dataset (19a190a1) Train	39.1s	-	sklearn	0.993865030...	lightgbm_d...

8.5 References

Wilkie, T. (2020, February). Retail Analytics: Estimating Censored Demand. Retrieved from [https://www.lancaster.ac.uk/stor-i-student-sites/tessa-wilkie/wp-content/uploads/sites/14/2020/03/Wilkie_RT1.pdf]

Mouw, T. (2018, October 8). LAB Color Values / Color Spaces. X-Rite. Retrieved from [<https://www.xrite.com/blog/lab-color-space>]

8.6 List of Tables & Figures

Figure 1 Product Lifecycle highlighting the buying process and Size Ratio Estimates. .	4
Figure 2 Project Plan showing POC, Model development and stakeholder management.....	8
Figure 3 M&S Architecture: Flow of M&S source systems to ADLS that can be access in Databricks and presented in Power BI	11
Figure 4 Example of curated features.	14
Figure 5 Heatmap of Department and Size Distribution.	15
Figure 6 Heatmap of Fabric and Size Distribution.....	15
Figure 7 Unsold Stock and Lost Sales by Size for an example article.	16
Figure 8 Lost Sales and FPST of Menswear.....	16
Figure 9 Unconstrained sales for an example variant.	17
Figure 10 Performance of unconstrained sales forecasts.....	18
Figure 11 Baseline performance of current size estimates.	18
Figure 12 Elbow method for Bisecting K-Means.	20
Figure 13 Performance of clustered estimates by model.	21
Figure 14 Average Size Ratios by K means clusters.	21
Figure 15 Decision Tree and Random Forest Model Accuracy.....	22
Figure 16 Power BI presenting recommended Size Ratios.	24