Pythia's Advice[1]

Floris Padt Ieke Le Blanc

May 11, 2025

[1]Mentor: {{}} - EAISI Academy, MASTERING DATA & AI, Technical University Eindhoven

Table of Contents

Chapter 1

Preface

Acknowledgments

I am grateful for the insightful comments offered by **Ieke le Blanc**.

These have improved this study in innumerable ways and saved me from many errors; those that inevitably remain are entirely my own responsibility.

1.1 How to Read This Book

This guide will help you navigate through this book and make the most of its features.

1.1.1 Book Structure

This book is organized into the following main sections:

- **Preface**: acknowledgments and how-to read this book
- **Main Research**: exploration of each phase of the CRISP-DM methodology in the Research Analysis
- **Appendix**: coding, references, and supplementary materials on each CRISP-DM phase.

1.1.2 Navigation Features

This book includes several features to enhance your reading experience:

Table of Contents

The sidebar on the left provides a full table of contents. Click on any section to navigate directly to it.

Search Function

Use the search icon ( ) in the top navigation bar to search for specific terms throughout the book.

Code Blocks

```
x <- 41   # code line 1
x <- 42   # code line 2
```

- Code blocks have a copy button ( ) in the top-right corner
- Some code blocks can be expanded to show more details

Cross-References

This book contains cross-references to:

- figures (Figure 20.1)
- definitions (def 1.1)

- sections (Research Analysis)

Interactive Elements

Some sections contain interactive elements:

- Click on images to enlarge them

- Interactive visualizations allow you to explore data
- Collapsible sections can be expanded for additional details

1.1.3 Reading Pathways

1. Start with the Research Analysis section which is designed to be read in a linear fashion following the CRISP-DM phases.
2. In the Appendix the related coding and detailed analysis can be found on each phase of CRISP-DM, e.g. Business Understanding

1.1.4 Best Viewing Experience

For the optimal experience:

- Use a modern web browser (Chrome, Firefox, Safari, or Edge)
- Enable JavaScript for interactive features
- Use the dark/light mode toggle for your preferred reading environment

1.1.5 Downloading and Sharing

- Access the source code and data in GitHub with the github icon.
- The pdf icon will open the book as a PDF.
- In section Chapter 28 the pptx and other resources can be locally saved( ).
- Use the share buttons to share specific chapters on social media.

def 1.1. DEF:

This is an example of a definition DEF = $\sqrt{mean(e_t^2)}$

Part I

Main Research

Chapter 2

Prophecy to Prediction

The Myth Inspiring Pythia's Advice

From Prophecy to Prediction: The Myth Inspiring Pythia's Advice

The *Pythia's Advice* project is dedicated to the art of sales forecasting. Drawing inspiration from the ancient myth of the Oracle of Delphi, where the *Pythia*—the high priestess—delivered cryptic prophecies, we harness the power of Data Science—our modern-day Pythia—to foresee future trends. Just as the Pythia's prophecies required careful interpretation, our statistical and machine learning techniques produce *Advice* that requires human expertise for effective application. By embodying this union of technology and human insight, we aptly named our project *Pythia's Advice*. This synergy ensures well-informed, data-driven decision-making and enhances the operational efficiency of the supply chain.

According to ancient myth, the Pythia inhaled vapors emanating from the remnants of the Python slain by Apollo—the Olympian god of the sun, music, and prophecy. These vapors induced a trance-like state that allowed her to channel his prophetic insights, revealing future events. Similarly, our project 'inhales' vast amounts of data—our 'Python's Vapors'—which, when processed with complex algorithms, enable us to decipher hidden patterns and unveil valuable foresight. However, these predictive insights must be carefully interpreted to avoid missteps—much like when King Croesus misread the Pythia's prophecy that *'a great empire will fall.'* He waged war and, in doing so, opened Pandora's Box—leading to the downfall of his own *great empire*—the Kingdom of Lydia.

Furthermore, *Pythia's Advice* emphasizes the crucial bond between humans and nature. By focusing on the sales forecast of *food for biodiversity*—products that support eco-

logical diversity—we aim to benefit human health and promote planetary well-being. This harmonious relationship mirrors the Pythia's sacred connection with the divine, symbolizing a balance between technological capabilities and the natural world.

Our logo, featuring a tree integrated with the *Pythia* through conductive traces against an enlightening sunset that symbolizes guidance and inspiration, beautifully encapsulates this synergy. This imagery symbolizes the fusion of natural wisdom and technological innovation. It reminds us that while we rely on cutting-edge technology to predict the future, our roots remain deeply intertwined with nature.

Through *Pythia's Advice*, we honor the legacy of the Oracle of Delphi, blending ancient wisdom with modern technology to illuminate the path forward.

Chapter 3

Research Analysis

TRUST - is all you need!

Chapter 4

Introduction

In an increasingly complex market environment, accurate demand forecasting is critical for ensuring the operational efficiency of supply chains, particularly in industries like wholesale, where products are perishable and sales are affected by seasonality and promotions. This research focuses on a wholesale company that buys, produces, and distributes biodiversity food to retailers, Out-Of-Home, discounters and E-commerce. To make informed decisions regarding production, buying, and negotiation with customers, accurate sales forecasts are required at different aggregation levels, depending on each use case.

While the company employs robust but dated statistical methods, such as moving averages and exponential smoothing, these techniques struggle to handle promotional impacts, capture complex sales patterns, and provide prediction intervals. Currently, a combination of bottom-up and top-down forecasting approaches is used, largely determined by the judgment of individual demand planners. However, the lack of clear guidance on which approach to apply in different scenarios has resulted in inconsistent and subjective forecasting outcomes. Furthermore, various use cases have distinct aggregation and time horizon requirements, adding complexity to maintaining consistent and accurate forecasts across the system.

The primary objective of this research is to determine the best way to generate a forecast, which methods to use and on what levels of granularity, enabling consistent, accurate forecasts that can be aggregated or dis-aggregated as required for different business needs. Specifically, the study aims to address the following key questions:

1. Which robust forecasting methods—statistical or data science techniques—should be used to meet the needs of each use case?
2. Which levels of aggregation provides the most accurate forecast for different use cases, and what when is an additional level justified by increased forecast accuracy.

3. What forecast accuracy improvements can be achieved through the integration of additional data, and what are the trade-offs or costs associated with these enhancements?
4. How should human resources be allocated effectively, given the large number of products (SKUs) and varying levels of importance across these products?

Ultimately, this research will provide an objective, data-driven strategy to improve forecast accuracy, reduce reliance on individual intuition, and ensure that forecasts are robust, consistent, and scalable across multiple aggregation levels, thereby enhancing overall decision-making efficiency.

4.1 Methodology

This research follows the **CRISP-ML methodology**, (Costa, 2022) as the guiding framework.

CRISP-ML is an acronym for "Cross-Industry Standard Process for Machine Learning." It is a systematic framework for organizing and executing machine learning projects. The methodology includes six key steps:

1. understanding the problem
2. preparing the data
3. selecting and tuning models
4. evaluating performance
5. deploying the solution
6. monitoring and maintaining the model.

This thesis focuses on the first 4 steps, the results will be taken as advice to be implemented in the current way of working.

Chapter 5

Business Understanding

where the business process is mapped and when the main indicators are identified, as well as when the business objectives are defined What is the business's need?

Figure 5.1: Why forecasting?
The forecast drives the business!

See Figure Figure 20.1 for the thumbnail overview.

Chapter 6

Data Undestanding

Chapter 7

Modeling

Chapter 8

Evaluation

1.2
floris@smart-r.nl
?meta:subtitle

> **Note**
>
> Note that there are five types of callouts, including: `note`, `warning`, `important`, `tip`, and `caution`.

> **Tip with Title**
>
> This is an example of a callout with a title.

> **Expand To Learn About Collapse**
>
> This is an example of a 'folded' caution callout that can be expanded by the user. You can use `collapse="true"` to collapse it by default or `collapse="false"` to make a collapsible callout that is expanded by default.

Chapter 9

Deployment

Chapter 10

Conclusion

Glossary

Chapter 11

Conclusions

Conclusions subtitle

3%

Part II

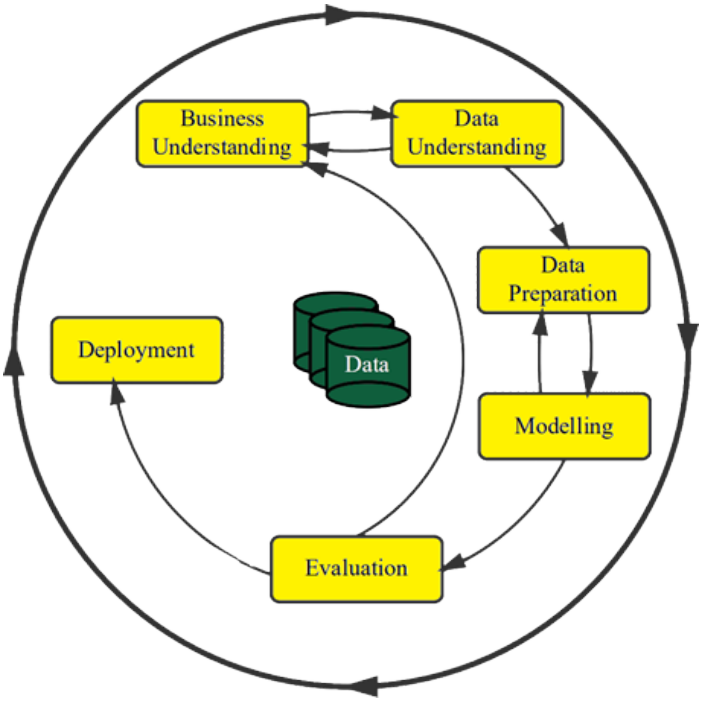
Appendix

Chapter 12

CRISP-DM

12.1 CRISP-DM

Figure 12.1: Project Plan



Chapter 13

Business Understanding

What is the business need?

Figure 13.1: Why forecasting?
The forecast drives the business!

See Figure Figure 20.1 for the thumbnail overview.

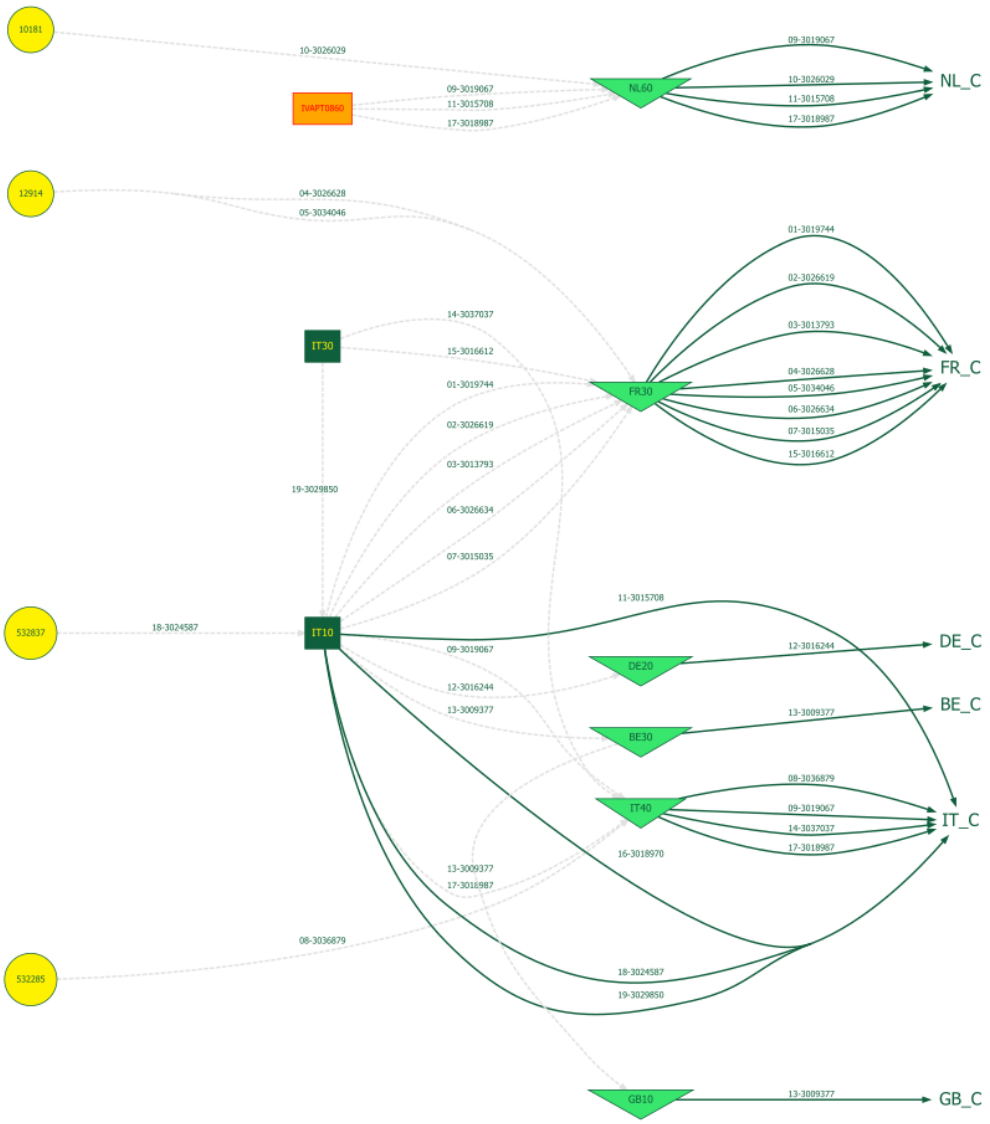
13.1 Defining the Scope of the ML Application

##Success Criteria - Business Success purpose and success criteria (reduction OOS&OOD) and other factors like customer & employee satisfaction - ML Success FA% and BIAS% - Economic Success impact of the project on the companies financial performance DAPS ## Feasibility availability, size, and quality of the training sample

13.1.1 Background

The company's forecasting requirements vary significantly across several key use cases, among others:

Figure 13.2: supply chain network (sample)



1. **Buying and Production**: Requires forecasts at the **material level** and **warehouse**, aggregated weekly, with a forecast horizon of **12 weeks**.
2. **Setting Safety Stock Profiles**: Requires forecasts at the **material level** and **warehouse**, aggregated weekly, with a forecast horizon of **26-52 weeks**, to determine safety stock profiles based on realized forecast accuracy.
3. **Negotiation with Customers**: Requires forecasts at a **monthly level**, ag-

gregated by **material categories** and **customer groups**, with a horizon of **6 months**.

4. **Production Cost Calculation**: Requires forecasts at a **quarterly level**, aggregated by material, with a horizon of **6 quarters**, grouped into annual time buckets.

The primary forecasting challenge is determining the optimal aggregation level and methods to use for generating forecasts that can meet the desired forecast accuracy of these different use cases while ensuring **consistency** and **accuracy**.

This inconsistency stems from a lack of objective criteria for selecting between forecasting methods—**bottom-up**, **top-down**, or **middle-out**—resulting in decisions that rely on the subjective judgment of demand planners, leading to variable performance and inefficiencies.

13.1.2 Business Objectives

The primary objective is to enhance operational efficiency by improving forecast accuracy across all key use cases. This will be achieved through the application of robust forecasting methods that can handle seasonality, promotions, and efficiently allocate scarce resources. Ultimately, these improvements will contribute to improved stock levels and a higher level of customer satisfaction, both key strategic goals for the company.

13.1.3 Business Success Criteria

Success will be measured by the following criteria:

1. **Improved Forecast Accuracy**:
   - **Tailored Forecasts for Each Use Case**: Enhanced accuracy across areas such as buying, production, safety stock settings, and customer negotiations, using methods that account for seasonality, promotions, and varying aggregation levels.
   - **Reduced Forecast Errors and Bias**: Measurable reductions in key error metrics like RMSE , MAPE , and BIAS will lead to better alignment between forecasted and actual sales. Increased accuracy will also foster trust among supply planners, reducing the need for excessive safety margins.
   - **Consistency Across Aggregation Levels**: Forecasts will remain consistent across different levels of aggregation, ensuring that detailed and aggregated forecasts are aligned.
2. **Informed Decision-Making for Forecasting System Functionality**:
   - **Selecting the Right Forecasting Methods**: This research will guide the selection of optimal forecasting methods (statistical, machine learning, or hybrid) to improve performance across various use cases.
   - **Efficient Resource Allocation**: A product classification scheme (e.g., ABC-XYZ) will be developed to help demand planners focus on high-impact

products, while low-impact products will be managed more efficiently through automated forecasts.

3. **Operational Benefits**:
   - **Improved Inventory Management**: More accurate forecasts will drive better purchasing and production decisions, leading to:
     - **Reduced Out-of-Stock (OOS) Incidents**: Ensuring product availability to meet customer demand and reduce penalties from stock-outs.
     - **Reduced Out-of-Date (OOD) Incidents**: Minimizing waste and ensuring the freshness of perishable goods, leading to lower storage costs and better inventory turnover.
   - **Optimized Safety Stock Levels**: Accurate forecasts will allow for better safety stock settings, reducing both overstocking and stock-outs.
   - **Cost Optimization**: Improved demand alignment will lower excess inventory and warehousing costs, minimize spoilage, and reduce costs associated with last-minute adjustments and overproduction.
4. **Enhanced Supply Chain Efficiency and Decision-Making**:
   - **Strategic Decision Support**: Consistent and accurate forecasts will improve decision-making across production planning, sales target setting, capacity management, and workforce optimization. These improvements will also support better customer negotiations and more precise COGS (Cost of Goods Sold) calculations, resulting in a more efficient supply chain.
   - **Increased Customer Satisfaction**: Improved product availability, fewer stock-outs, and fresher goods will not only reduce operational costs but will also enhance customer satisfaction and loyalty—a core strategic objective of the company.

By achieving these objectives, the company will see significant improvements in operational efficiency, cost-effectiveness, and most importantly, a higher level of customer satisfaction, which will ultimately drive long-term business success.

see upper part of the flow-down graph, (Section 25.1).

13.1.4 Inventory of Resources Requirements

Data sources:

The available data includes **historical sales**, **promotional data**, **stock quantities**, and **previous forecasts**. Stock-out periods will be adjusted for by inflating sales to reflect demand that would have occurred had stock been available.

Data Type	Details	Time Span
Sales	Customer Orders, Inter-company, Returns, Free-of-Charge and Missed	> 2020
Promotions	Promotions per material and customer	> 2020

Data Type	Details	Time Span
Stock	Daily stock level per material and distribution center	> 2020
Forecasts	generated forecast before- and after Demand Review	> Aug. 2023
Master data	Material, Customer and Organization, including hierarchy's	*N.A.*

Scoping

The scope of the project consist of two Marketing & Sales Organizations (MSO), with a focus on 6 brands, representing 5 out of the 6 categories, see Table 13.2. These materials are made-to-stock (MTS), requirements are planned on a weekly level, which represents 97% of the business.

This leaves us with about 1.000 materials out of 9.000 to focus on, making the assumption that these are representative for the rest of the products.

Table 13.2: Brands and Categories

Brand	Category
Bjorg	Dairy
Clipper	Tea / Coffee
Naturela	Tea / Coffee
Tanoshi	Meals
Alter Eco	Sweets
Zonnatura	Breakfast Cereals

13.1.5 Data Mining Goals

Predictive Model based on Correlations: Forecast Future Sales + Promotion effects

see lower part of the flow-down graph, (Section 25.1).

Data and Forecasting Methods

We will employ a combination of **statistical methods** and **machine learning techniques** to improve the forecasting process:

- **Current Methods**: The company's current forecasting system uses **moving averages**, **exponential smoothing** and **Box-Jenkins**, which face challenges with **Seasonality**, **Stationarity** and **promotional effects**. Forecasts are often inadequate in handling the impacts of promotions.

- **Proposed Methods**:
  - **ETS (Error, Trend, Seasonal)**: To replace exponential smoothing with a model that provides **prediction intervals** and better captures trends and seasonality.
  - **HTS (Hierarchical Time Series)**: To handle multiple aggregation levels, ensuring **forecast consistency** when forecasts are disaggregated or aggregated across levels.
  - **XGBoost**: A machine learning technique capable of handling **promotional impacts** and even identifying **cannibalization effects** of promotions between products.
  - **conformal prediction**: A method that provides **prediction intervals** and **calibrated forecasts**, ensuring that the forecasted values are within a certain confidence level.
  - **CatBoost and SHAP**: These will be used for feature analysis, determining which features (e.g., promotions, stock-outs) have the highest impact on forecast accuracy. This approach helps in choosing relevant input features to improve forecasts.

Consistency and Validation

One key requirement is ensuring that forecasts remain consistent across different levels of aggregation. This is where **HTS** will play a crucial role, reconciling forecasts generated at different aggregation levels to ensure that top-level forecasts align with aggregated lower-level forecasts. This will address the known issue of discrepancies between aggregated forecasts and directly forecasted aggregated levels.

Human Resources Allocation

The team consists of **X demand planners** working across **Y MSOs**. Currently, planners are involved in manually indicating promotional impacts and adjusting forecasts based on intuition, without systematic data cleaning for stock-outs or model tuning.

To optimize resource allocation, we plan to implement a **classification system** for the company's **9K SKUs**, utilizing an:

ABC-XYZ classification scheme:

- **ABC Classification**: Based on **sales volume**—focusing more resources on high-value items (A-class) and minimizing attention to low-volume items (C-class).
- **XYZ Classification**: Based on **sales variability**, indicating which products have stable demand versus those with erratic patterns.

- **Combined Classification**: These classifications will guide planners in determining where their efforts can have the most impact, focusing primarily on high-priority items while relying more on automated processes for low-priority ones.

13.1.6 Data Mining Success Criteria

The fiinal result will show an estimated relationship between forecast accuracy and the cost of resources required for the agreed scope. The cost of resources result from hardware, software and human resources needed for cleansing data, transforming data, model tuning and other manual activities needed.

To assess forecast accuracy, we will use a range of evaluation metrics tailored to each use case.

These metrics will help evaluate how well each approach performs against the unique requirements of each use case, ultimately guiding method selection and refinement.

- **RMSE (Root Mean Squared Error, def 13.1)**: The primary metric for measuring forecast accuracy, particularly useful for penalizing large errors.
- **MAPE (Mean Absolute Percentage Error)**: To align with existing company metrics.
- **MAE (Mean Absolute Error)** and **MASE (Mean Absolute Scaled Error)**: Additional metrics to provide a broader evaluation, addressing different aspects of forecast performance and comparing models to naïve baselines.

13.1.7 Project Plan

see Section 24.7

Initial Assessment of Tools and Techniques

see Literature review (Section 24.1)

def 13.1. RMSE:

Root Mean Squared Error

RMSE = $\sqrt{mean(e_t^2)}$

A Scale-dependent error, to compare forecast performances between data sets with the same unit.

Chapter 14

Data Understanding

Data Understanding subtitle

14.1 Data Collection Report

14.2 Data Description Report

14.3 Data Exploration Report

14.4 Data Quality Report

14.5 Identification of Data Sources

14.6 Data Quality Assessment

assessing the - completeness: impute missing data - accuracy: = consistency:

14.6.1 Data Description

A descriptive exploratory analysis describes the data by its statistical properties and metadata.

outliers?

14.6.2 Data Verification

Chapter 15

Data Preparation

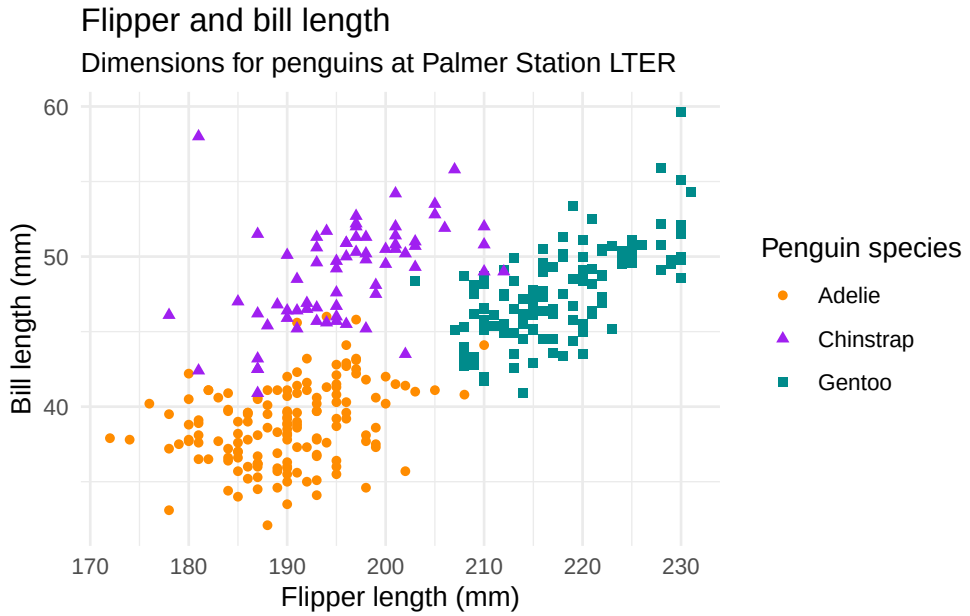
preparing the data for further analysis and modeling. This might involve cleaning and preprocessing the data, as well as transforming it into a format that is suitable for use in a machine-learning model.

5.1. Handling Missing Data

- imputing missing values, Imputation methods involve replacing missing values with estimated values.
- scaling numeric features
- encoding categorical features
- selecting a subset of the data to use in the model.

```
          uid              text   cnt
       <char>             <char> <int>
 1:         1       Electronics    11
 2:       1.4         Computers    11
 3:    1.4.10           Laptops    11
 4: 1.4.10.20        Ultrabooks    11
 5:    1.4.11          Desktops    11
 6: 1.4.11.21            Gaming    11
 7:    1.4.12           Tablets    11
 8: 1.4.12.22              iPads    11
 9:       1.5               TVs    11
10:    1.5.13               LED    11
11: 1.5.13.23           Samsung    11
12:         2              Home     9
13:       2.6           Kitchen     9
14:    2.6.14        Appliances     9
```

Figure 15.1: Bill length vs. depth for Palmer Penguins




```
15: 2.6.14.24             Ovens     9
16:    2.6.15           Cookware     9
17: 2.6.15.25              Pots     9
18:       2.7         Furniture     9
19:    2.7.16            Chairs     9
20: 2.7.16.26       OfficeChairs     9
21:         3            Garden     9
22:       3.8            Plants     9
23:    3.8.17           Flowers     9
24: 3.8.17.27             Roses     9
25:    3.8.18             Trees     9
26: 3.8.18.28              Oaks     9
27:       3.9             Tools     9
28:    3.9.19        PowerTools     9
29: 3.9.19.29            Drills     9
          uid              text   cnt
```

```
library(ggplot2)
library(data.table)
library(lubridate)
library(magrittr)
library(stringr)
```

```
library(openxlsx)

library(tidyverse)

LP0 <-
  function(x, width){
    # like CONVERSION_EXIT_MATN1_INPUT
    # only add leading zero's in case it is a number
    is_num <- grepl("^[0-9]+$", x)
    ifelse(
      is_num,
      stringr::str_pad(string = x, width = width, side = "left", pad = "0"),
      x
    )
  }

fOpen_as_xlsx <-
  function(pDT, pPath = "./Results", pFN, pasTable = TRUE){

    if (!dir.exists(pPath)) {
      dir.create(pPath)
    }

    if (missing(pFN) == TRUE) {
      pFN <- paste0("~", format(now(), "%Y%m%d-%H%M%S"), ".xlsx")
    }

    FFN <- file.path(pPath, pFN)
    write.xlsx(x = pDT, file = FFN, asTable = pasTable, tableStyle = "TableStyleMedium4")
    openXL(FFN)

  }
# PDAT <- file.path("C:", "PW", "OneDrive", "ET", "pythia", "dat")

PDAT <-
  switch(Sys.info()["nodename"],
         'TREX-TOAD' = file.path("U:", "floris"),
                       file.path("C:", "PW")
  ) %>%
  file.path("OneDrive", "ET", "pythia", "dat")

ET_CG <- "#0f5e3c"
ET_FG <- "#089b35"
```

```
MAT <- '000000000003036397'
MAN <- FALSE & substr(PDAT, 1,1) != "U"

BRNDS <-
  fread(text = "
  PRDH1, NAME
  07   , ALTER ECO
  08   , BJORG
  10   , CLIPPER (CUPPER)
  15   , ZONNATURA
  53   , TANOSHI
  65   , NATURELA"
  ) %>%
  .[, PRDH1:= str_pad(PRDH1, 2, pad = "0")]
```

15.1 SDMFRCAC FA

```
B4_SDMFRCAC2_A <-
  readRDS(file = file.path(PDAT, "B4_SDMFRCAC2_A.rds"))              %>%
  .[PLANT %chin% c("FR30", "NL60", "NL63")]

B4_MATERIAL_P  <-
  readRDS(file = file.path(PDAT, "B4_MATERIAL_P.rds"))               %>%
  .[MATL_TYPE %chin% c("FERT", "HALB")]

dtFA <-
  BRNDS[B4_MATERIAL_P[MATL_TYPE %chin% c("FERT", "HALB")],
        on = .(PRDH1 = PRDH1), nomatch = 0]                          %>%
  .[B4_SDMFRCAC2_A, on = .(SYSTID, CLIENT, MATERIAL), nomatch = 0]   %>%
  .[, .(
    ACT = sum(DEMND_QTY),
    FM1 = sum(DEMQTYM1)
  ), by = .(PRDH1, PLANT, MATERIAL, CALMONTH)
  ]                                                                  %>%
  .[, {
    E   = ACT - FM1
    E2  = E ^ 2
    AE  = abs(E)
    APE = 100 * (1 - AE / ACT)
    ECO = 100 * (1 - AE / FM1)
    .(
      PRDH1, PLANT, MATERIAL, CALMONTH,
      ACT  , FM1  , E, E2, AE, APE, ECO
```

```
      )
  }  ]

dtFA <-
  dtFA[
    !is.nan(APE) &
    !is.nan(ECO) &
    !is.infinite(APE) &
    !is.infinite(ECO) &
    APE > -100 &
    ECO > -100  ]
```

```
PYT <- file.path("C:", "PW", "OneDrive", "ET", "pythia", "upg", "data")

MATS <-
  fread(file.path(PYT, "MD_MATERIAL_SALES_ORG.CSV")) %>%
  .[, MATERIAL := LPO(V1, 18)]

TST <-
  MATS[V2 == 'NL10', .(MATERIAL, PLANT = 'NL60', PROMO = V17)][
    B4_SDMFRCAC2_A, on = .(MATERIAL, PLANT), nomatch = NA] %>%
  .[PLANT != 'FR30']
```

```
MAT <- LPO('10023', 18)

ggplot(
  data = dtFA[
    MATERIAL == MAT
    ], aes(x = CALMONTH, y = ACT)) +
  geom_col(fill = ET_CG) +
  theme_minimal()
```

```
ggplot(
  data = dtFA[
    CALMONTH >='202401'
    & APE > -50
    # & MATERIAL == MAT
  ],
  aes(x = APE, y = ECO)
) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  # geom_label(aes(label = round(PE, 1))) +
  # facet_wrap(~ PLANT + MATERIAL, scales = "free_y") +
  # scale_color_manual(values=c(ET_CG, ET_FG))+
```

```
  theme_minimal()
```

```
dtFA_melt  <-
  copy(dtFA) %>%
  melt.data.table(
    measure.vars  = c("APE", "ECO"),
    variable.name = "PE_TYPE",
    value.name    = "PE"
    ) %T>% setorder(CALMONTH)

ggplot(
  data = dtFA_melt[CALMONTH >='202401' & MATERIAL == MAT],
  aes(x = CALMONTH, y = PE, group = PE_TYPE, color = PE_TYPE)
  ) +
  geom_line(linewidth = 2) +
  geom_point() +
  geom_label(aes(label = round(PE, 1))) +
  # facet_wrap(~ PLANT + MATERIAL, scales = "free_y") +
  scale_color_manual(values=c(ET_CG, ET_FG))+
  theme_minimal()

# if(MAN){
#   if(rstudioapi::showQuestion(
#     title   = "SAC",
#     message = "Do you want to open SC-113B?"
#   ) == TRUE) {
#     browseURL(url = "https://wessanen.eu10.sapanalytics.cloud/link/SC113B")
#   }
# }
```

```
dtFA_melt  <-
  copy(dtFA) %>%
  melt.data.table(
    measure.vars  = c("APE", "ECO"),
    variable.name = "PE_TYPE",
    value.name    = "PE"
    )                                          %T>%
  setorder(PLANT, MATERIAL, CALMONTH)

ggplot(
  data = dtFA_melt[CALMONTH >='202401' & MATERIAL == MAT],
  aes(x = CALMONTH, y = PE, group = PE_TYPE, color = PE_TYPE)
  ) +
  geom_line(linewidth = 2) +
  geom_point() +
```

```
  geom_label(aes(label = round(PE, 1))) +
  # facet_wrap(~ PLANT + MATERIAL, scales = "free_y") +
  scale_color_manual(values=c(ET_CG, ET_FG))+
  theme_minimal()

# if(MAN){
#   if(rstudioapi::showQuestion(
#     title   = "SAC",
#     message = "Do you want to open SC-113B?"
#   ) == TRUE) {
#     browseURL(url = "https://wessanen.eu10.sapanalytics.cloud/link/SC113B")
#   }
# }
```

Chapter 16

Data Preparation

Data Preparation subtitle

Figure 16.1: Why forecasting?
The forecast drives the business!

See Figure Figure 20.1 for the thumbnail overview.

Chapter 17

Data Pipelines

Chapter 18

scenario:

Dataset Size:

- 400 MB in CSV format.
- 15 MB in Parquet format.

Data Access and Querying:

Happens once when the application starts. Needs to be quick to minimize user wait time. No data is written to disk during application use; data is only saved when the application exits. Data Manipulation:

Performed in-memory using R and Python. Requires fast execution. Consideration:

You like the idea of using Feather for its fast data loading. You're considering incorporating DuckDB to make the solution more scalable, despite minimal immediate benefits. Acknowledge that adding DuckDB introduces additional complexity.

Architecture: Architecture is about the design and setup, provides the framework and tools. This refers to the high-level structure of your system. It encompasses the selection of components, technologies, formats, and how they are organized and interact with each other. Architecture lays the foundation for your system's capabilities, scalability, and performance.

Workflow: Workflow is about the execution and processes, defines the processes and procedures using those tools. This pertains to the sequence of operations or processes that are carried out using the architectural components. It describes the specific steps taken to achieve your objectives, detailing how data and tasks flow through the system.

18.1 Work Flow & Architecture :

Need for a a robust, scalable and future-proof architecture with the flexibility to adapt as the data sets evolves, all while maintaining excellent performance and usability.

- Need for fast startup and in-memory processing.
- Scalable for future growth
- Minimal in complexity, leveraging familiar tools and lightweight components.

Vertically Scalable: The architecture leverages your node's 24 GB of memory effectively, with room for growth by upgrading hardware (e.g., adding more RAM or CPU cores).

Low Complexity Overhead:

DuckDB introduces minimal complexity since it integrates seamlessly with both R and Python. SQL provides a familiar and powerful tool for managing data subsets, making the workflow easy to maintain. Prepared for Growth:

The combination of Feather and DuckDB ensures you can handle current and moderate future data sizes efficiently. For substantial growth, DuckDB can work with Parquet files or other scalable formats with minimal changes. Optimized for Performance:

Feather ensures fast data access at startup. DuckDB allows efficient on-disk querying to reduce memory overhead, ensuring smooth performance even as datasets grow.

18.1.1

This workflow is ideal when working with datasets that are manageable in size and can be fully loaded into memory for analysis. Incorporating DuckDB adds the flexibility of SQL querying, which can be advantageous for preprocessing data before intensive computations. The use of Feather format ensures fast data loading, which is important for applications where user wait time during startup should be minimized.

If your dataset grows or your scalability needs change, you might need to adjust the workflow accordingly, possibly by switching to Parquet files and leveraging DuckDB's capabilities to handle larger datasets more efficiently.

Architecture

Scalability:

1. adjust your SQL queries to limit the data loaded into memory, ensuring your application remains performant.
2. scale vertically with increased RAM and CPU cores to manage larger Feather files and more complex algorithms without architectural changes.

Performance:

- Feather enables fast reads for smaller datasets during startup.

- DuckDB efficiently handles larger datasets by querying on disk and loading only the necessary data into memory.

18.1.2 Data Stageing

1. D00: contains the BW OpenHub Export, CSV - same as AL11

18.1.3 Data Ingestion:

Occurs once a month, converting CSV to Parquet (or another format). header is added and stored in D01

18.1.4 Data Access and Querying:

- Data is imported into R data.table objects via DuckDB, which reads the stored files.
- Queries are often simple SQL SELECT statements on one table or basic INNER JOINs.
- The same SQL queries are used to import data into Python using Polars via DuckDB.

18.1.5 Data Manipulation:

- Further data manipulation in R is done using data.table.
- In Python, data manipulation is performed using Polars.

Evaluation of Your Proposed Workflow

Given your familiarity with SQL and DuckDB, and your need for scalability, your proposed workflow is sound and aligns well with your goals.

18.1.6 Advantages of Your Workflow

1. **Scalability**
    - **Efficient Data Storage:** Parquet files are columnar, compressed, and optimized for performance, which is beneficial for large datasets.
    - **Query Optimization:** DuckDB is designed for efficient analytical queries, even on large datasets.
2. **Flexibility**
    - **Language Agnostic:** DuckDB can be used within R and Python, allowing seamless transition between languages.
    - **SQL Familiarity:** Your proficiency in SQL means you can leverage DuckDB's SQL interface effectively.
3. **Performance**

- **Fast Query Execution:** DuckDB is optimized for OLAP workloads and can execute complex queries quickly.
- **Direct Parquet Support:** DuckDB can read Parquet files directly without the need to load the entire dataset into memory.

4. **Integration with R and Python**

- **R Data.Tables:** You can fetch query results from DuckDB into R data.tables, integrating with your existing R codebase.
- **Python Polars:** Similarly, you can read data into Polars DataFrames in Python, which is efficient for data manipulation.

5. **Simplified Data Pipeline**

- **Unified Data Source:** Using DuckDB on top of Parquet files centralizes your data access, simplifying data management.

18.1.7 Potential Disadvantages

1. **Additional Complexity**

- **Library Dependencies:** Requires installation and management of additional libraries (DuckDB, Parquet support in R and Python).
- **Learning Curve for Integration:** Even though you're familiar with DuckDB, integrating it into R and Python workflows may require some setup and testing.

2. **Overhead of Data Conversion**

    - **Initial Conversion:** Migrating CSV files to Parquet adds an extra step in your data ingestion pipeline.
    - **Data Updates:** If your CSV data updates frequently, you'll need to automate the conversion process.

3. **Resource Usage**

- **Disk Space:** Maintaining both CSV and Parquet files (if not deleting the original CSVs) may consume additional storage.

Recommendations and Best Practices

Given your requirements and skills, your proposed workflow is suitable and offers several benefits in terms of scalability and flexibility. Here are some recommendations to optimize your workflow:

1. Automate the CSV to Parquet Conversion

- **Use Batch Processing:**
- Create scripts in R or Python to automate the conversion of CSV files to Parquet.
- **Leverage DuckDB for Conversion:**

  - DuckDB can read CSV files and write Parquet files, allowing you to perform the conversion within DuckDB.

```
COPY (SELECT * FROM 'your_data.csv') TO 'your_data.parquet' (FORMAT PARQUET);
```

18.1.8 2. Optimize DuckDB Usage

- **Indexing and Partitioning:**
  - While Parquet files do not support traditional indexing, consider partitioning your data to improve query performance.
- **SQL Query Optimization:**
  - Use DuckDB's advanced SQL features to optimize queries (e.g., window functions, common table expressions).

18.1.9 3. Efficient Data Retrieval into R and Python

- **In R:**
  - Use the `duckdb` package to execute SQL queries and fetch results into data.tables.

```
library(duckdb)
con <- dbConnect(duckdb::duckdb())

# Query data
result <- dbGetQuery(con, "SELECT * FROM 'your_data.parquet' WHERE conditions")

# Convert to data.table
library(data.table)
dt_result <- as.data.table(result)
```

- **In Python with Polars:**
  - Use DuckDB's Python API or integrate with Polars for efficient data handling.

```
import duckdb
import polars as pl

# Execute query and fetch result as Polars DataFrame
df = duckdb.query("SELECT * FROM 'your_data.parquet' WHERE conditions").to_df()
pl_df = pl.from_pandas(df)
```

18.1.10 4. Consider Data Volume and Hardware Resources

- **Memory Management:**

  - DuckDB processes data efficiently, but ensure your hardware resources are adequate for your data size.

- **Disk I/O:**
  - Using Parquet files reduces disk I/O due to compression, but be mindful of the storage subsystem performance.

18.1.11 5. Keep Libraries Updated

- **Stay Current:**
  - Ensure that you are using the latest versions of DuckDB, R packages, and Python libraries to benefit from performance improvements and bug fixes.

18.1.12 6. Handle Updates and Data Versioning

- **Incremental Updates:**
  - If your data updates incrementally, design your pipeline to handle partial updates rather than reprocessing entire datasets.
- **Data Version Control:**
  - Implement versioning for your Parquet files to track changes over time.

18.1.13 7. Monitor Performance and Adjust

- **Benchmarking:**
  - Regularly benchmark query performance to identify bottlenecks.
- **Adjust Strategy:**
  - Based on performance metrics, adjust partitioning schemes, query strategies, or hardware resources as needed.

Conclusion

Your proposed workflow of migrating CSV files to Parquet, using DuckDB for querying, and integrating with R and Python is a robust solution that addresses your needs for scalability and flexibility. Given your familiarity with SQL and DuckDB, and the fact that you possess the necessary skills to manage the additional complexity, this approach is well-suited to your situation.

Benefits of Your Workflow:

- **Scalable Data Processing:**
- Efficient handling of large datasets through Parquet and DuckDB.

- **Flexibility Across Languages:**
  - Seamless data access in both R and Python without data format conversion issues.
- **Performance Optimization:**
  - Faster query execution and data retrieval due to optimized storage and processing.

Key Actions Moving Forward:

1. **Set Up Automated Data Pipelines:**

- Automate the conversion from CSV to Parquet to ensure consistency and save time.

2. **Optimize Queries:**

- Utilize DuckDB's full SQL capabilities to write efficient queries for your analysis.

3. **Integrate Smoothly with R and Python:**
   - Establish standard functions or scripts in both environments to interact with DuckDB, minimizing repetitive code.
4. **Monitor and Iterate:**
   - Keep an eye on performance metrics and be ready to adjust your approach as your dataset grows or changes.

Additional Considerations

- **Community and Support:**
  - DuckDB is actively developed, and there is a growing community. Utilize resources like documentation and forums when needed.
- **Testing and Validation:**
  - As with any data pipeline, thoroughly test each component to ensure data integrity and correctness.
- **Security and Access Control:**
  - If working in a multi-user environment, consider how data access and permissions are managed.

Final Thoughts

Your willingness to embrace additional complexity due to your skill set positions you well to benefit from this workflow. By combining the strengths of Parquet's efficient

data storage and DuckDB's powerful query engine, you can achieve a scalable and flexible data analysis environment.

Should you need further assistance or have more questions as you implement this workflow, feel free to reach out. I'm here to help ensure your data processing is as efficient and effective as possible.

Chapter 19

Data Stageing

data lake design pattern: Bronze-Silver-Gold structure data according to its level of processing and readiness for use.

Bronze Layer (Raw Data): Contains raw data as it was ingested from the source systems. This data is typically in its original format and unprocessed.

Silver Layer (Cleaned and Conformed Data): Contains data that has been cleansed, filtered, and possibly enriched. This data is ready for further processing or analysis but is not yet aggregated or modeled.

Gold Layer (Curated Data): Contains data that has been transformed into business-level aggregates, models, or reports. This is the data used by analysts, data scientists, or applications for decision-making.

Chapter 20

Modelling

Modelling subtitle

Figure 20.1: Why forecasting?
The forecast drives the business!

See Figure Figure 20.1 for the thumbnail overview.

20.1 blackbox models

20.2 whitebox models

6.1. Literature Review / Similar Models - comprehensive overview of the state of the art in machine learning, including the latest algorithms, - insight into the performance of different algorithms and techniques on similar types of data. (avoid wasting time on models that are unlikely to perform well.)

20.2.1 Model Selection

different models have different strengths and weaknesses and are suitable for different types of data and problems. achieve the best possible performance and maximize the impact

avoid overfitting a trade-off between the simplicity and flexibility (robustness)

model performance - only article hierarchy - article and customer hierarchy

Cross-validation is a method for evaluating the performance of a model on a validation dataset. This involves dividing the data into k folds, training the model on k-1 folds, and evaluating the performance on the remaining fold. This process is repeated k times, with each fold serving as the validation set once, and the results are averaged to obtain a final performance score. This method provides a more reliable estimate of the model's performance, as it uses all of the data for training and evaluation.

algorithm - handle teh specific characteristics of teh data - balance performance and interpretability - memory and processing power

- validity -> fa & prediction intervals
- robustness ->
- transparency ->

Appendix

Chapter 21

Evaluation

Evaluation subtitle

21.1 Evaluate Results

- explainability

21.1.1 Assessment of Data Mining Results w.r.t. Business Success Criteria

- how well does the model work on new data?
- For whom does the model not work well?

21.1.2 Approved Models

21.1.3 Review Process

21.1.4 Review of Process

21.1.5 Determine Next Steps

21.1.6 List of Possible Actions

21.1.7 Decision

Appendix

Chapter 22

Deployment

Deployment subtitle

scalability - distributed techniques - maintenance & updating the model - monitoring the model - debugging the model

Chapter 23

model management

- model versioning GIT
- continuous integration/continuous deployment (CI/CD) pipelines.

Appendix

Chapter 24

Other

Other sub

other

24.1 Literature

The literature review focuses on the intersection of three fields, according to the now-ubiquitous Data Science Venn Diagram of Drew Conway.

24.2 Business Process

- SCOR MODEL A COMPLETE GUIDE - 2020 EDITION, (BLOKDYK, 2019)
- Inventory and production management in supply chains, (Silver et al., 2021)

24.3 Math & Statistics

- Forecasting: Principles and Practice (3rd ed), (Hyndman & Athanasopoulos, 2021)
- Demand forecasting best practices, (Vandeput, 2023)
- Data science for supply chain forecasting, (Vandeput, 2021)
- Inventory optimization: models and simulations, (Vandeput, 2020)
- Machine Learning With Boosting: A Beginner's Guide, (Hartshorn, 2017)
- Introduction to conformal prediction with Python, (Molnar, 2023)
- Practical guide to applied conformal prediction in Python, (Manokhin, 2023)
- SHAP with Python, (O'Sullivan, 2024)
- Introduction to SHAP with Python, (O'Sullivan, 2023)
- Ensemble methods for machine learning, (Kunapuli, 2023)

24.4 Programming

24.4.1 Python

- Python Crash Course, (Matthes, 2019a)
- Lightning fast forecasting with statistical and econometric models, (*Nixtla/statsforecast*, 2024)

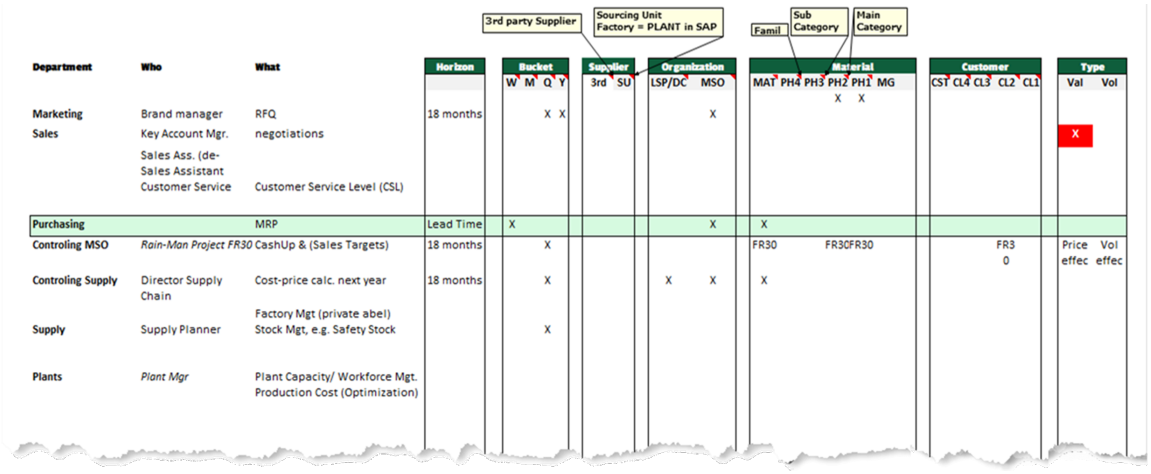
- Scalable Machine Learning for Time Series Forecasting, (*Nixtla/mlforecast*, 2024)
- Probabilistic hierarchical forecasting with statistical and econometric methods, (*Nixtla/hierarchicalforecast*, n.d.)
- TS Features, calculates various features from time series data, (*Nixtla/tsfeatures*, 2024)

24.4.2 R

- R for Data Science, (Wickham, n.d.)
- R forecasting package, (Hyndman [aut, cre, cph, Athanasopoulos, Bergmeir, et al., 2024)
- R feasts package, (O'Hara-Wild, Hyndman, Wang, Cook, et al., 2024)
- R fable package, (O'Hara-Wild, Hyndman, Wang, implementation), et al., 2024)

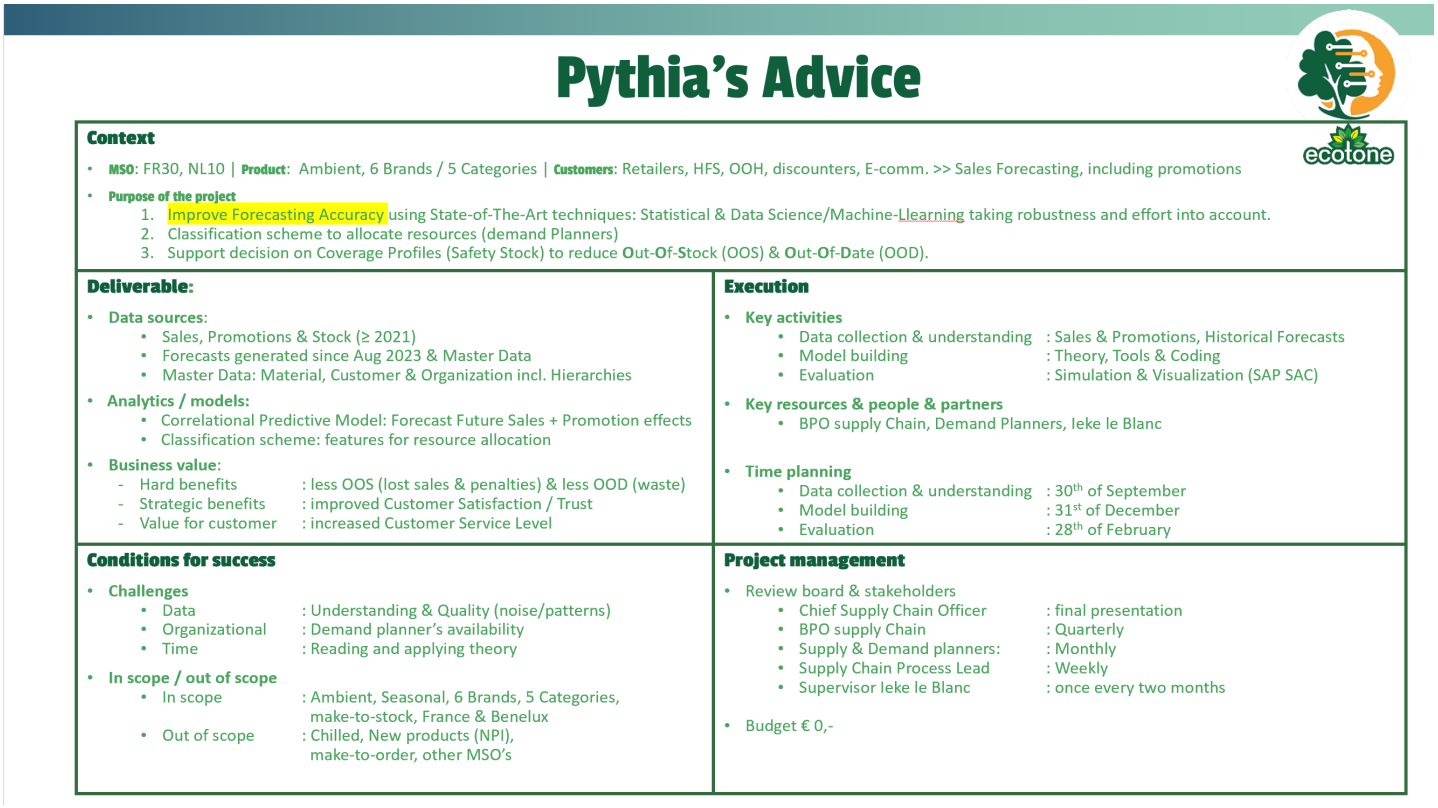
24.5 Forecast Use Cases

Figure 24.1: Forecast Use Cases



24.6 Project Charter

Figure 24.2: Project Charter



Pythia's Advice

Context

- **MSO**: FR30, NL10 | **Product**: Ambient, 6 Brands / 5 Categories | **Customers**: Retailers, HFS, OOH, discounters, E-comm. >> Sales Forecasting, including promotions
- **Purpose of the project**
  1. Improve Forecasting Accuracy using State-of-The-Art techniques: Statistical & Data Science/Machine-Llearning taking robustness and effort into account.
  2. Classification scheme to allocate resources (demand Planners)
  3. Support decision on Coverage Profiles (Safety Stock) to reduce **O**ut-**O**f-**S**tock (OOS) & **O**ut-**O**f-**D**ate (OOD).

Deliverable:

- **Data sources**:
  - Sales, Promotions & Stock (≥ 2021)
  - Forecasts generated since Aug 2023 & Master Data
  - Master Data: Material, Customer & Organization incl. Hierarchies
- **Analytics / models:**
  - Correlational Predictive Model: Forecast Future Sales + Promotion effects
  - Classification scheme: features for resource allocation
- **Business value**:
  - Hard benefits : less OOS (lost sales & penalties) & less OOD (waste)
  - Strategic benefits : improved Customer Satisfaction / Trust
  - Value for customer : increased Customer Service Level

Execution

- **Key activities**
  - Data collection & understanding : Sales & Promotions, Historical Forecasts
  - Model building : Theory, Tools & Coding
  - Evaluation : Simulation & Visualization (SAP SAC)
- **Key resources & people & partners**
  - BPO supply Chain, Demand Planners, Ieke le Blanc
- **Time planning**
  - Data collection & understanding : 30th of September
  - Model building : 31st of December
  - Evaluation : 28th of February

Conditions for success

- **Challenges**
  - Data : Understanding & Quality (noise/patterns)
  - Organizational : Demand planner's availability
  - Time : Reading and applying theory
- **In scope / out of scope**
  - In scope : Ambient, Seasonal, 6 Brands, 5 Categories, make-to-stock, France & Benelux
  - Out of scope : Chilled, New products (NPI), make-to-order, other MSO's

Project management

- Review board & stakeholders
  - Chief Supply Chain Officer : final presentation
  - BPO supply Chain : Quarterly
  - Supply & Demand planners: : Monthly
  - Supply Chain Process Lead : Weekly
  - Supervisor Ieke le Blanc : once every two months
- Budget € 0,-

24.7 Project Plan

Figure 24.3: Project Plan

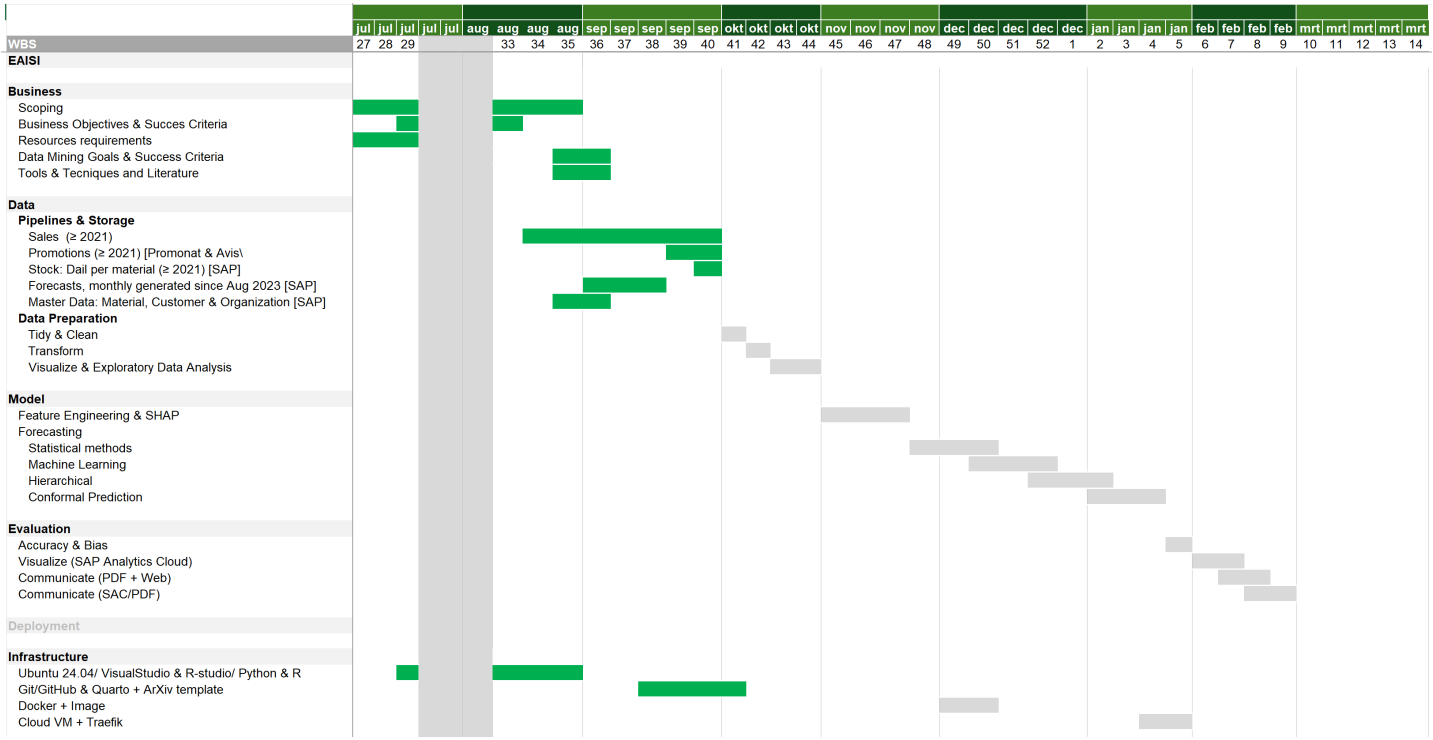
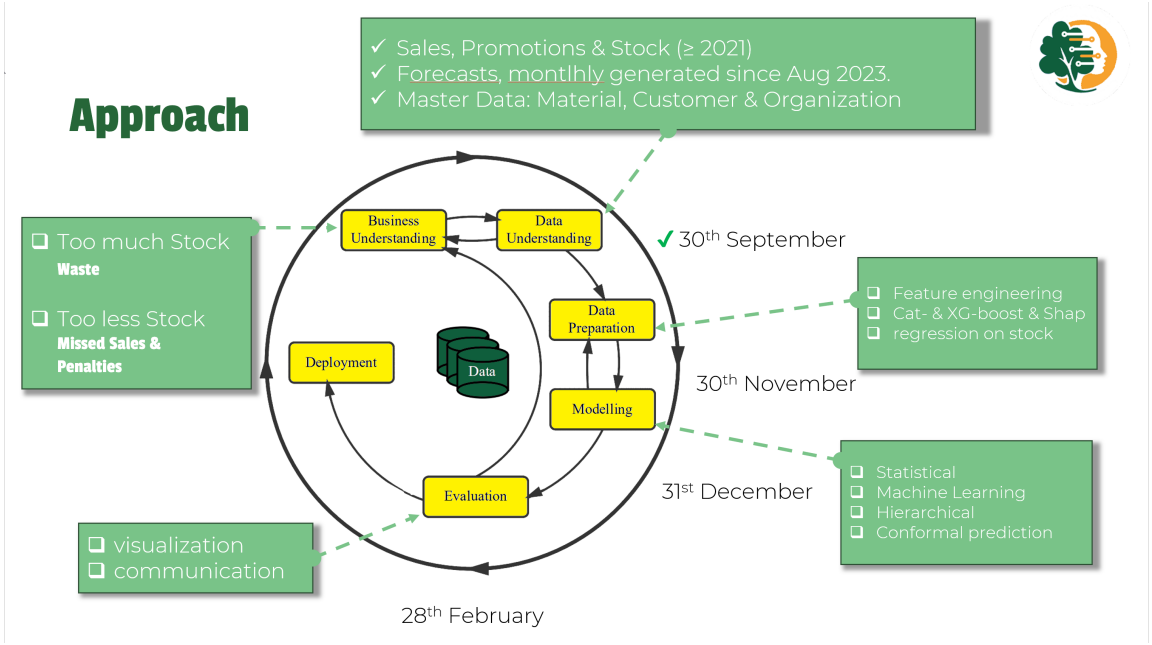


Figure 24.4: Project Plan



24.8 EAISI

24.8.1 Deliverables & requirements

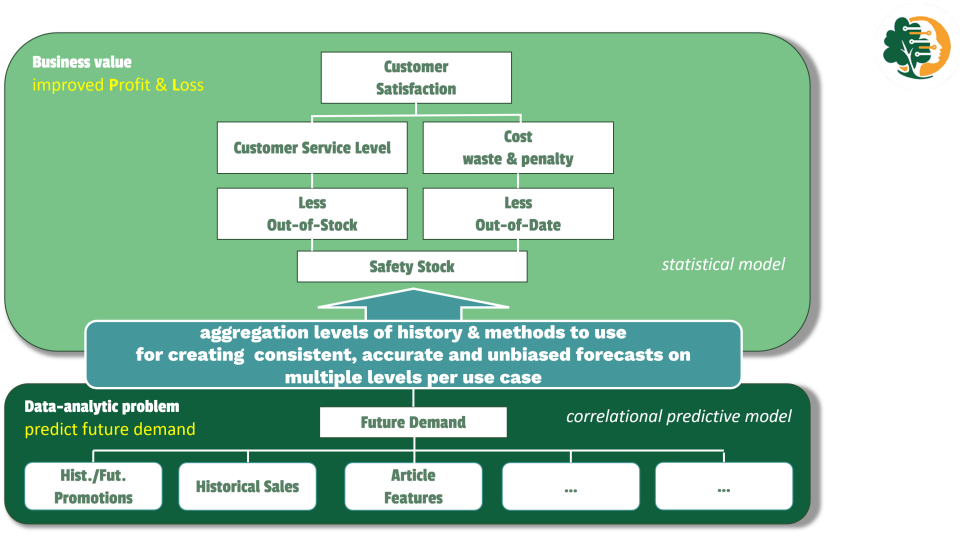
- Clear description of the need / problem / opportunity
- Clear description of the long-term ambition
- Clear focus on what part of this long-term ambition will be taken on in the project
- Description of the data science project with the help of a flowdown chart and project charter
- Clear description of how the model outcome / analysis results will be used (e.g. who, at what time, in which process, in what way, will use the model outcome / analysis results to make a better decision about what?) and how this translates into improved KPIs as outlined in the flowdown chart
- Substantiated choice for a 'performance metric' (i.e. what should the model be good at)
- Clear description of a business case / cost-benefit (no cost!!) analysis
- Realistic plan / timeline for how to execute the consecutive CRISP-DM phases

Chapter 25

Flow Down

25.1 Flow-down

Figure 25.1: Flowdown Graph



Chapter 26

References

References

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Schwendinger, B., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., … Krylov, I. (2024). *Data.table: Extension of 'data.frame'*. https://cran.r-project.org/web/packages/data.table/index.html

BLOKDYK, G. (2019). *SCOR MODEL A COMPLETE GUIDE - 2020 EDITION*. 5STARCOOKS.

Core Team}, {R. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Costa, R. (2022). *The CRISP-ML Methodology: A Step-by-Step Approach to Real-World Machine Learning Projects*. Independently published.

DeBruine, L. (2023). *Glossary: Glossaries for Markdown and Quarto Documents*. https://cran.r-project.org/web/packages/glossary/

Guja, A., & Siwiak, M. (2025). *Generative AI for Data Analytics*. Manning Publications Co. LLC 2025.

Hartshorn, S. (2017). *Machine Learning With Boosting: A Beginner's Guide*.

Hyndman [aut, R., cre, cph, Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., Kuroptev, K., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., Garza, F., Girolimetto, D., Ihaka, R., R Core Team, Reid, D., Shaub, D., … Zhou, Z. (2024). *Forecast: Forecasting functions for time series and linear models*. https://cran.r-project.org/web/packages/forecast/index.html

Hyndman [aut, R., cre, cph, Athanasopoulos, G., O'Hara-Wild, M., Palihawadana, N., Wickramasuriya, S., & RStudio. (2024). *fpp3: Data for "Forecasting: Principles and Practice" (3rd Edition)*. https://cran.r-project.org/web/packages/fpp3/index.html

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd ed)*. https://otexts.com/fpp3/

Kunapuli, G. (2023). *Ensemble methods for machine learning*. Manning.

Manokhin, V. (2023). *Practical guide to applied conformal prediction in Python: Learn and apply the best uncertainty frameworks to your industry applications*. <packt>.

Matthes, E. (2019b). *Python crash course: A hands-on, project-based introduction to programming* (2nd edition). No Starch Press.

Matthes, E. (2019a). *Python crash course: A hands-on, project-based introduction to*

programming (2nd edition). No Starch Press.

Molnar, C. (2023). *Introduction to conformal prediction with Python: A short guide for quantifying uncertainty of machine learning models* (First edition). Chistoph Molnar c/o MUCBOOK, Heidi Seibold. https://christophmolnar.com/books/conformal-prediction/

Nixtla/hierarchicalforecast: Probabilistic hierarchical forecasting with statistical and econometric methods. (n.d.). https://github.com/Nixtla/hierarchicalforecast

Nixtla/mlforecast. (2024). Nixtla. https://github.com/Nixtla/mlforecast

Nixtla/statsforecast. (2024). Nixtla. https://github.com/Nixtla/statsforecast

Nixtla/tsfeatures. (2024). Nixtla. https://github.com/Nixtla/tsfeatures

O'Hara-Wild, M., Hyndman, R., Wang, E., Cook, D., features), T. T. (Correlation., & method), L. C. (Guerrero's. (2024). *Feasts: Feature extraction and statistics for time series*. https://cran.r-project.org/web/packages/feasts/index.html

O'Hara-Wild, M., Hyndman, R., Wang, E., implementation), G. C. (NNETAR., Bergmeir, C., Hensel, T.-G., & Hyndman, T. (2024). *Fable: Forecasting models for tidy time series*. https://cran.r-project.org/web/packages/fable/index.html

O'Sullivan, C. (2023). Introduction to SHAP with Python. In *Introduction to SHAP with Python*. https://towardsdatascience.com/introduction-to-shap-with-python-d27edc23c454

O'Sullivan, C. (2024). SHAP with Python [Course]. In *SHAP with Python*. https://adataodyssey.com/course/

Porter, L., & Zingaro, D. (2024). *Learn AI-assisted Python programming: With GitHub Copilot and ChatGPT*. Manning.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2021). *Inventory and production management in supply chains* (Fourth edition, first issued in paperback). CRC Press, Taylor & Francis Group.

Vandeput, N. (2020). *Inventory optimization: Models and simulations*. De Gruyter.

Vandeput, N. (2021). *Data science for supply chain forecasting* (Second edition). De Gruyter.

Vandeput, N. (2023). *Demand forecasting best practices*. Manning Publications Co.

Wickham, H. (n.d.). *R for data science (2e)*. https://r4ds.hadley.nz/

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., Brand, T. van den, Posit, & PBC. (2024). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://cran.r-project.org/web/packages/ggplot2/index.html

Chapter 27

Resources

Chapter 28

Resources

- Download the Business Understanding notebook (QMD)
- Download the Data Understanding notebook (QMD)
- Download EAISI Graduation Presentation (PPTX)
- Download all notebooks (ZIP)