FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Big Data Processing and Applications

RETAIL ANALYTICS WITH BIG DATA USING APACHE SPARK

SHAKIB POLOCK (2307673)

K H M BURHAN UDDIN (2307264)

AKASH SHINGHA BAPPY (2307938)

MITUN PAUL (2307325)

May 6, 2024

1 Introduction

Big data accounts for mammoth amounts of data both structured and unstructured sourced from diverse channels such as social media, sensors, mobile devices, and transaction records. Its significance in the contemporary era is growing exponentially due to several factors. The digital revolution has spurred an unprecedented surge in data generation, with platforms like social media and IoT devices contributing extensively [1]. This abundance of data necessitates robust strategies for effective utilization. Big data's capabilities are multifaceted, notably its capacity for high-speed analysis of large datasets. Advanced analytics techniques like machine learning enable real-time extraction of insights, facilitating tasks such as predictive analytics for businesses [2].

Big data holds immense potential for addressing various challenges across industries. In healthcare, for instance, it facilitates personalized medicine and predictive analytics, optimizing patient care and reducing costs [3]. Similarly, in finance, big data analytics aids in fraud detection, credit risk assessment, and investment strategy optimization [4]. The benefits of leveraging big data extend to informed decision-making, operational efficiency enhancement, and innovation stimulation. By relying on data-driven insights, organizations can optimize strategies, allocate resources effectively, and improve overall performance. Moreover, big data streamlines processes, identifies bottlenecks, and fuels innovation by uncovering market trends and opportunities [5].

In the retail sector, big data encompasses a plethora of transactional sales data gathered from different sources like social media interactions, online transactions, customer feedback, and supply chain operations. Its indispensability for supermarkets and grocery stores stems from the inadequacy of traditional data processing methods to handle the sheer volume of data generated. Big data technologies enable real-time storage, processing, and analysis of massive datasets, offering actionable insights at scale. Predictive analytics plays a pivotal role, leveraging historical sales data and other factors to accurately forecast demand and tailor inventory and promotions accordingly [6].

Moreover, big data facilitates personalized marketing strategies by analyzing past purchases, browsing patterns, and demographic information to offer targeted promotions and product recommendations. It also optimizes supply chain operations by identifying inefficiencies and streamlining logistical processes based on supplier performance, inventory levels, and transportation routes. Overall, big data revolutionizes retail operations, enhancing efficiency, customer engagement, and profitability.

2 Related Works

Use cases for big data tools and applications in retail has attracted considerable attention, driven by the rise of connected devices and mobile apps. Current literature and expert insights highlight key areas of focus in retail logistics [7]. Granular sales data improves choices about availability and assortment, while historical sales data and loyalty programs provide insightful customer information for operational planning. External data sources like competitor prices and weather patterns offer opportunities for demand forecasts and pricing strategies.

Despite its advantages, the adoption of big data in retail encounters obstacles including lack of expertise, difficulties in IT integration, and managerial reservations about data sharing. To overcome these issues, a suggested data maturity profile for retail enterprises has been developed to direct future research and address these challenges. [8]. The literature emphasizes big data's transformative potential in assortment planning, pricing strategies, and store layout

design to enhance operational efficiencies. Micro-segmentation of customers for assortment optimization and dynamic pricing based on real-time analysis show promise. The integration of e-tailing and multi-channel fulfillment presents both opportunities and logistical complexities navigable through advanced data analytics [9].

However, implementing big data solutions in retail operations isn't without risks. Privacy concerns, data credibility issues, and a shortage of analytical talent pose significant challenges. Cultural shifts within organizations towards data-driven decision-making processes are necessary [10]. A study on big data analytics in online retail, utilizing a Hadoop-based framework for managing large e-commerce datasets, is highlighted [11]. Key technologies such as HDFS, Sqoop, MySQL, HBase, Pig, Hive, and Spark are employed to address business queries like product analysis and financial performance evaluation, showcasing big data's transformative impact on retail decision-making.

3 Project Description

The retail industry is undergoing a significant transformation due to advance technology and changing consumer preferences. In this digital age, retail chains are generating vast amounts of data encompassing customer behaviour, purchasing patterns, and market dynamics. Leveraging this data effectively holds the key to unlocking competitive advantages and driving sustainable growth in the retail sector. In our project, we want to explore how big data tools can help in retail chains to generate insight form customer data, aiming to derive descriptive and predictive analysis that can reshape the very dimension of the superstore industry. With the advancement of big data technologies, retailers now possess an unprecedented opportunity to tap into the wealth of information surrounding consumer behaviour, purchasing trends, and market dynamics. Our overarching goal is to navigate this vast landscape, translating raw data into actionable intelligence that drives strategic decision-making and propels business growth.

The motivation behind this project arises from the recognition of the immense potential that big data analytics holds for revolutionizing retail operations. With the exponential growth of data sources and the availability of sophisticated analytical tools, retailers are presented with a huge opportunity to gain deep insights into their business processes and consumer interactions. Retailers may stay ahead of the competition in today's highly competitive market landscape by utilizing these big data technologies to enhance customer experiences, customize advertising efforts, as well as maximize inventory control.

For the given dataset, we start exploration of the dataset through comprehensive exploratory data analysis. Where we will investigate the various patterns and trends in the data. For example, profit and sales performances, product performance, month on month sales trend, top performing product categories and individual products, profit and sales that are deriving from products and so on. This analysis will also help to find rooms for improvements from the business perspective and will support the decision makers in taking actions. Additionally, we will provide predictive analytics using PySpark and MLlib. This will entail implementing machine learning models for inventory stock prediction in a supermarket, focusing on predicting stock levels for certain food categories on specific days. After that we will evaluate the results of the trained model with proper visualizations.

4 Data Description

The dataset for our analysis comes from the "Retail Analytics Trends" [12] project on Kaggle, which is publicly accessible under the CC0: Public Domain license. The dataset encompasses

comprehensive product information from leading UK supermarkets, such as Aldi, ASDA, Morrisons, Sainsbury's, and Tesco, providing a rich source for market trend analysis. It comprises over 2.1 million records, organized into different files corresponding to the aforementioned supermarkets. The total file size is approximately 182.3 MB, with the distribution of data varying across each supermarket. This dataset offers a diverse range of product-related information, including pricing, product names, and categories.

File Name	File Size (MB)	Row Count
Aldi	9.20	104.06 K
ASDA	46.09	538.74 K
Sains	54.11	591.41 K
Morrisons	31.44	380.93 K
Tesco	41.46	489.21 K
Total	**182.30**	**2.10 M**

Table 1: File size and row count from dataset files

The data includes a temporal element, capturing sales information across various dates. The dataset provides valuable insights into pricing across different supermarkets. The lowest recorded price is £0.01, while the highest is £479.99. The average product price is around £5.23, with a median of £2.99 and a mode of £2.00. The standard deviation, reflecting price variability, is approximately £7.56. Table 2 provides a detailed schema outlining their corresponding descriptions.

Table 2: Data Schema

Field Name	Description	Type	Example
Store_name	Name of the supershop	String	Aldi
Price	Price of product	Float	2.30 (£)
Price_unit	Price per unit	Float	5.50
Measure_unit	Measurement unit of sold product	String	Kg
Product_name	Name of sold products	String	Xbox Series X Console
Date	Date of sale	Date	2023-11-01
Category	Product category	String	Frozen
Own_brand	Whether the product is an own brand	Boolean	Yes

5 Methodologies and Tools

Initially, the dataset consisted of five large CSV files containing data for different supermarkets. These files were read through Spark and merged into a single DataFrame. To address the machine learning problem, the DataFrame was aggregated using Spark SQL. Subsequently, feature engineering was applied to the aggregated data. For model training, a Decision Tree Regressor from PySpark MLlib was used, and visualizations of the actual and predicted results were provided.

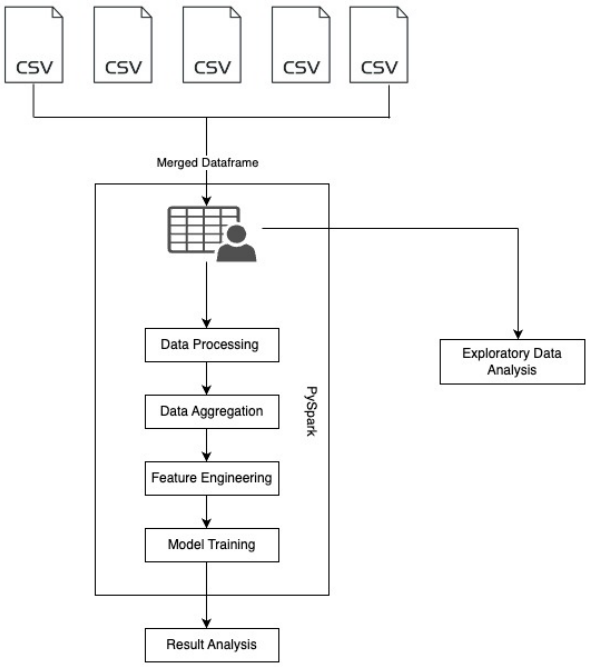


Figure 1: Workflow

5.1 Data Processing

The dataset was almost clean; however, a few columns didn't serve our machine learning purpose, so we dropped them for faster processing. Moreover, the date column was in string format, so we cast it as a date type and transformed it into the 'dd-mm-yyyy' format using PySpark.

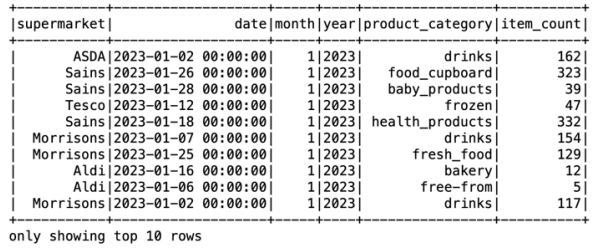
5.2 Data Aggregation

The dataset contained transactional data of customers checking out from the store, with each data point representing a single product entry into the system. To predict the amount of product needed to be stocked on a particular day of the month for a specific category in a specific supermarket, we aggregated the data based on the supermarket, date, month, and product category, counting each product category on a specific date. The count represents the amount of stock that needs to be preserved within a particular day of the month in a supermarket's product category.

Below are excerpts highlighting the schema, aggregated data, and summary for the machine learning case. In this context, we are considering the item count as the predictive variable.

```
root
 |-- supermarket: string (nullable = true)
 |-- date: date (nullable = true)
 |-- month: integer (nullable = true)
 |-- year: integer (nullable = true)
 |-- product_category: string (nullable = true)
 |-- item_count: integer (nullable = true)
```

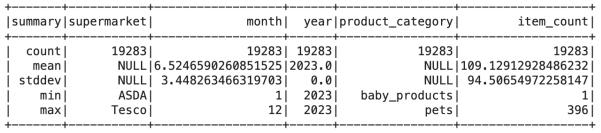
Figure 2: Schema for aggregated data



supermarket	date	month	year	product_category	item_count
ASDA	2023-01-02 00:00:00	1	2023	drinks	162
Sains	2023-01-26 00:00:00	1	2023	food_cupboard	323
Sains	2023-01-28 00:00:00	1	2023	baby_products	39
Tesco	2023-01-12 00:00:00	1	2023	frozen	47
Sains	2023-01-18 00:00:00	1	2023	health_products	332
Morrisons	2023-01-07 00:00:00	1	2023	drinks	154
Morrisons	2023-01-25 00:00:00	1	2023	fresh_food	129
Aldi	2023-01-16 00:00:00	1	2023	bakery	12
Aldi	2023-01-06 00:00:00	1	2023	free-from	5
Morrisons	2023-01-02 00:00:00	1	2023	drinks	117

only showing top 10 rows

Figure 3: Aggregated data for predicting stocks



summary	supermarket	month	year	product_category	item_count
count	19283	19283	19283	19283	19283
mean	NULL	6.5246590260851525	2023.0	NULL	109.12912928486232
stddev	NULL	3.448263466319703	0.0	NULL	94.50654972258147
min	ASDA	1	2023	baby_products	1
max	Tesco	12	2023	pets	396

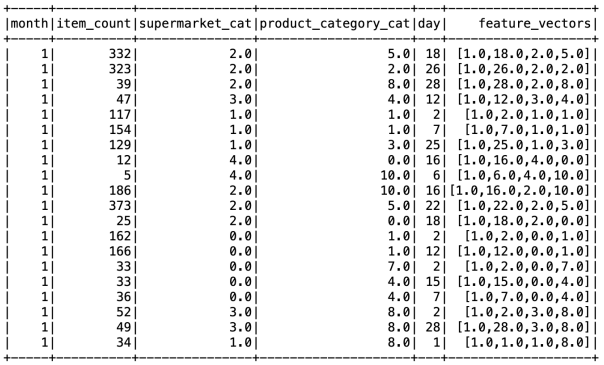
Figure 4: Summary of the aggregated data

5.3 Feature Engineering

The dataset contained two columns, *supermarket* and *product_category*, which included categorical data. These columns were encoded into numerical values using PySpark's *StringIndexer* function. Additionally, a feature named *day* was created from the *dd-mm-yyyy* date column by extracting the day (*dd*). The year was not considered, as the data pertained to a single year.

To make the data suitable for model training, a *VectorAssembler* was applied to create a vector space incorporating all features, resulting in a column named *feature_vectors*.

Below is an example of the transformed data in PySpark.



month	item_count	supermarket_cat	product_category_cat	day	feature_vectors
1	332	2.0	5.0	18	[1.0,18.0,2.0,5.0]
1	323	2.0	2.0	26	[1.0,26.0,2.0,2.0]
1	39	2.0	8.0	28	[1.0,28.0,2.0,8.0]
1	47	3.0	4.0	12	[1.0,12.0,3.0,4.0]
1	117	1.0	1.0	2	[1.0,2.0,1.0,1.0]
1	154	1.0	1.0	7	[1.0,7.0,1.0,1.0]
1	129	1.0	3.0	25	[1.0,25.0,1.0,3.0]
1	12	4.0	0.0	16	[1.0,16.0,4.0,0.0]
1	5	4.0	10.0	6	[1.0,6.0,4.0,10.0]
1	186	2.0	10.0	16	[1.0,16.0,2.0,10.0]
1	373	2.0	5.0	22	[1.0,22.0,2.0,5.0]
1	25	2.0	0.0	18	[1.0,18.0,2.0,0.0]
1	162	0.0	1.0	2	[1.0,2.0,0.0,1.0]
1	166	0.0	1.0	12	[1.0,12.0,0.0,1.0]
1	33	0.0	7.0	2	[1.0,2.0,0.0,7.0]
1	33	0.0	4.0	15	[1.0,15.0,0.0,4.0]
1	36	0.0	4.0	7	[1.0,7.0,0.0,4.0]
1	52	3.0	8.0	2	[1.0,2.0,3.0,8.0]
1	49	3.0	8.0	28	[1.0,28.0,3.0,8.0]
1	34	1.0	8.0	1	[1.0,1.0,1.0,8.0]

Figure 5: Transformed dataset after feature engineering

5.4 Model Training

Given that this is a regression problem, we applied a Decision Tree regressor from PySpark MLlib. While other models are available in MLlib, we selected this model to understand the implementation of machine learning in a big data context. The data was split into 80% for training and 20% for testing. As a parameter for the tree, we used 'maxDepth = 6' to prevent the tree from growing excessively, which could lead to overfitting. Exploratory data analysis has described in next chapter.

6 Data Analysis

We utilized PySpark to aggregate a dataset, generating key insights and transforming raw data into meaningful information that elucidates sales distributions and market dynamics within the supermarket sector. After preparing the final dataset, we used pandas, Matplotlib, and Seaborn to create visualizations. This big data processing and analysis allowed us to effectively summarize and visualize crucial business metrics.

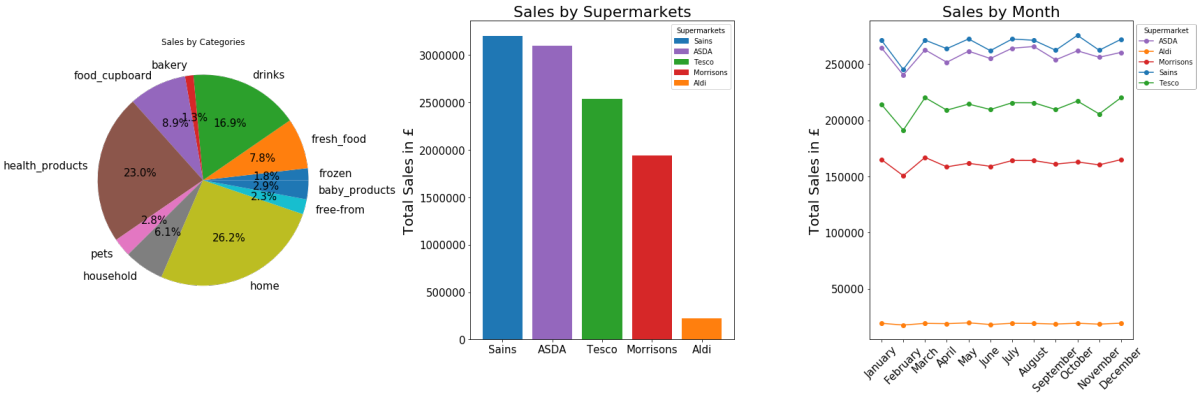


Figure 6: Sales Analysis

The resulting visualizations distinctively highlight supermarket sales. The pie chart shows that health products and home goods are the most popular, commanding approximately 23% and 26.2% of the market respectively. Drinks also make a significant contribution, while bakery items have the smallest share of 1.3%. In the bar chart, Sains leads with total sales of approximately 3.20 million, closely followed by ASDA at around 3.10 million, with Aldi at the lower end with about 0.25 million. The line chart of monthly sales trends reveals consistent performance from ASDA and Sains, peaking in December, indicative of seasonal shopping increases. Conversely, February shows a dip across all the markets, underscoring fluctuating consumer spending habits 6.

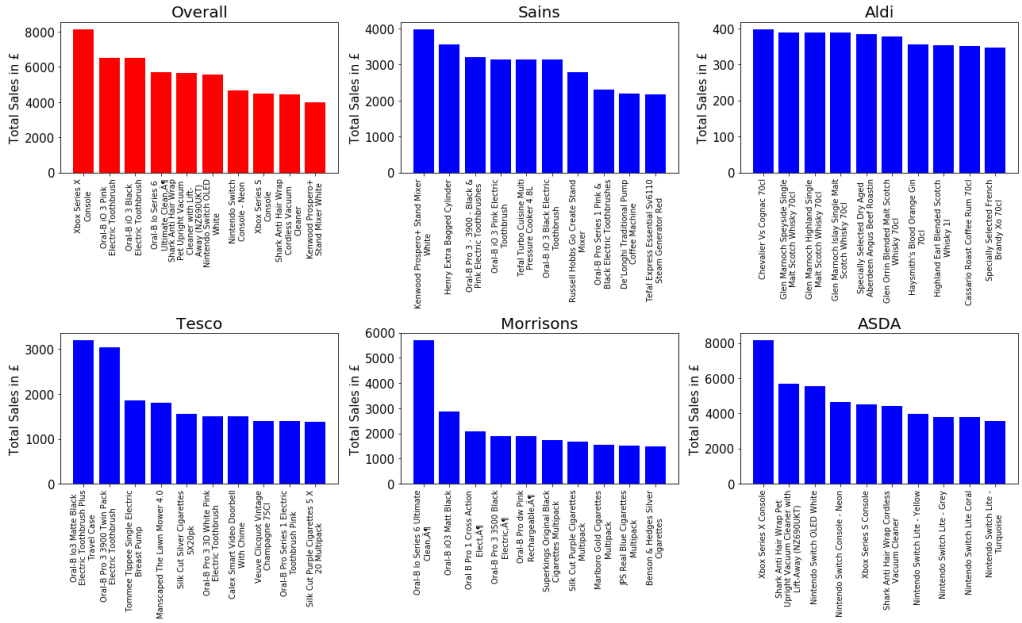


Figure 7: Top Products

The top 10 selling products overall and across specific supermarkets reveal distinct consumer preferences, reflecting both overall trends and individual supermarket demographics. Globally, high-end electronics, personal care items, and home cleaning products dominate, with items like the Xbox Series X Console, Oral-B toothbrushes, and Shark vacuum cleaners leading the sales 7. In terms of product categories, the 'Home' category leads globally with sales approximately £2.89 million, followed by 'Health Products' and 'Drinks' at £2.53 million and £1.86 million respectively. Each supermarket caters to its unique customer base: Sains's customers favor premium home appliances and health products, leading their sales at over £900k; Aldi's shoppers show a preference for everyday items like 'Fresh Food' and 'Drinks'. Tesco, Morrisons and ASDA emphasize 'Home' as their top-selling category, indicating strong demand for home goods with a sales of approximately £700K, £480K and £1M respectively. 8

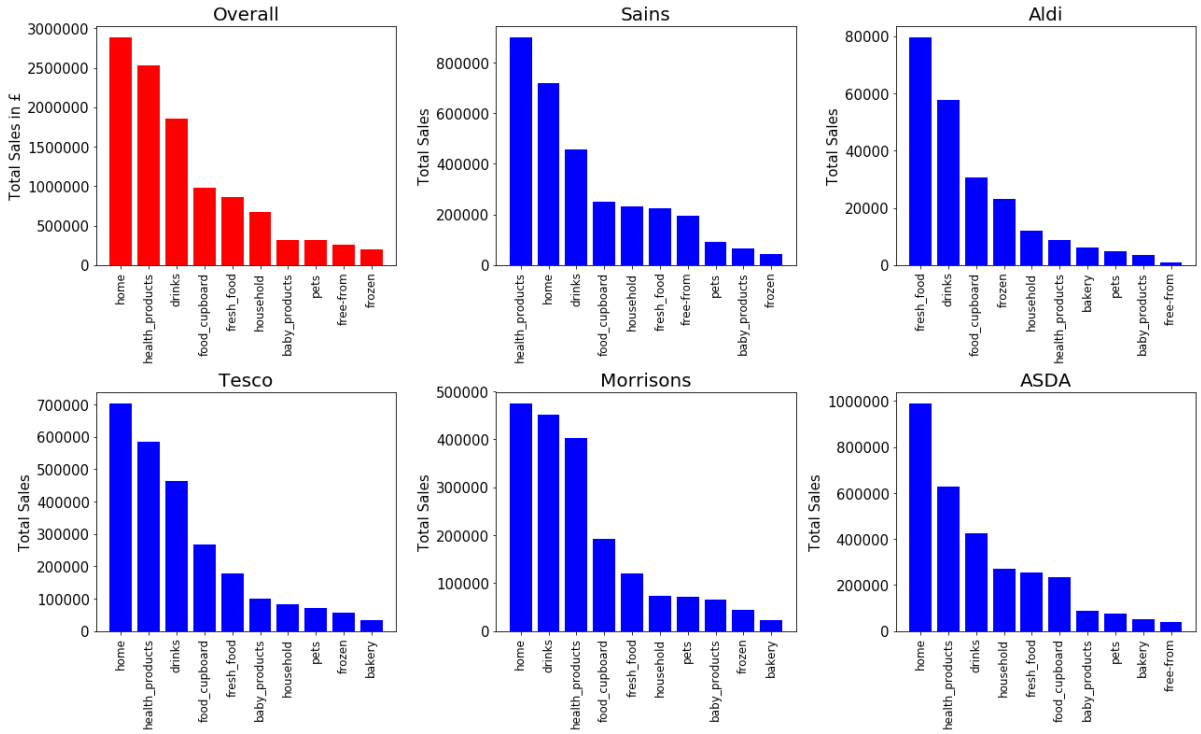


Figure 8: Top Product Categories

The sales data reveals a dominant preference for non-own brand products across various supermarkets, accounting for around 82% of total sales, compared to about 18% for own brand products. This trend is almost consistent across major retailers like Sains, Aldi, Tesco, and Morrisons, where non-own brands significantly outperformed own brands. Despite this, own brands maintained a notable presence, indicating their substantial role in the market due to factors like price sensitivity and perceived value among consumers. 9

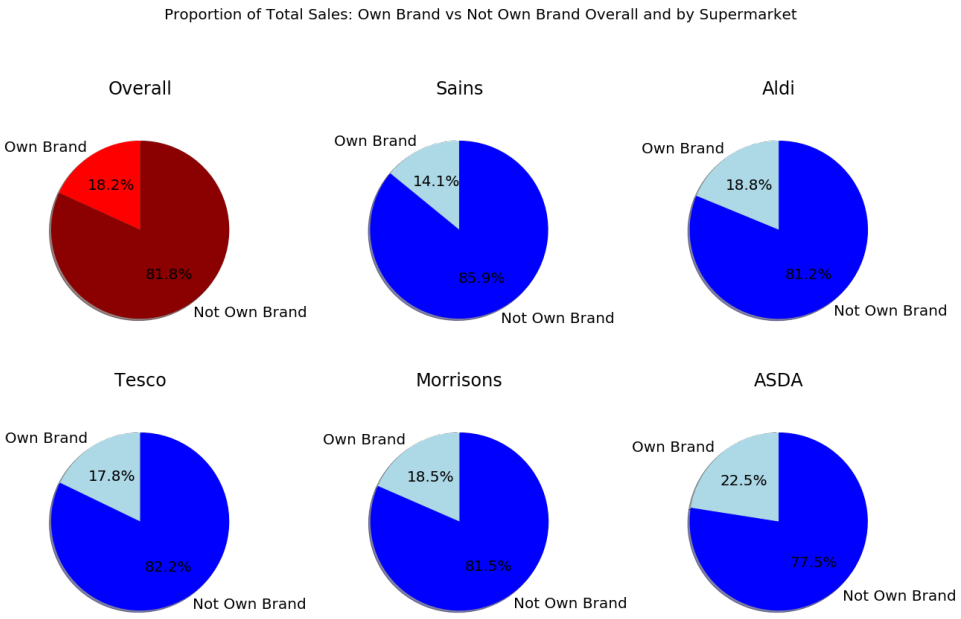


Figure 9: Own vs other brands sell across super markets

The sales volume heatmap underscores distinct purchasing patterns across supermarkets and product categories. Both ASDA and Tesco leads notably in 'Home' and 'Health Products', while Sains excels in 'Health Products'.' Aldi shows strong sales in 'Fresh Food' and 'Drinks,' despite its smaller size, and Morrisons competes closely in 'Drinks.' The 'Free-from' category is significant in Sains, indicating a specialized market focus, whereas Tesco records no sales in this category. This data highlights key areas for targeted marketing and inventory strategies for each supermarket. 10

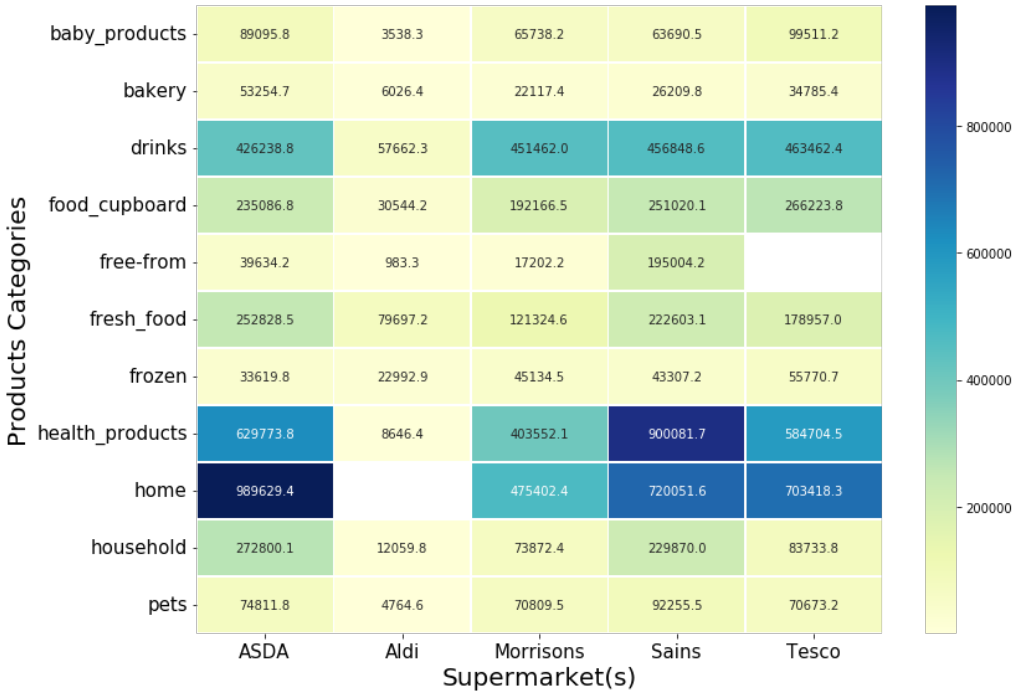


Figure 10: Heatmap of sales volume by product category by supermarket

7 Results

We employed a number of key performance indicators (KPIs) to assess our model's effectiveness on the test dataset. Strong fit was demonstrated by the remarkable 0.9861 obtained by the R-squared (R^2) statistic, which measures the amount of variance in the dependent variable that is explained by the independent variables. The average departure of the predictions from the actual data was indicated by the Root Mean Squared Error (RMSE) of 11.2828. The average squared difference was measured by the Mean Squared Error (MSE) of 127.301, while the average absolute deviation was shown by the Mean Absolute Error (MAE) of 8.12435. The map displaying the alignment between projected and actual values demonstrates the great prediction accuracy and precision of our model, which are highlighted by these measures.

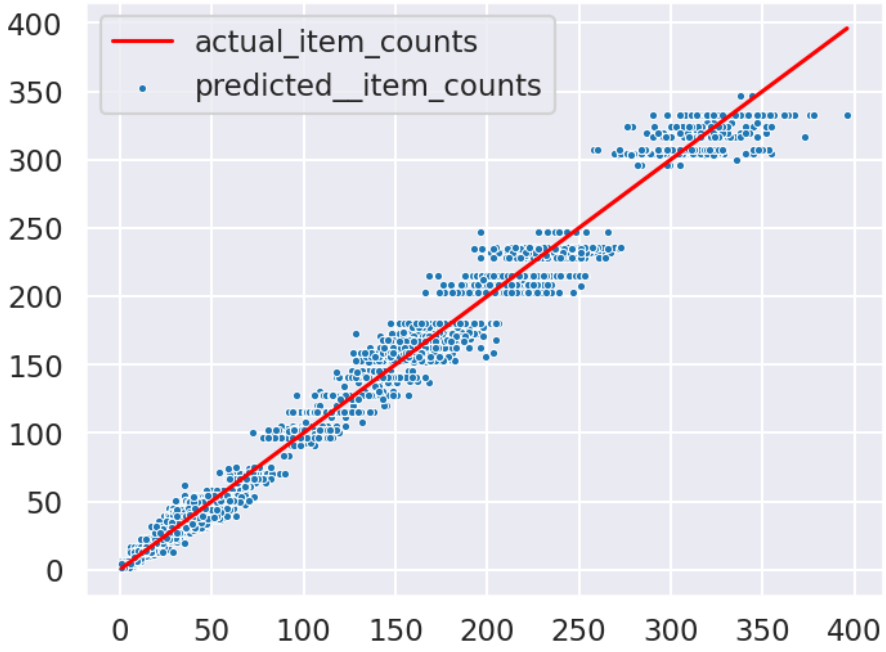


Figure 11: Evaluating Model Performance

As we discussed earlier, besides data analysis and visualization, we also conducted sales forecasting based on historical data, which is rare in similar works. While the model's key performance indicators (KPIs) are promising, several limitations must be acknowledged. The dataset had limited attributes, restricting the depth of analyses. The aggregation of data for stock prediction reduced the dataset to around 20,000 rows for training, which hindered pattern learning and generalization. Additionally, the lack of hyperparameter tuning and cross-validation limited the regression model's effectiveness. Implementing these techniques could enhance accuracy and robustness. Using only one MLlib regression model restricted our understanding of how other models might perform. Experimentation with diverse regression models could improve predictive performance. Lastly, our limited experience with PySpark and MLlib hindered our ability to utilize advanced functionalities. Gaining more experience with these technologies could facilitate sophisticated model development and analysis.

Given our merged dataset from five distinct CSV files containing transactional data, we have substantial potential for sales forecasting at various temporal intervals, such as quarterly, monthly, or weekly. This allows us to leverage time series analysis to predict future sales trends, aiding in inventory management, resource allocation, and strategic decision-making for supermarkets. Additionally, exploring customer segmentation, market basket analysis,

and promotional effectiveness can enhance sales strategies. By utilizing advanced analytics techniques like clustering, association rule mining, and predictive modeling, we can identify distinct customer segments, understand their purchasing behavior, and tailor marketing initiatives to boost customer engagement and loyalty.

8 Contribution Report

In this project, all team members collaborated across various stages, with each focusing on specific areas, outlined below:

Data Processing Shakib Polock and Akash Bappy handled data sourcing from Kaggle, preliminary cleaning, and integration, ensuring consistency and preparing it for analysis.

Data Visualization Mitun Paul and K H M Burhan Uddin conducted exploratory data analysis (EDA) and developed visualizations using PySpark, pandas, Matplotlib, and Seaborn to illustrate key insights.

Machine Learning Shakib Polock and Akash Bappy processed and transformed data using PySpark, aggregated it with Spark SQL, led machine learning efforts with a Decision Tree Regressor using MLlib, and evaluated the model using various metrics.

Documentation, and Report Writing K H M Burhan Uddin, Mitun Paul documented the project, and compiled and formatted the final report.

Literature Review Akash Bappy completed the literature review.

Each member's contributions were vital to the project's success, and this division of responsibilities allowed the team to leverage individual strengths effectively.

9 Conclusion

In this project, we utilized big data analytics in retail using Apache Spark, focusing on UK supermarkets. Our analysis highlighted key sales trends, customer behavior, and potential for predictive modeling in inventory management. Despite limitations like dataset constraints and the lack of hyperparameter tuning, our model showed strong accuracy. Future work includes exploring customer segmentation and market basket analysis. This project demonstrated the value of data-driven decision-making in retail, highlighting the potential for improved operations and enhanced customer experiences.

References

[1] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. Social media big data analytics: A survey. *Computers in Human Behavior*, 101:417–428, 2019. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2018.08.039. URL https://www.sciencedirect.com/science/article/pii/S074756321830414X.

[2] Mahya Seyedan and Fereshteh Mafakheri. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal of Big Data*, 7(1): 53, Jul 2020. ISSN 2196-1115. doi: 10.1186/s40537-020-00329-2. URL https://doi.org/10.1186/s40537-020-00329-2.

[3] Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. Big data analytics for preventive medicine. *Neural Computing and Applications*, 32(9):4417–4451, May 2020. ISSN 1433-3058. doi: 10.1007/s00521-019-04095-y. URL https://doi.org/10.1007/s00521-019-04095-y.

[4] Xueqi Cheng, Shenghua Liu, Xiaoqian Sun, Zidong Wang, Houquan Zhou, Yu Shao, and Huawei Shen. Combating emerging financial risks in the big data era: A perspective review. *Fundamental Research*, 1(5):595–606, 2021. ISSN 2667-3258. doi: https://doi.org/10.1016/j.fmre.2021.08.017. URL https://www.sciencedirect.com/science/article/pii/S2667325821001722.

[5] Marta Fernandes, Alda Canito, Verónica Bolón-Canedo, Luís Conceição, Isabel Praça, and Goreti Marreiros. Data analysis and feature selection for predictive maintenance: A case-study in the metallurgic industry. *International Journal of Information Management*, 46:252–262, 2019. ISSN 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2018.10.006. URL https://www.sciencedirect.com/science/article/pii/S0268401218304699.

[6] A. H. P. K. Putra, K. M. Rivera, and A. Pramukti. Optimizing marketing management strategies through it innovation: Big data integration for better consumer understanding. *GRMILF*, 3(1): 71–91, Jan 2023. doi: https://doi.org/10.52970/grmilf.v3i1.398.

[7] Emel Aktas and Yuwei Meng. An exploration of big data practices in retail sector. *Logistics*, 1(2), 2017. ISSN 2305-6290. doi: 10.3390/logistics1020012. URL https://www.mdpi.com/2305-6290/1/2/12.

[8] A. Seetharaman, I. Niranjan, V. Tandon, and A. S. Saravanan. Impact of big data on the retail industry. *Corporate Ownership & Control*, 14(1-3):506–518, 2016. doi: 10.22495/cocv14i1c3p11. URL https://doi.org/10.22495/cocv14i1c3p11.

[9] Shubham Lekhwar, Shweta Yadav, and Archana Singh. *Big Data Analytics in Retail: Proceedings of ICTIS 2018, Volume 2*, pages 469–477. 01 2019. ISBN 978-981-13-1746-0. doi: 10.1007/978-981-13-1747-7_45.

[10] Marnik G. Dekimpe. Retailing and retailing research in the age of big data analytics. *International Journal of Research in Marketing*, 37(1):3–14, 2020. ISSN 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2019.09.001. URL https://www.sciencedirect.com/science/article/pii/S016781161930062X.

[11] Aashish Prasad. Analysing online retail transactions using big data framework, 04 2019.

[12] willian oliveira gibin and DECLAN MCALINDEN. Retail analytics trends, 2024. URL https://www.kaggle.com/dsv/7584159.