DATA ANALYTICS IN RETAIL INDUSTRY WITH AWS

Harshal Shinde, Devendra Dahale, Siddhant Jain

Abstract: As retail market becomes extensively competitive, the ability to optimize on serving business processes while satisfying customer expectations has never been more important. Therefore, managing and channelizing data to work towards customer delight as well as generate healthy profits is crucial to survive prosperously. In the case of big retail players internationally as well as in India, data or rather big data analytics is now being applied at every stage of the retail process - tracking emerging popular products, forecasting sales and future demand through predictive simulation, optimising product placements and offers via customer heat-mapping and many more. Alongside this, identifying the customers likely to be interested in particular product types based on their previous purchase behaviours, working out the best way to approach them through targeted marketing efforts and finally working out what to sell them next is what forms the core of data analytics. This article is the outcome of a descriptive research on the past, present and future of retail industry. . Analytics comes to the rescue and with the aid of Cloud Computing it aids to explore this paradigm to a previously unimaginable extent.

Key Words: *Retail Industry, Retail Analytics, AWS, Spark, TidyVerse, Prediction, Recommendation, Association*

1. INTRODUCTION

Analytics is the discovery and communication of meaningful patterns in data.

Ours is a data centric world. Organizations around the globe are looking for ways to exploit the propulsive growth of data to find ways of exploring previously hidden insights so as to publish new revenue streams, gaining operational efficiencies and understanding customer needs better. The advent of the digital age has led to rise of data with every passing day. Organisations are producing and storing large amounts of data for gaining insights over data from various aspects. The increasing volume and detail of information, the rise of multimedia and social media and the Internet of Things are fuelling exponential data growth. Data Analytics refers to automating insights into a dataset .It requires the use of queries and data aggregation procedures. It establishes various dependencies between input variables. The data is extracted and categorized so as to identify and analyse behavioural patterns in the associated data. The techniques used for analytics vary according to the requirements of the The very notion of what might attract the new customers is to a large extent resolved by the use of analytics. By recognizing existing patterns in data, data analytics helps serve organizations in better serving their existing customers which is practical and more cost effective than establishing a whole new business. It gives business organizations the edge in recognizing changing climates of their markets so that they can initiate appropriate actions to stay competitive Industry can identify the current trends, re-order supplies for hot-selling items, adjust the prices in real time and also manage and control product distribution across different stores to channelize their sales in more effective manner. This provides retail industry with entirely different perspectives of looking towards the datasets available at their disposal. By collating these organisational datasets with social media data streams, they can also use it for better sales predictions, designing relevant campaigns to suit their profitable customers and thereby ensuring customer satisfaction. The benefits of AWS in the modern cloud are huge. Data protection, regulatory compliance flexibility, cost-effectiveness, multiple storages, auto-scaling, access to the data anytime, data-centric encryption, high-performance processing are few benefits of AWS cloud.

. We employed machine learning algorithms on Spark to uncover more complex customer behaviour patterns, like which products are frequently purchased together.

2. METHODOLOGY

The process was divided in Four Steps.

1. Data Collection
2. Environment setup
3. Cleaning the Data
4. Exploratory analysis
5. Employ machine learning algorithms
6. Developing interactive web application
7. Deploying the application on AWS

The purpose of the analysis is to extract insights with business value for an online retailer using R, tidyverse, sparklyr, and Spark.

Data Collection: Data analysed is a table containing real online retail sales data downloaded from the UCI machine learning repository. It contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Environment Setup: - Using tidyverse for data manipulation and visualization, and Spark for data processing. R sparklyr package will enable us to do data analysis using the typical tidyverse approach, while leveraging the power of Spark in the background. Using a local Spark instance, which comes with the sparklyr package.

Cleaning the Data: - Regardless of the type of analysis or data visualizations you need, data cleaning is a vital step to ensure that the answers you generate are accurate. When collecting data from several streams and with manual input from users, information can carry mistakes, be incorrectly inputted, or have gaps. Incorrect or inconsistent data leads to false conclusions. And so, how well you clean and understand the data has a high impact on the quality of the results.

Exploratory analysis: - Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Results of Exploratory Analysis

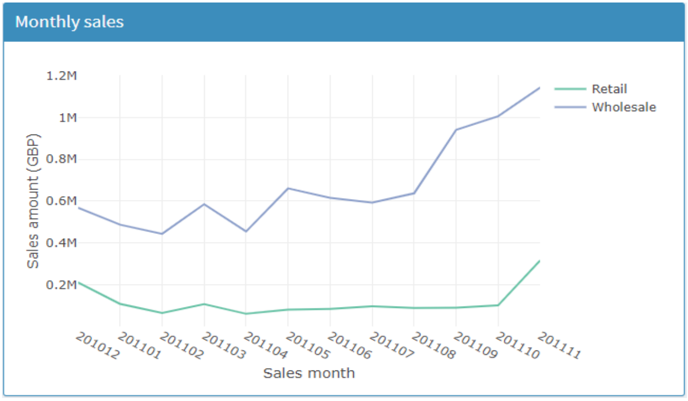


Fig 1: Monthly sales for retail and wholesale

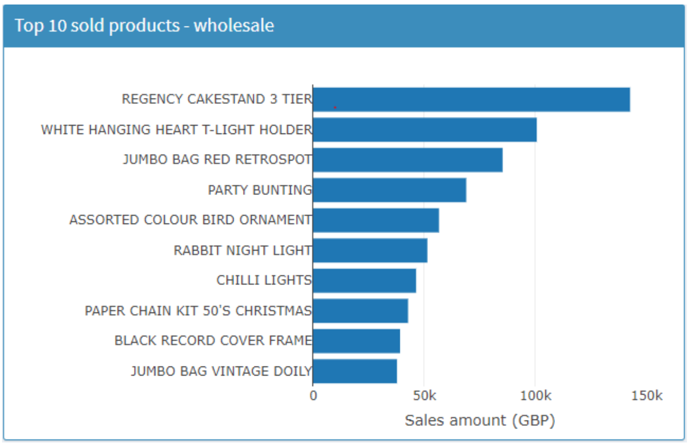
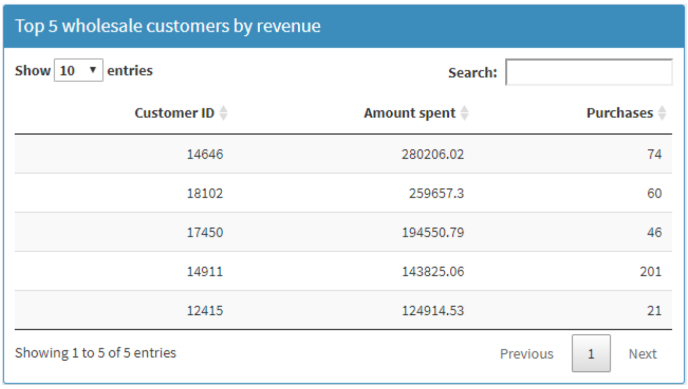


Fig 2: Top 10 sold products – wholesale


Top 5 wholesale customers by revenue

Customer ID	Amount spent	Purchases
14646	280206.02	74
18102	259657.3	60
17450	194550.79	46
14911	143825.06	201
12415	124914.53	21

Showing 1 to 5 of 5 entries

Fig 3: Top 5 wholesale customers by revenue

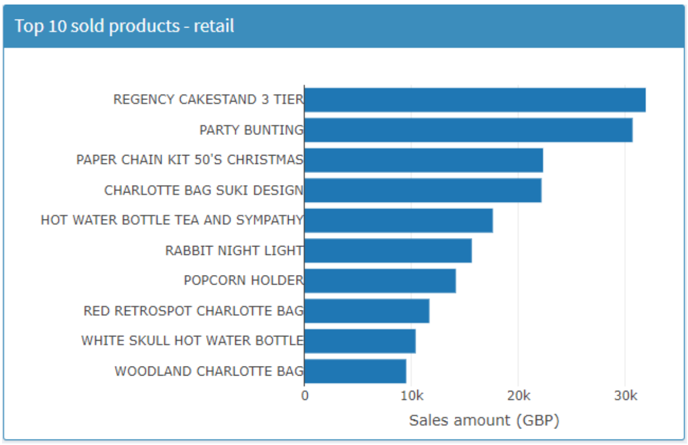


Fig 4: Top 10 sold products - retail

Employ machine learning algorithms: FP-Growth algorithm is a development of Apriori, the deficiency of the Apriori algorithm improved by the FP-Growth algorithm. In Apriori a generate candidate is required to get frequent itemsets. However, FP-Growth generate candidate algorithm is not done because FP-Growth uses the concept of tree development in search of the frequent itemsets. This is what causes the FP-Growth algorithm is faster than the Apriori algorithm. In this particular application we have extensively used FP growth to predict the pairs or triplets of product combination. FP-Growth algorithm is more effective in generating the frequent and informative association rules to find out the consumer spending patterns compared to the Apriori Algorithm. The related products are placed together to give the customers a logical view to select items they might buy

Developing interactive web application: It is easier to visualize complex data and relationships than deciphering them from spreadsheets / tables. We have used shiny for visualization of different graphs and patterns. Shiny is an open package from RStudio, used to build interactive web pages with R. It provides a very powerful way to share your analysis in an interactive manner with the community. The ease of working with Shiny has what popularized it among R users. These web applications seamlessly display R objects (like plots, tables etc.) and can also be made live to allow access to anyone. Shiny provides automatic reactive binding between inputs and outputs. It also provides extensive pre-built widgets which make it possible to build elegant and powerful applications with minimal effort.

Deploying the application on AWS: Amazon Web Services is the umbrella term for the range of Amazon’s cloud computing offerings. We be using Amazon Elastic Compute Cloud (EC2), a service where we rent virtual computers in the cloud to run applications. AWS EC2 offers a free tier so we can deploy without spending a rupee. We used port number 8787 for R Studio and Port number 3838 for Shiny Server.

3. IMPLEMENTATION

3.1 Setup

Let's start with basic initialization - using tidyverse for data manipulation and visualization, and Spark for data processing. R sparklyr package will enable us to do data analysis using the typical tidyverse approach, while leveraging the power of Spark in the background.

Using a local Spark instance, which comes with the sparklyr package. Compared to using a cluster syntax - wise, only the connection string is different. Starting with the analysis by reading the data into Spark DataFrame from the Parquet file

From a quick look its clear all the columns from the description are there, and their datatypes are in line with the descriptions. It is also clear that the records are actually transaction details - every record corresponds to one-line item from the invoice. To get a general overview of the online retailer sales, it would make sense to aggregate the data to invoice level first.

3.2 Data Quality

first look at general data quality using summarise function makes it clear that all of the records have an invoice identifier set. This will make aggregating data to invoice level easy. Only 25.900 invoices with 541.909-line items - an average of 20.92-line items per invoice. It looks like wholesaler orders are dominating this dataset. I'll have to check this assumption later, by looking at the distribution of invoices by the number of line items. Most interestingly, 1/4 of records have no CustomerID - retail customers or just unregistered retail customers. In this case, we will look at the distribution of invoices by the number of line items and by having/not having CustomerID. If invoices without CustomerIDs really belong to retail customers, their distribution of invoices by line items should be radically different than the one with CustomerIDs (the wholesalers). All of the line items have all of the necessary data filled: StockCode, Description, Quantity and UnitPrice. Number of distinct stock descriptions is a bit larger than the number of distinct stock codes, implying that some stock codes have multiple descriptions - probably due to name corrections during the year. This is something to check later, when the focus will be on stock items.

3.3 Invoice level analysis

Now We are ready to aggregate the data to invoice level. We also save the invoice level data to Parquet file for persistence. This step should be executed only once - for new analysis runs Spark can read directly from the Parquet file. We will also add an indicator column for cancellations We found later that there are cases where not all line items in one invoice have the same InvoiceDate - working around that by taking the last timestamp for each invoice. also later detected large cancelled invoices, so making sure they are marked here. Loading the invoice data into Spark Memory and check for unique invoice identifiers. We've found the cause of the duplicates - not all invoice line items have the same datetime stamp. This can be easily fixed in the aggregation to invoice - simply aggregate the InvoiceDate and take the latest From the dataset description we know that there are some cancellations of invoices.

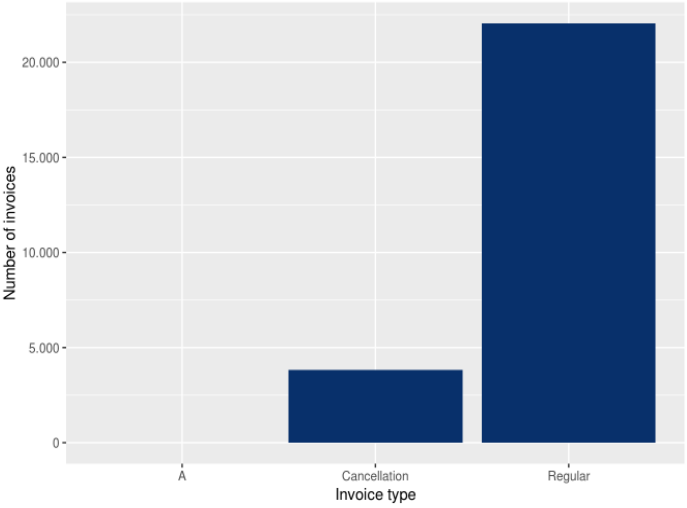


Fig 5: Invoice types according to no. of invoice

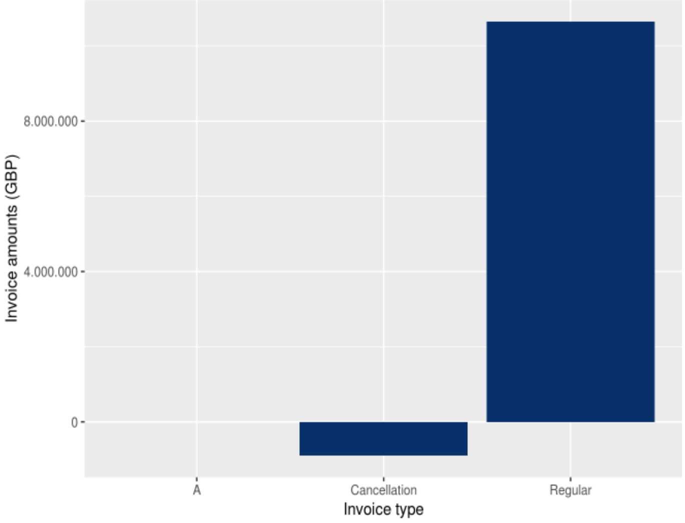


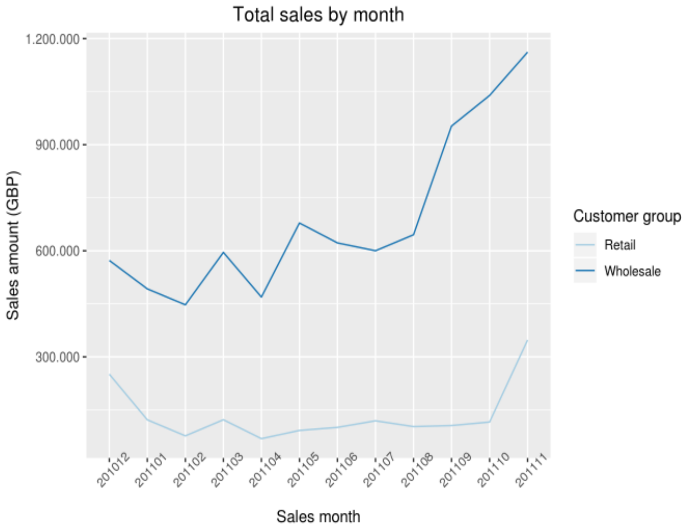
Fig 6: Invoice types according to invoice Amount

Number of cancellations is relatively high compared to regular invoices - it means every 6th invoice has been cancelled! A rather surprising insight - it may indicate something is wrong with the sales process or inventory management. On the invoice amounts side the situation is not so dramatic, but serious - cca 9% of total invoice amounts have been cancelled.

3.4 Invoice level analysis according to customer group

The customers are classified into two types: Retailers and Wholesalers. Retail part of the business is quite small, roughly 1/6 of the total business (invoices issued AND revenue). This means that business optimization should be focused on the wholesale side of the business to maximize impact on revenue. In a real-life project, this decision would be discussed with the businesspeople.

Let's look at total sales in a year, per month and customer group:



For wholesalers, peak months are those leading to season changes - March (start of spring), May (summer is coming), September (autumn is coming) and of course Christmas and New Year. This is not surprising, as the company sells all-occasion gifts - they change with the seasons of the year. Sales uptick comes prior to season change, as wholesale customers are stocking with the coming season merchandise. There is also significant sales growth visible towards the end of the year, not caused by seasonality only - sales in November 2011 is twice the amount from December 2010. It will be interesting to see whether there are some particular items or customers driving that growth. For retail customers, seasonality is visible mainly in November and December (shopping season leading up to Christmas). Let's look at contribution of countries –

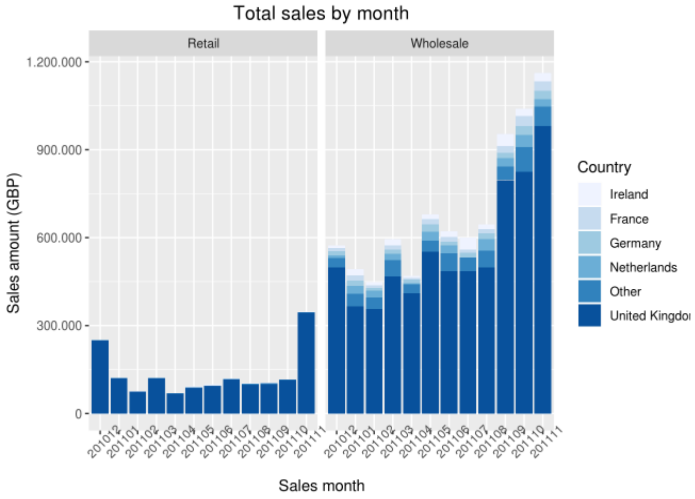


Fig 7: Contribution of countries in sales

The charts reveal a couple of insights: UK sales heavily dominates in wholesale, roughly 85% of total sales comes from the UK - it's the store's home market. Other than UK, sales is mostly coming from neighbouring countries. Rest of the world is almost negligible. Sales growth is fuelled exclusively by UK market. In the retail business, sales are coming almost exclusively from the UK. This implies that any marketing actions should be focused on the domestic retail market.

3.5 Product level analysis

I'll take only regular invoices, as in the invoice's analysis. They are variation in product description, variation naming is not consistent. It would be difficult to extract core product names from variation names.

In addition, not all variations are handled by product suffixes

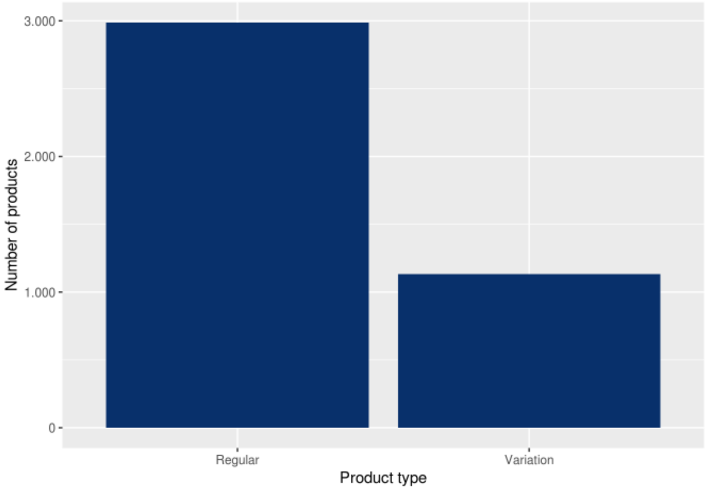


Fig 8: Variation in Invoice

3.6 The market basket analysis (frequent itemsets):

Using the Spark's MLlib machine learning library, which contains a parallelised implementation of the frequent itemsets algorithm, FPgrowth. We have to prepare the data first - fpgrowth expects one row per purchase, and all products in that purchase collapsed in a list.

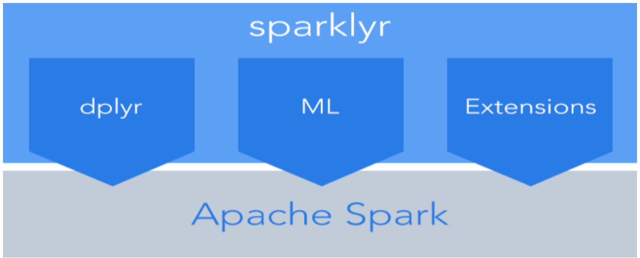


Fig 8: Spark Modules

Now running the FPgrowth. with already tuned the hyperparameters to a sensible value for this case. Here they are - pairs of items frequently purchased together. This indicates in how many purchases does the product pair appear. It would be easier to see the itemsets arranged relative to frequency of appearance - here is one way to do it - a wordcloud:

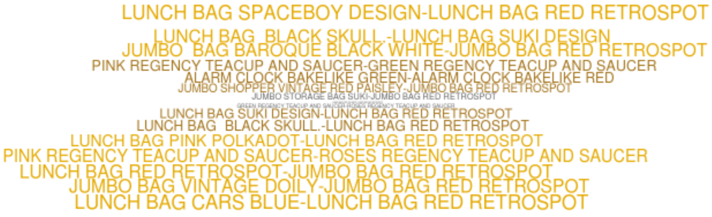


Fig 9: Word Cloud

Text colour and size are proportional to pair frequency. Notice that some products appear in more than one pair. That means it would be much more revealing to display the data in network form. It would also be helpful to see association directions between pair members - indicating the purchase of which product leads to purchase of another, the FPgrowth algorithm provides us with just that - association rules Let's extract them and display them in network form. Using the D3 based network visualization.

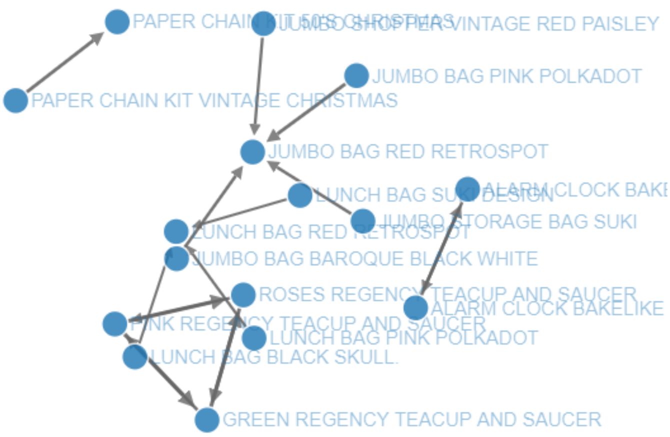
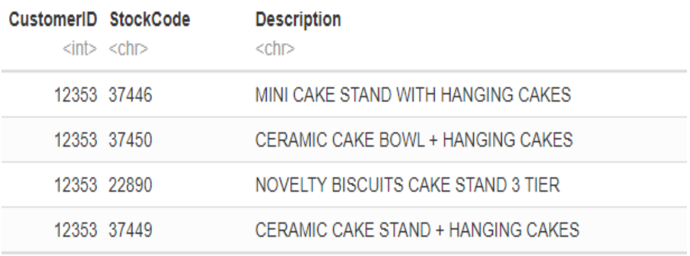

PAPER CHAIN KIT 50'S CHRISTMAS
JUMBO SHOPPER VINTAGE RED PAISLEY
JUMBO BAG PINK POLKADOT
PAPER CHAIN KIT VINTAGE CHRISTMAS
JUMBO BAG RED RETROSPOT
LUNCH BAG SUKI DESIGN
ALARM CLOCK BAKE
JUMBO STORAGE BAG SUKI
LUNCH BAG RED RETROSPOT
JUMBO BAG BAROQUE BLACK WHITE
ROSES REGENCY TEACUP AND SAUCER
ALARM CLOCK BAKELIKE
PINK REGENCY TEACUP AND SAUCER
LUNCH BAG PINK POLKADOT
LUNCH BAG BLACK SKULL.
GREEN REGENCY TEACUP AND SAUCER

Fig 10: D3 based network visualization.

I can now clearly see that items which are frequently bought together appear in clusters of variations of the same product. This behaviour is a signature of wholesale customers - when they order a product, they frequently order multiple variations of that product, to add variety to their gift shop offering. It would be even more interesting to see how the itemsets change with seasons - since I already know from exploratory analysis that gifts are a seasonal business. That would be interesting for the client businesspeople too

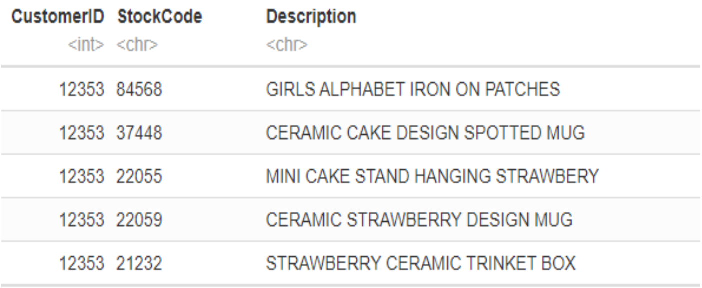
3.7: The recommendation engine

In the previous step We've detected products frequently purchased together over the whole customer base. To make an operational impact, we need to make next best product recommendations on individual customer level. Using the Spark's MLlib machine learning library, which contains an ALS (alternating least squares) recommender engine. We will be focusing on wholesale customers, because we can identify them, unlike retail customers. We have to prepare the data first. A recommender engine expects rating information for each customer-product pair - however, we only have the purchase information. Since this is often the case, it is possible to supply alternate information, such as purchase quantity - it is used as a measure of the strength of the relationship. We also have to generate a replacement ID for StockCode, as Spark ALS implementation currently accepts only integer values for product IDs, and some StockCodes contain letters. We also need to aggregate all purchases of the same product by the same customer, to enable ALS to construct a customer-product relationship matrix.



CustomerID <int>	StockCode <chr>	Description <chr>
12353	37446	MINI CAKE STAND WITH HANGING CAKES
12353	37450	CERAMIC CAKE BOWL + HANGING CAKES
12353	22890	NOVELTY BISCUITS CAKE STAND 3 TIER
12353	37449	CERAMIC CAKE STAND + HANGING CAKES

Fig 11: Customer 12353 Purchase history



CustomerID <int>	StockCode <chr>	Description <chr>
12353	84568	GIRLS ALPHABET IRON ON PATCHES
12353	37448	CERAMIC CAKE DESIGN SPOTTED MUG
12353	22055	MINI CAKE STAND HANGING STRAWBERY
12353	22059	CERAMIC STRAWBERRY DESIGN MUG
12353	21232	STRAWBERRY CERAMIC TRINKET BOX

Fig 12: Customer 12353 Purchase recommendation

4. RESULT

From the examples Its Evident that the recommender is making sensible recommendations. In a real-world project, the next step would be to make the recommender operational - by exposing an API for the webshop application to use, or to deliver data via batch jobs to an interface database table, or any other way of integration suitable for the webshop system. In this case We will demonstrate how the recommender performs by displaying its output in a webapp (refer to the figures given below). The examples of EBay, Amazon and so many others, who have successfully reaped great benefits of data analytics, goes on to prove that retailers, large and small would definitely be able to harness the miraculous benefits of analysing not only structured, but also unstructured data on consumer behaviour

5. CHALLENGES IN DATA ANALYTICS IN RETAIL INDUSTRY

Retailers have already started putting data analytics at the heart of their operations across the value chain - procurement, supply chain, sales and marketing, store operations, and customer management. However, they now need to establish a big data ecosystem, which processes multiple terabytes of new data and petabytes of historical data, which will help them improve their revenues via analytics-based decision-making. While this may sound really exciting, big data management and analysis comes with its own set of challenges.

Several issues will have to be kept in mind to optimize the full capabilities of big data. Privacy, security, intellectual property, and even liability policies need to be stringent in terms of big data. Since big data encapsulates high end analytics, specially trained professionals need to be added to the team to utilize and functionalize the big data.

Companies need to integrate information from multiple data sources, often from third parties, as well as deploy an efficient data to aid such an environment. Many times, companies fall in short-sightedness, failing to implement insights from analytics. However, this could be fixed by continuous alterations of retail styles where a certain team is allotted for task of arrangement of insights and their implementation.

One of the most challenging tasks in the retail industry is pricing. Price it too high, and you may lose a customer, price it too low, and it hurts your margins. Then, there is a host of other pricing problems including, pricing decisions across different channels, geographies, pricing benchmarks, and mark-ups.

Most of the products in the retail industry are seasonal, and demand forecasting is usually done on a historical basis. But the method is outdated and won't be sufficient in an age where demand is highly fluctuating, trends are rapidly changing, and consumer preferences are dynamic.

However, the pace of development and adoption of innovation is increasing rapidly at present times due to the advent of data analytics and automation. Today, managers have access to a large stream of data, and decision-making on the basis of gut-feeling, the rule of thumb, and guesswork are largely eliminated due to the advent of data analytics.

6. CONCLUSION

Retailing is at the platform for more data-driven disruption because the quality of data available from internet purchases, social-network conversations, and recently, location-specific smart phone interactions have emerged into a new entity for digital based transactions. Improved performance, better risk management, and the ability to unearth insights that would otherwise remain hidden, are the benefits organisations reap through utilization of big data management. Retailers can benefit immensely form a structured analytics-driven approach that will help them understand how their customers are using their products and services, how their operations and supply chain are performing, how to manage their workforce and how to identify key risks - insights that they then can then act upon. The pace and the dexterity with which micro data is collected, gives the retailers immediate insights on the shopping trends. This analysis on the move allows them to adjust their prices and add to the lure by announcing on the spot discounts on the sales floor based on their current and previous shopping patterns. This data, often collected through interactive mobile devices in stores, provides the retailer an understanding of the buyers needs and give insights into making smarter decisions about product placement in the store. Data capture and analytics usage certainly have come a long way in the last ten years, and it is interesting to look back on how trends in data analytics have affected the marketplace. As the Internet of Things expands further and our world becomes even more connected, this space will continue to evolve. Analytics plays an important role for marketers as they work to achieve the goal of understanding customers. Mobile devices have a prominent place in the expanding Internet of Things (IoT) ecosystem, and businesses shall be leveraging analytics to collect the rich data they provide. Once consumers have agreed to “opt in,” retailers can learn quite a bit from how they use their devices to interact with a brand. For example, what products are they most interested in browsing and buying? How often are purchases made and are there developing patterns? If a shopper is buying the same box of baby diapers once every two weeks, for example, they might appreciate a reminder to buy, notifications of sales or an automated purchase renewal option. Analytics give retailers the power to identify these patterns and adjust their offerings to better cater to users, in turn enhancing the convenience, customization and commerce of mobile shopping. Analytics plays the most crucial role in this process as, it is the very thing that drives those ‘understand what makes the customer decide’, policies of the retail companies. Through the various functions that it carries forward, it creates insights such as, how would a company be able to increase margins at the product level, it also provides insights in what the customer is like, or why the customer would want to buy a certain product. This is called Market Basket Analysis. The analytics help the companies also identify those items, which a customer would be very likely of purchasing together, what promotions and offers would work the very best for which products and personalized offers for every individual customer. These insights are totally customer based, but there are also those that are fully company-based insights. Analytics is able to give the insights in terms of how much spending would a company have to do, store wise product mix, optimal pricing that would get them more buyers, efficient stock strategies and many more. The facilities are trembling up the computing world in a similar technique that Amazon is varying in the retail industry. By assessing its cloud very less expensive, Amazon can offer reasonable and scalable facilities to everybody from the latest start-up to a Fortune 500 company. Also, as we mentioned earlier the flexibility and scalability are the key reasons why AWS cloud is different from any other cloud platform. In AWS cloud the collections of objects can work together and distinctly.

REFERENCES

[1] R. Gangurde, D. B. Kumar, and D. S. D. Gore, “Building Prediction Model using Market Basket Analysis,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, no. 2, pp. 1302–1309, 2017.

[2] Shmueli G, Koppius O (2011). Predictive Analytics in InformationSystems Research. MIS Quarterly, 35(3), pp. 553 – 572.

[3] Camm J, Cochran J, Fry M, Ohlmann J, Anderson D (2014). A Categorization of Analytical Methods and Models. In Essentials of Business Analytics, pp. 5 – 7.

[4] Demirkan, Haluk, and Dursun Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud., Decision Support Systems 55.1 (2013): 412-421.

[5] Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi, Big data and cloud computing: current state and future opportunities., Proceedings of the 14th International Conference on Extending Database Technology. ACM, 2011.

AUTHORS PROFILE

Harshal Shinde, Information Technology, MIT Academy of Engineering, Pune, Maharashtra, India

Devendra Dahale, Information Technology, MIT Academy of Engineering, Pune, Maharashtra, India

Siddhant Jain, Information Technology, MIT Academy of Engineering, Pune, Maharashtra, India