SMCP x EDHEC

Business Data Management

Group 7

The Team

66778

Manto Ip

The handsome analyst

66850

Anmol Bhoj

The cool analyst

66827

Santosh Barik

The amazing analyst

79740

Rohit Dharmadhikari

The fun analyst

The Objective

Provide guidance to the design and supply teams of SMCP by predicting the **Future Sales Volume** of products for upcoming seasons using:

1. **Previous Sales Volume**
2. **Product Characteristics**

At a Glance

Recovered from pandemic in 2021

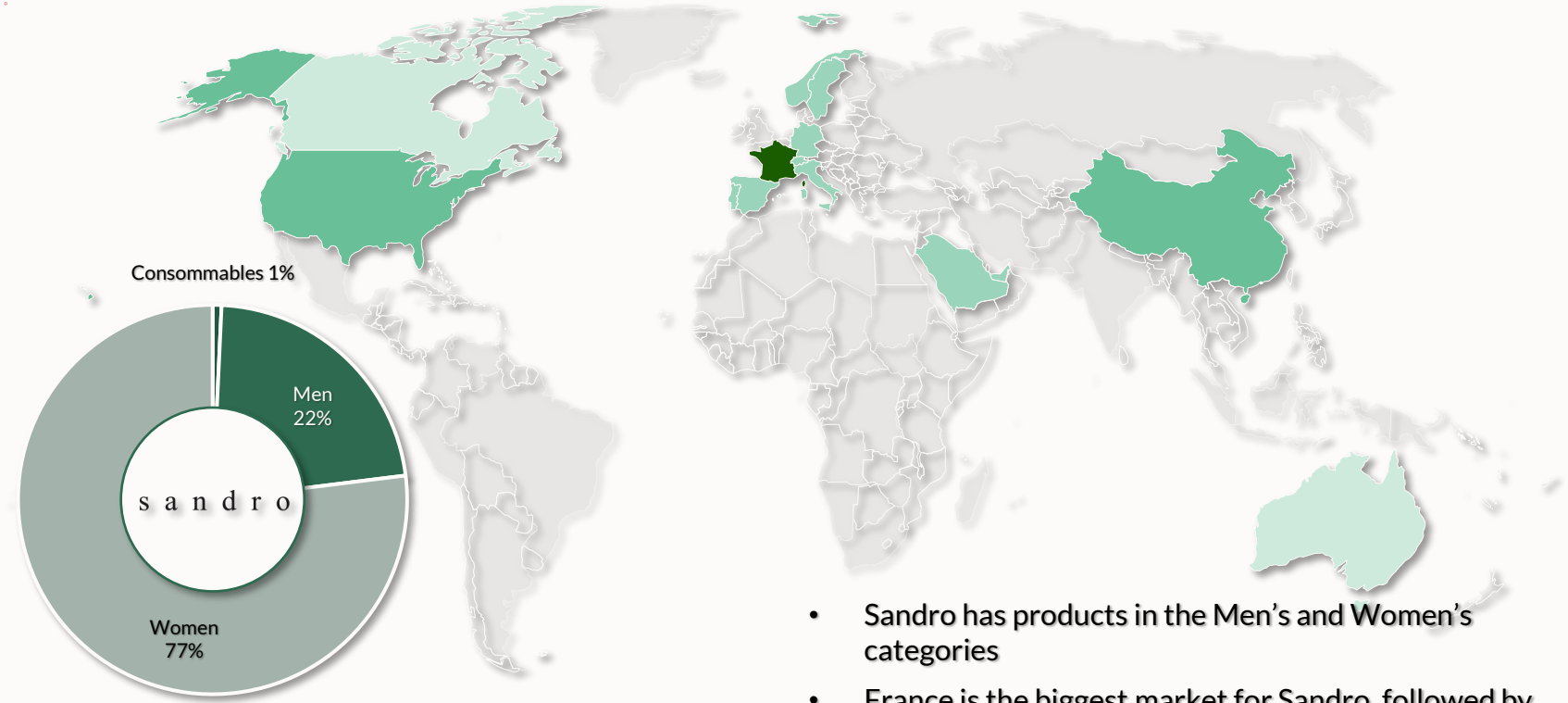
France continues to be the best region

Higher sales to date in Summer 2021

01

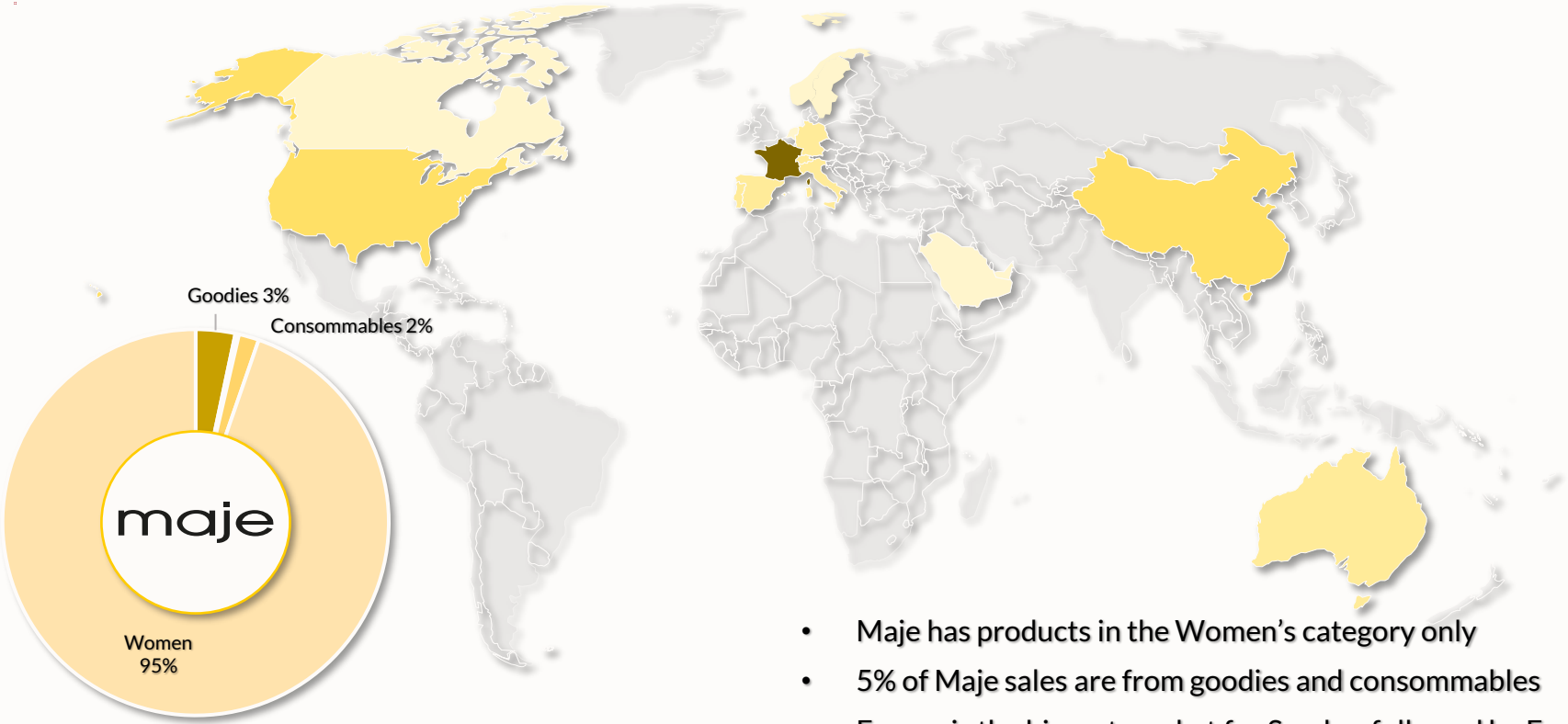
Exploratory Data Analysis

A Closer Look - Sandro



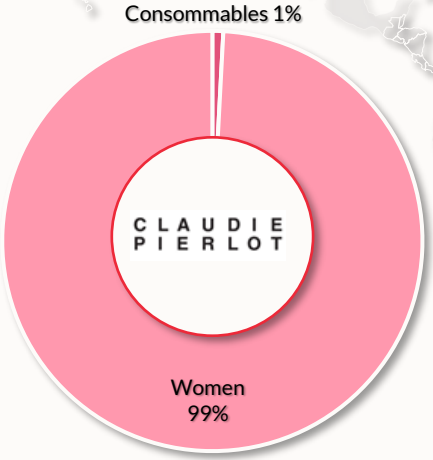
- Sandro has products in the Men's and Women's categories
- France is the biggest market for Sandro, followed by Europe

A Closer Look - Maje



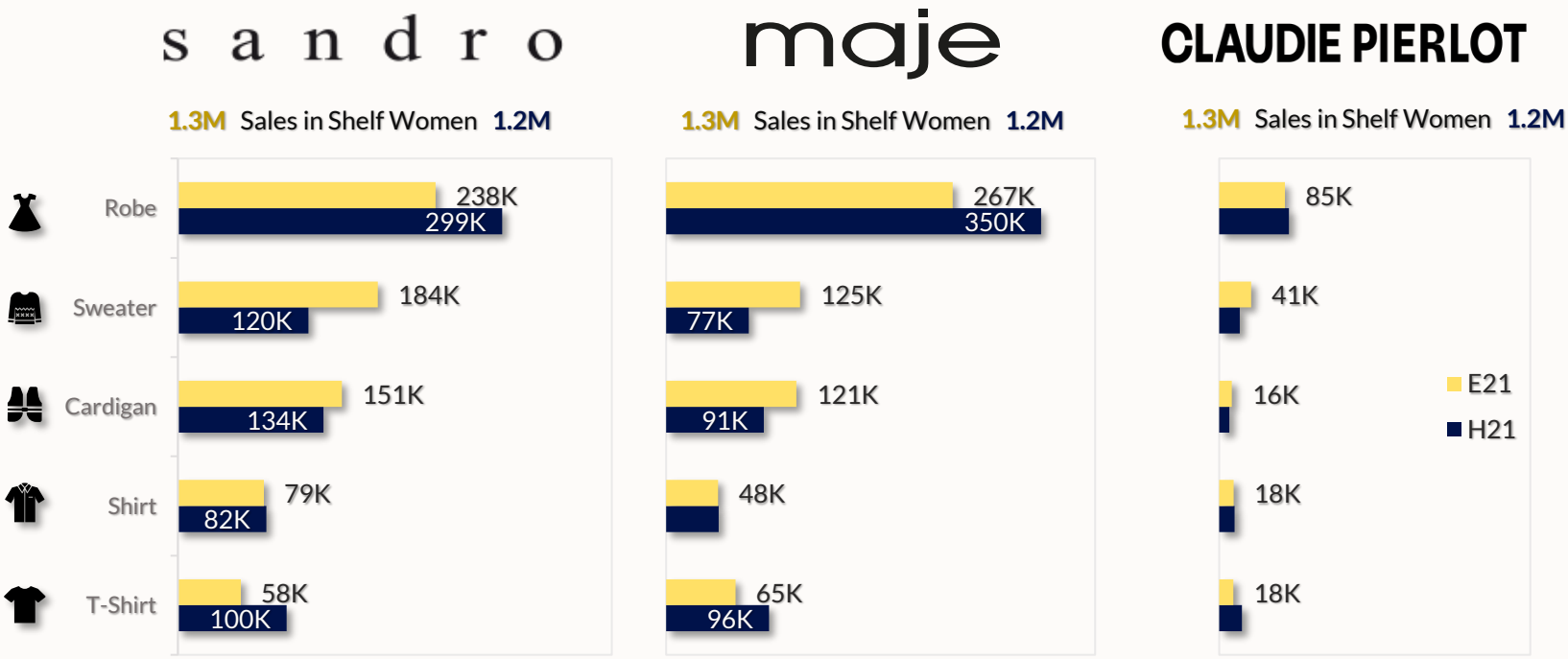
- Maje has products in the Women's category only
- 5% of Maje sales are from goodies and consommables
- France is the biggest market for Sandro, followed by Europe

A Closer Look - Claudie



- Claudie is a smaller brand compared to Maje and Sandro
- It is not present in the North American market
- 99% of its sales are from the Women's category

The Bestsellers



More than 25% of volume sales across S.M.C.P comprise of Robes.

02

Introduction to the Data

Our Data In A Nutshell

	Train Data	Test Data
# Products	22.4K	16.1K
Seasons	E17 – H20	E21 - H21
# Products-Season-Region	244K	64.9K
Total Volume Sales	26.7M	6.7M

Under the Hood

Unique ID	Brand Name	Product ID	Region	Sales Season	Fresh Product Flag	Volume Sales	Predicted Sales
NA-H21-1-00034695	Sandro	1-00034695	North America	H21	Fresh	110	111
FRANCE-H21-3-00013186	Claudie	3-00013186	France	H21	Fresh	1	-1
ASIA-E21-2-00016554	Maje	2-00016554	Asia	E21	Fresh	2	4
FRANCE-E21-1-00023227	Sandro	1-00023227	France	E21	Fresh	2	7
ASIA-E21-2-00017212	Maje	2-00017212	Asia	E21	Repeat	2	1
EUR-E21-1-00018492	Sandro	1-00018492	Eur	E21	Repeat	1	1
EUR-E21-1-00025781	Sandro	1-00025781	Eur	E21	Repeat	1	5
ASIA-H21-2-00021221	Maje	2-00021221	Asia	H21	Repeat	450	450
EUR-E21-1-00031980	Sandro	1-00031980	Eur	E21	Repeat	10	8
EUR-H21-1-00034803	Sandro	1-00034803	Eur	H21	Repeat	102	107

*Representative Data

Dedicated Models for Fresh and Repeat Products

03

Feature Engineering

The Features

Laundry Variables

- Dry clean Difficulty
- Washing Difficulty
- Bleach Difficulty
- Dry Difficulty
- Care Label

Product Characteristics

- Shelf
- Product Family
- Department
- Product Sub-Family

Logistics

- Storage Type
- Supply Type

Product Style

- Product Material
- Pattern* (Checks/Stripes)

*Not Applicable for fresh products

The Past Sales

Product Sales*

Aggregated Sales

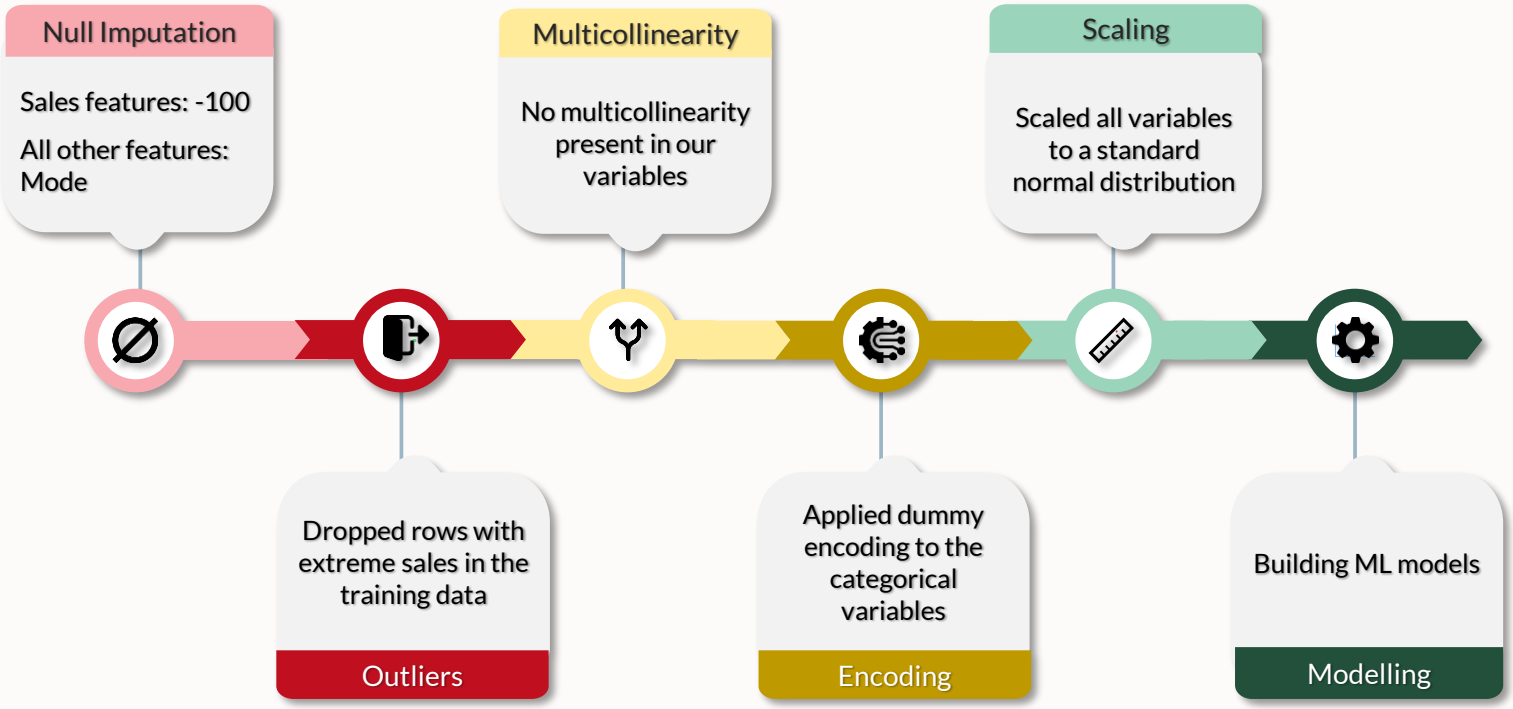
Last Season Shelf Sales	Last Off-Season Shelf Sales	Last to last Season Shelf Sales
Last Season Family Sales	Last Off-Season Family Sales	Last to last Season Family Sales
Last Season Sub-Family Sales	Last Off-Season Sub-Family Sales	Last to last Season Sub-Family Sales

*Not Applicable for fresh products

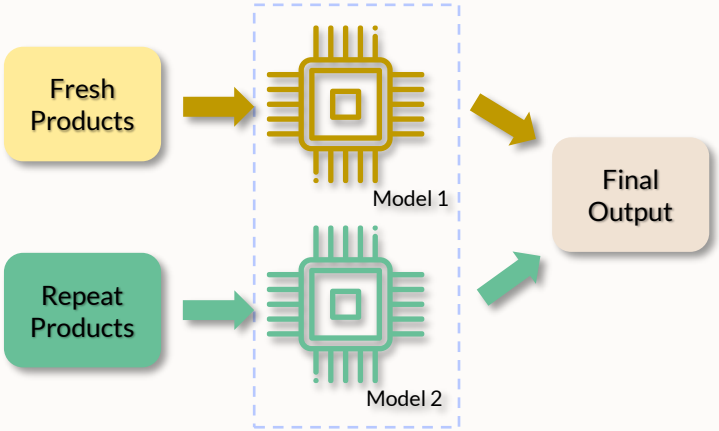
04

Modelling

The Flow


Null Imputation
Sales features: -100
All other features: Mode
Outliers
Dropped rows with extreme sales in the training data
Multicollinearity
No multicollinearity present in our variables
Encoding
Applied dummy encoding to the categorical variables
Scaling
Scaled all variables to a standard normal distribution
Modelling
Building ML models

The Approach



- Error Metric: **Mean Absolute Error**

- It is difficult to predict sales volume of Fresh products as they have never entered the market.
- Variables like historical product sales and pattern are unavailable

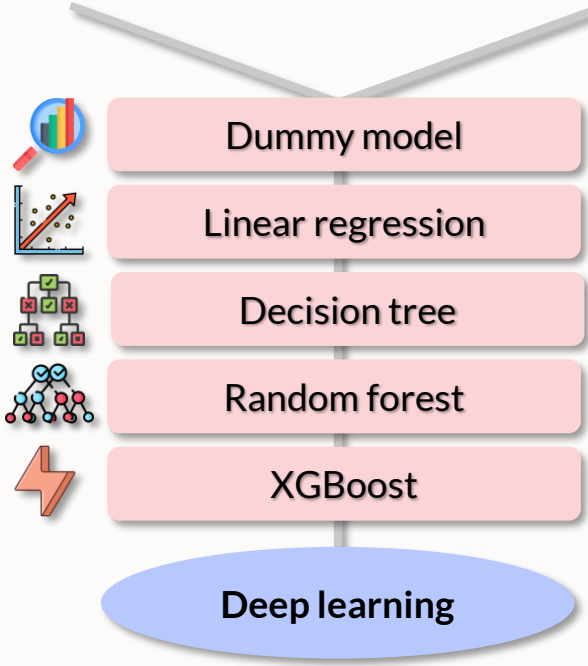
- Repeat products have historical sales data, which is one of our best predictor
- Splitting the data into 2 models helps our models concentrate on the relevant information

- Final Output includes prediction for both the datasets

The Models

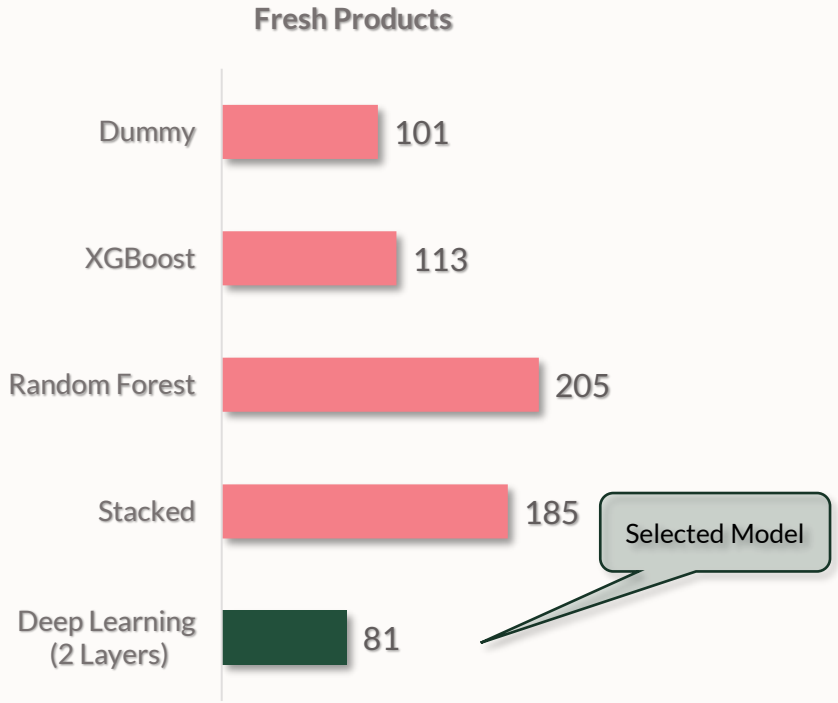
Repeated products

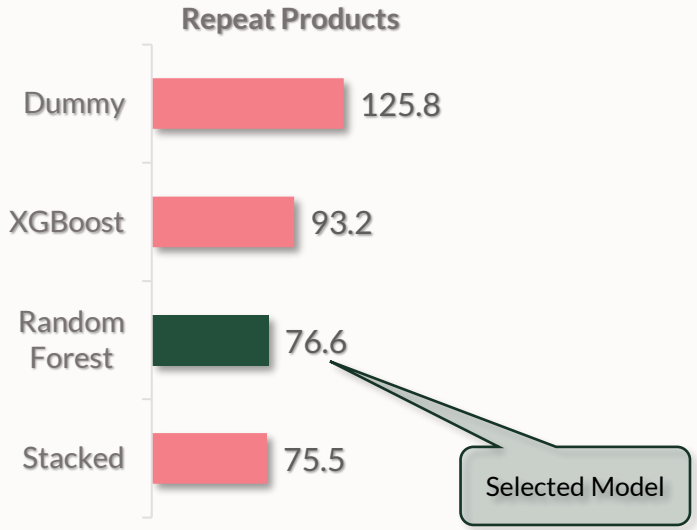
Fresh products



Stacked model

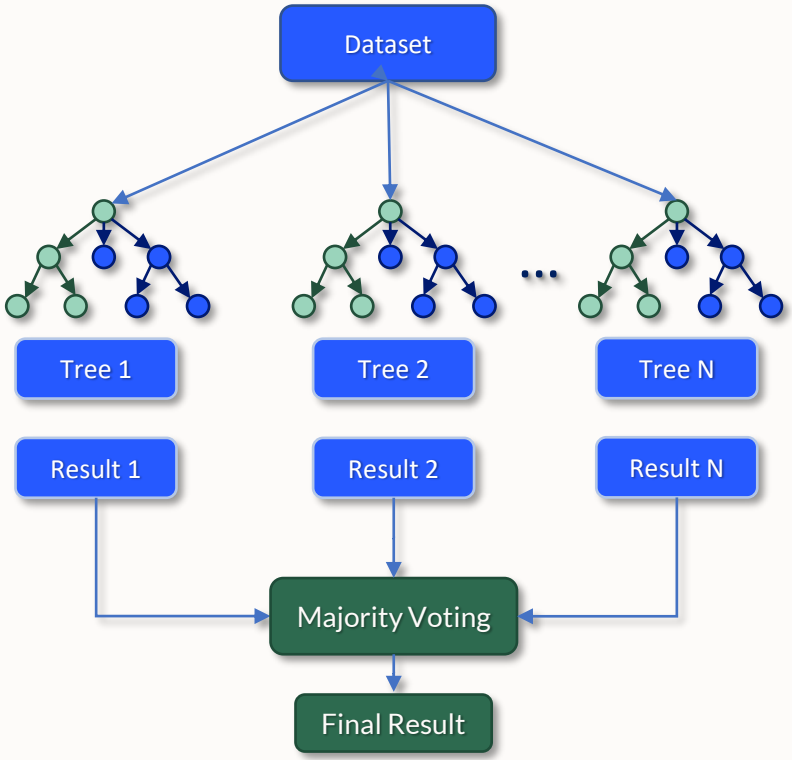
Results





Mean Absolute Error: Lower is better

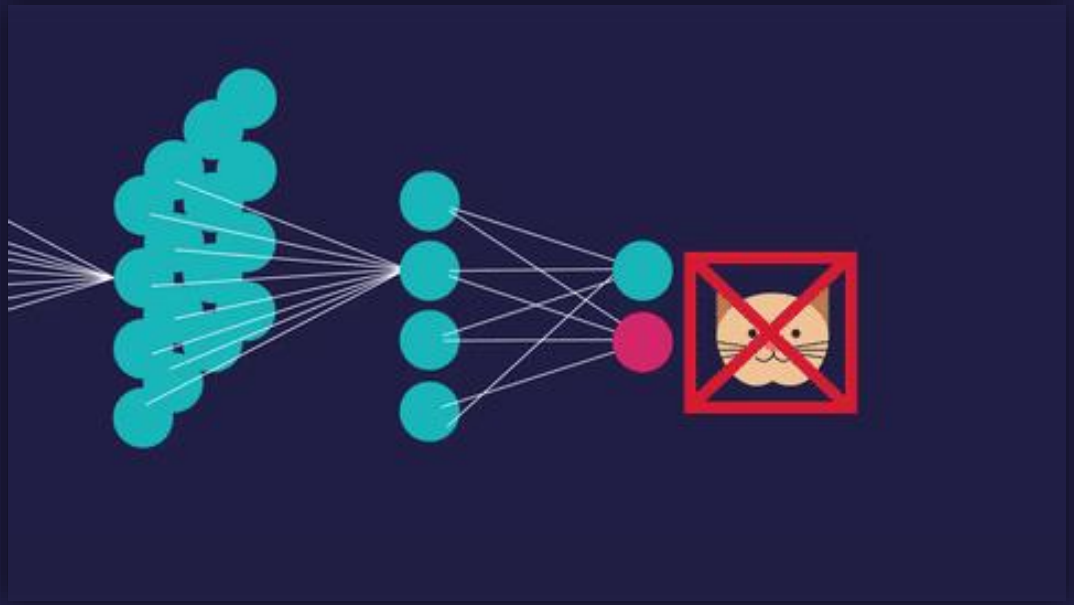
Model - Random forest



We selected Random forest as our model for the repeated product because:

- We want a **simple model** to capture the relationship between previous sales and future sale
- The model is **easily interpretable**
- Ability to check the **significant variables** easily

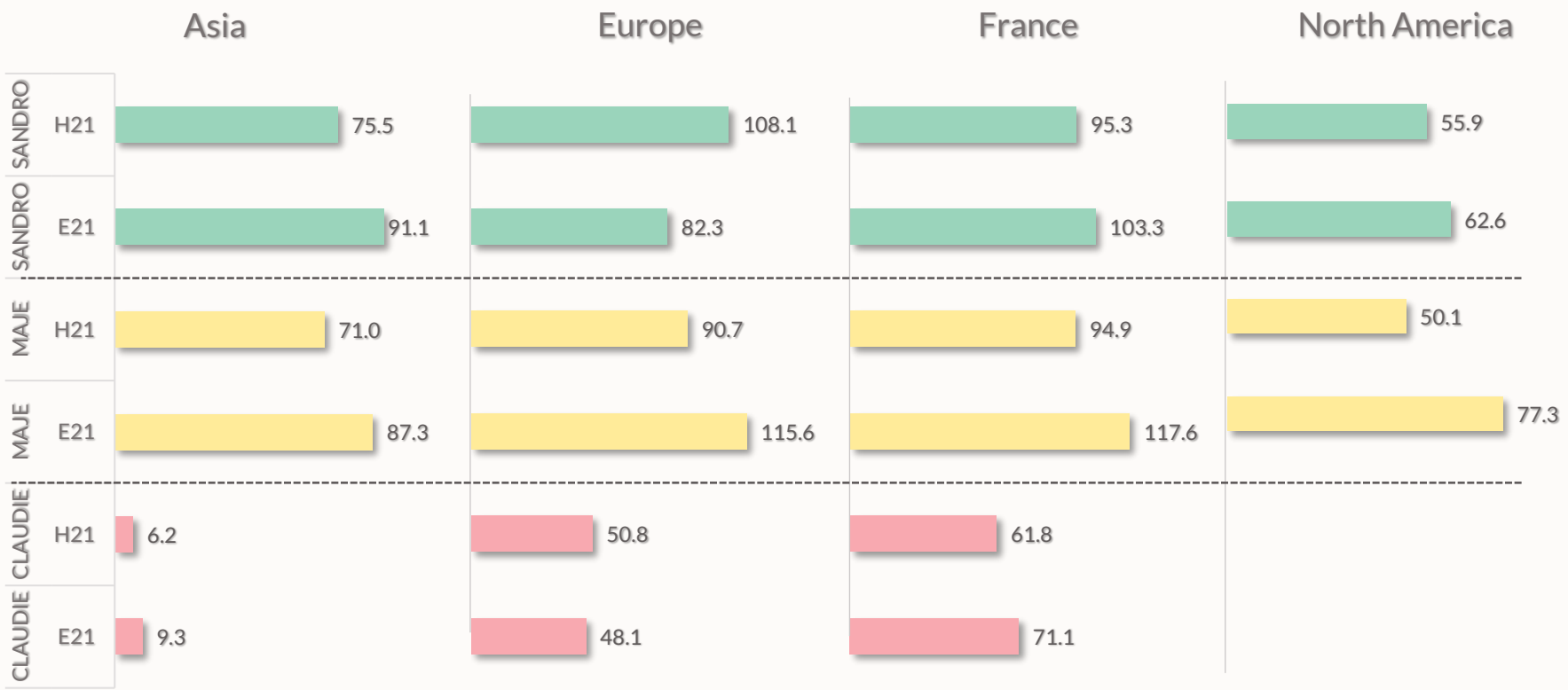
Model – Deep Learning



We selected Deep Learning (2 layers) as our model for the **fresh product** because:

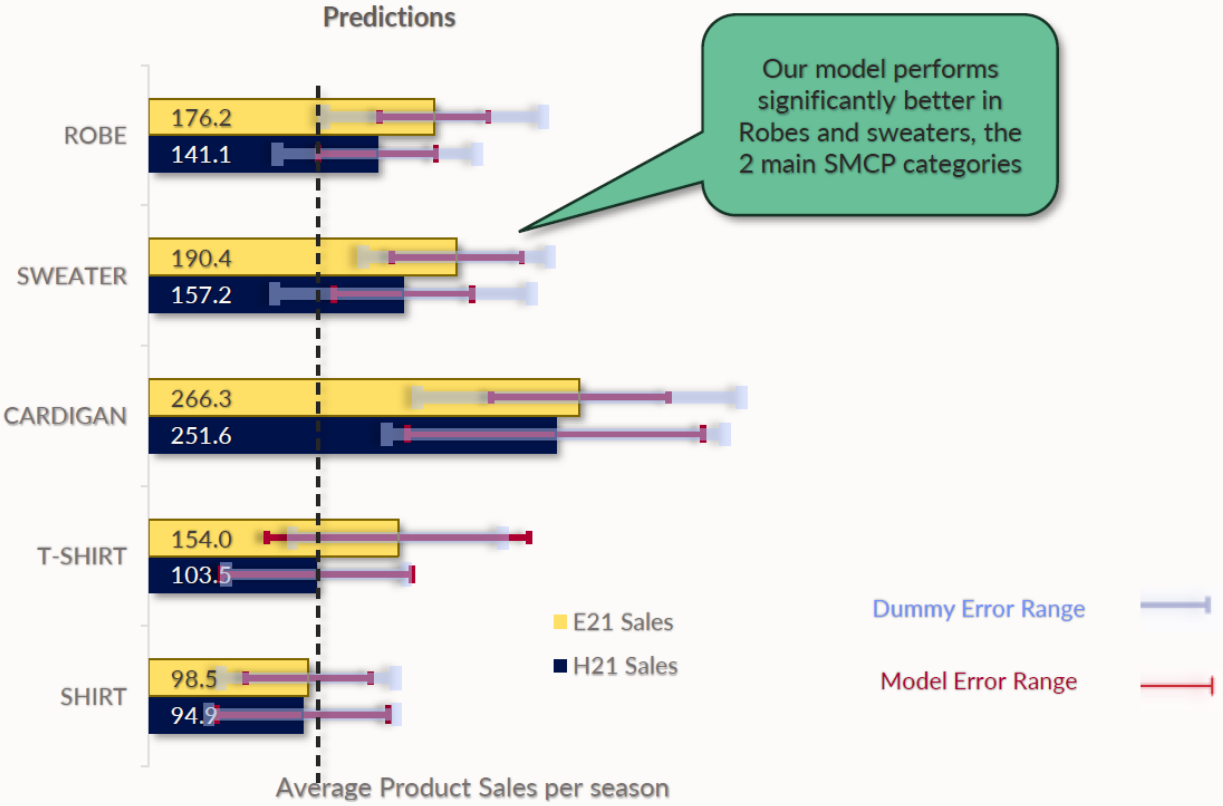
- Required a **complex model** for prediction as these products are not seen before
- Sacrificed interpretability for **better accuracy**

Models - Performance



Mean Absolute Error: Lower is better

Models - Performance



Models – Feature Importance

Existing Products

Random Forest

Explainable

VS

Repeated Products

Shallow Neural Network

Black Box

Next steps

1

Past Sales Volume

As our model suggests that past sales volume is a strong predictor to future sales volume. SCMP should keep the clean collection of past sales volume.

2

Data Collection

We can observe that some product characteristics in the dataset are not well managed. SCMP can consider to re-construct the data collection process for some features like pattern/ material/ colors etc.

3

Deep Learning

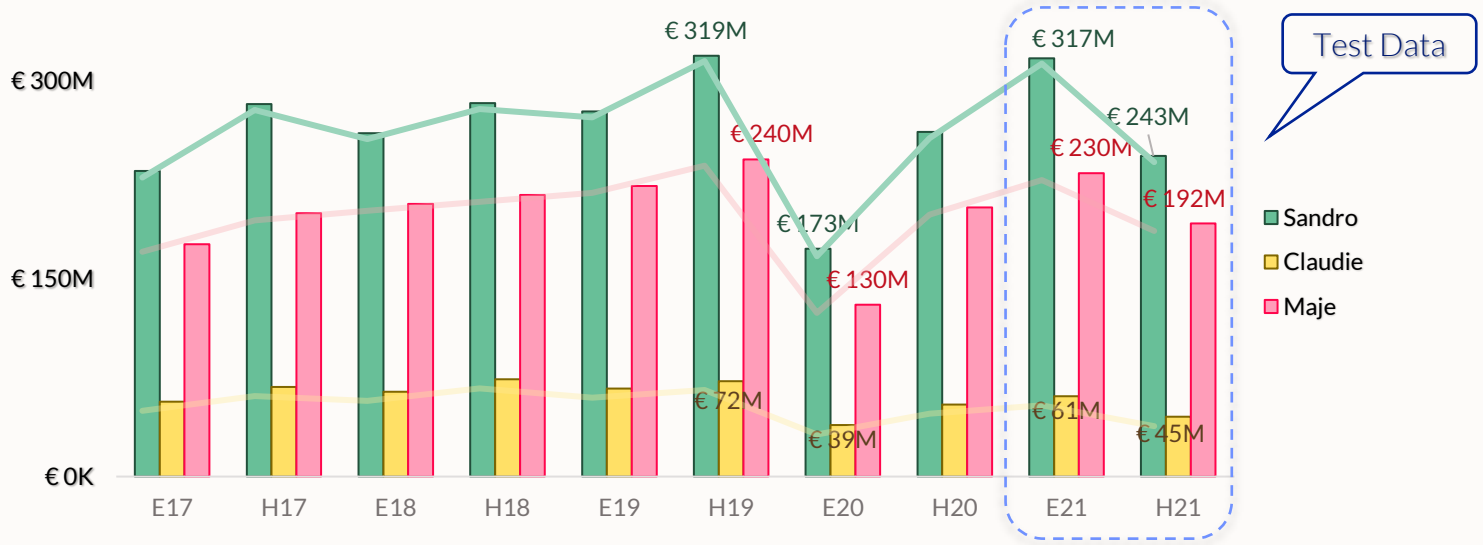
We have observed that deep learning indeed performed well in predicting sales volume. SCMP can consider to further develop this area and provide trainings to employees in this topic.

05

Q&A

Thank You!

Evolution Over Time



45% Season over Season drop in sales during E20 due to the pandemic, but since then SMCP has recovered its revenue.

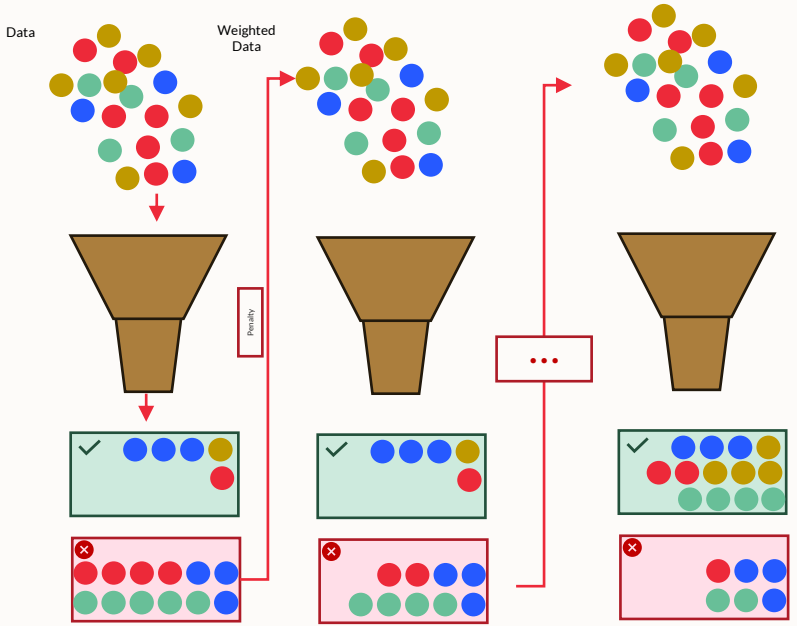
*Revenue calculated using the column TURNOVEREURO

Modeling

- Boosting is a technique used to train datasets when there is a high bias.
- In gradient boosting, we perform an initial classification, based on a simple decision tree
- The values that were incorrectly predicted, the weight of the datapoint is increased.
- We repeat the process till we get better results.
- We will be using ***LightGBM*** and ***XGBoost*** for our predictions.

Other models we tested

- Linear Regression
- Light GBM



Gradient Boosting