Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

ul. Długa 44/50, 00-241 Warszawa

Studia Podyplomowe

Data Science w zastosowaniach biznesowych

Warsztaty z wykorzystaniem programu R

Maciej Łecicki

Examples of the practical application of machine learning in supply chain management

Praca wykonana pod kierunkiem:

dr hab. Piotr Wójcik

Zakład Finansów Ilościowych WNE UW

Warszawa, październik 2020

Oświadczenie kierującego pracą:

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki określone dla prac dyplomowych.

Data:

Oświadczenie autora pracy:

Mając świadomość odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiazujacymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem świadectwa studiów podyplomowych lub tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data:

Table of contents

Introduction

Rising customer's expectations for instant service and highly customized products, global network, short life cycle of products, business strategies based on growth through innovation and high demand volatility put enormous pressure on Supply Chain Management forcing it to adapt to rapidly changing environment.

SC professionals search for ways to respond to these challenges and one of the answers is to leverage data and predictive analytics to improve and speed up decision making process. This is possible thanks to recent technological boom, specifically increase in computational power of computers and development of various open source software which massively increased popularity and ease of application of Data Science methods to complex problems.

The purpose of this thesis is to present examples of the practical application of Machine Learning in SCM to tackle some of the typical challenges in Supply Chain area. The concept of SC and its challenges will be described in first two chapters of this thesis and the aim of last t chapters is to find solutions using various Data Science techniques.

Chapter 1 describes what Supply Chain is and brings to life role and importance of Supply Chain Management. We will cover some of the key concepts and components of Supply Chain and the purpose of measurement of SCM performance and describe some of the KPIs[1]. Additional purpose of this introductory chapter is to bring in SC vocabulary for readers unfamiliar with SCM as it be used in further chapters of the thesis.

Chapter 2 briefly describes the company, structure of SCM, supply chain network and available data which will be used in our research. Although we often hear that in today's world data is omnipresent this may not be the case in business environment. We will deep dive into this problem and explain typical issues and challenges related to missing or incomplete data and how this can be resolved taking into consideration various data types. Additional SCM definitions and abbreviations will be introduced too.

[1] KPI – Key Performance Indicator.

Chapter 3 defines research aim and methodology. This is where we will connect the dots between typical SCM problems and available data to ask ourselves some questions if (and how) specific components of incredibly wide concept of Data Science could be applied in SCM to help find answers to the most burning issues reflected in selected KPIs from Chapter 1.

Chapters 4 and 5 will take use through results of research in scope of this thesis. We will:

- perform Explanatory Data Analysis. EDA is commonly used by statisticians to explore data in a systematic and graphical way and we will use it to investigate relationship between downstream stock availability and upstream supply chain and manufacturing constraints,
- investigate possibility for building predictive model for stock-out [2]of products on the basis of available supply chain parameters and known manufacturing constraints,
- use Unsupervised Machine Learning to provide proposal for products' segmentation as alternative to classic ABC anlaysis,

This thesis concludes with a summary to capture key findings from the research.

As part of the introduction it also needs to be mentioned that all analyses and models used in this thesis had been made in R on anonymized data from a large manufacturing company from FMCG (fast moving consumer goods) sector.

[2] The concept of stock-out will be introduced in further chapters of this thesis.

1. Role of Supply Chain and Supply Chain Management

Gartner[3] defines Supply Chain Management (SCM) as "the processes of creating and fulfilling demands for goods and services. It encompasses a trading partner community engaged in the common goal of satisfying end customers" [4].

More comprehensive definition, which also captures Supply Chain and its physical and non-physical aspects, can be found in latest APICS Dictionary from ASCM[5]. SCM is "the design, planning, execution, control, and monitoring of supply chain activities with the objective of creating net value, building a competitive infrastructure, leveraging worldwide logistics, synchronizing supply with demand, and measuring performance globally".

On the basis of both definitions we can draw below conclusions:

- Supply Chain include people and infrastructure involved in SCM activities and can be described as a network connecting individual nodes within it,
- Goals of SCM are internal, where the focus is on processes optimization and cost reduction, and external where the key objective is customer satisfaction from fulfillment of their demand or gaining competitive advantage over direct competitors which leads to business growth from increased market share,
- Important element of SCM is performance measurement.

Above conclusions will be used as a starting point to familiarize the reader with some of the key elements of Supply Chain, functional components of SCM and Key Performance Indicators (KPIs) used to measure SCM performance.

[3] Gartner is a leading research and advisory company famous for releasing annual Supply Chain Top 25 ranking, identifying supply chain leaders.
[4] gartner.com
[5] ASCM – Associate For Supply Chain Management.

1.1. Supply Chain – Material and Information Flows

Supply Chain is "the global network used to deliver products and services from raw materials to end customers through an engineered flow of information, physical distribution and cash"[6].

This simple definition can be further simplified with network visualization as shown in Figure 1.

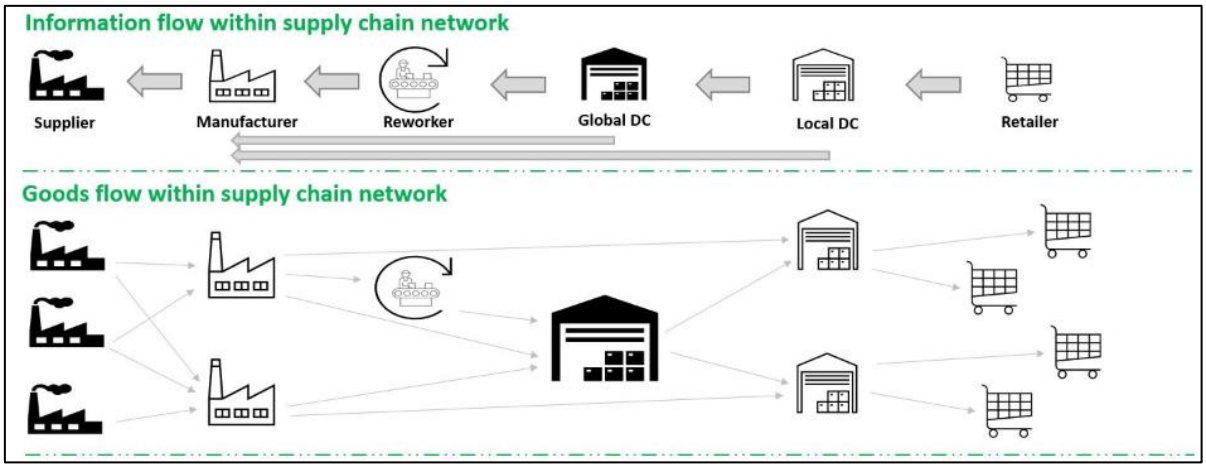

Figure 1: Scheme of Supply Chain Network. Materials (goods) flow downstream (from Suppliers to Retailers, which in this case is the final customer) to satisfy demand of the final customer. Information flows upstream (from Retailers, who submit their demand, to Suppliers to deliver raw materials to manufacturing sites).

SC network scheme in Figure 1 can be of course extended depending on business nature, supply chain strategy or product type. For example we could also include final customers who buy products from retailers.

One additional point to mentioned is that typically when we mention SC network we mean global SC network where distance (and therefore transit time) between suppliers and manufacturers or global distribution center[7] (DC) and local DC can be significant (and usually is measured in days or weeks).

Services mentioned in the definition of Supply Chain can include On Time In Full (OTIF) deliveries of raw materials from Suppliers to Manufacturer or Stock Availability of final

[6] APICS Dictionary. The essential supply chain reference. Sixteenth edition.
[7] Distribution center (DC) - A location used to store inventory. Decisions driving warehouse management include site selection, number of facilities in the system, layout, and methods of receiving, storing, and retrieving goods; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

product for Retailers at Local DC. The level of services is measured and these measures are called Key Performance Indicators. More on that in chapter 1.3.

1.2. Functional components of SCM

Responsibilities behind Supply Chain Management are as wide and complex as Supply Chain itself. Depending on literature we can define between five to eight[8] key components of SCM:

1. Planning,
2. Information,
3. Source (of raw materials and packaging),
4. Inventory,
5. Production (of final goods),
6. Location,
7. Transportation,
8. Return of goods.

From my professional experience I would indicate Planning as the most important component of SCM with Information being its backbone. As much as my opinion can be biased taking into account my planning background, I think most SCM professionals and literature would agree with me.

The reason behind it is that, first of all, all other component of SCM are the outcome of various types of Planning processes distinguished on the basis of planning horizon (short to long) or frequency (daily, weekly or monthly)[9].

---

[8] https://www.iqualifyuk.com/library/business-management-section/the-eight-components-of-supply-chain-management/

References:

Ballou, R.H., 2007. Business Logistics/supply Chain Management: Planning, Organizing, and Controlling the Supply Chain. Pearson Education India.

Christopher, M., 2016. Logistics & Supply Chain Management. Pearson UK.

Lambert, D.M. & Cooper, M.C., 2000. Issues in Supply Chain Management. Industrial Marketing Management.

Tan, K.C., 2001. A framework of supply chain management literature. European Journal of Purchasing & Supply Management.

[9] Typically, taking into consideration both factors, planning activities can be operational, tactical and strategic.

Secondly, simply due to its nature and complexity planning is most difficult part of SCM due to challenges like demand uncertainty or downstream and upstream constraints that we will describe in more details in one of next chapter.

Thirdly, planning can span over other components of SCM. We can distinguish source (material) planning, inventory planning, production planning etc.

Last but not least, one other reason why planning is the key element of SCM is that deterioration of any SCM KPIs (regardless of SCM component, maybe except for Information) can be traced back to planning. In other words, from practical point of view and my professional experience planning tends to be blamed for poor SCM KPIs.

1.3. Measurement of Performance of SCM

"The importance of measurement and control of supply chain costs arises directly from the objectives and tasks of supply chain management. (...)

The indicators to evaluate the performance capacity or the performance of an organization should cover both the financial sector as well as the operations, since the aim is to achieve customer satisfaction at lower costs and ensure the long-term competitiveness. In this sense, performance indicators are not only intended to contribute to the continuous improvement of the performance of the supply chain, but also to control the business and competitive strategy." [10]

There are many ways and KPIs to measure performance of SCM. The number of different measures is overwhelming and very often it depends on the business or a company, however recently with influence of organizations like Gartner or ASCM there is a level of standardization introduced to this area of SCM. The idea is that by using standard KPIs SCM performance of companies within same sector can be compared and best SCM can be identified and used as a benchmark for others.

[10] The Quintessence of Supply Chain Management; What You Really Need to Know to Manage Your Processes in Procurement, Manufacturing, Warehousing and Logistics, page 61; Rolf G. Poluha.

For the purpose of bringing to life the idea behind SCM KPIs we can use ASCM's SCOR model.

SCOR model "describes the business activities associated with satisfying a customer's demand, which include plan, source, make, deliver, return, and enable. Use of the model includes analyzing the current state of a company's processes and goals, quantifying operational performance, and comparing company performance to benchmark data. SCOR has developed a set of metrics for supply chain performance, and ASCM members have formed industry groups to collect best practices information that companies can use to evaluate their supply chain performance."[11]

As we can read on ASCM website[12] "there are over 250 SCOR metrics that are organized in a hierarchical (and codified) structure from organization level 1 to process level 2 to diagnostic level 3. The metrics are categorized in five performance attributes: reliability, responsiveness, agility, costs and asset management efficiency. The first three attributes are considered customer-focused; the latter two are internally focused."

Naturally there is no point focusing on all KPIs[13] taking into account available data (more on that in Chapter 3). Instead we will focus on one – Perfect Order Fulfillment which directly translates into service for final customer.

Perfect Order Fulfillment is defined as "a measure of organization's ability to deliver a perfect order, which is and order in which the 'seven Rs' are satisfied: the right product, the right quantity, the right condition, the right place, the right time, the right customer and the right cost".[14]

To simplify this measure we will focus on its completeness part assuming all other conditions are met. In other words will focus on stock availability for Retailer (or rather its antonym, stock-outs) and all details will be explained in Chapter 3.

[11] APICS Dictionary. The essential supply chain reference. Sixteenth edition.
[12] http://www.apics.org/apics-for-business/benchmarking/scormark-process/scor-metrics
[13] KPIs are used interchangeably with metrics or measures.
[14] APICS Dictionary. The essential supply chain reference. Sixteenth edition.

2. Overview of the Company and Data

2.1. Brief overview of the company and its SCM

The subject of this thesis is SCM of a leading manufacturer from FMCG[16] sector. It is one of the Fortune 500 companies with revenue of over ten billion USD. Distribution network of the company consists of several producing plants and distribution centers (global, located close to key manufacturing sites which supply all regions and local, supplying specific countries). There are over 250 suppliers of packaging and raw materials located on each continent and their transit time to producing plants varies from 1 day to 6 weeks.

For the purpose of this research (and due to limited data availability) we will only be look into part of the global network: 5 production sites, 2 rework facilities, 1 central (global) DC and 3 local (in-country) DCs and therefore a subset of full data. Such significant reduction in analyzed data should not be of any concern as many SCM mechanisms are similar across various locations and selected SC nodes provide good representation in terms of variability of supply chain and manufacturing variables.
The number of suppliers (unknown in subset of the data) is irrelevant either as instead we will focus on availability of the leading component used for production.

The simplified scheme of the supply network is presented in Figure 2. Please note that not all above mentioned SC nodes have been included as the main purpose, similar to Chapter 1 is to visualize information and material flows and, additionally, a list of available variables.

[16] FMCG – fast moving consumer goods.

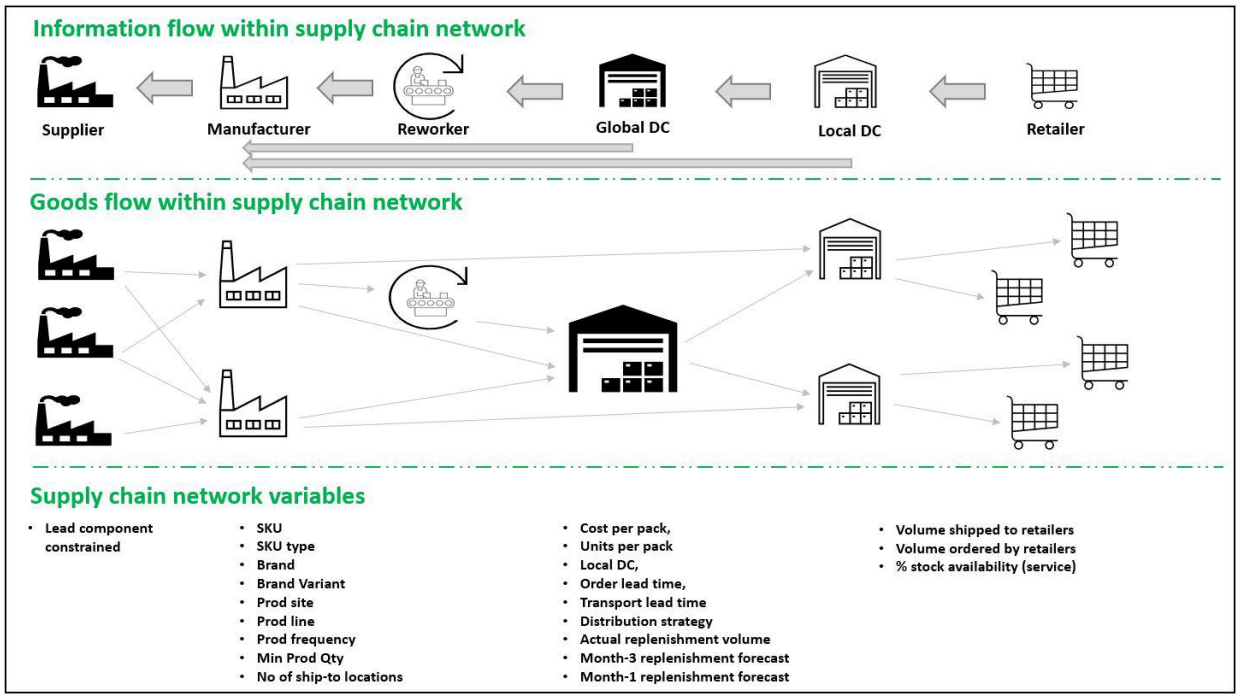

Figure 2: Scheme of company's information and goods flow within supply network with a list of available variables corresponding to relevant node in the network.

Information and material flows have already been briefly described in Chapter 1, however it is time to fully explain this concept. It is also a good opportunity to introduce a few more definitions from SCM to familiarize the reader with SCM jargon.

Information stream (commonly called demand signal) that triggers all SCM activities flows from retailers through all upstream nodes to suppliers. All nodes are connected through ERP system[17] and all nodes excluding retailers and suppliers are connected through Advanced Planning System (APS).

Local DCs submit 104-week statistical sales forecast on the basis of sales orders history and all commercial activities informed by retailers in advanced. That forecast is then converted in APS into replenishment signal at global DC, rework facilities or manufacturing sites depending on DRP[18], replenishment type (direct dispatch or consolidation) or manufacturing process (product requires to be reworked or not).

[17] ERP (Enterprise Resource Planning) system - framework for organizing, defining, and standardizing the business processes necessary to effectively plan and control an organization so the organization can use its internal knowledge to seek external advantage. An ERP system provides extensive databanks of information including master file records, repositories of cost and sales, financial detail, analysis of product and customer hierarchies, and historic and current transactional data; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

[18] DRP (Distribution Requirements Planning) - The function of determining the need to replenish inventory at branch warehouses. A time-phased order point approach is used where the planned orders at the branch

Replenishment signal from global DC and rework facilities triggers production plan at manufacturing site which is then cascaded through Bill of Materials[19] (BOM) explosion to material planning and suppliers.

All these activities respect total lead time[20] made of ordering lead time and planning horizons[21] set up at individual nodes (and reflected in APS).

The role of SCM is to offer best possible service to retailers which is as described in Chapter 1 as Perfect Order Fulfillment.

Goods (materials) flow in opposite direction (from Suppliers to Retailers). All nodes are connected through transportation lanes in APS, and one of the key attributes of each transportation lane is transit (transport) lead time[22].

One last thing worth mentioning to increase SCM awareness of the reader and, to emphasize in the context of the research, is that parameters like transit lead time, planning horizon create sets of SCM rules which in fact constrain supply chain and need to be taken into consideration by planners. Colloquially speaking, we can consider them as 'rules of the game' which objective is to deliver best service (Perfect Order Fulfillment) at acceptable cost to maximize profits.

Naturally, there are more parameters like these. Supply chain data collected for the purpose of this research represents key supply chain variables (parameters) used in planning and manufacturing process. All variables and their meaning will be introduced in Chapter 2.2.

---

warehouse level are "exploded" via MRP logic to become gross requirements of the supplying source; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

[19] Bill Of Materials (BOM) - 1) A listing of all the subassemblies, intermediates, parts, and raw materials that go into a parent assembly, showing the quantity of each required to make an assembly. It is used in conjunction with the master production schedule to determine the items for which purchase requisitions and production orders must be released. A variety of display formats exists for bills of material, including the single-level bill of material, indented bill of material, modular (planning) bill of material, transient bill of material, matrix bill of material, and costed bill of material. 2) A list of all the materials needed by a contract manufacturer to make one production run of a product's piece parts/ components for its customers. The bill of material may also be called the formula, recipe, or ingredients list in certain process industries; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

[20] lead time – 1) A span of time required to perform a process (or series of operations); APICS Dictionary. The essential supply chain reference. Sixteenth edition.

[21] planning horizon - a point in time denoted in the planning horizon of the master scheduling process that marks a boundary inside of which changes to the schedule may adversely affect component schedules, capacity plans, customer deliveries, and cost. Outside the planning time fence, customer orders can be booked and changes to the master schedule can be made within the constraints of the production plan. Changes inside the planning time fence must be made manually by the master scheduler. Syn: planning fence. See: cumulative lead time, demand time fence, firm planned order, planned order, planning horizon, time fence; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

[22] transit lead time - The time between the date of shipment (at the shipping point) and the date of receipt (at the receiver's dock); APICS Dictionary. The essential supply chain reference. Sixteenth edition.

2.2. Supply Chain Data

2.2.1. Source and Structure of Data

In this chapter we will take a closer look at available data collected for one fiscal year and explain key challenges data-related challenges. Some of them can be considered as universal while other are 'research specific'. Both cover typical issues for a data scientist who needs to rely on data quality.

First concern related to data is that it is scattered across 8 different source files (5 .csv files, 2 .xlsx files and 1 .txt file).
Type of files and brief explanation of their content is provided below:

- cost_of_goods.csv – provide details of cost of SKUs based on amount of direct materials, direct labor, and allocated overhead associated with the products sold during a fiscal year (in scope of this research); source – ERP system,
- orders_type.csv – information about logistics strategy applied to each SKU on the basis of orders from one fiscal year,
- prod_parameters.csv – key manufacturing parameters for each SKU set up in ERP system,
- stock_availability.csv – details of total quantity ordered and shipped to Retailers from local DCs during the course of one fiscal year; internal report based on BI system,
- Australia_SS.csv – information about Safety Days and Order Lead Time for Australian distribution center,
- forecast_raw.xlsx – time series data for one fiscal year consisting of monthly snapshots of actual replenishment quantity shipped to local DCs and their anticipated replenishment requirements from local DCs taken 3 months and 1 month prior to actual shipments,
- portfolio.csv – supply chain parameters set up in advanced planning system (APS),
- ship_to_count.txt – count of local DCs each SKU is shipped to; internal report.

Whilst having various sources of data is not a major problem (although it adds workload and complexity to data wrangling process) it perfectly reflects actual challenges of analyst who works in a department where various IT systems are used for data maintenance and processing (ERP, APS). Very often the root cause of this is business growth through acquisition and its side effect, namely 'inheritance' of IT infrastructure. Additionally, reporting side of data processing often require customized BI[23] solutions as business requirements tend to exceed standard, pre-defined reports in BI tools.

Another reason, research specific this time, is that:

- data is grouped around specific supply chain nodes and is scattered across SCM components (planning, sourcing, location),
- it is multi-department (cross-functional) data (Supply Chain and Manufacturing) which means its source is aligned to departments' structure and their IT infrastructure.

The solution is data consolidation and standardization of its formatting. Since all analysis and data modeling will be performed in R I will stick to 'tidyverse' [24] principles which aim to organize it into 'tidy data'.

"There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell."[25]

Figure 3[26] presents graphic representation of these rules. The practicalities of tidy data approach from data manipulation point of view usually mean changing data layout from wide to long (or the other way around, however it is less common).

Above three rules, could be also supplemented with additional rule that each observation needs to be complete.

[23] BI – Business Intelligence.
[24] tidyverse – a group of R packages mostly developed by Hadley Wickham supporting the concept of 'tidy data'.
[25] R for Data Science, page 149; Hadley Wickham, Garrett Grolemund.
[26] R for Data Science, page 149; Hadley Wickham, Garrett Grolemund.

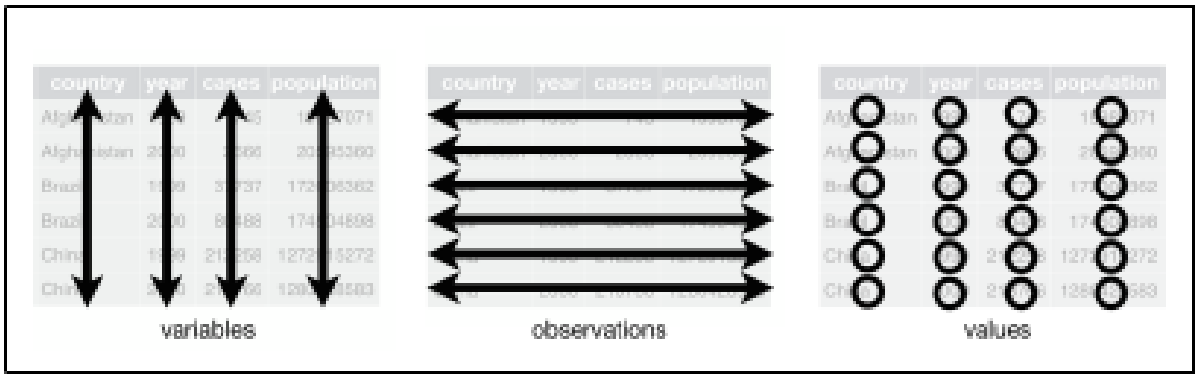

Figure 3: The following rules make a data set tidy: variables are in columns, observations are in rows, and values are in cells[27].

The outcome of first data consolidation (all source file excluding replenishment forecast – more on that in Chapter 2.2.2) is 990 observations and 19 variables.

Variables name simplification and alignment to naming convention standards (snake case[28] alike[29]) required renaming some of the variables so the reader should not be surprised seeing a change in variables names listed in Figure 2.

In the next sub-chapter we will take a closer look at each variable and its meaning.

2.2.2. Variables and Their Meaning

All the variables collected for this research span over key supply chain and manufacturing parameters, which as explained earlier can be considered as 'rules of SCM game' and represent specific planning and operational constraints.

Full list of variables and their meaning[30] is as follows:

- Lead_comp_constr (lead component constrained) – indication if most important (leading) component of BOM used in production process was in short supply[31] at any time during the fiscal year (due to supplier manufacturing capacity constraints, commodity availability etc.),

[27] R for Data Science, page 149; Hadley Wickham, Garrett Grolemund.
[28] Snake case refers to naming style in which each space is replaced by an underscore character and the first letter of each word is written in lowercase.
[29] Snake case standard is not fully respected, as not all first letters of each word are written in lowercase.
[30] As some of the names of variables changed (as explained in Chapter 2.2.1) previous names are captured when necessary to guide the reader through the changes.
[31] Short or long term unavailability of Material (in this context Packaging used in production process).

- SKU[32] – unique code of a single product at a specific location,
- SKU_reworked (SKU type) – indication if product require any post-production rework at 3rd party or internally before shipping to Retailer,
- Brand - Product's Brand name, highest aggregation level in Products hierarchy which disregard its technical specification described in BOM,
- Brand_Variant (Brand Variant) – intermediate aggregation level for SKUs based on Brand and Product's technical characteristics of packaging configuration (like Units per pack) or Raw Materials (like flavor),
- Prod_site (Production site) – name of manufacturing site[33] which is part of company's manufacturing infrastructure (replenishing inventory at global or local distribution centers,
- Prod_line - code of production line (resource) [34] on which SKU is manufactured,
- Number_of_periods (Prod frequency) – time interval (in weeks) between productions of a SKU; it can be used to calculate the number of productions per year (frequency of 2 weeks means 26 productions a year),
- Min_Lot_Size_base_unit (Min Prod Qty) – minimum production batch in base unit of measure (unit) of a Product or Brand Variant defines by physical or technological constraints of manufacturing process,
- No_of_ship_to_markets (No of ship-to locations) – number of local DCs product is shipped to; SKUs shipped to 1 local DC are 'Market Specific' and those to more than 1 local DCs are 'General Export' (or 'Genex' for short),
- Cost_per_pack – cost of goods sold (COGS)[35] of SKU in US dollars per pack,
- Units_per_pack – number of units (of SKU) per pack,
- Order_LT (Order lead time) – time duration which indicates how many days in advance customer order needs to be placed by local DC (before actual dispatch date

[32] APICS Dictionary defines SKU as stock keeping unit, which is a Material (code) at a specific location; we will simplify this terminology and in the thesis term SKU will be used interchangeably with Material or Product (without the location part) which focuses on its physical components (Bill of Materials) which differentiate it from other Products.
[33] Manufacturing site and Producing Plant are considered synonyms and are used interchangeably.
[34] Production line and Resource are considered synonyms and are used interchangeably.
[35] cost of goods sold (COGS) - an accounting classification useful for determining the amount of direct materials, direct labor, and allocated overhead associated with the products sold during a given period of time; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

from manufacturer site or global DC); order lead time is defined in SLA[36] between the company and retailers,

- Transport_LT (Transport lead time) – time (in days) between shipping date (at global DC or Producing Plant) and the date of receipt (at local DC),
- Market (Local DC) – geographical location of distribution center used to store inventory and replenish requirements of Retailers; there are three Markets in scope of this research: Australia, UK and Spain,
- Min_order_qty – An order quantity modifier, applied after the lot size has been calculated, that increases the order quantity to a pre-established minimum[37]; UoM[38] is pack,
- Safety_Days – the quantity of buffer inventory in local DC (Market) to protect against fluctuations in demand or supply measured in quantity of demand in daily buckets,
- Order_Type (Distribution strategy) – defines shipping rules and shipping point to local DC:
    - DPL - replenishment from production site directly after production (1 product per order),
    - CPL - consolidated dispatch from production site (2 or more products per shipment),
    - DNPL - replenishment from global DC (1 product per shipment); additional SCM comes from Safety Stocks kept at global DC),
    - CNPL - consolidated shipment from global DC (multiple products per shipment),
- Shipped_to_Customer (Volume shipped to retailers) - annual quantity (in packs) shipped from local DC to retailers,
- Ordered_by_Customer (Volume ordered by retailers) - annual quantity (in packs) ordered by retailers aggregated to local DC,
- perc_SA (% stock availability[39]) - fulfillment of retailers orders calculated on single SKU level as (Volume shipped to retailers / Volume ordered by retailers) * 100.

[36] SLA (Service Level Agreement) – a document that represents the terms of performance for organic support; APICS Dictionary. The essential supply chain reference. Sixteenth edition.
[37] APICS Dictionary. The essential supply chain reference. Sixteenth edition.
[38] UoM – Unit of Measure.
[39] This measure can be considered as part of Perfect Order Fulfillment measure focusing on right quantity aspect (and hence Products availability); the scenario of shipping too much volume is not considered in the research.

The data that has been consolidated so far does not include time series data from 'raw_forecast.xlsx' file with below variables:

- Actual replenishment – snapshot of monthly volume per SKU shipped from a manufacturer or Global DC to a local DCs (Market); UoM is pack,
- Month-3 replenishment forecast – monthly snapshot of replenishment requirements of a local DC taken three months in advance of actual replenishment request; as explained in Chapter 1 replenishment requirements triggers master production plan and procurement of packaging and raw materials); UoM is pack,
- Month-1 replenishment forecast – monthly snapshot of replenishment requirements of a local DC taken one month in advance of actual replenishment request; this snapshot is a basis for final production plan and packaging and raw materials call off for production; UoM is pack.

This fact requires explanation and is linked with the research aim which will be the scope of Chapter 3.

First of all, it is important to understand that this data represents two things:

- actual stock movement quantity from producing plants, rework plants and global DC to local DCs (Markets),
- anticipated quantity of stock movement from producing plants, rework plants and global DC to local DCs (Markets) which on one hand triggers planning and manufacturing activities and on the other hand (if incorrect) might have negative impact on inventory level at local DC and therefore availability of stock for Retailers (represented by perc_SA variable).

With that in mind it becomes obvious to investigate the quality of this anticipation (forecast) based on the actual quantity and (potentially) its correlation with perc_SA (more on that in Chapter 3).

Forecast quality can be measured with some standard measures like WAPE[40] or BIAS[41] but since both measures will annualized (to align with perc_SA as this measure is only available for full fiscal year, not in monthly buckets), we will create a few other measures to capture time series nature of initial variables. Additionally, Total Replenishment will be converted

[40] WAPE – Weighted Average Percentage Error.

[41] BIAS - A consistent deviation from the mean in one direction (high or low). A normal property of a good forecast is that it is not biased; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

to sum of monthly quantities to match Shipped_to_Customer and Ordered_by_Customer variables (captured at local DCs). What in fact it means is for each SKU additional 'synthetic' variables will be created on the basis of original ones which then can be consolidated with the rest of the dataset.

Additional (synthetic) measures with their description:

- Total_replenishment – sum of Actual Replenishment quantity per SKU shipped from a manufacturer or Global DC to a local DCs (Market); UoM is pack,
- WAPE_Month_1 - Weighted Average Percentage Error based on Actual Replenishment and Month-1 replenishment forecast; this measure takes values between 0 and 100 (%), is aggregated to full fiscal year and weighted by percentage share of actual replenishment quantity of the month in Total_replenishment,
- WAPE_Month_3 - Weighted Average Percentage Error based on Actual Replenishment and Month-3 replenishment forecast; this measure takes values between 0 and 100 (%), is aggregated to full fiscal year and weighted by percentage share of actual replenishment quantity of the month in Total_replenishment,
- BIAS_Month_1 - cumulative difference for full fiscal year between Actual replenishment and Month-1 replenishment forecast quantity; UoM is pack,
- BIAS_Month_3 - cumulative difference for full fiscal year between Actual replenishment and Month_3 replenishment forecast quantity; UoM is pack,
- Count_positive_BIAS_Month_1 - number of months within fiscal year when Month_1 replenishment forecast quantity was lower than Actual replenishment quantity,
- Count_positive_BIAS_Month_3 - number of months within fiscal year when Month_3 replenishment forecast quantity was lower than Actual replenishment quantity,
- Count_Month_1_FA_zero - number of months within one fiscal year when WAPE_Month_1 was 0 (no replenishment forecast or BIAS_Month_1 quantity exceeded Actual replenishment quantity); measure takes values between 0 and 12,
- Count_Month_3_FA_zero - number of months within one fiscal year when WAPE_Month_3 was 0 (no replenishment forecast or BIAS_Month_1 quantity exceeded Actual replenishment quantity); measures takes values between 0 and 12.

Now when data from all source files has been consolidated dataset consist of 30 variables and 990 observations. In the next chapter we will investigate completeness of data.

2.2.3. Missing Data and Data Imputation

Quality of data is one of the key success factors in any data science project and in this chapter we will investigate key attribute of data quality, namely its completeness. At some point we will also take a decision whether to exclude variables from the dataset or impute observations and decide on imputation technique on the basis of variable type.

At this stage, removal of variables should not be mistaken with de-selection of variables for modelling as any decisions regarding that will only be based on completeness rate or expert knowledge. Variables selection for modeling will be covered in Chapter 4. Figure 4 shows a summary of missing data on Market level.

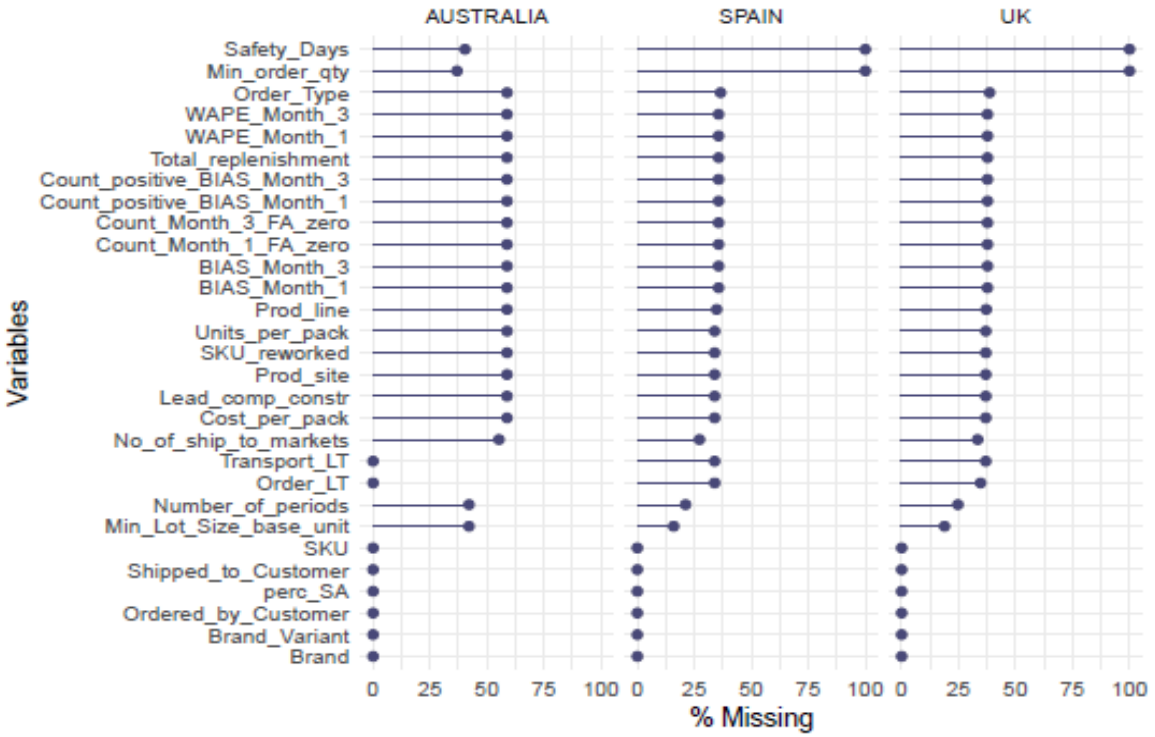

Figure 4: % of missing values per variable by Market. We can clearly see that variables are clustered in line with their source.

First of all, we can see that variables are clustered which can be explained by different sources of data. Secondly, there is similar completeness rate of variables within each cluster. Thirdly, overall fraction of missing data is high, which is worrying. There are only seven variables (including Market) without missing observations. What is more, Safety_Days and Min_order_qty are only available for Australia, and even for this Market fraction of missing

data is high (approx. 40%). This indicates that already at this stage both variables need to be removed from dataset as there is no possibility for any kind of data imputation. Finally, general conclusion from this summary is that overall data completeness rate is low (there is only 48% of complete observations). The solution to this is data imputation, otherwise observations with incomplete data need to be excluded.

Before we proceed with data imputation, let's identify potential root cause of this state on the basis of personal experience and facts that we have known so far:

- it's multi-department data coming from many source systems that is scattered across eight different files so we cannot exclude human error in data gathering process; indirect root cause of this is that departments in the company operate in silos, which also means lack of common planning and reporting platform or standardized data management process,
- surprisingly, some key planning parameters are not set up in ERP or APS as data maintenance responsibilities within planning community are often neglected and there is no governance to control that or reports to bring this issue to life for management; this can be driven by company's culture and resulting in low data literacy or low awareness of data importance.

Data imputation process need to be aligned with type of variable – numeric and categorical.

"A numeric variable is one whose observations are naturally recorded as numbers. There are two types of numeric variables: continuous and discrete. A continuous variable can be recorded as any value in some interval, up to any number of decimals (...) A discrete variable, on the other hand, may take on only distinct numeric values—and if the range is restricted, then the number of possible values is finite. (...) Like some discrete variables, categorical variables may take only one of a finite number of possibilities. Unlike discrete variables, however, categorical observations are not always recorded as numeric values. There are two types of categorical variables. Those that cannot be logically ranked are called nominal. (...) Categorical variables that can be naturally ranked are called ordinal."[42]

Having that in mind let's review imputation options for variables in scope of this research. Let's also take into consideration known interdependencies between variables on the basis of expert knowledge and knowledge of SCM processes. The sequence of

[42] The Book of R, page 242; Tilman M. Davies.

imputation can also play a significant role and again, this is where expert knowledge plays invaluable role.

There are a couple of variables where imputation should be straightforward:

- Order_LT is consistent for Prod_site and Market so we can replace missing data with correct values,
- For Number_of_periods and Min_Lot_Size_base_unit we can take assumption that all 0 and NA values can be changed to 1 as all these values generate the same output in APS. This simply means that SKUs can be produced every week and there is no physical constraints regarding the size of production batch; from that point of view these two parameters simply do not need to be maintained in APS other than for data tidiness reason.

Data imputation of remaining variables needs to be done in a specific order to capture interdependencies between them. The order is based on below set of rules:

- Prod_line, Lead_comp_constr, No_of_ship_to_markets and Cost_per_pack can be imputed on the basis of Brand_Variant or Brand (in that order),
- Prod_site, SKU_reworked and Order_Type can be imputed on the basis of Prod_line,
- Transport_LT and Order_Type can be imputed on the basis of Prod_site,
- Units_per_pack can be imputed on the basis of Brand_Variant and Prod_line,
- All variables related to replenishment forecast accuracy can be imputed on the basis of Brand_Variant and Market (to mirror the process of submission of statistical sales forecast),
- Total_replenishment will be imputed on the basis of Shipped_to_Customer, Ordered_by_Customer and Markets variables.

In addition to above rules and order, imputation of qualitative variables has been done using mode, numeric discrete variables using median, and imputation of numeric continuous variables using mean or linear regression.

Important aspect of data imputation is its impact on distribution, however R is equipped with data imputation packages[43] which also allow visual (post-imputation) inspection of imputation outcome and comparison to distribution of complete observations.

[43] List of all R packages used in this research is provided in Appendix.

Example of such graphical inspection is presented in Figure 5.

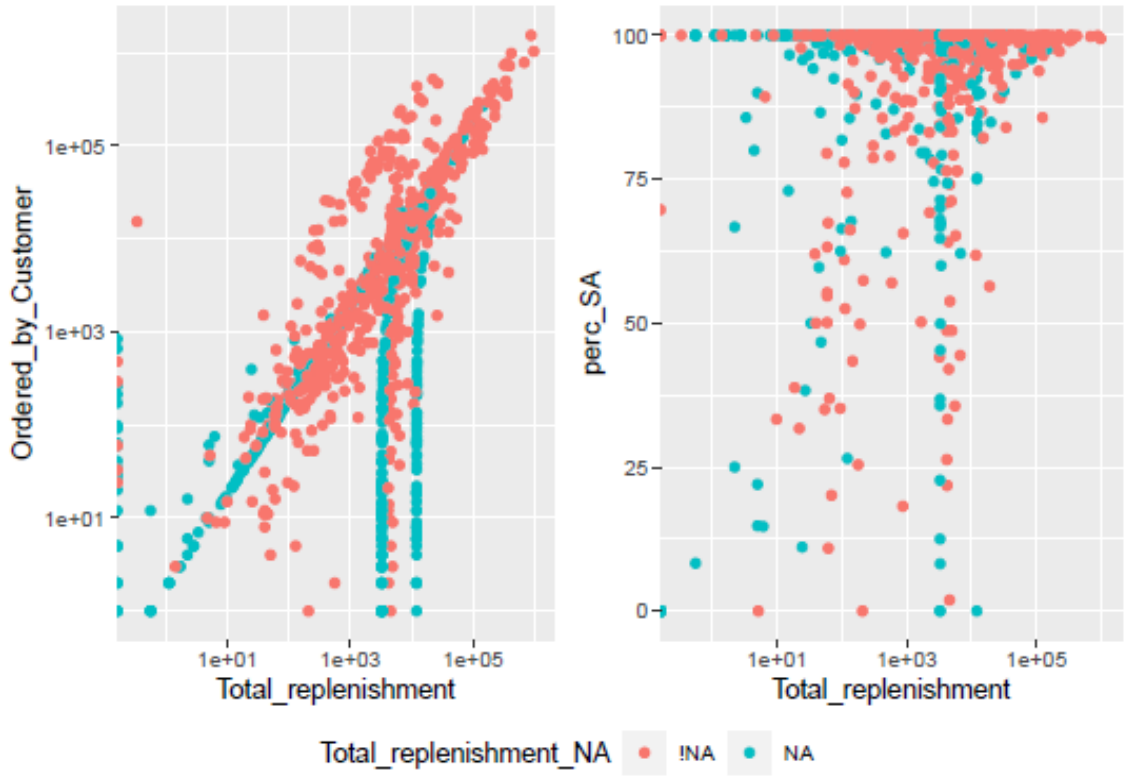

Figure 5: Imputation results for Total_replenishment variables plotted together with Ordered_by_Customer variable (left-hand side) and perc_SA variable (right-hand side). Imputed observations (NA) are blue and complete (!NA) are red. Position of imputed observations mirror position of complete observations in both plots suggesting that overall, data imputation using linear regression has been successful and there is no major impact on distribution of Total_replenishment variable.

Now that imputation process is complete, let's visualize its outcome.

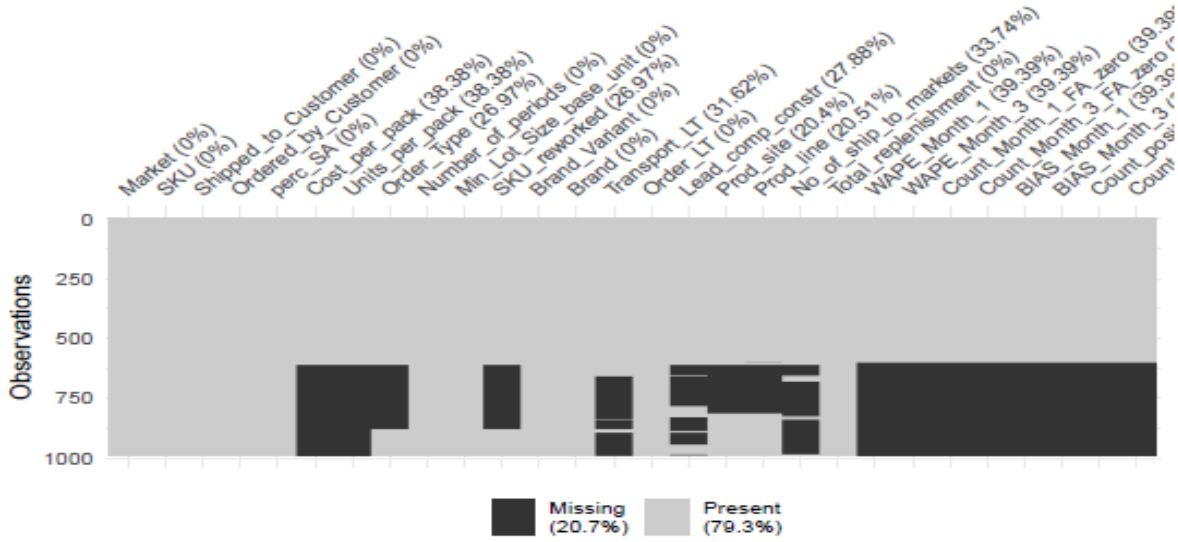

Figure 6: Summary of data imputation process. Although completeness rate has been improved with data imputation, there is still 20.7% observations with missing values. % values in the top show % rate of missing observations for each variable.

Positive result of imputation process is that the rate of missing data reduced from 48% to 20.7%, however the bad news is that after excluding missing data final number of complete observations reduced from 990 to 600. Additionally, the number of variables reduce from 30 to 28.

Summary of data imputation scale by variable is presented in Figure 6.

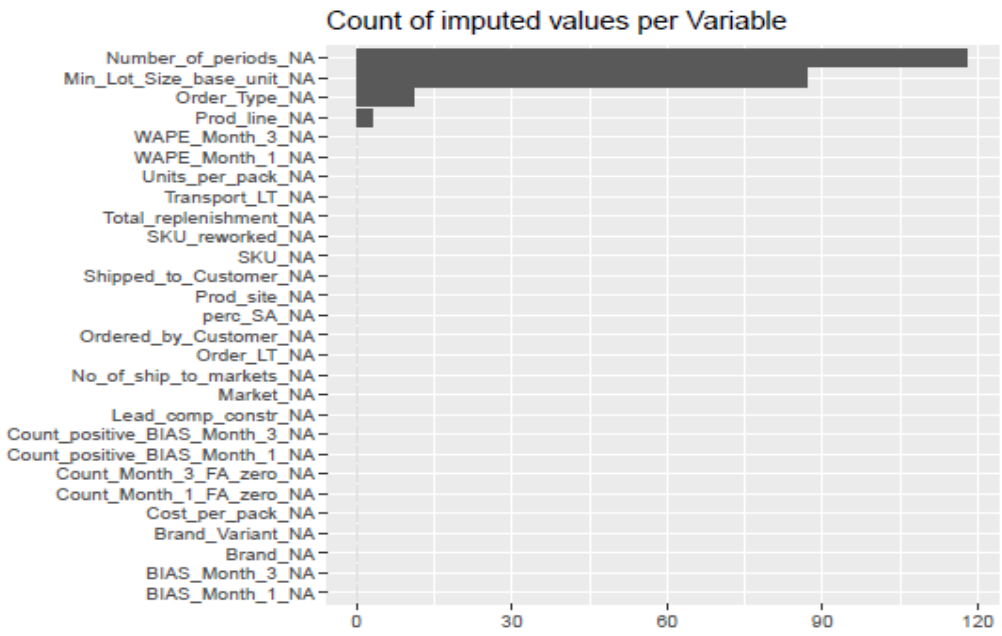

Figure 7: Summary of imputation by variables. Please note that imputation for none of replenishment accuracy measures could be performed and together with 390 removed observations, all imputed observations for Total_replenishment have been removed which is a good evidence of how much workload can go down the drain due to high rate of data incompleteness. It also means that final imputation rate per variable is not known until full imputation process will have been completed.

This outcome shows a great importance of data maintenance and challenges for data scientists driven by data incompleteness. Although they are equipped with better and better tools, sometimes imputation is simply impossible and a decision need to be made to remove observations to keep the right balance between data completeness and data accuracy.

A few minor remaining data wrangling activities will be tackled in Chapter 2.2.4.

2.2.4. Final fine-tuning of Data

There is still a few final checks that needs to be made before we can close data preparation subject down. We will:

- check units of measure consistency across all numeric variables related to quantity,
- inspect categorical variables to:
  - invest number of levels per factor[44],
  - check if there are any ordinal variables that need to be recoded
- investigate if other fine-tuning of data is required.

Among all numeric variables describing quantity, all but one use pack as a unit of measure. Min_Lot_Size_base_unit is in units so let's convert it using Units_per_pack variable to reflect packs and align it with other variables. New variable will be called Min_Lot_Size and Min_Lot_Size_base_unit can be dropped from the final dataset.

To investigate ordinals we need a bit of expert knowledge. Looking at a list of factors Order_Type can definitely be classified as ordinal as its four levels can be ordered from least (DPL) to most expensive (CNPL) from logistics cost point of view.
Let's reflect this feature and recode this variable to set its level in that order.

Next step is to check is the number of levels per factor for categorical variables. We can visualize it as shown in Figure 8.

[44] "Factors are R's most natural way of representing data points that fit in only one of a finite number of distinct categories, rather than belonging to a continuum"; The Book of R, page 79; Tilman M. Davies.

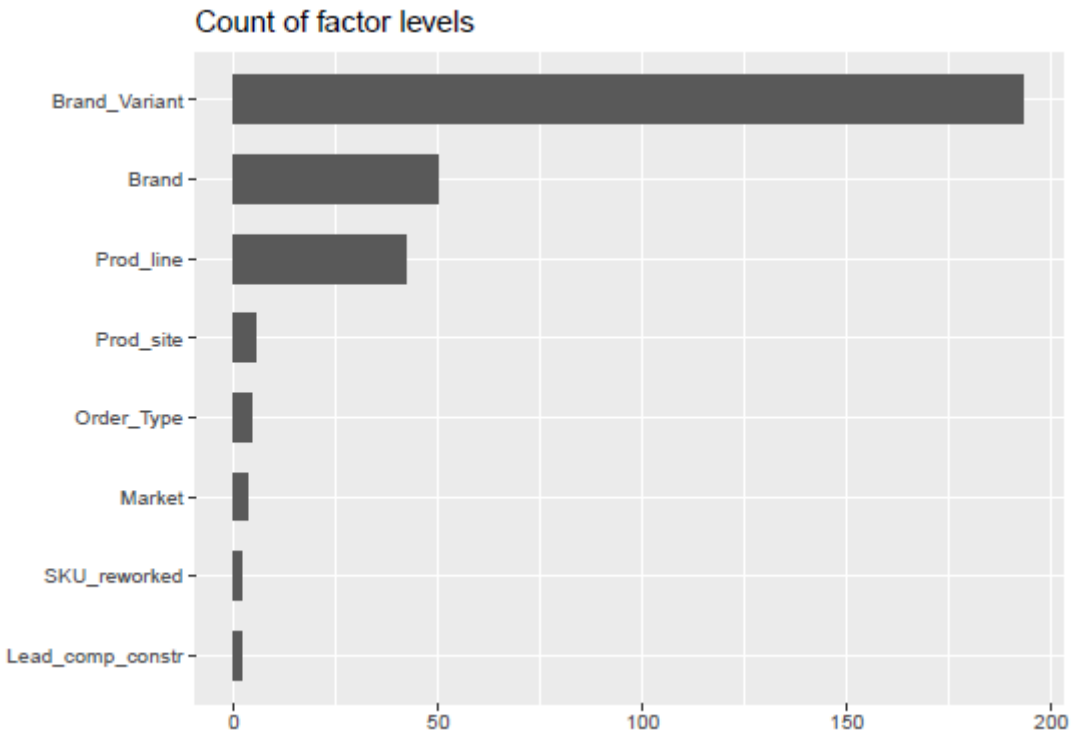

Figure 8: Visual representation of number of levels per factor. Variables with high number of levels can be considered as 'problematic' from modeling point of view.

Taking into consideration number of levels Brand_Variant, Brand and Prod_line variables can be 'problematic' for below steps and issues related to modeling:

- data split for train and test sets (namely, keeping similar balance of levels in both sets),
- cross-validation,
- overfitting the data [45].

Figure 9 presents distribution of these three 'problematic' variables. We can see that there are many levels with unique observation. What we can do in this situation is to group levels with low number of observations. At this point it is important to mention that with this we are entering a 'gray area' as we can only use arbitrary threshold (three observations) to do that. Alternatively we could remove these variable from our dataset but in my opinion it is too early to do that, especially that at later stage we will be selecting variables for modeling using proper statistical tools, tests and measures.

[45] Overfitting the data essentially means they (models – author's footnote) follow the errors, or noise, to closely; An Introduction to Statistical Learning with Applications in R, page 22; G. James, D. Witten, T. Hastie, R. Tbishirani.

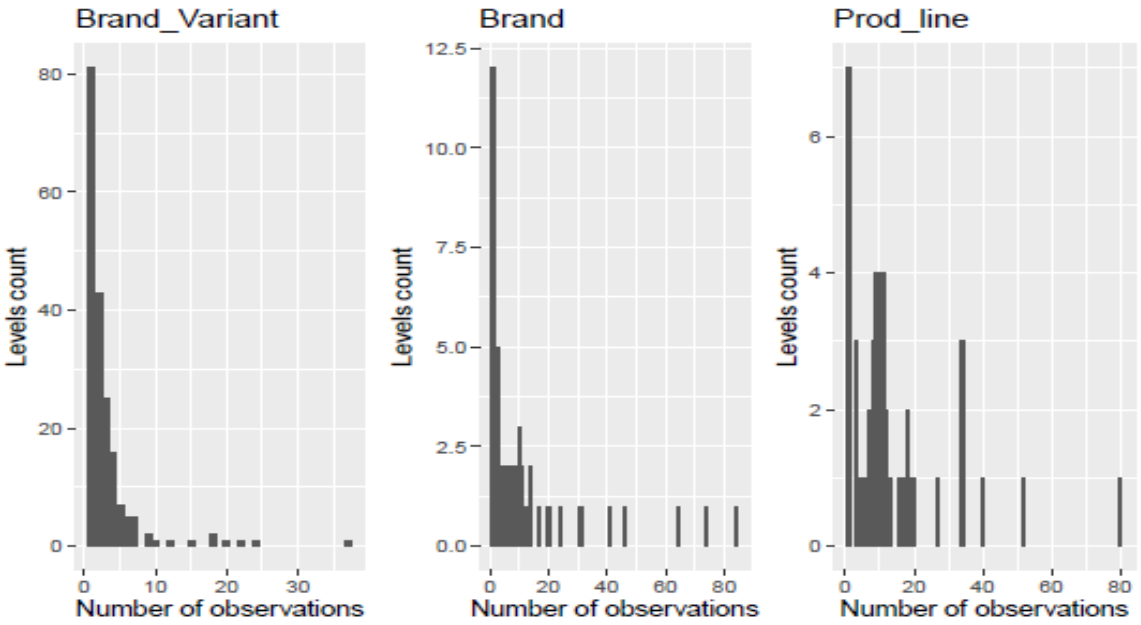

Figure 9: Number of observations per factor level for Brand_Variant, Brand and Prod_line. In extreme situation (left-hand side) there are over eighty levels that hold unique observation.

Figure 10 shows number of levels per factor after levels grouping have been done (for multi-level variables). Please note that corrected variables have '_g' suffix to distinguish them from original variables.

We can see that, although situation improved, the number of levels for those three variables is still significant but let's leave this issue for later.

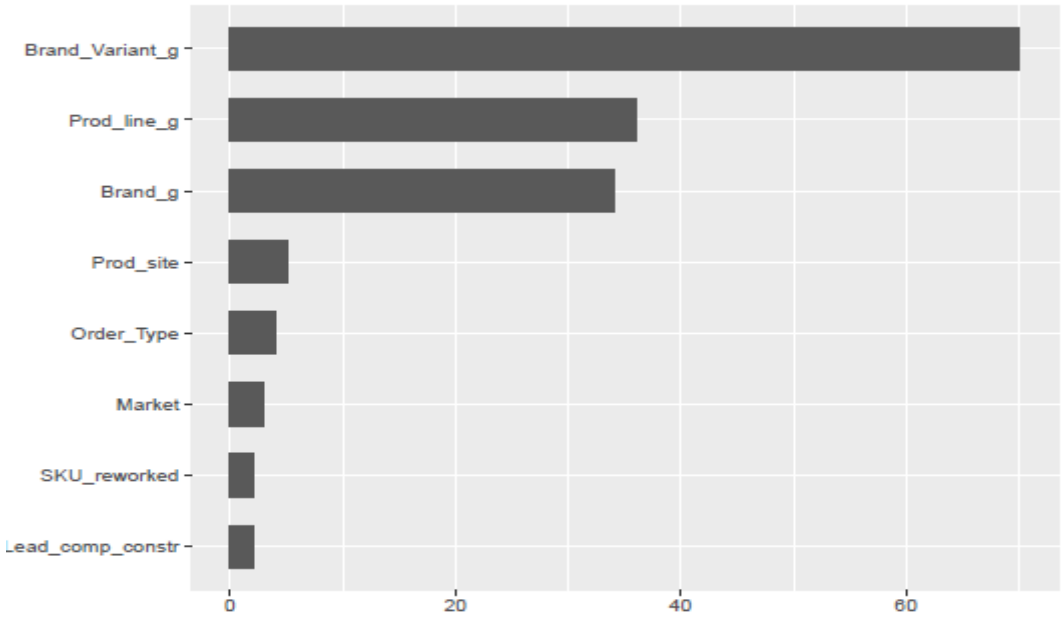

Figure 10: Number of levels per factor post reduction after levels grouping in multi-level variables. Brand_Variant_g variable still has over sixty levels.

Remaining fine-tuning of the dataset includes:

- Recoding No_of_ship_to_markets to Product_type, a factor with 2 levels:
  - market specific - with 1 ship-to market,
  - genex - 2 or more ship-to markets.
- Creating two additional variables on the basis of perc_SA:
  - out_of_stock - factor with 2 levels, Yes - for perc_SA < 1 and No - for perc_SA equals to 1,
  - service_level - ordered factor which cuts perc_SA into 8 thresholds (0 to 25, 25 to 50, 50 to 75, 75 to 90, 90 to 95, 95 to 98.5, 98.5 to 99.5 and 99.5 to 100).
- Removing Min_Lot_Size_base_unit, Brand, Brand_Variant, Prod_line and No_of_ship_to_markets variables from the dataset.

Final dataset consist of 30 variables and 600 observations.

I am conscious we have spent a lot of time on data manipulation, however this should not be a surprise to the reader. There is a 80/20 rule saying that up to 80% of the time in any data science project is needed to tidy up data and within that "80% of the time it is routine and boring, and the 20% of the time it is weird and frustrating."[46].

Good news for us is we have it all complete and we can move on to the most interesting (and fun) part of this thesis.

[46] R for Data Science, preface; Hadley Wickham, Garrett Grolemund.

3. Research Aim and Methodology

At this stage the reader should have good understanding of what Supply Chain (SC) and Supply Chain Management (SCM) are and be familiar with information, service and goods flows (and their direction) in Supply Chain network. The role of SCM and its components should be clear too, as well as the need for measurement of SCM performance. We have also introduced supply chain vocabulary and a wide set of variables which describe physical attributes of SKUs, manufacturing and supply chain constraints or actual and anticipated replenishment quantity for one fiscal year which trigger all planning activities and most operations within SCM. Variables that capture replenishment quantity have been used to introduce 'synthetic' variables that help summarize and measure how well-anticipated actual replenishment quantity had been during the fiscal year.

Equipped with all above information we will research two main areas of SCM, namely performance of SC and portfolio management.

Since both topics are wide the scope of the research will be narrowed to:

1. study of relationship between stock-out[47] (out_of_stock variable) and physical aspects of SKUs, upstream supply chain and manufacturing constraints and quality of demand signal cascaded from local DC (Market) down the supply chain,
2. investigation if stock-outs can be predicted on the basis of available variables,
3. development of portfolio segmentation strategy as an alternative to ABC classification.

The aim of the first research could be boiled down to a question like "does the fact that one of the SKU is produced every week at site_11_line_14 and is directly shipped to local DC after production mean that manufacturer could offer higher service (or guarantee no stock-outs at local DC) than for a SKU produced on a different line, less frequently?" Naturally, we could embed other variables in question like this, but that raises another question. Are all variables relevant? Maybe value of our response variable depends only on some of the predictor variables or none of them?

These kind of questions could be tackled with various methods from Data Science toolbox including Exploratory Data Analysis (EDA) which is used by statisticians to 'interrogate'

[47] Stock-out occurs when Retailer's requirement cannot be fulfilled due to temporary unavailability of a product (SKU) at local DC (Market).

the data and find correlations between variables on the basis of their type and statistical tests and measures. What is more, this can be done visually which is very appealing to end-user. There is also another important benefit from EDA. "EDA is an important part of any data analysis, even if the questions are handed to you on a platter, because you always need to investigate the quality of your data"[49]. That is right, we may come to a conclusion that there is something wrong with our data and we could brainstorm possible reasons and measures to improve its quality and potential results.

Stock-out prediction is a very ambitious task taking into account the nature of supply chain, however we will investigate if such model can be built with Machine Learning[50] and Deep Learning (based on concept of Neural Networks[51]). Both methods will be covered in Chapter 4 where more details and various set of tools embedded in those methods will be uncovered (tailored to specific problem in scope of the research). This will be also a perfect opportunity to touch base on one of key drawback of both methods often referred as 'black boxes' [52] and new and dynamically developing branch of Data Science called XAI[53] which aims for transparency in terms of how models work.

High customization of products results in portfolio growth. The number of SKUs has been increasing and this trend is unlikely to change. This is one of the challenges for SCM called out at in the introductory chapter and aim of the third research.

SC professionals diversify portfolio management efforts in line with ABC classification[54] based 80/20 rule[55]. Although it helps manage workload, this classification method has been used since 1950s to bring structure in times where there were no ERP or APS systems. In

[49] R for Data Science, page 81; Hadley Wickham, Garrett Grolemund.

[50] Machine Learning – "provides a set of tools that use computers to transform data into actionable knowledge"; Machine Learning in R, page 6; Brett Lantz.

[51] "Neural networks use (artificial – author's footnote) neurons defined in this way as building blocks to construct complex models of data"; Machine Learning in R, page 208; Brett Lantz.

[52] In the case of machine learning, the black box is because the underlying models are based on complex mathematical systems and the results are difficult to interpret.; Machine Learning with R, page 205; Brett Lantz.

[53] XAI – Explainable Artificial Intelligence.

[54] ABC classification - The classification of a group of items in decreasing order of annual dollar volume (price multiplied by projected volume) or other criteria. This array is then split into three classes, called A, B, and C. The A group usually represents 10 percent to 20 percent by number of items and 50 percent to 70 percent by projected dollar volume. The next grouping, B, usually represents about 20 percent of the items and about 20 percent of the dollar volume. The C class contains 60 percent to 70 percent of the items and represents about 10 percent to 30 percent of the dollar volume. The ABC principle states that effort and money can be saved through applying looser controls to the low-dollar-volume class items than to the high-dollar-volume class items. The ABC principle is applicable to inventories, purchasing, and sales. Syn: ABC analysis, distribution by value.; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

[55] 80-20–A term referring to the Pareto principle. The principle suggests that most effects come from relatively few causes; that is, 80 percent of the effects (or sales or costs) come from 20 percent of the possible causes (or items).; APICS Dictionary. The essential supply chain reference. Sixteenth edition.

today's world this approach seems to be outdated and even adding XYZ elements to ABC classification (which bring demand variability factor resulting in ABC/XYZ matrix with 9 clusters of products from A/X - most valuable/least volatile to C/Z - least valuable/most volatile) does not change this fact.

The reasons for this are:

- ABC segmentation is based on historic data and from that point of view it does not add any additional information about the product itself (volume, cost and even volatility are known),
- it does not capture physical aspect of goods,
- it does not capture any manufacturing or upstream supply chain constraints,
- it does not clearly answer a question about optimal number of clusters or thresholds used for the split.

With this in mind, in Chapter 5 we will investigate alternative segmentation methodology using Unsupervised Machine Learning[56], evaluate outcome, compare with classic ABC (or at least approximation of ABC classification as we do not have this information in our dataset) and brainstorm pros and cons behind this methodology or its deployment in SCM.

As Unsupervised Machine learning is considered "part of Exploratory Data Analysis"[57] even if we reject this idea we will learn something new about our data.

That last sentence is important as the reader need to understand that it is in the nature of Data Science that not all ideas come to life. This is what H. Wickham writes in 'R for Data Science' book: "During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will hone in on a few particularly productive areas that you will eventually write up and communicate to others." [58].

Although he refers to EDA, this is applicable to Data Science in general.

[56] "Unsupervised Machine Learning task (...) automatically divides the data into clusters, or groupings of similar items. It does this without having been told what the groups should look like ahead of time. As we may not even know what we're looking for, clustering is used for knowledge discovery rather than prediction. It provides an insight into the natural groupings found within data."; Machine Learning with R, Brett Lantz.

[57] An Introduction to Statistical Learning with Applications in R , page 374; G. James, D. Witten, T. Hastie, R. Tbishirani.

[58] R for Data Science, page 81; H. Wickham, Garrett Grolemund.

4. Can Stock-outs Be Predicted?

Stock-out is reported when Retailer's requirement cannot be fulfilled due to temporary unavailability of a product (SKU) at local DC (Market). In other words there is no inventory to cover Retailer's demand represented by order. Stock-out in our dataset is represented by out_of_stock variable which has two levels: 'Yes' and 'No'. The variable is calculated on the basis of perc_SA which indicates percentage of quantity shipped to Retailers out of total quantity ordered by Retailers over the course of one fiscal year. If perc_SA equals to 100% then out_of_stock is 'No' and 'Yes' otherwise.

Both variables are closely related to the concept of Perfect Order Fulfillment which has been explained in Chapter 1.3. As a reminder, Perfect Order is based on 'seven Rs', one of which is right quantity and it is the quantity that links Perfect Order with our response variables. We are not interested in Perfect Order itself although with out_of_stock variable we can indicate which SKUs did not deteriorate it ('No' value of the variable). We are interested to research if there is any relationship between physical features of SKUs or upstream supply chain and manufacturing constraints and SKUs availability for orders placed by Retailers. This way we can bypass important feature of Perfect Order – it may consist of many SKUs and this information is not available in our dataset.

The key response variable in our research will be out_of_stock, however perc_SA will help us define final list of predictors (Chapter 4.1).

4.1. Selection of Predictors

The choice of predictors is an important step of modelling. Selection of variables can be done it two stages:

- pre-modelling, which mostly aims to exclude variables with low correlation to response variable and remove effect of collinearity[59] which can skew the results of a model; this point is valid mostly for linear models,
- modelling/post-modelling, where a model is 'trimmed' on the basis of statistical significance of predictors or penalty parameter; in this case stepwise selection or regularized models[60] can be used.

There are various benefits from reduction of the number of predictors. The key one is simplicity as the model is easier to interpret and maintain (hence costs associated with such model are lower too).

The scope of this chapter is pre-modelling step and we will start with investigation of distribution of perc_SA variable. For this purpose we will convert its values from 0% to 100% to <0; 1> range. Already on the basis of this information we can tell that it is non-normal distribution. Let's log-transform it and evaluate distribution type visually with Frey and Cullen graph.

$$\log\_perc\_SA = \log\left(1 - perc_{SA} + 0.01\right)$$

Results are presented in Figure 11 and our log-transformed perc_SA has beta distribution. Let's take into consideration fact that response variable does not have normal distribution when investigation correlation between response and predictor variables (we will use Spearman's method rather than Pearson's).

[59] Collinearity refers to the situation in which two or more predictor variables are closely related to one another (...) the presence of collinearity can pose problems (...) since it can be difficult to separate out the individual effects of collinear variables on the response.; An introduction to Statistical Learning with R, page 99; G. James, D. Witten, T. Hastie, R. Tbishirani.
[60] Example of a regularized model with be presented Chapter 4.3.

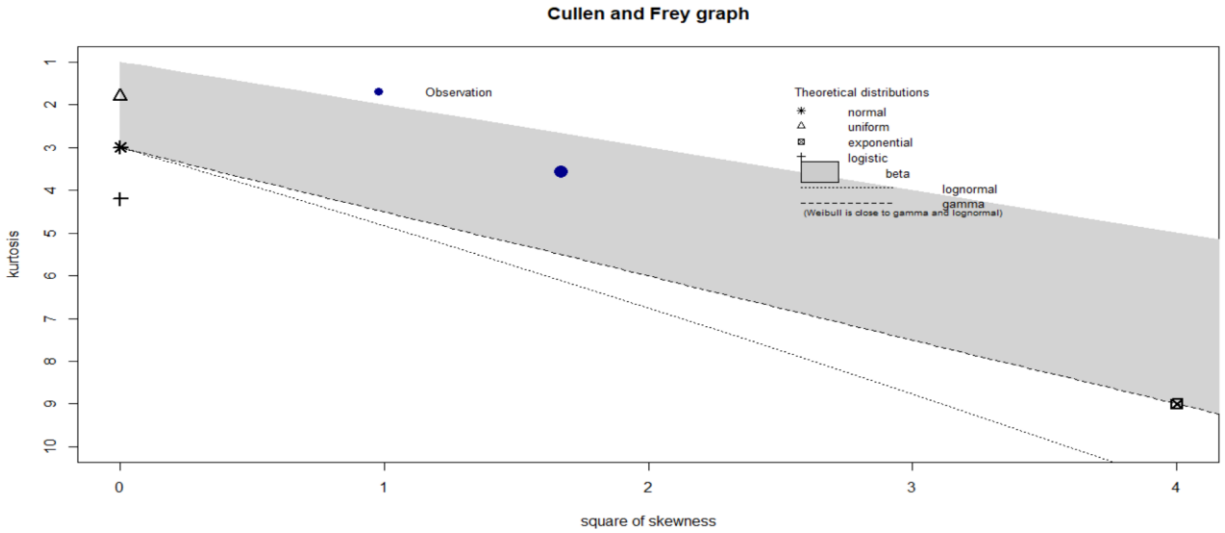

Figure 11: Cullen and Frey graph, on the basis of kurtosis and square of skewness suggests that log-transformed perc_SA has beta distribution.

Correlation and collinearity will be investigated separately for numeric and qualitative variables as different methods and tests are used depending on variable type.

Correlation matrix for numeric variables is presented in Figure 12.

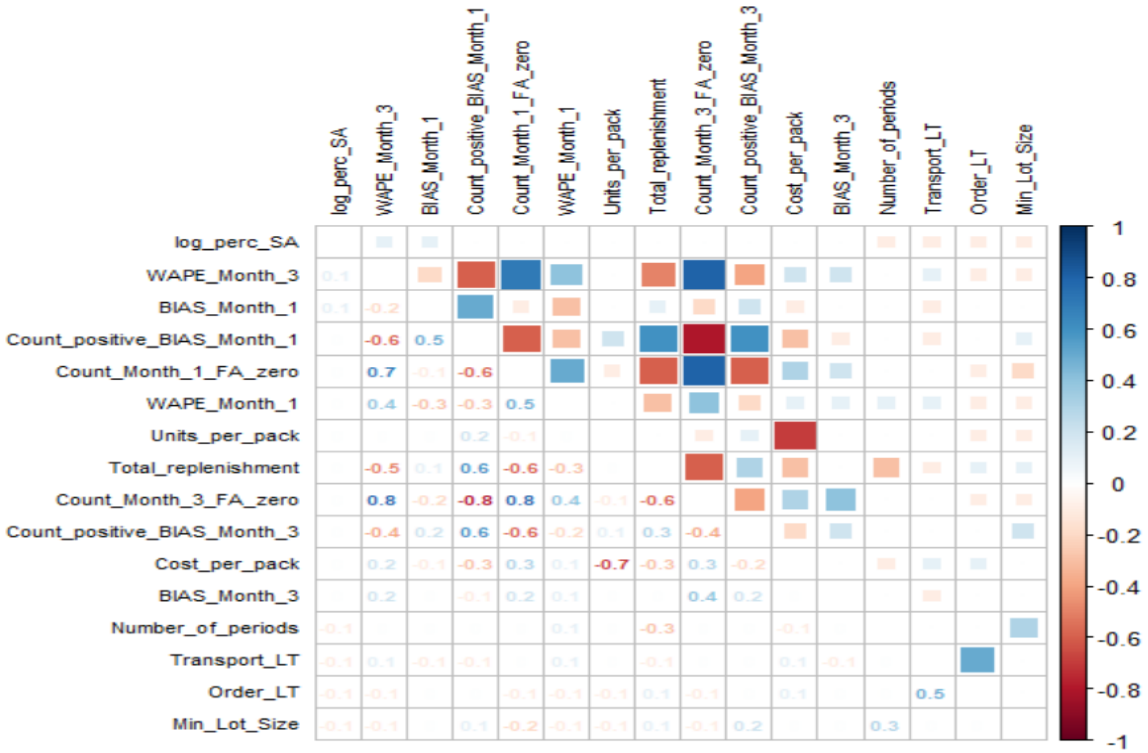

Figure 12: Correlation matrix between response and numeric variables based on Spearman's method. Lower part marked with diagonal line shows correlation in numeric format, upper half in illustrative format. Shades of blue represent positive correlation whereas shades of red negative correlation.

There are two key findings from the correlation matrix for numeric variables:

- there is low or no correlation between log_perc_SA and numeric predictors; it is important to recognize this fact as it can be very challenging to build a stable model of high predictive power on this foundation,
- some predictors are highly correlated which suggests the presence of collinearity effect.

The last point (collinearity effect) needs to be actioned. With arbitrary threshold of 0.7 Count_Month_3_FA_zero predictor should to be removed to reduce pair-wise correlations in our dataset.

Correlation between numeric response and qualitative variables can be investigated with anova[61] test. Let's see if all predictors differentiate the mean of of log_perc_SA.

Variable	AOV results
Replenishment_Type	10.97
Lead_comp_constr	8.44
Prod_site	3.43
SKU_reworked	3.18
Brand_g	2.63
Prod_line_g	1.88
Market	1.75
Brand_Variant_g	1.72
Product_type	1.63

Table 1: Anova test for correlation between response variable and categorical predictors. Outcome suggest that based on Anova test, all variables should be kept.

As we see in Table 1 the answer to this question is 'yes'.

Example of monotonic relationship between log_perc_SA and Replenishment_Type is presented in Figure 13.

[61] Anova (AOV) – Analysis of Variance.

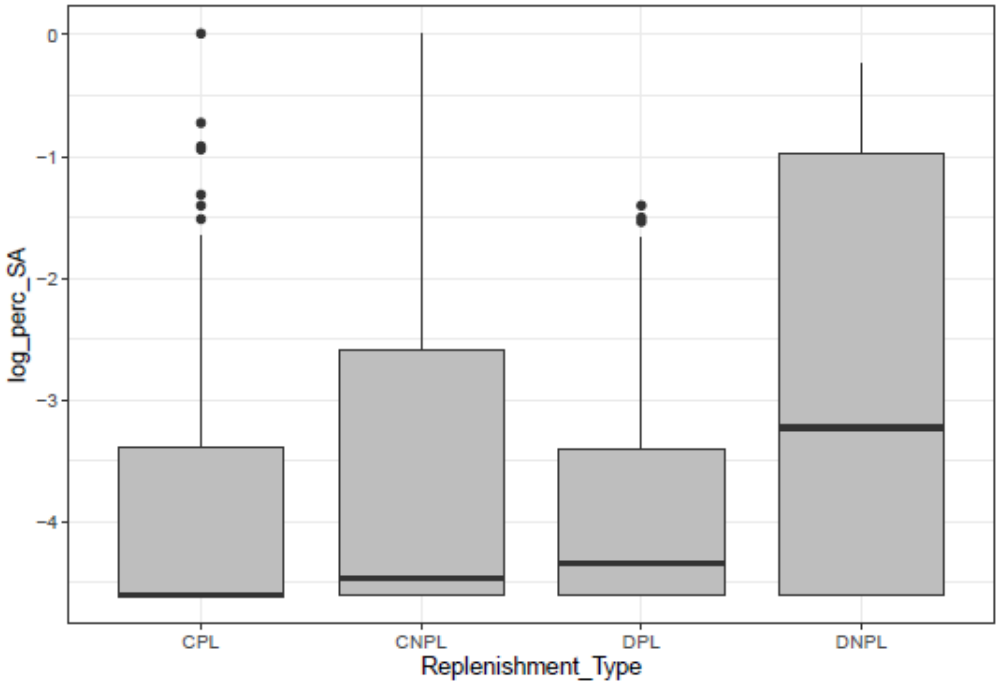

Figure 13: Clear monotonic relationship between log_perc_SA (y-axis) Replenishment_Type (x-axis).

Collinearity between qualitative variables can be done analogically to numeric variables, however their levels need to be first recoded into separate (integer) variables which take two values: 0 and 1. There are ten collinear variables (Figure 14) and as we can see all of them are levels of a 'parent' variable.

```
[1] "MarketSPAIN"                "Brand_Variant_gbrand_S_100"
[3] "Brand_Variant_gbrand_C_1200" "Brand_Variant_gbrand_B_200"
[5] "Brand_Variant_gbrand_P_100"  "Brand_Variant_gbrand_H_101"
[7] "Brand_Variant_gbrand_K_200"  "Prod_line_gsite_13_line_RWK"
[9] "Prod_line_gsite_11_line_33"  "Prod_line_gsite_21_line_12"
```

Figure 14: List of collinear categorical variables. Most of them come from two parent predictors: Brand_Variant and Prod_line.

Collinearity between categorical variables, due to the nature of this test, should not be of concern, however we know that both Prod_line_g and Brand_Variant_g have high number of levels. Both of them also represent a level in Product hierarchy (Brand_Variant_g) and Manufacturing resource hierarchy (Prod_line_g). With this in mind and the outcome of the test we can remove them from the dataset as product and manufacturing resource hierarchy is already represented by Brand_g and Prod_site_g variables.

The last thing we can do, regardless of variable type, is 'near zero variance' analysis. It "diagnoses predictors that have one unique value or predictors that have very few unique values relative to samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large"[62]. In a nutshell those predictors can be removed from modelling as they are non-informative. As per Table 2, we do not have such predictor variables.

variables	freqRatio	percentUnique	zeroVar	nzv
WAPE_Month_3	30.800000	54.0000000	FALSE	FALSE
log_perc_SA	14.105263	25.6666667	FALSE	FALSE
Min_Lot_Size	11.941176	14.5000000	FALSE	FALSE
Order_LT	11.765957	0.3333333	FALSE	FALSE
Lead_comp_constr	9.526316	0.3333333	FALSE	FALSE
SKU_reworked	6.058823	0.3333333	FALSE	FALSE
Brand_Variant_g	4.513514	11.6666667	FALSE	FALSE
WAPE_Month_1	4.250000	57.3333333	FALSE	FALSE
Units_per_pack	2.906780	2.8333333	FALSE	FALSE
Total_replenishment	2.800000	92.6666667	FALSE	FALSE
Prod_site	2.617188	0.8333333	FALSE	FALSE
BIAS_Month_1	2.500000	94.5000000	FALSE	FALSE
Count_Month_3_FA_zero	1.586667	2.1666667	FALSE	FALSE
Prod_line_g	1.538461	6.0000000	FALSE	FALSE
Market	1.490991	0.5000000	FALSE	FALSE
Product_type	1.380952	0.3333333	FALSE	FALSE
BIAS_Month_3	1.333333	97.0000000	FALSE	FALSE
Transport_LT	1.304813	0.8333333	FALSE	FALSE
Count_positive_BIAS_Month_3	1.244186	2.0000000	FALSE	FALSE
Count_positive_BIAS_Month_1	1.240000	1.8333333	FALSE	FALSE
Count_Month_1_FA_zero	1.227273	2.1666667	FALSE	FALSE
Replenishment_Type	1.172093	0.6666667	FALSE	FALSE
Number_of_periods	1.140127	1.8333333	FALSE	FALSE
Brand_g	1.135135	5.6666667	FALSE	FALSE
Cost_per_pack	1.000000	45.1666667	FALSE	FALSE

Table 2: Near zero analysis for all predictor variables. Value 'FALSE' in last two columns for all predictors suggest that there are no variables with unique values or near zero values.

This last step concludes selection of variables. To summarize, Brand_Variant_g, Prod_line_g and Count_Month_3_FA_zero have been excluded from the dataset based on all analyses and tests performed in this chapter.

[62] Documentation of caret library in R.

4.2. Study of relationship between Stock-out and SC and Manufacturing Constraints with Exploratory Data Analysis

Selection of variables, or rather some of the steps we have taken as part of it (like correlation analysis), have already given us some fragmented information regarding the nature of relationship between response variable (perc_SA) and predictor variables. In this chapter we will deep dive into this subject, but this time with out_of_stock as response variable.

Out_of_stock is a factor with two levels: "No" and "Yes". Visual summary of classes (levels) is presented in Figure 15.

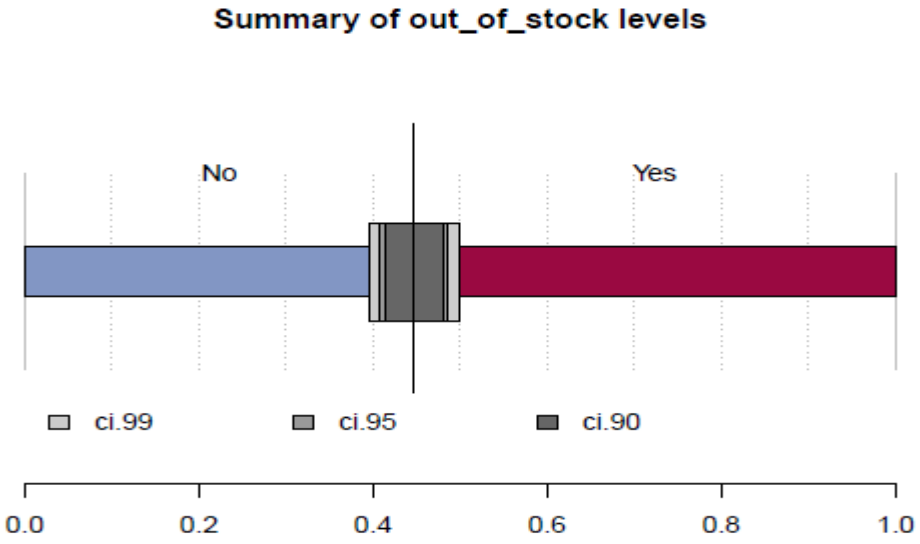

Figure 15: Summary of out_of_stock levels. 268 observations (44.7% of total) take value 'No' and 332 observations (55.3% of total) take value 'Yes'.

Percentage rate (55.3%) of dominant class ('Yes') will be important in the context of validation of model's quality. In that context it is No Information Rate[63] and can be considered as benchmark for evaluating predictive power of a model. Since we know percentage split of classes, if a sample is big enough, assigning 'Yes' value to all

[63] No Information Rate – "is the accuracy rate that can be achieved without a model (...) and (...) is the percentage of the largest class in the training set"; Applied Predictive Modelling, page 255; Max Kuhn, Kjell Johnson.

observations should give us prediction accuracy rate of 55% (without having to build any model).

Exploratory Data Analysis (EDA) is best performed visually as this approach is most appealing to end-user. Simple graphical tools like boxplots, bar charts or scatterplots can vastly simplify EDA and help deliver the message to the audience. Important is to tailor plot types to types of variables. Aesthetic factor is also very important and so is the right amount of information that is presented. Plots can be combined but only if it does not overcomplicate the picture (and message it brings).

We are going to investigate potential relationship between out_of_stock and both categorical and numeric variables. Let's tackle this from various angles and summarize key findings. This will help estimate if a model of a decent quality can be built for out_of_stock prediction.

Key findings from each chart will be captured in its caption. Most important ones will be summarized at the end of this chapter.

1. Relationship between out_of_stock and Market

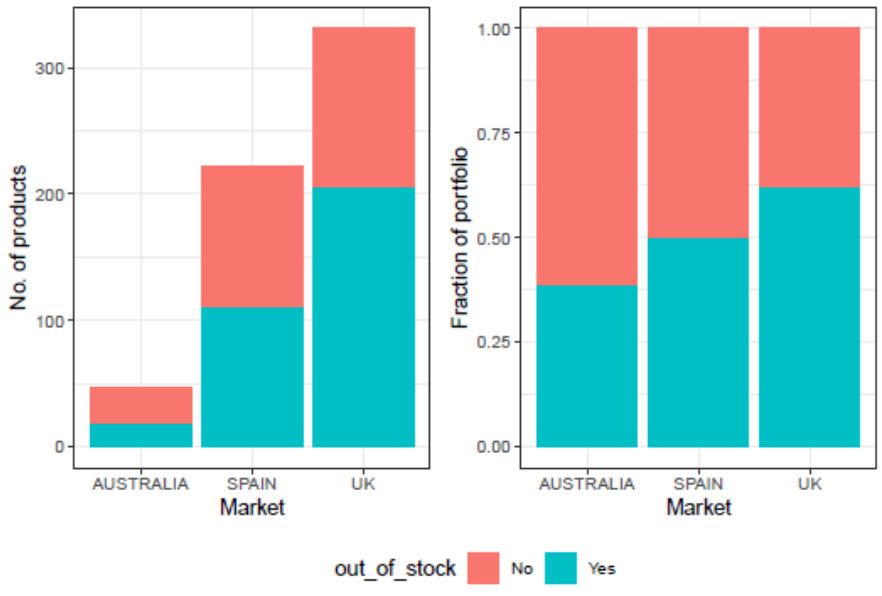

Figure 16: Relationship between the size of portfolio (no of products, left-hand side and percentage of total, right-hand side) and Market.

Key findings: The bigger portfolio size the higher fraction of stock-out SKUs.

2. Relationship between out_of_stock and Replenishment_Volume and Cost_per_pack

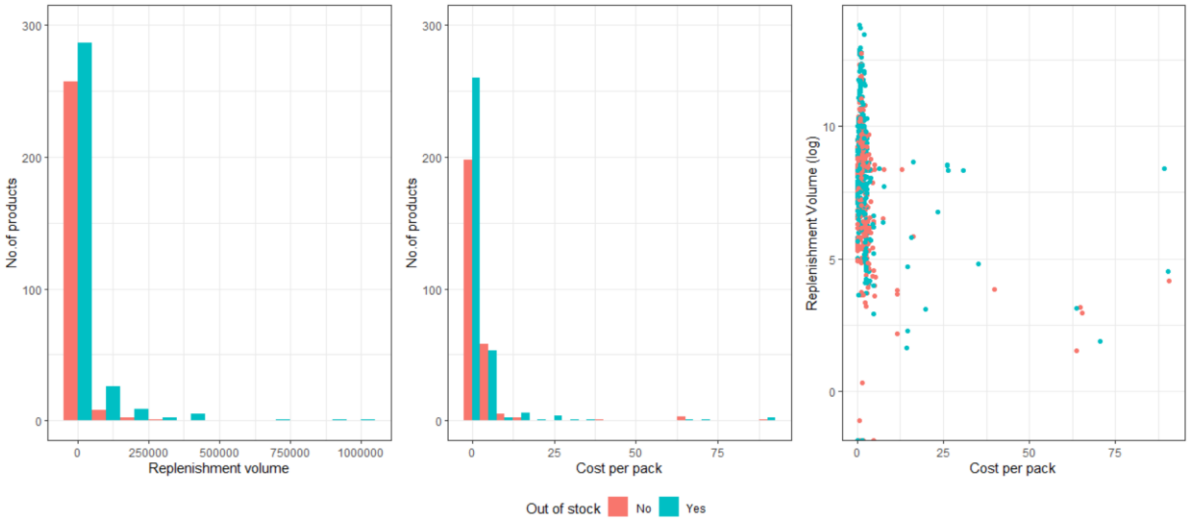

Figure 17: Reletionship between out_of_stock and Replenishment_Volume and Cost_per_pack.

Key findings: stock-outs had been reported across full range of SKUs regardless of Replenishment_volume or Cost_per_pack.

3. Relationship between out_of_stock and Replenishment Requirements error (WAPE).

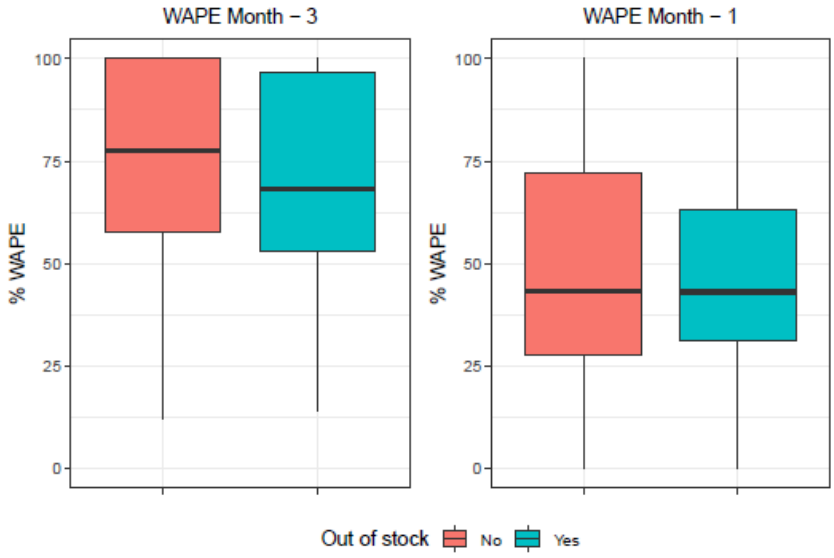

Figure 18: Relationship between out_of_stock and Replenishment Requirements error (WAPE).

Key findings:

- *WAPE_Month_3 is lower for SKUs with stock-out; this is surprising as logical expectation is opposite,*
- *As expected, WAPE_Month_1 is significantly lower than WAPE_Month_3; its median is equal for both groups of products*

4. Relationship between out_of_stock and Replenishment_type

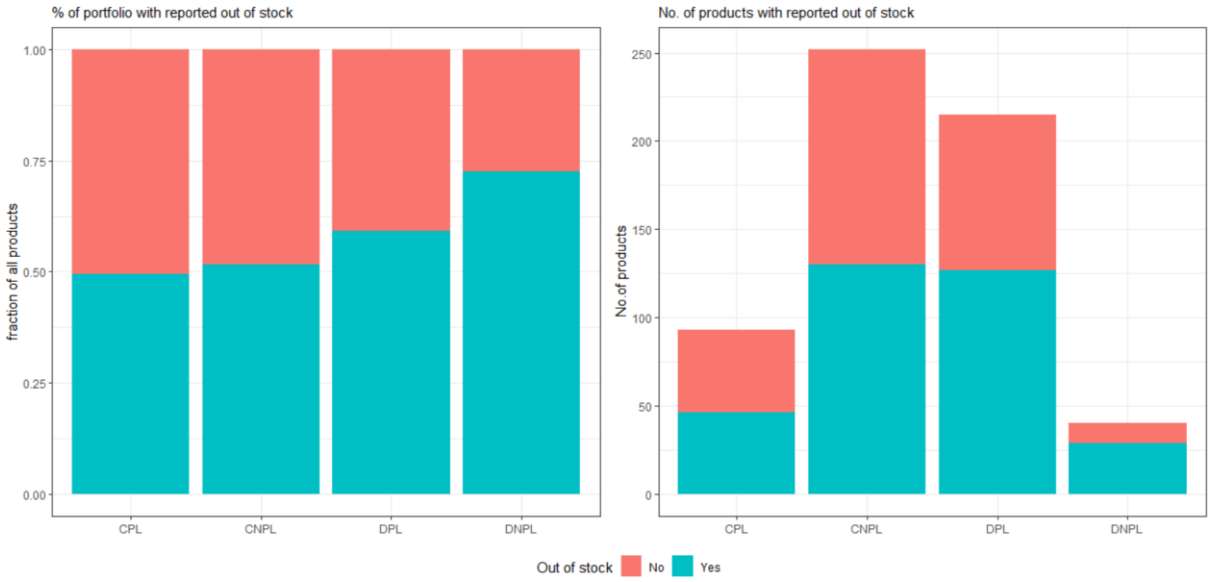

Figure 19: Relationship between out_of_stock and Replenishment_Type.

Key findings: Percentage of reported stock-outs is lower among consolidated shipments. This is important finding as usually Safety Stocks are kept at shipping point for consolidated orders, however information regarding Safety Stock is not available in our dataset.

5. Relationship between out_of_stock and Prod_site

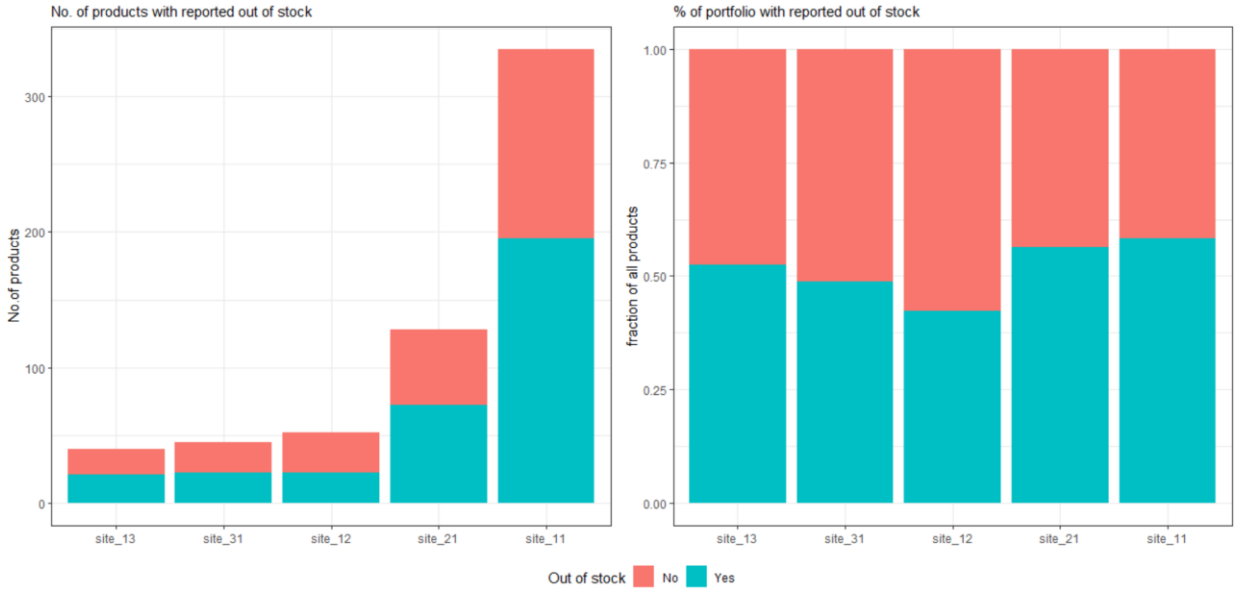

Figure 20: Relationship between out_of_stock and Prod_Site.

Key findings: SKUs from Production Sites with bigger (manufacturing) portfolio tend to have slightly higher stock-out rate at Local DC.

6. Relationship between out_of_stock and Number_of_periods (Production frequency)

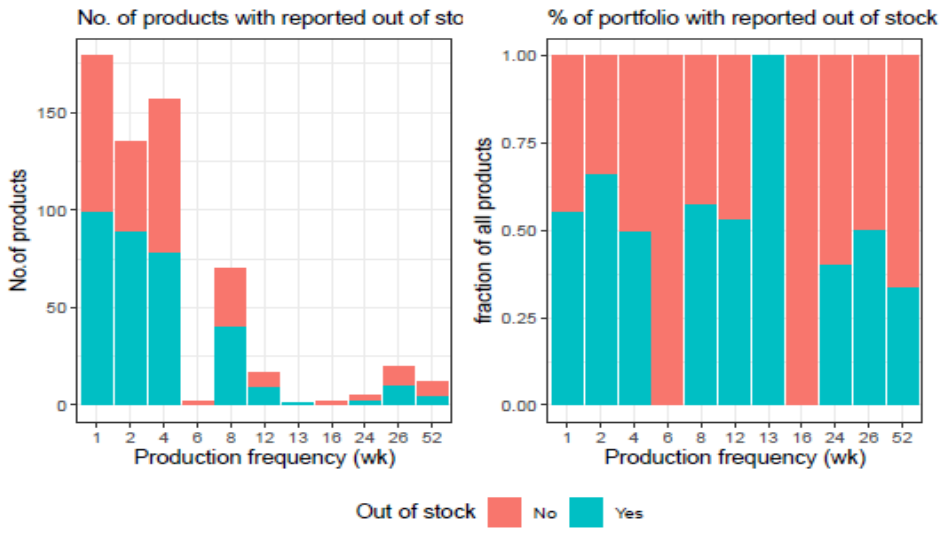

Figure 21: Relationship between out_of_stock and Number_of_periods.

Key findings:

- *Vast majority of SKUs are produced every 1 to 4 weeks,*
- *Higher frequency of production does not guarantee lower out_of_stock rate*

7. Relationship between out_of_stock and availability of the main component (Lead_comp_constr)

Figure 22: Relationship between out_of_stock and limited availability of main component.

Key finding: There is no evidence of impact from limited availability of the main component on out_of_stock rate.

8. Relationship between out_of_stock and complexity of manufacturing process

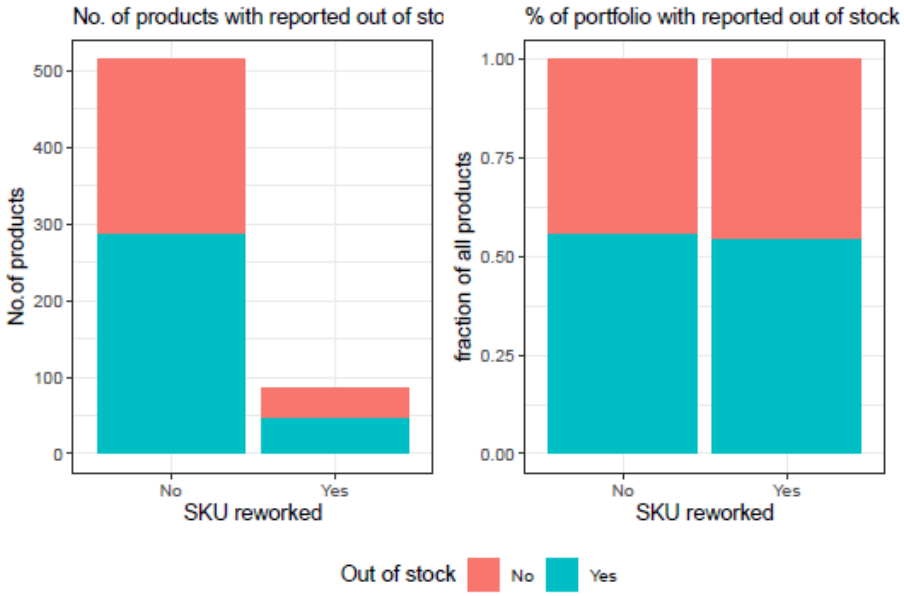

Figure 23: Relationship between out_of_stock and complexity of manufacturing process.

Key findings: Additional step in a manufacturing process (rework) does not seem to be impacting stock availability in Local DCs.

9. Relationship between out_of_stock and Product_type

Figure 24: Relationship between out_of_stock and Product_type. Key findings: Fraction of out_of_stocks for General Export and Market Specific products is on similar level. Slightly higher rate for Market Specific products could be attributed to higher number of such SKUs.

Conclusions and summary from EDA:

1. Some of the findings are in line with expectations:
   - it is easier to manage smaller portfolio in terms of stock-out mitigation both from manufacturing (Prod_site) and distribution perspective (Market),
   - replenishment requirements error (WAPE) measured one month before actual shipping date (WAPE_Month_1) is lower than error measured three months before actual shipping date (WAPE_Month_3).

2. Other findings are rather surprising:
   - Median of WAPE_Month_3 is significantly lower for products with reported out_of_stock (which could trigger a question regarding value added from long term forecasting in the environment of reactive supply chain),
   - higher production frequency does not guarantee less stock-outs in the market so selecting production cycle should be considered as a trade-off between manufacturing KPIs like OEE (Overall Equipment Effectiveness[64]) and SC inventory KPIs like Average Inventory Turns[65],
   - direct dispatch does not improve stock availability; on the contrary, less stock-outs was observed for consolidated products (it might be driven by Safety Stock kept at shipping point, however data to validate this hypothesis is not available).

Findings from this chapter might suggest the presence of additional factors which are impactful on out_of_stock variable and yet have not been captured in this dataset.
On the basis of expert knowledge we can make a hypothesis that these could include safety stock and sales forecast accuracy (measured in Local DCs). Additionally, there is a lack of information regarding the number of stock-out occurrences and their data stamp which makes it difficult to correlate it with replenishment requirements accuracy measures.

Stock-out prediction is the subject of next sub-chapter where we will investigate how Machine Learning models can deal with this challenge on the basis of available data.

[64] Overall Equipment Effectiveness (OEE) - Measuring the effectiveness of all of the equipment of a company based on usage, performance, and production quality; APICS Dictionary. The essential supply chain reference. Sixteenth edition.
[65] Inventory Turns KPI measures the number of times inventory is sold during one fiscal year.

4.3. Stock-out Prediction with Machine Learning

So far, by using descriptive statistical methods like correlation analysis or EDA, we have been trying to uncover relationship between response variable and predictors. Our findings suggest low correlation and unclear nature of relationship between predictors and out_of_stock variable. This is rather discouraging fact in the light of the task for this chapter, namely prediction of stock-outs, however let's investigate how Machine Learning models will deal with this challenge.

4.3.1. Overview of Classification Algorithms

Out_of_stock is a qualitative variable, which takes values "Yes" and "No". "Predicting a qualitative response for and observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class[66]".

"Classification models usually generate two types of predictions. Like regression models, classification models produce a continuous valued prediction, which is usually in the form of a probability (i.e., the predicted values of class membership for any individual sample are between 0 and 1 and sum to 1). In addition to a continuous prediction, classification models generate a predicted class, which comes in the form of a discrete category. For most practical applications, a discrete category prediction is required in order to make a decision.[67]"

There is various different classification methods under the umbrella of Machine Learning and in this chapter we will describe briefly techniques selected for stock-out prediction together with their advantages and disadvantages. Selection of classification algorithms has been arbitrary, however it covers entire spectrum of classification techniques to try achieve best results and to make comprehensive comparison of available methods.

[66] An Introduction to Statistical Learning with Applications in R, page 127; G. James, D. Witten, T. Hastie, R. Tbishirani.
[67] Applied Predictive Modelling, page 247; M. Kuhn, K. Johnson.

The list of classification techniques and algorithms used in this research, is shown below.

1. Logistic Regression (glm) [68],
2. Generalized Linear Model - Ridge Regression and the Lasso (glmnet),
3. Naive Bayes (naive_bayes),
4. K-Nearest Neighbors (knn),
5. Decision Tree
    a. CART (rpart),
6. Bagged Decision Tree
    a. Random Forest (ranger),
7. Boosted Decision Tree
    a. Stochastic Gradient Boosting (gbm),
    b. Extreme Gradient Boosting (xgbTree),
    c. Adaptive Boosting (adaboost)
8. Support Vector Machine with Polynomial Kernels (svmPoly),
9. Neural Networks
    a. Multilayer Perceptron (MLP)

Logistic Regression can be considered as a baseline for models comparisons. It is easy to interpret as provides with coefficient estimates and does not have any hyperparameters. It also assumes linearity between response and predictors which is one of the key disadvantages, especially taking into consideration our findings so far. Ridge Regression and Lasso (glmnet) is not either free of this drawback, however it should generate better results through shrinkage of coefficient estimates towards zero and reduction of their variance.

Naive Bayes has been selected due to high number of qualitative predictors in our dataset. It uses Bayes' theorem about prior and posterior probability for classification. Its key strengths and weaknesses have been summarized well by B. Lantz[69]. It is fast and effective and it is easy to obtain the estimated probability for a prediction. Its key disadvantage is often-faulty assumption of equally important and independent features.

K-Nearest Neighbors (knn) algorithm is also fast and uses hyperparameter K – number of neighbors. Predictor variables create p-dimensional space where observations are

[68] Values in parenthesis for points 1 to 8 describe method (or a function/package) name in R that uses the algorithm and will be used interchangeably with classification technique.
[69] Machine Learning with R, page 95; Brett Lantz.

positioned. Class of observation is assigned on the basis of classes of others, K-nearest, observations (majority voting rule is applied) and proximity to these observations is usually measured in Euclidean distance[70]. KNN is easy to implement and explain to stakeholders, however it may struggle with high dimensional data and high number of categorical variables (especially those with multiple levels). Small K may also lead to data overfitting.

Next group of methods are decision trees. Classification trees use recursive binary partitioning with a goal to maximize homogeneity[71] (in reference to the class of response variable) of each result node of the tree. One of the biggest advantages of decision trees is their ease of interpretation, especially that results can be visualized in a legible way which can result in a clear set of classification rules. What is more, they can detect non-linear and complex relationships between independent and dependent variables and decision trees algorithms are fast and do not require data preparation (standardization, imputation etc). We will start with a single decision tree (rpart) which can be vulnerable to overfitting (despite control mechanisms like minimal number of observation in each result node, minimal number of observations in a leaf, maximal number of leaves or maximal depth of a tree). The problem of overfitting can be handled by bagging and we will use its special version – Random Forest (ranger), where both observations and predictors are bootstrapped[72]. Prediction will be made not on the basis of one decision tree but on the basis of multiple trees (each independent from other trees – author's footnote) which are combined to yield a final prediction. "Combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss in interpretation.[73]"
Boosted decision trees (algorithms gbm, xgbTree and adaboost) also use multiple trees and from that point of view are similar to bagged decision trees, however what differentiate boosted methods from bagged method is that boosted trees grow in a sequential way and use

[70] Since there is more than one way to calculate distance, calculation method can be considered as a second hyperparameter (next to number of neighbors, K).
[71] Homogeneity can be measured using percentage of observations in a branch that does not belong to the most common class of the dependent variable, Gini factor (used in CART algorithm) or entropy (used in C5.0 algorithm).
[72] "Bootstrap sampling, the bootstrap, or bootstrapping for short. Generally speaking, these refer to statistical methods of using random samples of data to estimate properties of a larger set. When this principle is applied to machine learning model performance, it implies the creation of several randomly-selected training and test datasets, which are then used to estimate performance statistics. The results from the various random datasets are then averaged to obtain a final estimate of future performance. (...) the bootstrap allows examples to be selected multiple times through a process of sampling with replacement."; Machine Learning with R, page 322; Brett Lantz.
[73] An Introduction to Statistical Learning with Applications in R, page 303; G. James, D. Witten, T. Hastie, R. Tbishirani.

information from prior trees. Having evaluated a certain model of a decision tree, we build new one on residuals from this model. Each of such trees can be small.
In a nutshell, the idea behind this approach is to "boost the performance of weak learners to attain the performance of stronger learners[74]". Weak learner is a simple model which classification error is slightly below 0.5. Classification error of strong learner is close to 0.
Since boosting is slow, Stochastic Gradient Boosting (gbm) allows selection of random sub-sample of training set which increase the speed of a learning process. Extreme Gradient Boosting (xgbTree) and Adaptive Gradient Boosting (adaboost) algorithms are extensions of Stochastic Gradient Boosting (gbm) and bring additional features (hyperparameters) to improve its performance and prediction accuracy. These include (among others) randomization parameters, intelligent trees pruning, parallel computation, or assigning weights to individual observations (incorrectly classified observations get higher weights). Boosted methods, similarly to bagged methods are not easy to interpret and additionally, boosting is vulnerable to outliers.

Last two methods are often considered as "Black Box Methods"[75] however vast majority of base methods' extensions which use hyperparameters (like bagging or boosting – extensions of single decision tree) could fall into this category due to difficulty of model interpretation. Nevertheless, both Supply Vector Machine with Polynomial Kernels (svmPoly) and Neural Networks are definitely most complex models used in this research and they interpretation require using XAI[76].

The purpose of Support Vector Machine is to find (in a multidimensional space created by predictors) a hyperplane that separates observations belonging to different groups (classes). "The Support Vector Machine is a generalization of a simple and intuitive classifier called the maximal margin classifier (...) which unfortunately cannot be applied to most data sets, since it requires the classes to be separable by a linear boundary. (...) support vector classifier, an extension of the maximal margin classifier can be applied in a broader ranges of cases (...)[77]". Support Vector Machine, a further generalization of support vector classifier accommodates non-linear class boundaries[78].

[74] Machine Learning with R, page 343; Brett Lantz.
[75] Machine Learning with R, chapter 7; Brett Lantz.
[76] XAI – Explainable Artificial Intelligence which will be the subject of Chapter 4.4.
[77] An Introduction to Statistical Learning with Applications in R, page 337; G. James, D. Witten, T. Hastie, R. Tbishirani.
[78] An Introduction to Statistical Learning with Applications in R, page 337; G. James, D. Witten, T. Hastie, R. Tbishirani.

Problem of non-linear class boundaries can be resolved by increasing the number of dimensions however it increases computational complexity. To bypass this problem SVM uses so called 'kernel trick' which allows transforming data in such a way as if new features has been added to the model, but without increasing the dimensionality of data (and hence computational complexity). This is achieved through kernel function and in our research we will be using polynomial kernel (svmPoly) which allows polynomial transformations of the dataset.

The main idea behind Neural Networks (NN) is extracting information hidden in linear combinations of input data and then modelling non-linear relationship between them and response variable. This takes place in a system where input signals are connected to the output signals through a network made of neurons, layers, activation function and weights. The value of weights is determined by the training algorithm. In this research, Multilayer Perceptron (MLP) algorithm will be used and its architecture and parameters will be shared in the next chapter. Neural Networks can replace complex analytical processes, are scalable and comprehensive. Trained NNs can be used to other analytical tasks or can be extended by new data without the need to re-estimate the entire model on the full dataset. Key disadvantage, as mentioned earlier, is the effect of Black Box. Additionally, there is no golden rule for setting its architecture and parameters. It is usually a matter of trial and error to set it up.

As mentioned earlier, although classification methods have been selected arbitrary, they cover a spectrum of available techniques and therefore should give us a good understanding if stock-outs can be predicted (or accuracy of such prediction).

Machine Learning methods can also be combined to improve prediction power and therefore their accuracy. This is models ensemble procedure, however it is out of scope in this research.

4.3.2. Comparison of Predictive Models

All models presented in this chapter have been built in R and purely due to technical aspect of this process we will first review results of first ten models (which have been built using 'caret' package) and lastly we will take a look into Neural Network model (built with 'keras' library) to see if it delivers any improvements related to prediction accuracy.

Technical side of modelling in R, although not in scope of this thesis, still requires brief explanation as some aspects refer to modelling process itself (and its best practices).
Key points worth mentioning:

- dataset has been split into training set and test set in proportion 4:1 (in other words, 80% of randomly selected observations fall into training set and 20% into test set),
- models results have been validated using ten cross-validation folds,
- reference level of response variable 'out_of_stock' is set to 'Yes',
- seed has been set for reproducibility and comparability of models,
- ROC metric has been selected as summary metric to define optimal model, which measures the trade-off between detection of true positives while avoiding false positives and which curve can be interpreted as per Figure 25[79]

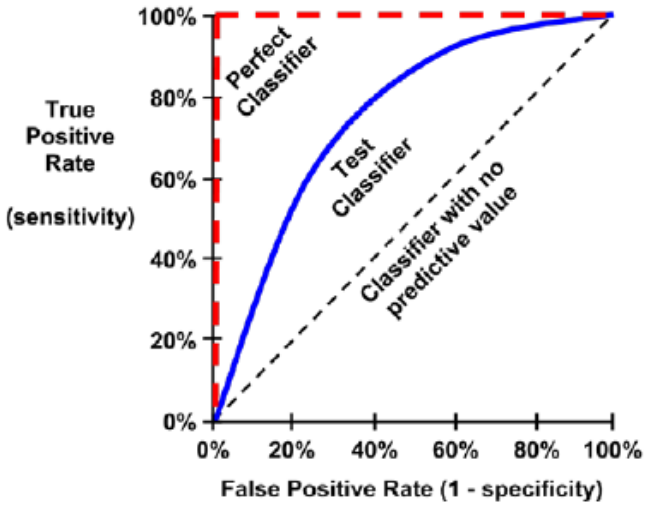

Figure 25: ROC curve tells how much model is capable of distinguishing between. It is plotted with True Positive Rate (Sensitivity) against False Positive Rate (1 – specificity). Intuitive interpretation of the plot is the more ROC curve resembles red line the higher predictive value of the model. On the other hand, the flatter the curve the less class distinguishing capability of the model. Black dotted line indicates model with no predictive value.

[79] Machine Learning with R, page 312; Brett Lantz.

- in addition to ROC, key measures from Confusion Matrix (listed below) have been taken into consideration within a range of probability cut-off points from 0.15 to 0.85 (as a reminder, No Information Rate is 0.55):
    - Accuracy – calculated as the number of all correct predictions divided by the total number of observations,
    - Sensitivity (Recall or True Positive Rate) – calculated as the number of correct positive predictions divided by the total number of positives ("Yes"),
    - Specificity (True Negative Rate) – calculated as the number of correct negative predictions divided by the total number of negatives ("No"),
    - Positive Predictive Value (Precision) – calculated as the number of correct positive predictions divided by the total number of positive predictions,
    - Negative Predictive Value – calculated as the number of correct negative predictions divided by the total number of negative predictions,
    - Balanced Accuracy – calculated as Sensitivity plus Specificity divided by 2; it is a much better metric than Accuracy for unbalanced values of response variable (in other words when No Information Rate is significantly different than 0.5),
    - F1 – harmonic mean of Precision and Recall.
- models have been run on normalized data (range from zero to one) due to inclusion of qualitative variables,
- methods' hyperparameters listed in Table 3 have been tuned automatically (option offered by caret package).

Model	Method	Tuning Parameters
Logistic Regression	glm	
Ridge Regression and the Lasso	glmnet	alpha, lambda
Naive Bayes	naive_bayes	laplace, usekernel, adjust
k-Nearest Neighbors	knn	class
Decision Tree - CART	rpart	cp
Random Forest	ranger	mtry, splitrule, min.node.size
Stochastic Gradient Boosting	gbm	n.trees, interaction.depth, shrinkage, n.minobsinnode
eXtreme Gradient Boosting	xgbTree	nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample
AdaBoost Classification Trees	adaboost	nIter, method
Support Vector Machines with Polynomial Kernel	svmPoly	degree, scale, C

Table 3: Classification models and their tuning parameters. Parameters tuning has been done automatically with a goal to maximize ROC metric.

Figure 26 presents ROC curve plots for training and test datasets for all models. There are a couple of clear conclusions from both plots.

First of all, there is significant difference in shape of ROC curves between training and test dataset. This suggests data overfitting for all models with extreme example of Adaptive Gradient Boosting. What is more, ROC curves for test data are nearly flat which means that all models have poor or no discriminatory power.

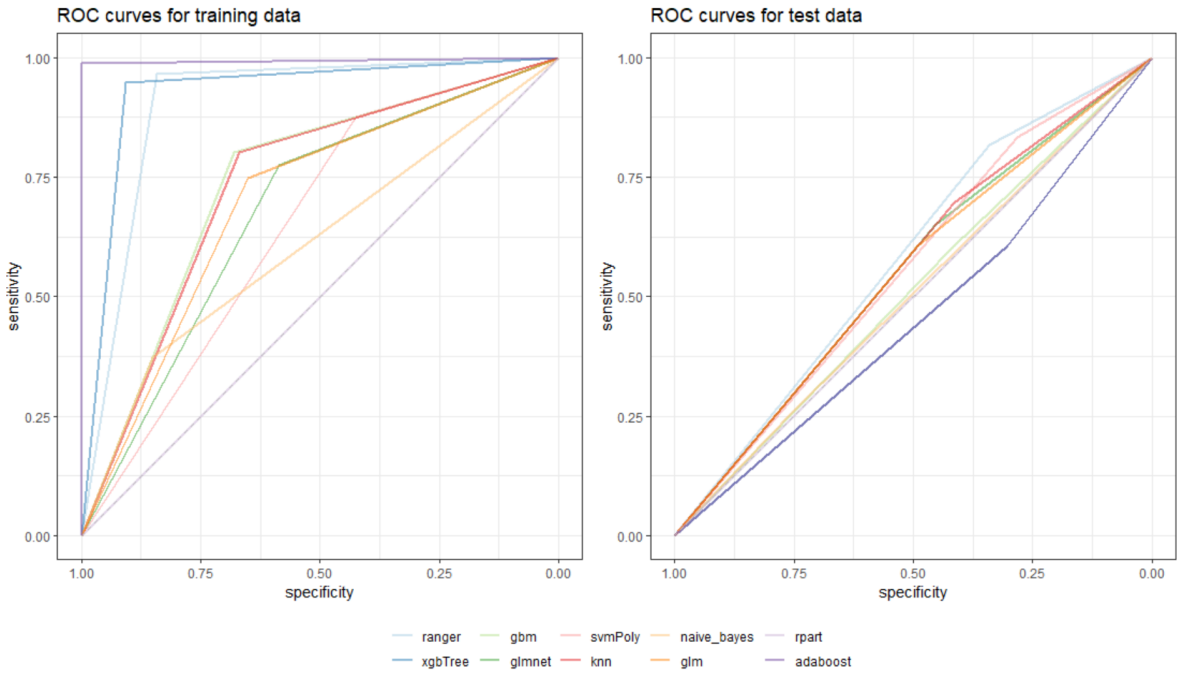

Figure 26: Comparison of models' ROC curves for training and test dataset. Left-hand side plot shows significant data overfitting for Adaptive Gradient Boosting where AUC (Area Under the Curve) is close to 1. Right-hand side plot shows little discriminatory power of all ML models on test dataset.

Discriminatory power can be also checked with AUC (Area Under the Curve) score.

A convention for interpreting AUC scores uses a system similar to academic letter grades[80]:

- 0.9 – 1.0 = A (outstanding)
- 0.8 – 0.9 = B (excellent/good)
- 0.7 – 0.8 = C (acceptable/fair)
- 0.6 – 0.7 = D (poor)
- 0.5 – 0.6 = F (no discrimination)

[80] Machine Learning with R, page 313; Brett Lantz.

Results of AUC score for all models on test dataset is presented in Figure 27. This confirms conclusions from visual inspection of ROC curves. Unfortunately all models according to AUC score have no discriminatory power.

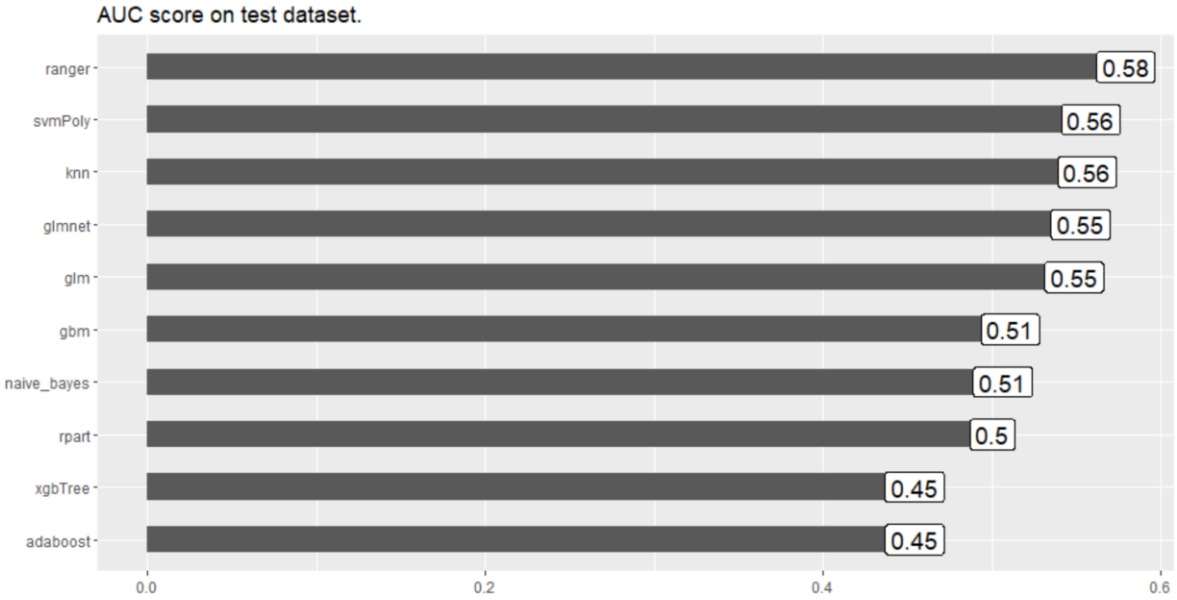

Figure 27: AUC score results for all models calculated on test dataset. Unfortunately taking into account this metric all models have no discriminatory power.

We can also take a look at balanced accuracy metric calculated validated using 10-fold cross-validation. This is another option to asses predictive power of our models as we can always use No Information Rate as a benchmark.

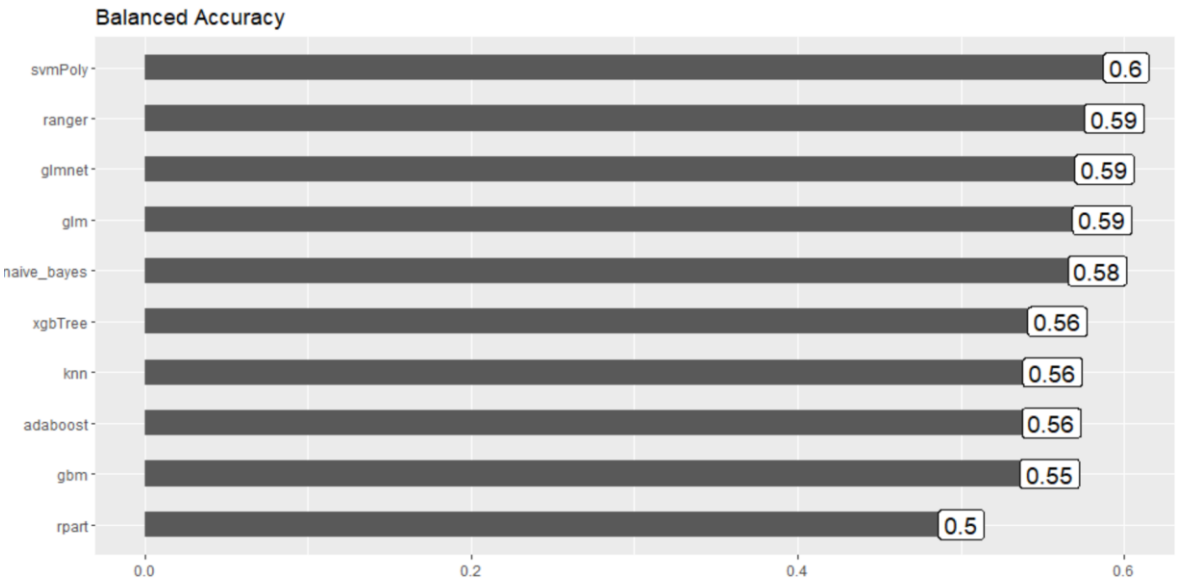

Figure 28: Balanced Accuracy calculated using 10-fold cross-validation. Value of this metric is still small, however for some models it is slightly higher than No Information Rate (0.55).

Values of Balanced Accuracy measures confirms that all models have little (or no) predictive power. Positive news is that Balanced Accuracy of some models is higher than No Information Rate (0.55) which means using model svmPoly (which has highest BA) delivers better results than a blind guess.

Figure 29 presents results for all other key measures for probability cut-off range between 0.15 and 0.85. Balanced accuracy 0.6 is highest for probability cut-off point 0.56.
0.57 looks like a better cut-off point as Balanced Accuracy is on similar level and Sensitivity is equal to Specificity. In addition to that, values of all measures except for Negative Predictive Value are higher than No Information Rate.

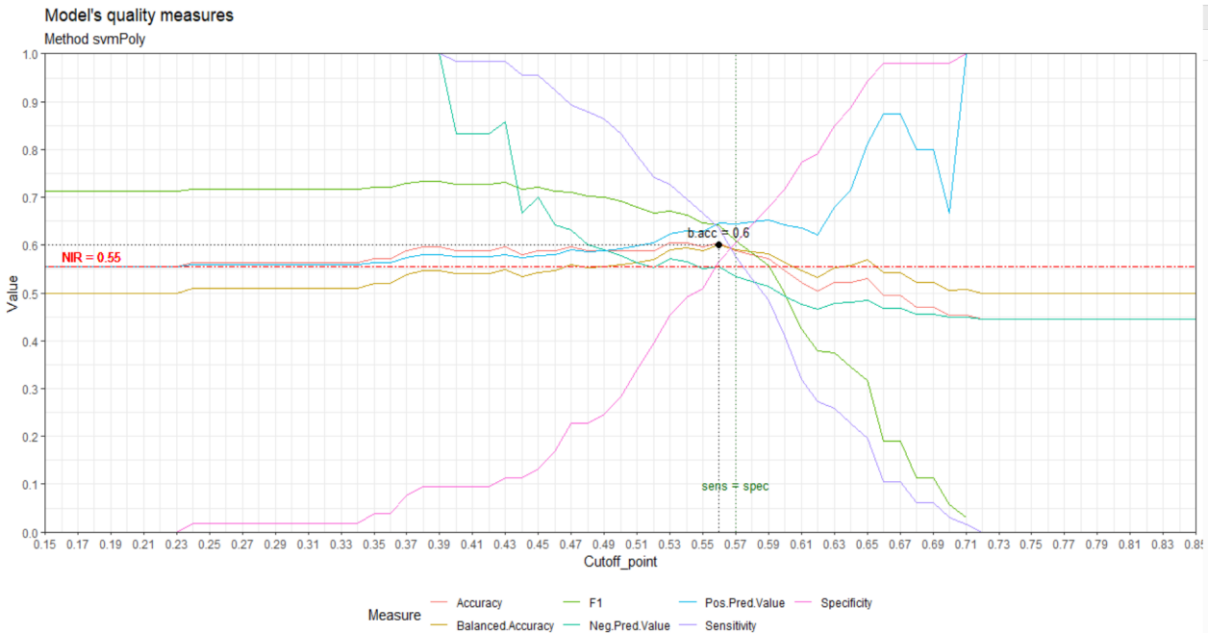

Figure 29: Summarized values of key measures for Support Vector Machine with Polynomial Kernel captured for probability cut-off points between 0.15 and 0.85. This model delivers better results than random guess.

Taking into account that measures' values presented in Figure 31 have been calculated using cross-validation results on test dataset should be similar. Confirmation of that can be found in Figure 30 which presents Confusion matrix and Statistics for test dataset. For cut-off point 0.57 balanced accuracy and all other key measures are on the level of results from 10-fold cross-validation. For Negative Predictive Value we can even see significant improvement as it is now above No Information Rate.

```
Confusion Matrix and Statistics

          Reference
Prediction Yes No
       Yes  42 23
       No   24 30

               Accuracy : 0.605
                 95% CI : (0.5113, 0.6934)
    No Information Rate : 0.5546
    P-Value [Acc > NIR] : 0.1552

                  Kappa : 0.202

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.6364
            Specificity : 0.5660
         Pos Pred Value : 0.6462
         Neg Pred Value : 0.5556
             Prevalence : 0.5546
         Detection Rate : 0.3529
   Detection Prevalence : 0.5462
      Balanced Accuracy : 0.6012

       'Positive' Class : Yes
```

Figure 30: Confusion matrix and Statistics for svmPoly model generated on test dataset at cut-off 0.57. Values of all key metrics are similar to outcome from 10-fold cross-validation. It's also worth to note that Negative Predictive Value is higher than NIR (which provides improvement comparing to cross-validation results).

So far, although we have slightly managed to improve prediction of out_of_stock class (comparing to NIR), even the best model (svmPoly) is still of a low quality and reliability. What is more AUC score suggest it does not have discriminative power.
SVM with polynomial kernel model can be used as a benchmark for comparison with Neural network model – the last one we will review in this chapter. Let's see if NN can outperform SVM and show decent level of reliability and predictive power measured on the basis of some metrics.

As we did with first ten models (built using framework of "caret" library) we will start with brief explanation of model assumptions, architecture and all key steps taken in the modelling process (using "keras" package):

- dataset has been split into training and test sets,
- all values of categorical predictors have been mapped to integers (one-hot encoding),
- dataset has been normalized to a range between 0 and 1,
- levels of response variable ("Yes" and "No") have been recoded into 1 and 0,
- dataset has been converted into a matrix (as required in keras),

- Multilayer Perceptron (MLP) model has been used with below architecture and parameters[81]:
    - 3 hidden layers with 32, 16 and 16 neurons and activation function ‘relu’,
    - dropout 0.2 as regulating function,
    - activation function ‘sigmoid’ in output layer,
    - optimization algorithm ‘adam’,
    - loss function ‘binary cross-entropy’,
    - quality metrics ‘AUC’[82],
    - validation split 0.2,
    - 100 epochs,
    - batch size 100.

Figure 31 presents results for loss function (cross-entropy) and quality metric (AUC) by epoch. We can conclude that similar to previous ten models predictive power of MLP model is low and it shows data overfitting tendency as number of iteration (epochs) increases (AUC values for training data are significantly higher than validation data and loss function values are significantly lower for training data than validation data).

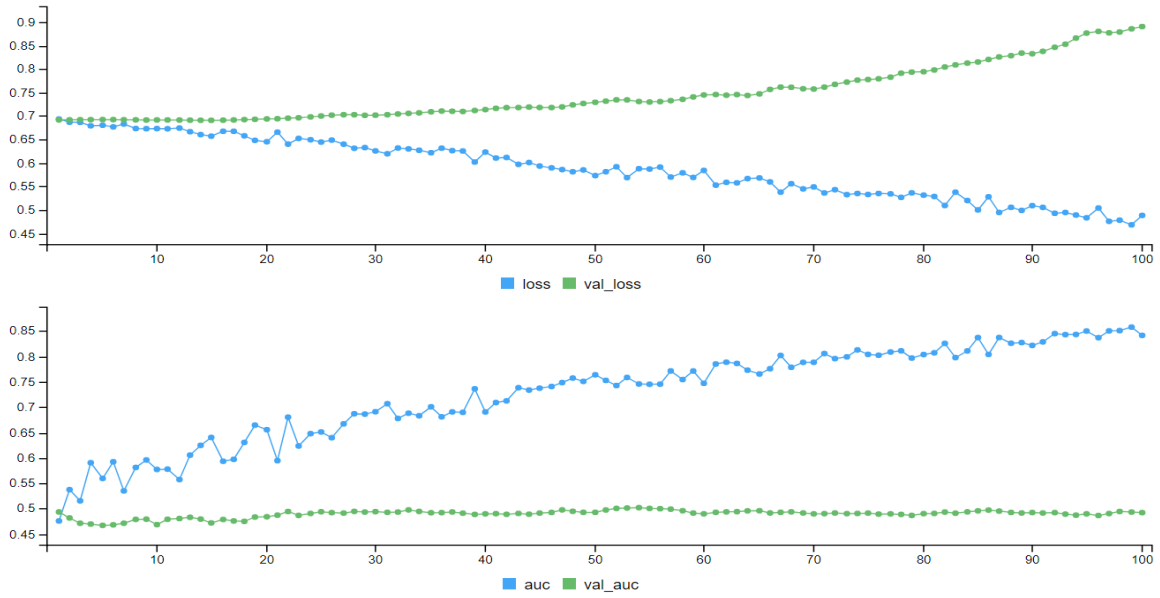

Figure 31: Plots of loss function (binary cross-entropy) and quality metric (AUC) for train (blue line) and validation (green line) datasets. Lines in both plots diverge which suggests data overfitting.

[81] As explained in chapter 4.3.1, this has been ‘trial and error’ process and above architecture and parameters level delivered best set of results out of all trials.
[82] Analogically to first ten models where ROC metric has been used to keep approach consistency.

Figure 32 confirms that. This time we compare training and test datasets and we can see clear gap between ROC curves of both datasets (or AUC score displayed in the title of the plot). What is worth mentioning is that AUC score for test data is 0.6 which puts it one level up comparing to svmPoly ('no discrimination'), to 'poor' category (which refers to discriminative power of the model).

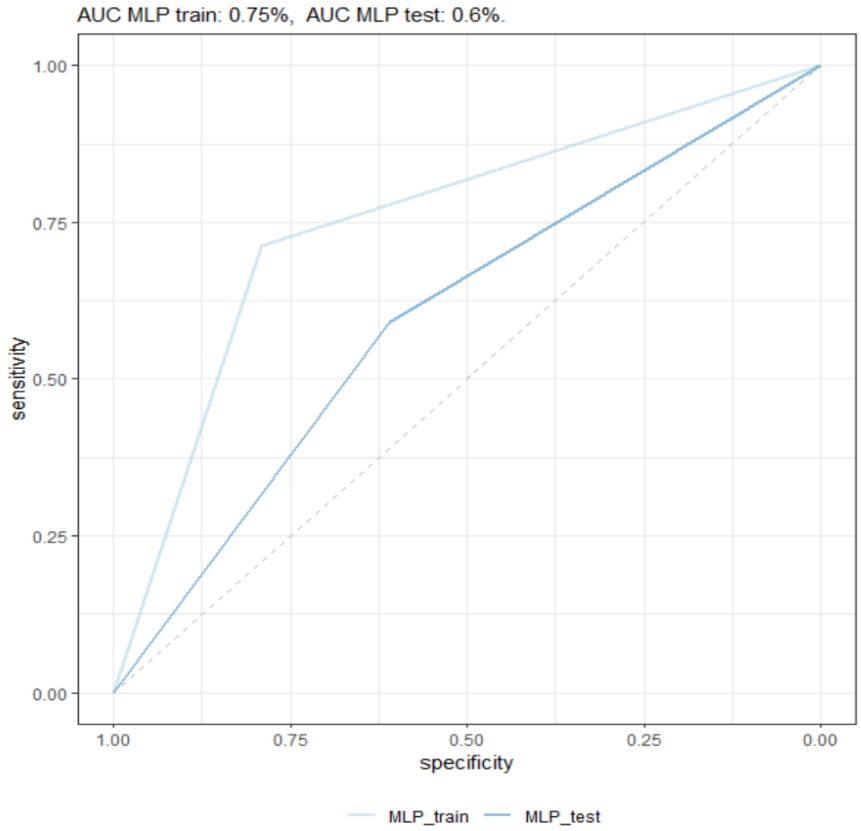

Figure 32: Comparison of ROC curves and AUC scores of MLP model measured for training and test datasets. Similar to previous models we can observe overfitting tendency. AUC of 0.6 for test data suggests 'poor' discriminative power of the model, however this is better result than best of the previous models (svmPoly).

Comparison between MLP and previous models can be concluded by checking Confusion Matrix and Statistics for test data (Figure 33). For consistency with other models, probability cut-off point of 0.57 has been used.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  0
         1 36 23
         0 25 36

               Accuracy : 0.6
                 95% CI : (0.5066, 0.6883)
    No Information Rate : 0.5083
    P-Value [Acc > NIR] : 0.02729

                  Kappa : 0.2002

 Mcnemar's Test P-Value : 0.88523

            Sensitivity : 0.5902
            Specificity : 0.6102
         Pos Pred Value : 0.6102
         Neg Pred Value : 0.5902
             Prevalence : 0.5083
         Detection Rate : 0.3000
   Detection Prevalence : 0.4917
      Balanced Accuracy : 0.6002

       'Positive' Class : 1
```

Figure 33: Confusion Matrix and Statistics for MLP model calculated on test dataset. We can see that most key measures are balanced and their values are between 0.59 and 0.61.

Balanced accuracy of MLP model (approx. 0.6) is on the same level with svmPoly model to third decimal place. What is interesting about this model is it delivers balanced set of metrics for positive and negative predictions (Sensitivity/Specificity and Positive Predictive Value/Negative Predictive Value) which values range from 0.59 to 0.61. Unfortunately, although these values are higher No Information Rate (0.55) they are still low which indicates low discriminative power of the model.

4.3.3. Conclusions From Comparison of Predictive Models

In the previous chapter (4.3.2) we saw results for eleven different predictive models which span over several different Machine Learning techniques from simple ones to those that are considered as 'Black Boxes'. Arguably best results have been achieved using Neural Network (Multilayer Perceptron), however even with this advanced approach discriminative power, using AUC score, can be described as 'poor'. Positive fact is that we have managed to improve predictive power comparing to No Information Rate by approx. five percentage points, however reliability of such model is still considerably low.

This outcome, taking into account descriptive analyses done so far like correlation or EDA where we did not see clear relationship between response and predictors, should not be a surprise and the root cause of this should lie elsewhere than ML methods as the only potential options that has not been explored is models ensemble.

We could then hypothesize what is the root cause of low predictive power of our models. Based on expert knowledge I could indicate three main areas:

- reactive nature of supply chain planning,
- data collection methodology,
- unavailability of important predictors.

The first area leads us to 'rules of the game' which have been mentioned in Chapter 2. Mitigation of out_of_stocks is one of the key goals for supply chain planning which often turns on 'reactive mode' where some of the constraints (like agreed production frequency or sequence) or SC parameters like shipping forecast accuracy ("let's forget there was no forecast and protect the service to customer") are simply disregarded. This often leads to positive outcome, however it can also start a 'domino effect' and lead to a situation where mitigation of stock-out for one product results in a stock-out for another product (often it is also a conscious business decision).

The second point indicates that data collection methodology might not have been correct. Good example could be production frequency (Number_of_Periods variable). We know how ERP system is set up (as this is what this variable tells us) but we do not have information about actual production frequency in that fiscal year which could totally change the picture. Perhaps replacing this variable with number of actual makes per fiscal year would be a better

idea. This way we could also identify any outliers, like products produced only once for a specific event.

Third area is lack of other key predictors. It has been already mentioned that although we have information about stock-outs and their cumulative volume, we are missing time stamp of stock-out which makes it impossible to correlate with shipping forecast quality measures. It is important fact since we have not differentiated regular occurrences of stock-outs from those that happened only once. Sales forecast accuracy could be also helpful with prediction of stock-outs as it triggers entire supply chain. Another good example of valuable and unavailable predictors is Safety Stocks which should be correlated with stock-outs (the purpose of keeping safety stocks is to mitigate service risks hence stock-outs).

The last two areas are improvement opportunities in terms of increasing predictive power of the models and once again highlight importance of good quality data in any Machine Learning project. On the other hand, reactive nature of supply chain cannot be changed.

4.4. What is in the (Black) Box?

This chapter concludes modelling using supervised Machine Learning methods.
The purpose is to touch a serious problem related to those advanced techniques, namely their difficulty of interpretation, which turn them into so-called 'black boxes'. What that means, in a nutshell, we can assess predictive power of a model based on measures like Accuracy, Sensitivity, Specificity or AUC score, however it is difficult to understand how probability behind prediction is calculated and what are the key variables used for prediction. This problem becomes even more serious in business environment where we would like to build a cross-functional process based on such model or business sector we operate in is set in environment regulated by law (good example is banking and customer's 'right to know'[83]).
Challenges related to 'black box'-like models are tackled in dynamically growing branch of Machine Learning, Explainable Artificial Intelligence (XAI).
XAI approach to Machine Learning is model transparency at various stages of modelling process: model performance validation, selection of best model (from a range of models) or model's outcome explanation (how model works).
In this chapter we will focus on that last point to show that 'black box' can indeed be turned in transparent 'glass box'. We will explain:

- importance of predictors,
- effect of variables on prediction, including "what if" scenarios.

It is worth to mention that we will be using DALEX library[84] which offers model-agnostic diagnostic methods which separate explanations from Machine Learning model. It is an important fact (and key benefit of such approach) as "typically, not just one, but many types of machine learning models are evaluated to solve a task, and when comparing models in terms of interpretability, it is easier to work with model-agnostic explanations, because the same method can be used for any type of model."[85]
The model that will be explained is Multilayer Perceptron (MLP) from chapter 4.3.2.

[83] Amongst other rights of customer set by The Federal Trade Commission (FTC) an stated in Equal Credit Opportunity Act (ECOA) is the right to know why load application has been declined.
[84] More on this package can be found in article: "DALEX: Explainers for Complex Predictive Models in R", Przemysław Biecek, Journal of Machine Learning Research 19.
https://jmlr.org/papers/volume19/18-416/18-416.pdf
[85] Interpretable Machine Learning. A guide for Making Black Box Models Explainable, chapter 5; Christopher Molnar.
https://christophm.github.io/interpretable-ml-book/agnostic.html

Importance of predictors can be visualized with Feature Importance plot which tells us how much value of loss function (1-AUC for classification) would change if values of variables had been changed to random. Based on that we can draw a conclusion that the bigger the increase of loss value, the higher the importance of a given predictor. Feature Importance plot for Multilevel Perceptron is presented in Figure 34.

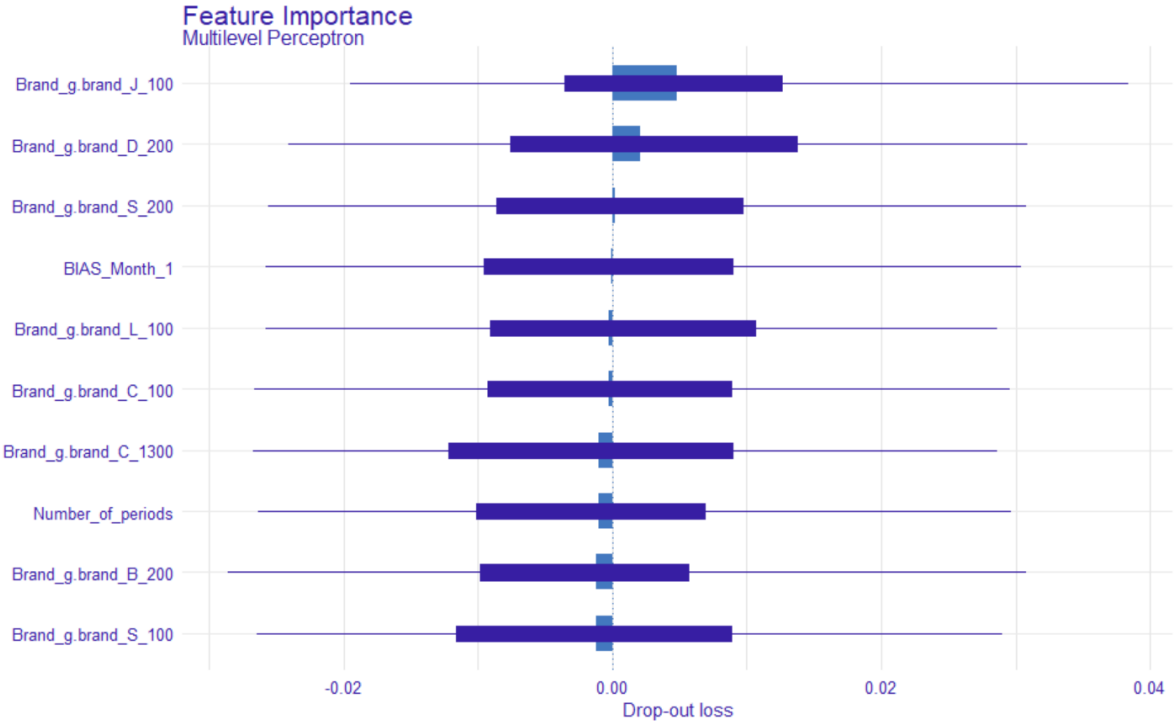

Figure 34: Feature importance plot of type 'difference' for Multilayer Perceptron (MLP) using 1-AUC as the loss function and 20 permutation rounds. Light blue bars are means of variable-importance measures for top 10 explanatory variables. Boxplots (dark blue bar and whiskers) show distribution of drop-out loss results for all bootstrap samples.

Visual analysis of feature importance suggests that Brand_g variable is the most important feature used for class prediction in MLP model, however certainty of features importance reduces significantly taking into consideration width of their boxplots and the fact they span over value 0 ('no difference').

Features importance plot can be also used to evaluate model's stability. The lower general drop-out loss the higher stability of the model. Additionally, lack of predictor with significantly high value of loss measure (in comparison to loss value of other variables) suggests less dependency on that particular predictor[86].

[86] This is important from 'out-of-time' perspective and can extend life duration of the model.

What if scenario can be run on the basis of feature effect analysis which tells us how probability of stock-out could change upon the change in value of a particular variable (while other variables remain constant).

Visualization of such scenario can be found in Figure 35 where Partial Dependence profile plot is presented for WAPE_Month_1[87] predictor.

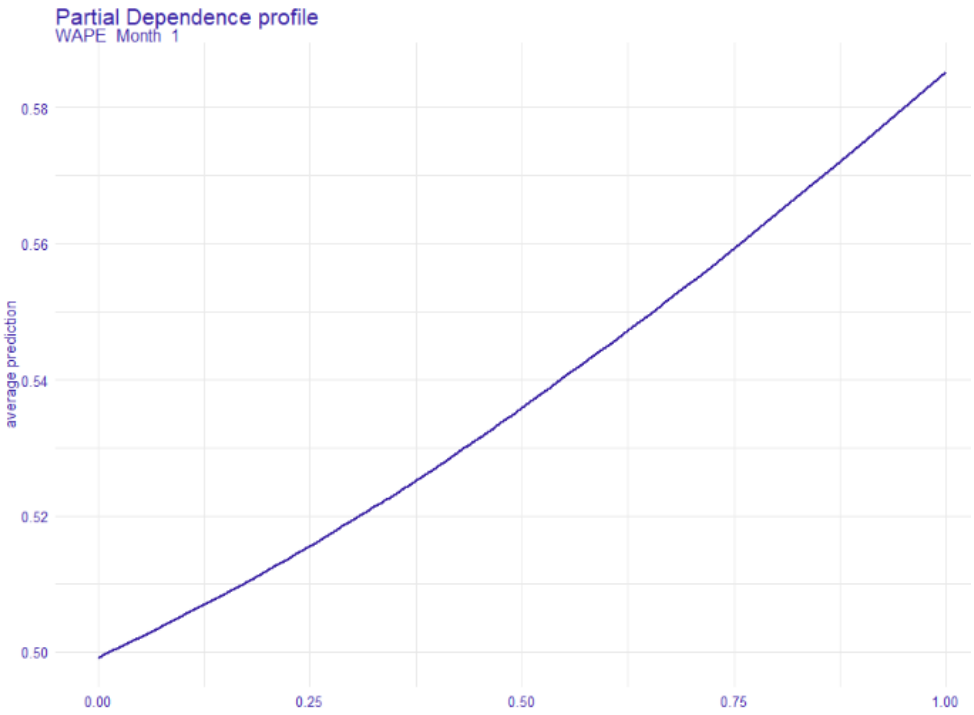

Figure 35: Partial Dependence profile plot for WAPE_Month_1. With other predictors hold constant, change in value of WAPE_Month_1 from 0% to 100% increases probability for stock-out by 8 percentage points (from 50% to 58%).

Findings from Partial Dependence profile plot for WAPE_Month_1 are in line with Feature Importance plot. Although this predictor is in top 3 important features, it is impact on prediction is rather low as a significant increase in its value (from min 0% to max 100%) increases probability for stock-out on average only by eight percentage points.

Explanation of prediction for a single observation can be visualized with a Break Down profile plot. We will simulate this process for randomly selected observation from our test dataset[88]. The value of this observation is 0 which means no stock-out.

[87] WAPE_Month_1 predictor has been selected for ease of interpretation - it is a numeric variable which normalized values are 1/100 of original values and still can interpreted as percentage error.
[88] For traceability, it is observation with record number 78.

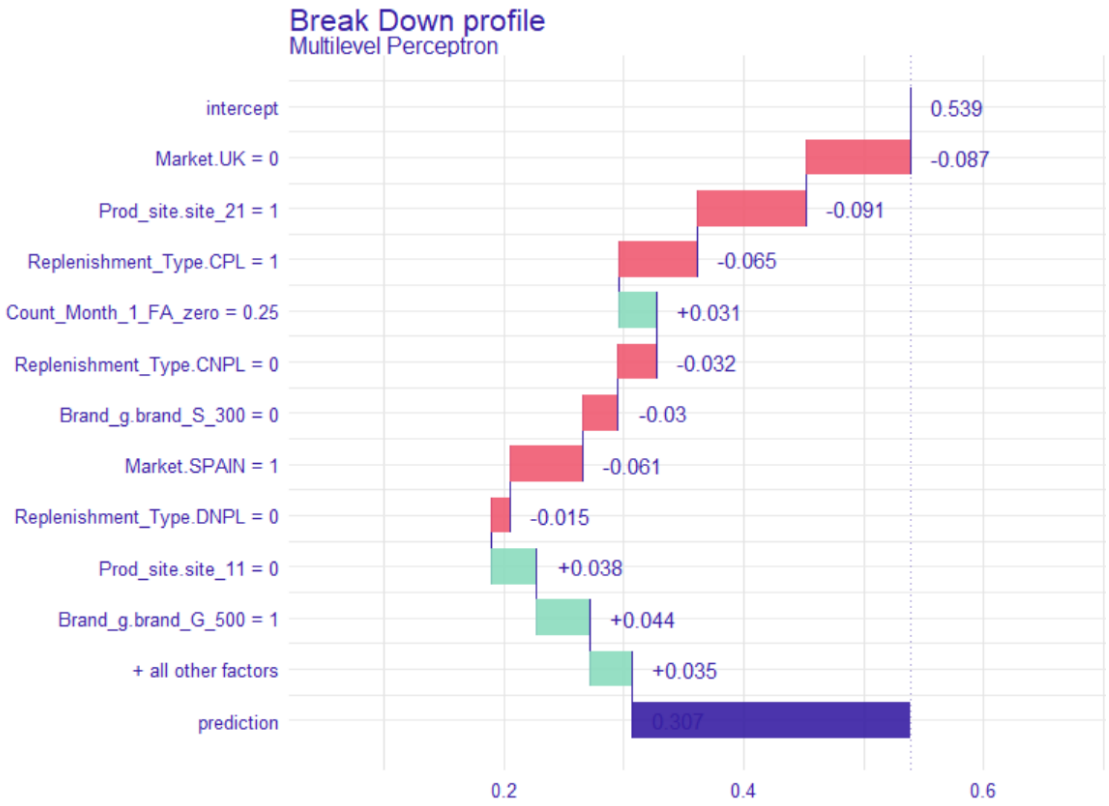

Figure 36: Break Down profile plot for a single observation. intercept can be interpreted as No Information Rate and prediction describes probability value for success (stock-out is 'Yes'). Y-axis include all key variables and their contribution towards total probability (0.307).

Figure 36 presents outcome of class prediction for selected observation. First of all, break-down plot tells us total probability for stock-out (30.7%) and secondly describes contribution of key variables towards this result. Intercept can be interpreted as No Information Rate (NIR) and is a starting point. We can see for example total effect from Market variable (approx. -0.15). Just because Market is not UK reduces stock-out probability by approx. 8.7% and Market Spain further decreases it by another 6.1%. This plot can be a great tool for final user who, based on expert knowledge, can also evaluate performance and quality of the model.

Feature Importance, Partial Dependence profile and Break Down profile plots naturally do not exhaust model-agnostic methods from DALEX package but it also was not a goal of this chapter to present all available methods. Even on the basis of these three diagnostic methods data scientist and final user can get comprehensive understanding of decision making process embedded in the model.

5. Portfolio Clustering with Unsupervised Machine Learning as an Alternative to ABC Classification

The concept of ABC classification, its application in Supply Chain Management together with its advantages and disadvantages has already been presented in Chapter 3.
The key points worth refreshing are:

- ABC classification (and its ABC/XYZ extension) is based on historic data and from that point of view it does not add any additional information about the product itself (volume, cost and even volatility are known),
- it does not capture physical aspect of goods,
- it does not capture any manufacturing or upstream supply chain constraints,
- it assumes fixed number of clusters and thresholds used for split of products into groups.

All above points will be taken into consideration in this hands-on chapter where we will go through clustering process using Unsupervised Machine Learning (UML) methods to provide potential alternative to ABC classification.

We will start with selection of variables and finding optimal number of clusters, then we will build clustering models using various UML methods and we will compare their results. Chapter will conclude with a summary where we will brainstorm advantages and downsides of model deployment to Supply Chain Management.

5.1. Selection of Variables and Number of Clusters

One of the downsides of ABC/XYZ classification is limited selection of variables. It only uses information about the volume/value of products and volatility of their demand. The problem with this approach is that it takes several periods to evaluate realistic volatility of demand which means ABC/XYZ cannot be used for new or newly launched products, unless certain assumptions regarding demand volatility are taken.

In order to take this valid point into consideration we will take a different approach. First of all we will expand the selection of variables to include those that describe physical aspects of products and manufacturing and supply chain constraints (used in the planning process). Secondly we will only use variables that are known at all stages of product life-cycle including replenishment volume which is thoroughly evaluated throughout all stages (or gates as they are called) of New Product Implementation (NPI) process[89].

Taking this into consideration the final list of variables used for clustering include:

- Cost_per_pack,
- Units_per_pack,
- Number_of_periods,
- Total_replenishment,
- Replenishment_Type,
- SKU_reworked,
- Lead_comp_constr,
- Prod_site,
- Product_type.

As we can see both numeric and categorical variables will be used which can be considered as yet another differentiator between our approach and ABC classification.

Although certain criteria have been met in terms of variables selection, it was still an arbitrary decision based on expert knowledge. Contrary to this, selection of optimal number of clusters can be performed using various statistical tests and measures, however before we move to that step we should ask ourselves a fundamental question, namely can data be clustered.

To answer this question we can use Hopkins statistics, which is a cluster tendency measure and it acts as statistical hypothesis test where Null hypothesis (H0) is that the data is uniformly randomly distributed and hence cannot be clustered. Hopkins statistics score take values from 0 to 1, where a value close to 1 indicate that data is highly clustered and value close to 0 suggests that data cannot be clustered.

[89] Although the author is conscious that actual replenishment volume may change, NPI estimates hold until actual volume is known and are used across end-to-end supply chain processes.

As we perform this test on our data we get value 0.9957582 which means we can reject Null hypothesis in favor of Alternative hypothesis that data can be clustered. We can do a complementary visual check of clustering tendency using Ordered Dissimilarity Matrix (ODM) presented in Figure 37. Clear presence of readable rectangular shapes confirms outcome from test based on Hopkins statistics.

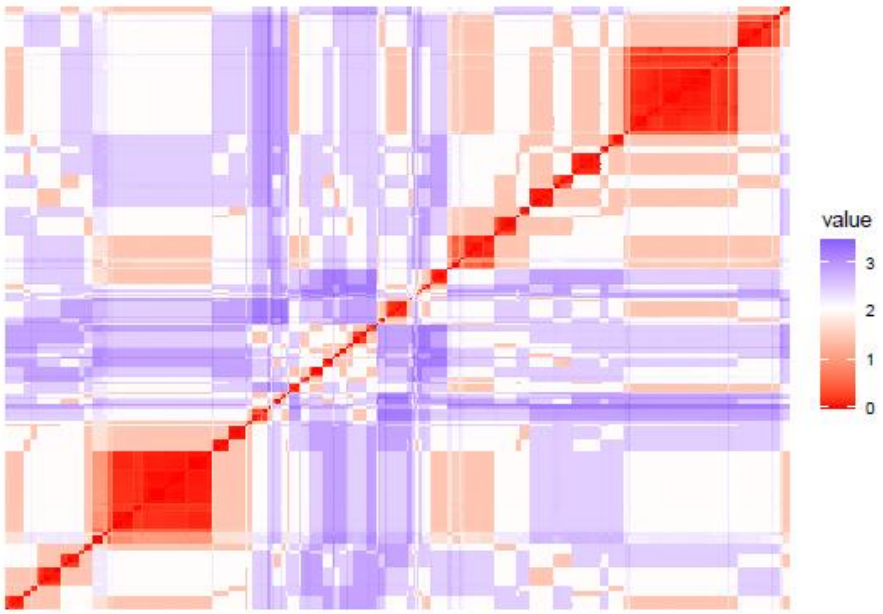

Figure 37: Ordered Dissimilarity Matrix (ODM) for selected variables. Since the more readable rectangles the higher tendency for clustering we can confirm that our dataset can be clustered.

Having checked that data can be clustered we can proceed with selection of optimal number of clusters. From technical perspective in R, this requires using one of clustering algorithms (methods) which will be presented closely in Chapter 5.2. To simplify narrative, at this stage I will only mention that 'K-means' algorithm has been used to complete this task and we will focus on measure and tests which identify optimal number of clusters.

Those measures and tests include:

- Silhouette measure,
- Total Within Sum of Square measure (also known as elbow plot),
- Gap statistic,
- Calinski-Harabasz index (CH) which is very helpful when above measures return not unanimous results; CH index performs a test on different numbers of centers (clusters) and the highest value indicates best fit (optimal number of clusters).

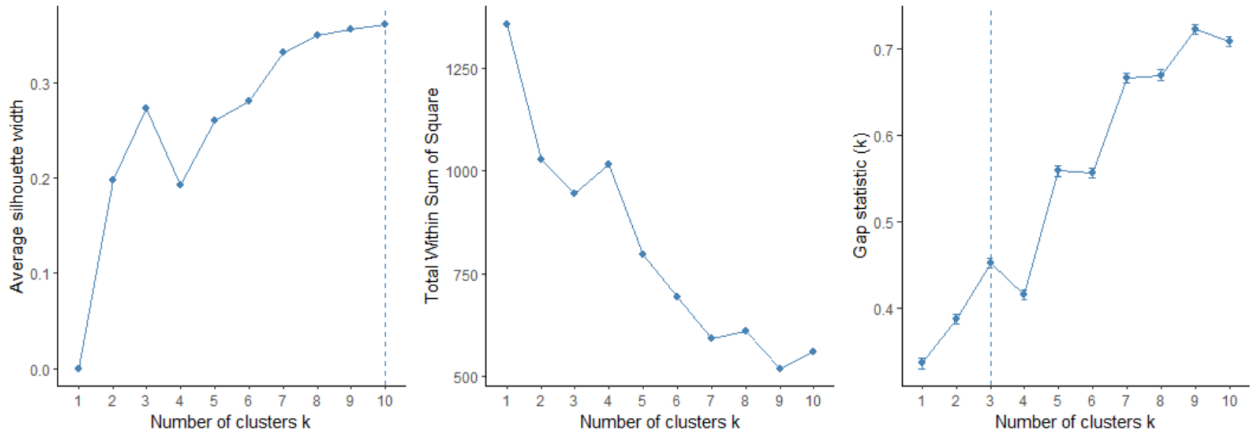

Figure 38: Indication for optimal number of clusters using three different measures: (Average silhouette width – Left, Total Within Sum of Square – Middle and Gap statistic – Right). Depending on measure, optimal number of clusters varies from 3 to 10.

Figure 38 summarizes results of first three measures. Optimal number of clusters on the basis of Silhouette measure is 10 (although 3 is also strong contender), Gap statistic is three and elbow plot (middle) suggests 7 or 9.

Since results are not unanimous let's perform Calinski-Harabasz test for 3, 7 and 10 clusters. Test results are as follows:

- 3 centers (clusters) - 127.1,
- 7 centers (clusters) - 109.0,
- 10 centers (clusters) -115.9

This concludes that there clusters are optimal as value of the test is highest (127.1).

5.2. Clustering

"There are two best-known clustering approaches: K-means clustering and hierarchical clustering. In K-means clustering, we seek to partition the observations into a pre-specified number of clusters. On the other hand, in hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n."[90]

The goal of K-means is to find K centroids (K is known a priori) which represent centers of clusters and assign each observation to the nearest centroid. The algorithm (for K clusters) is as follows:

- place K points (which will serve as the cluster centers) in the feature space,
- allocate each observation to the nearest center (centroid) based on Euclidean distance (or other distance),
- recalculate the position of centroids on the basis of the mean of all the points assigned to that centroid's cluster,
- repeat last two points until centroids stop changing position; this will suggest that the sum of distances of individual observations from centroids is as small as possible.

A variation of K-means is K-medoids and its common realization, Partitioning Around Medoids (PAM) algorithm. Steps of both algorithm are similar, although the main difference is that PAM uses medoids (instead of centroids) which usually are K of n observations in the first step.

Both algorithms divide dataset into subgroups which is a top-down procedure. Contrary to this, hierarchical clustering is a bottom-up process. On the basis of Euclidean distance (or other) and a linkage method (complete, average or single) observations are grouped together. The key benefit of hierarchical is it does not require to pre-specify the number of clusters, however we already know optimal number of clusters (as per Chapter 5.1.)

All three algorithms will be used in clustering process.

[90] An Introduction to Statistical Learning with Applications in R, page 386; G. James, D. Witten, T. Hastie, R. Tbishirani.

Prior to clustering data required some pre-processing, this include:

- normalization of numeric variables to the <0; 1> range to remove effect from different units of measures; this normalization method was selected since we also use categorical variables,
- order has been removed from ordinal factors,
- levels of qualitative predictors have been recoded as 0 and 1.

Figure 39 presents clustering outcome using K-means algorithm with 3 clusters and Euclidean distance reduced to two dimensions using Principal Components Analysis (PCA)[91]. First two Principal Components explain over 30% of variance in the model.

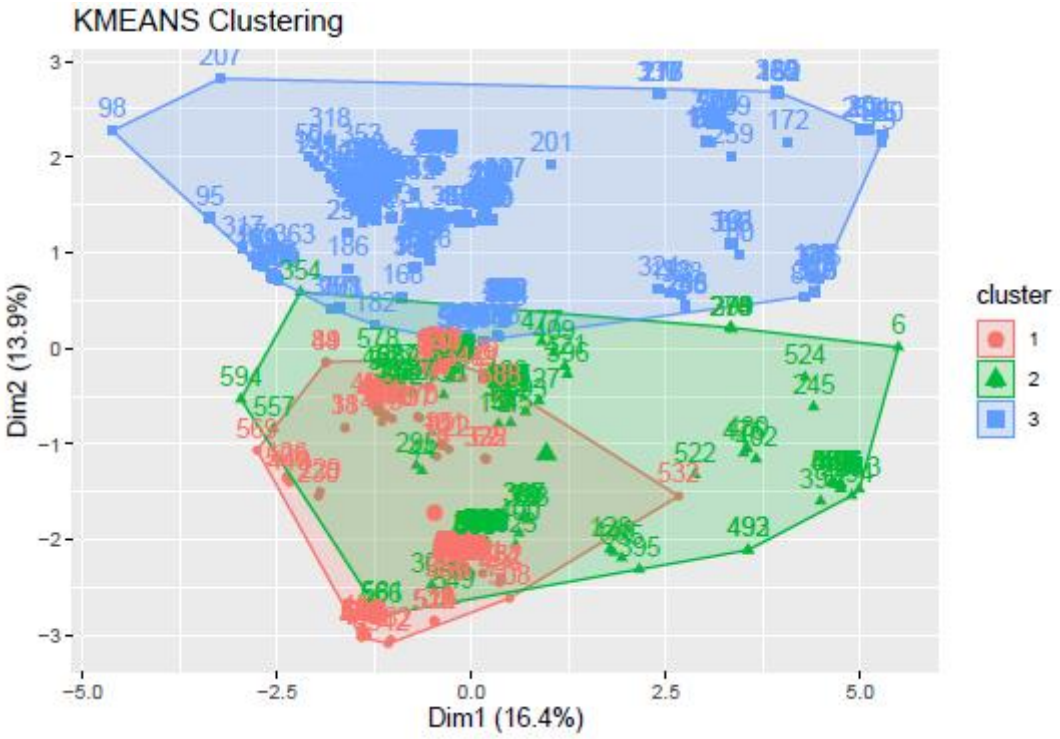

Figure 39: Visual representation of clustering outcome using K-means algorithm with 3 clusters and Euclidean distance reduced to two dimensions with PCA.

[91] PCA „finds a low-dimensional representation of a dataset that contains as much as possible of the variation. The idea is that each of the *n* observations lives in *p*-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are interesting as possible, where the concept of *n*-interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the *p* features."; An Introduction to Statistical Learning with Applications in R, page 375; G. James, D. Witten, T. Hastie, R. Tbishirani.

K-means clustering diagnostic can be performed with Average silhouette width measure which has been presented in Figure 40.

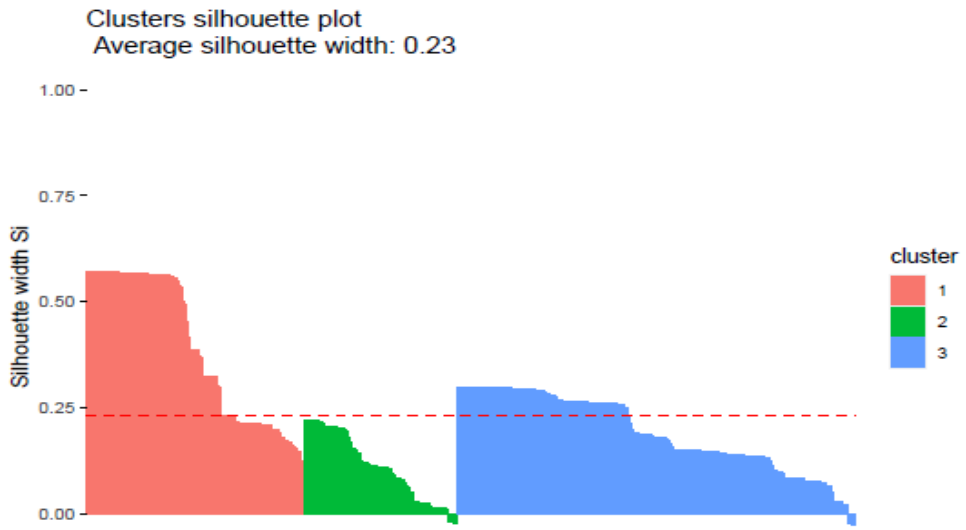

Figure 40: Clusters silhouette plot for K-means algorithm with 3 clusters and Euclidean distance.

Average silhouette width is 0.23 and the score for individual clusters (together with clusters size) is as follows.

```
cluster size ave.sil.width
      1  170          0.39
      2  119          0.11
      3  311          0.19
```

Above results although not perfect (silhouette width 1) are acceptable. Figure 42 shows that there are only a few observations (from cluster 2 and 3) which fall into wrong cluster (silhouette width < 0).

Let's investigate if PAM algorithm delivers better results. Clustering outcome and diagnostic has been performed in analogical way to K-means clustering and are displayed in Figures 43 and 44.

Average silhouette width for PAM algorithm with 3 clusters and Euclidean distance is 0.21 and its score for individual clusters (together with their size) is as follows.

```
cluster size ave.sil.width
      1  249          0.18
      2  123          0.18
      3  228          0.27
```

Although PAM algorithm delivers a more balanced clustering outcome, both in terms of size of individual clusters and score of average silhouette width (ranging from 0.18 to 0.27), overall value of this metric is 0.21 and it is lower than K-means which is preferred top-down clustering method.

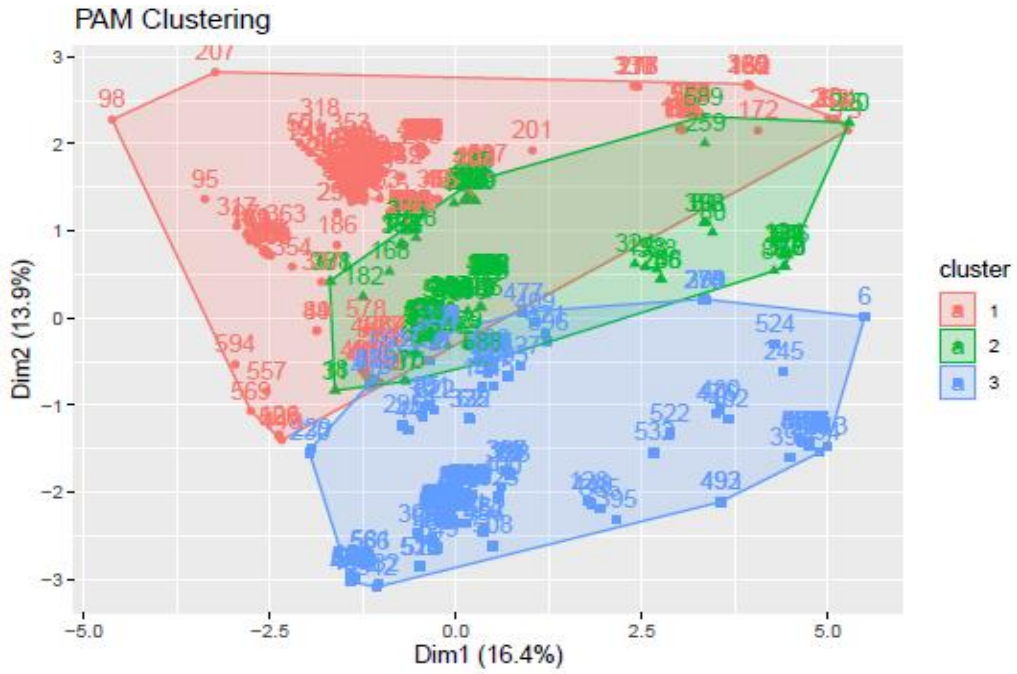

Figure 41: Visualization of PAM algorithm with 3 clusters and Euclidean distance reduced to two dimensions with PCA. Similar to K-means, PCA covers over 30% of variance in the model.

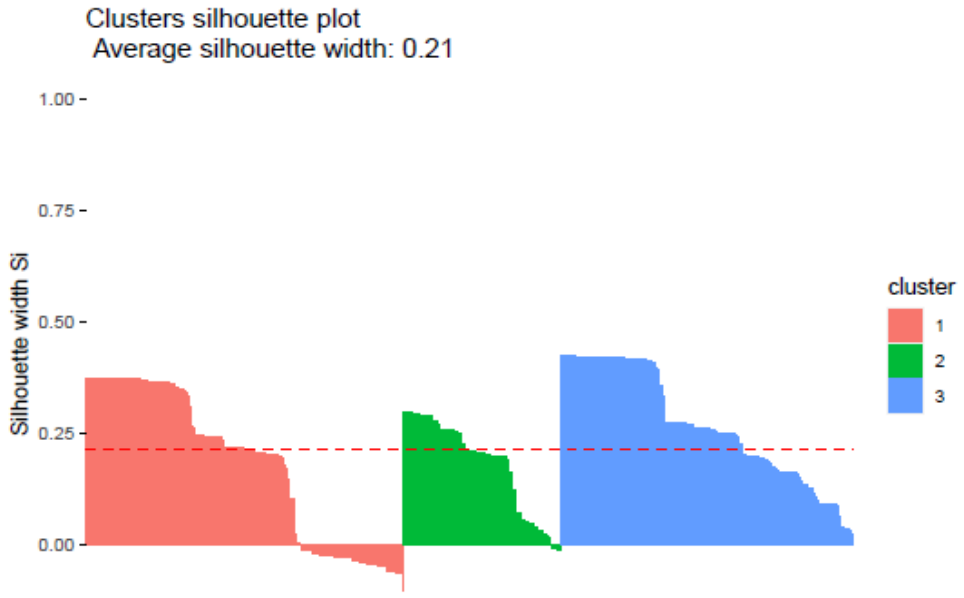

Figure 42: Clusters silhouette plot for PAM algorithm with 3 clusters and Euclidean distance. Overall score of silhouette width is 0.21 which is worse result than K-means. We can also see significant number of observations is cluster 1 with negative score of silhouette width which suggest they have been clustered incorrectly.

Hierarchical clustering has been made on the basis of Euclidean distance and complete[92] linkage. Its dendrogram is presented in Figure 43.

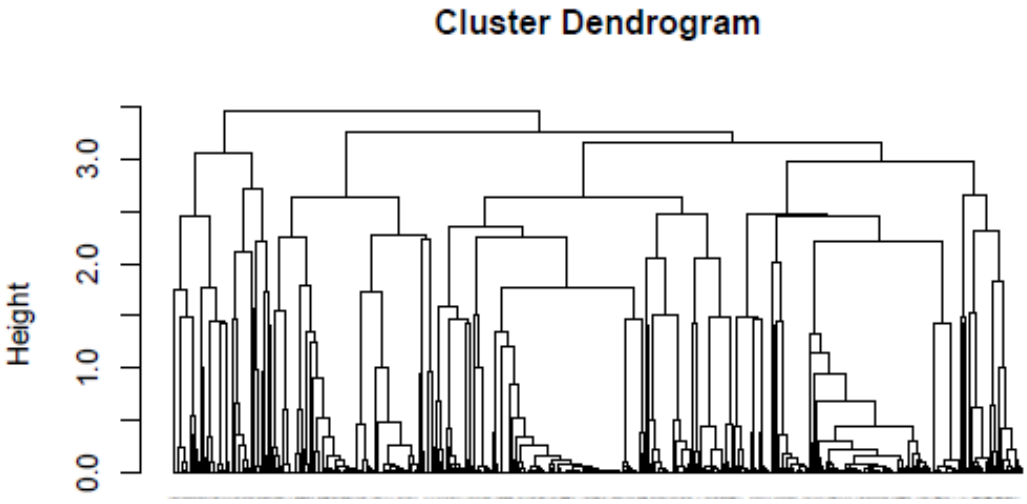

Figure 43: Cluster dendrogram for hierarchical clustering with Euclidean distance and complete linkage.

Cutting dendrogram in three distinct clusters results in clusters of below sizes.

```
comp_clusters
  1   2   3
414 117  69
```

Cluster 1 with 414 observations holds approx. 70% of all observations which is very unbalanced comparing to top-down methods. From that point of view, top-down K-means algorithm is still our preferred clustering method.

[92] „Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities."
An Introduction to Statistical Learning with Applications in R, page 395; G. James, D. Witten, T. Hastie, R. Tbishirani.

5.3. Overview and Comparison of Clusters

Let's summarize visually each of 3 clusters based on K-means algorithm.

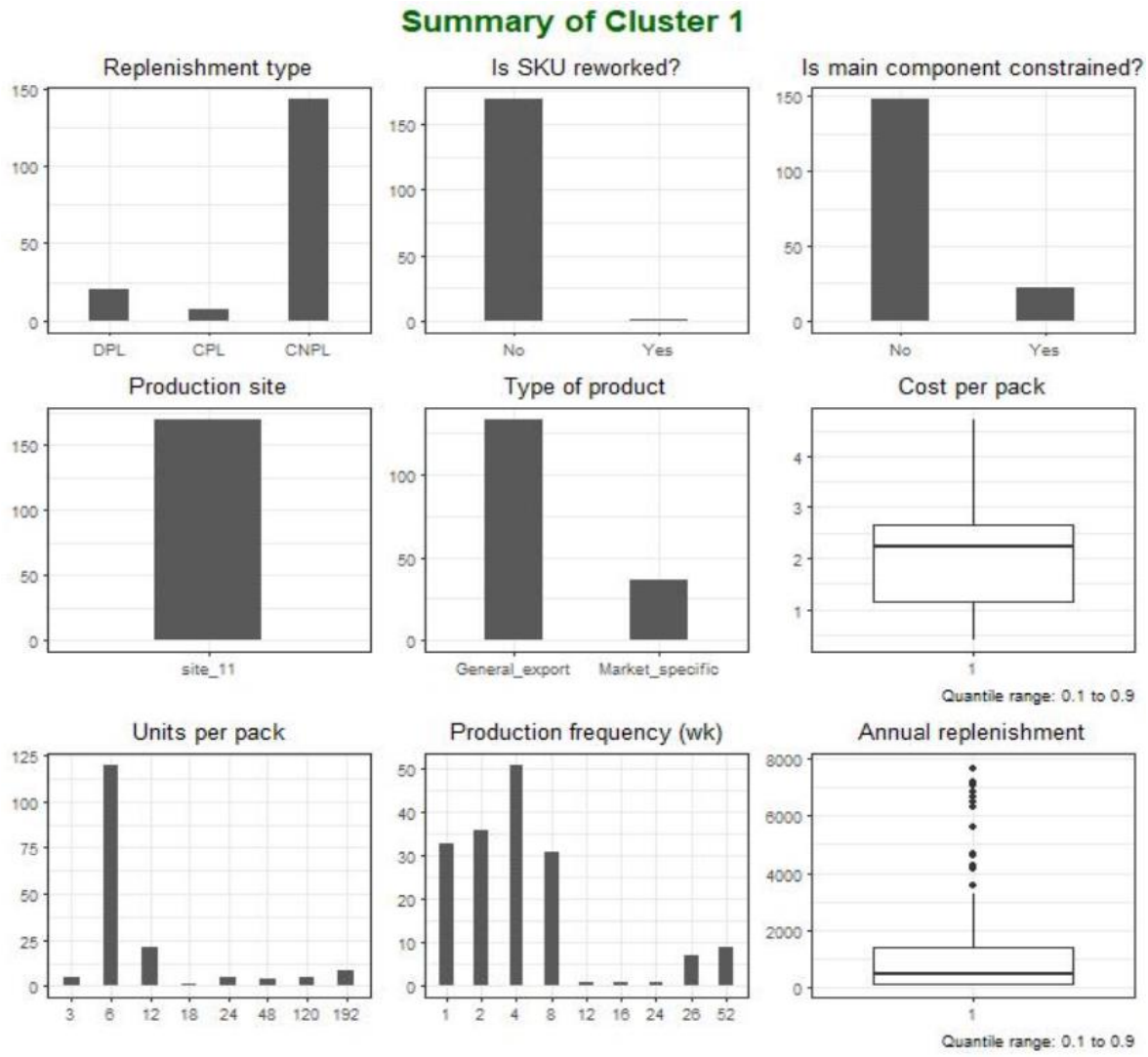

Figure 44: Key characteristics of observations in cluster 1.

Cluster 1 can be summarized as observations:

- exclusively produced in site_11,
- (with small exceptions) with one stage of manufacturing process (no rework is required),
- typically produced for General Export,
- mostly 6-packs with median of replenishment volume of approx. 1k,
- consolidated outside of production location for distribution to final customers.

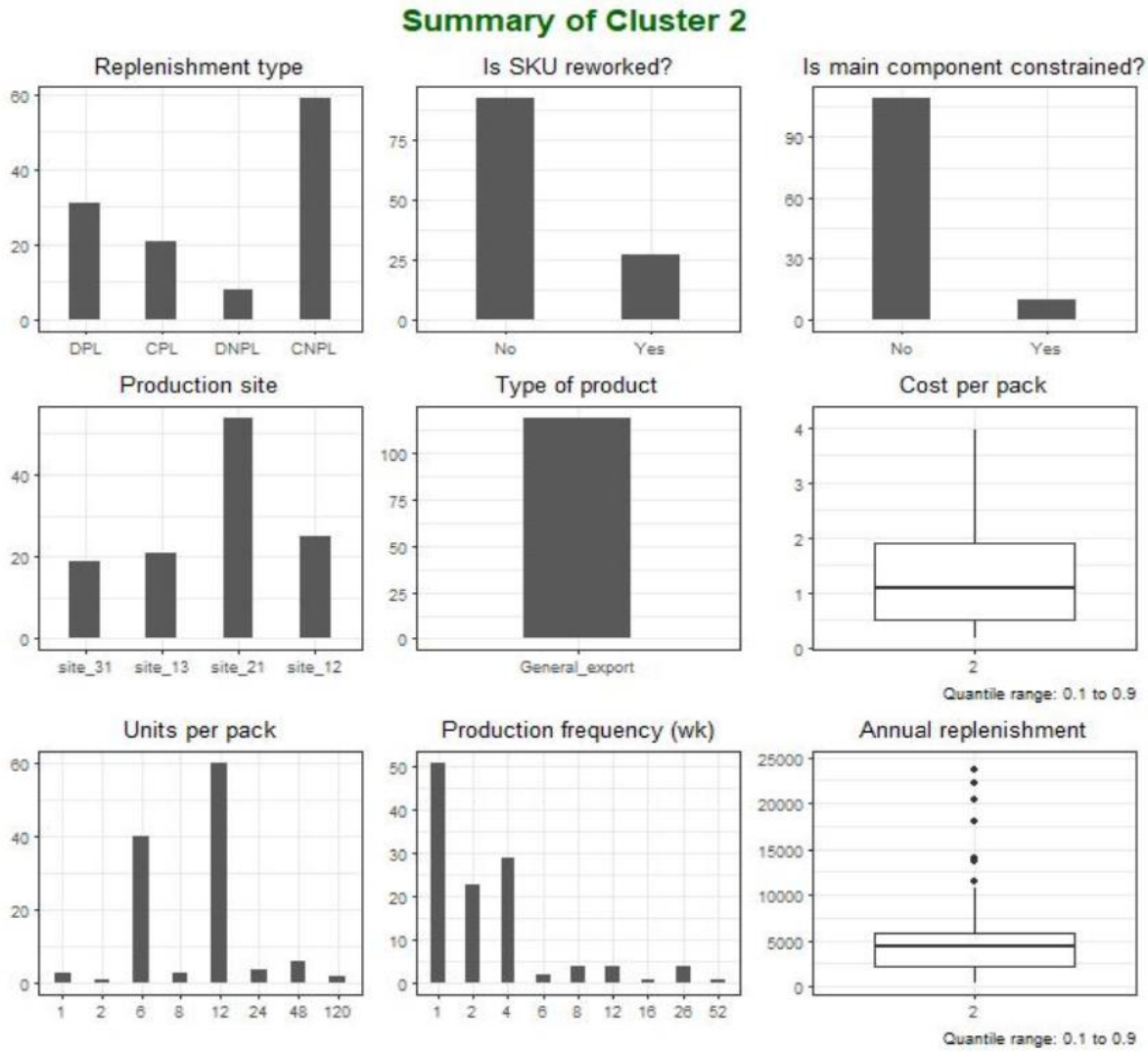

Figure 45: Key characteristics of cluster 2.

Observations in cluster 2 are:

- exclusively produced for General Export,
- mostly consolidated for shipment outside of production location,
- median of annual replenishment below 5k packs,
- produced frequently (every 1 to 4 weeks),
- produced mostly in site 21.

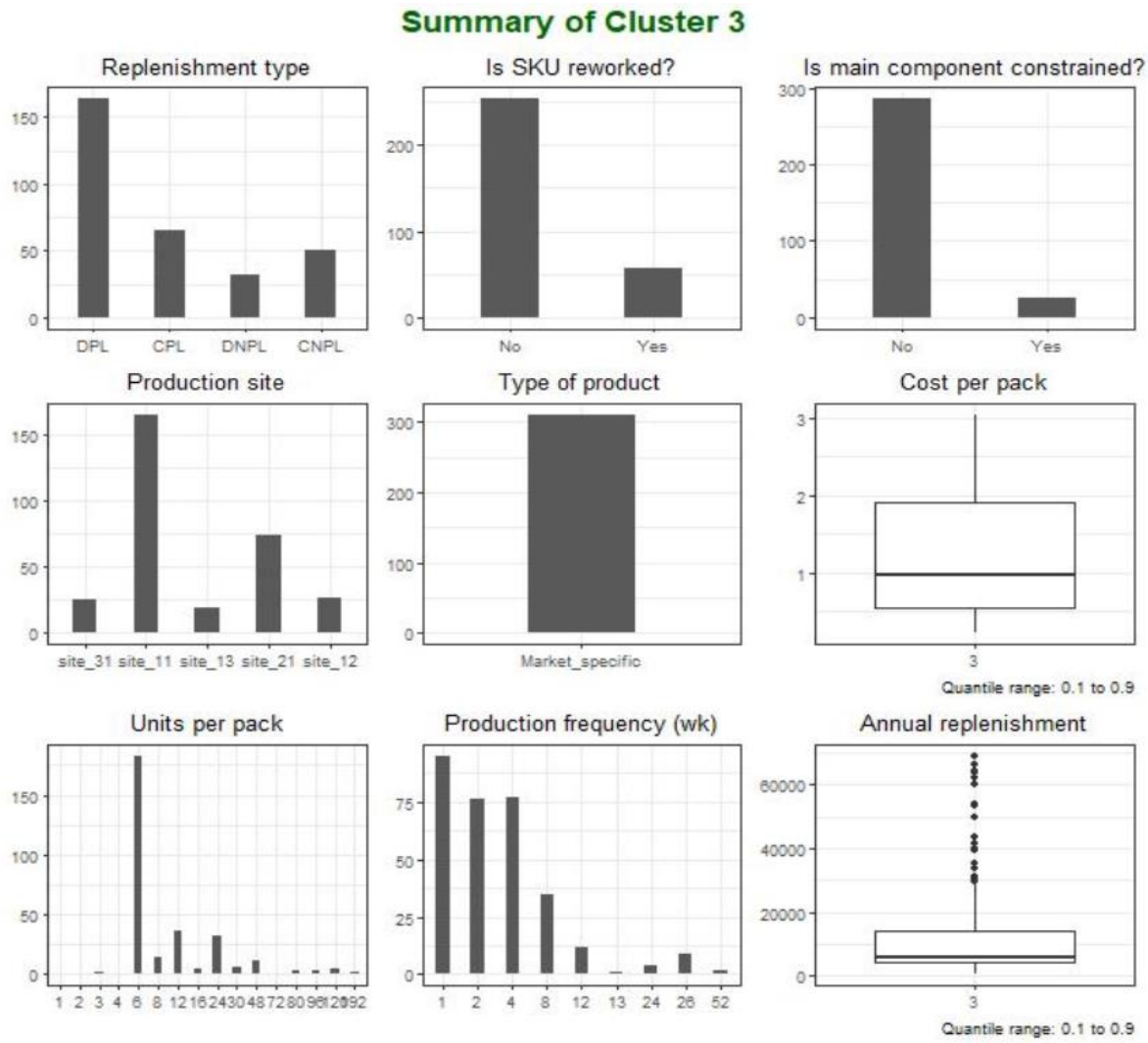

Figure 46: Key characteristics of cluster 3.

Key features of observations in cluster 3:

- exclusively Market specific products (observations),
- mostly 6-packs produced in site_11 every 1 to 4 weeks,
- dispatched mostly directly from production location (Replenishment_type DPL).

Graphical comparison of clusters is presented in Figure 47.

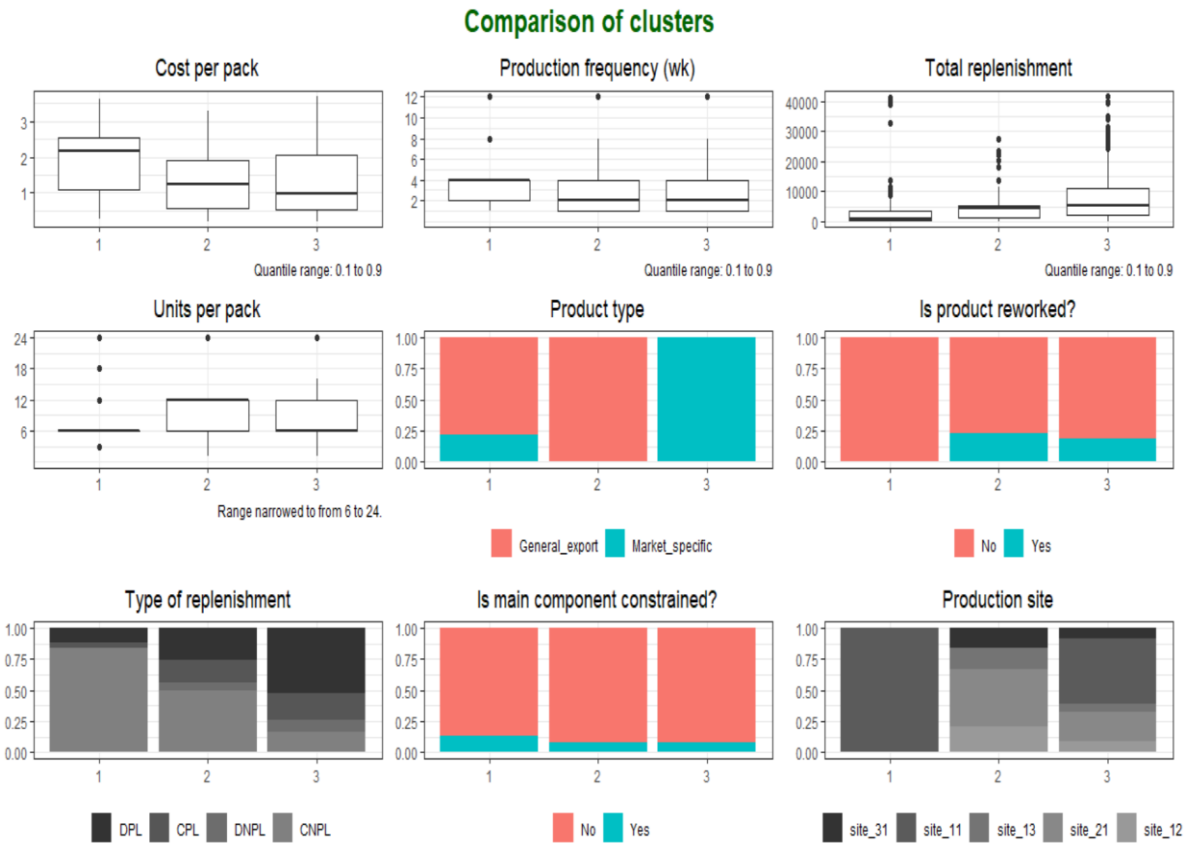

Figure 47: Comparison between K-means clusters. Key differentiators are Product type, Rework and Production site as some clusters consist of homogenous observations taking into account one of these features.

Our clusters can be also plotted on a classic ABC scale where Cost of goods sold (Cost_per_pack) is displayed on Y-axis and Annual replenishment volume (Total_replenishment) on X-axis.

Such plot is presented in Figure 50 and with quick glance we can tell that our alternative approach, which captures additional features, has generated significantly different results to classic ABC classification based on volume and cost.

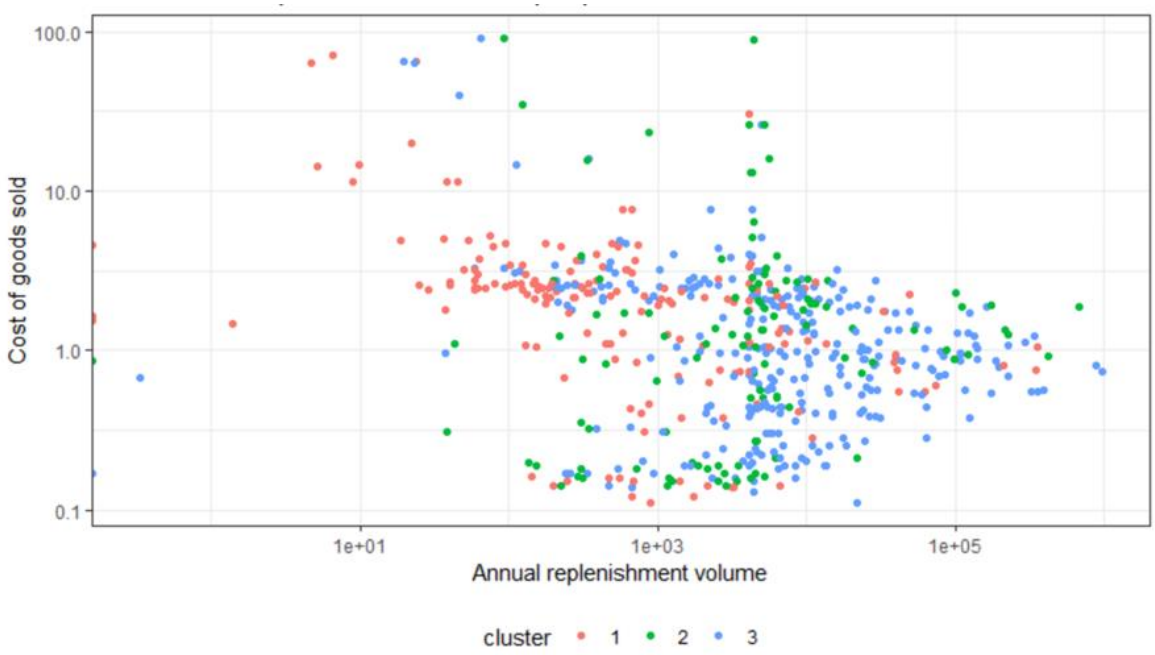

Figure 48: K-means clusters on Cost_per_pack (Y-axis)/Total_replenishment (X-axis) plot. This simple graphical comparison of both approaches tells that since newly created clusters overlap and are not grouped in line with ABC classification thresholds (based on volume/value of products) our alternative approach delivers significantly different clustering outcome comparing to ABC.

5.4. Conclusions and Summary

Clustering methods under the umbrella of Unsupervised Machine Learning can be successfully used for portfolio segmentation. They give answers to key questions like can data be clustered and what is optimal number of clusters. Approach taken to products classification in this chapter can also be applied to all products, regardless of their life cycle stage as we have limited the use of features based on historic data. In addition to that we have included features that describe physical aspects of products and manufacturing and (upstream) supply chain constraints. This is a more mature and cross-functional approach which is a good alternative to ABC classification (or at least can enrich this method) in terms of setting up correct replenishment strategy and service offering for each product.

Practical aspects of deployment of our method can be however problematic. It is more complex approach than ABC (which strength lies in its simplicity) and requires high maturity and data-oriented culture of the planning team and the management which often require structural and mindset changes in the organization.

These changes are though necessary in today's world where business growth comes from innovations and high customization of products (which means more and more value is created by outliers). In this case, replenishment strategy maximizing service needs to capture all bottlenecks across Plan, Make and Move processes. Ideally, customization of products should be paired with customization of replenishment strategy which would lead to situation where each product is a cluster on its own. Such approach requires however advanced planning system embedded in cross-functional processes, high maturity of the planning team, data-oriented mindset and strong data foundations.

Alternative approach to ABC classification presented in this chapter can be considered as intermediate solution, which captures cross-functional challenges related to portfolio management.

Summary

The purpose of this thesis was to merge two worlds: Data Science and Supply Chain by showing examples of the practical application of Machine Learning in Supply Chain Management on the basis of actual (but randomized) data from a FMCG company that operates on a global scale.

We started with the concept of Supply Chain Management, SC vocabulary, its role and goals and performance metrics (KPIs). Key SCM challenges related to performance of SC and portfolio management have been reflected into three research areas:

1. relationship between stock-out and physical aspects of SKUs, upstream supply chain and manufacturing constraints and quality of demand signal cascaded from local DC (Market) down the supply chain,
2. prediction of stock-outs on the basis of available predictor variables,
3. development of portfolio segmentation strategy as an alternative to ABC classification.

Investigation of relationship between stock-out and predictor variables has been done with Exploratory Data Analysis (EDA).

Some of the findings are in line with expectations:

- it is easier to manage smaller portfolio in terms of stock-out mitigation both from manufacturing (Prod_site) and distribution perspective (Market),
- replenishment requirements error (WAPE) measured one month before actual shipping date (WAPE_Month_1) is lower than error measured three months before actual shipping date (WAPE_Month_3).

Other findings are rather surprising:

- Median of WAPE_Month_3 is significantly lower for products with reported out_of_stock (which could trigger a question regarding value added from long term forecasting in the environment of reactive supply chain),
- higher production frequency does not guarantee less stock-outs in the market so selecting production cycle should be considered as a trade-off between

manufacturing KPIs like OEE and SC inventory KPIs like Average Inventory Turns,

- direct dispatch does not improve stock availability; on the contrary, less stock-outs was observed for consolidated products (it might be driven by Safety Stock kept at shipping point, however data to validate this hypothesis is not available).

Above findings can suggest the presence of additional variables which are impactful on out_of_stock and yet have not been captured in research dataset.

In Chapter 4 we also learned that even sophisticated Machine Learning models (so-called 'black boxes') struggled with stock-out prediction.
Comparison of various methods (linear regression, regularized linear regression, naive bayes, knn, decision tree, random forest, bagged trees, boosted trees, support vector machine and neural network) showed that MLP model (based on neural network) has delivered best yet rather disappointing results. Balanced Accuracy of MLP model is approx. 5 percentage points above No Information Rate which means its predictive power is higher than random guess, however AUC score of this model falls in 'poor' category which makes it unreliable. We tried to brainstorm root cause of this issue and we have identified a few key areas, namely:

- reactive nature of supply chain planning,
- data collection methodology,
- unavailability of important predictors.

Last two points can be considered as improvement areas which could result in better models, also in a form of 'black box' which, as shown in Chapter 4, can be made transparent for end-user.

Development of portfolio segmentation as an alternative to ABC classification (in the scope of Chapter 5) proved to be much more successful. Various improvement opportunities have been identified and embedded in the modelling process.

Key ones are:

- statistical tests confirming that data can be clustered,
- identifying of optimal number of clusters on the basis of statistical tests and measures,

- inclusion of predictor variables that describe physical aspects of products and manufacturing and upstream SC constraints,
- exclusion of predictor variables that describe historic data
- identifying clustering method that delivers best segmentation results.

As a result of that and we got three clusters applicable to all products regardless of the stage of their life-cycle.

The final conclusion from this chapter was that our approach (although more mature than ABC classification), in the absence of advanced planning system supporting cross-functional processes, should be considered as intermediate step towards creation of ultimate replenishment strategy that considers each product as individual cluster which could be paired up with products customization to maximize service and profits.

Bibliography

1. Association For Supply Chain Management (2019), APICS Dictionary. The essential supply chain reference. Sixteenth edition.
2. Association For Supply Chain Management, SCOR Metrics, http://www.apics.org/apics-for-business/benchmarking/scormark-process/scor-metrics, (accessed 01.09.2020).
3. Biecek Przemysław (2018), DALEX: Explainers for Complex Predictive Models in R., Journal of Machine Learning Research 19., https://jmlr.org/papers/volume19/18-416/18-416.pdf, (accessed 25.09.2020).
4. Davies Tilman M. (2016), The Book of R. A first course in programming and statistics.
5. iQualifyUK, https://www.iqualifyuk.com/library/business-management-section/the-eight-components-of-supply-chain-management/, (accessed 03.09.2020).

   References:

   Ballou, R.H., 2007. Business Logistics/supply Chain Management: Planning, Organizing, and Controlling the Supply Chain. Pearson Education India.

   Christopher, M., 2016. Logistics & Supply Chain Management. Pearson UK.

   Lambert, D.M. & Cooper, M.C., 2000. Issues in Supply Chain Management. Industrial Marketing Management.

   Tan, K.C., 2001. A framework of supply chain management literature. European Journal of Purchasing & Supply Management.

6. James Gareth, Witten Daniela, Hastie Trevor, Tbishirani Robert (2015), An Introduction to Statistical Learning with Applications in R.
7. Kuhn Max, Johnson Kjell (2013), Applied Predictive Modelling.
8. Lantz Brett (2013), Machine Learning in R.
9. Molnar Christopher (2020), Interpretable Machine Learning. A guide for Making Black Box Models Explainable., https://christophm.github.io/interpretable-ml-book/agnostic.html, (accessed 24.09.2020).
10. Poluha Rolf G. (2016), The Quintessence of Supply Chain Management; What You Really Need to Know to Manage Your Processes in Procurement, Manufacturing, Warehousing and Logistics.
11. Wickham Hadley, Grolemund Garrett (2017), R for Data Science.

Appendix - List of R libraries

Chapter 2:

simputation, tidyverse (ggplot2, tidyr, dplyr, readr, purrr, tibble, stringr, forcats, readxl, rvest, lubridate, magrittr), naniar, ggpubr, skimr, broom, patchwork, glmnet, caret, corrplot, fitdistrplus

Chapter 4:

skimr, patchwork, gridExtra, tidyverse (ggplot2, tidyr, dplyr, readr, purrr, tibble, stringr, forcats, readxl, rvest, lubridate, magrittr), verification, caret, tictoc, pROC, keras, DALEX, knitr

Chapter 5:

clustertend, fpc, psych, factoextra, dendextend, patchwork, wesanderson, caret, tidyverse (ggplot2, tidyr, dplyr, readr, purrr, tibble, stringr, forcats, readxl, rvest, lubridate, magrittr), patchwork