Hi35!

SIAHAAN–GENSOLLEN Rémy, MARINHO FERNANDES DANILO, PELTER PAUL, SIGNE TALLA FRANCK

Hi!ckaton #4 - Work overview

1 Overview

Hi, we're Hi35!, aa start-up with the ambition to develop a set of tools that are essential for supervising a supply chain. These tools include:

- A prediction model based on deep neural networks, whose performance increases as time series are enriched.
- An interactive dashboard that displays sales trends and predictions based on regions, storage sites and production lines storage sites.

The end result is the completion of a number of intermediate stages. This project encompasses data mining, a reformatting of the data for imputation of historical values and finally a prediction model to fill in the missing values in a time series corresponding to product sales. Each step is treated as an individual Web application.

2 Data understanding

We conduct a thorough exploration of the data through the analysis of variable correlations, pairplots, and conditioners on categorical variables. This enables us to gain insights into the distribution of data and understand the impact of individual features, facilitating more informed decision-making.

Our findings include a comprehension of the supply chain and how sales are distributed according to variables such as division and country, as represented in Figure 4.

3 Data preprocessing

Our data preprocessing approach comprises several essential steps to enhance the quality and utility of our dataset:

Data Cleaning: Addressing issues such as type inconsistencies and missing values is pivotal. We meticulously handle problems within our dataset, excluding columns that contribute little value to our analysis and potentially introduce noise.

Data Enrichment: To provide a comprehensive perspective for each product and date entry, we augment our dataset by incorporating columns that represent product prices across all available months. We have identified that, when merging the train and test datasets, every product had available data for each of the dates (except for May 2023, which was completely missing), allowing to reconstruct the temporal evolution of sales on the last 3 years, with gaps on each 4th month. This augmentation was useful to boost our models' confidence in sales estimations by leveraging additional data points and enabling them to project into the future rather than relying solely on past data. In addition to that, we include external ecnomic indicators such as GSPCI, inflation and energy price data into our final dataset.

Standardization: Prior to inputting data into our models, we standardize numerical columns. This crucial step ensures a consistent and stable behavior during model training, optimizing the overall performance of our predictive models.

4 Modelling Development

Before delving into complex models, we have explored linear regression methods, which, in addition to being simple and fast, are very powerful for our task, where the predicted month sales are largely explained by other month sales

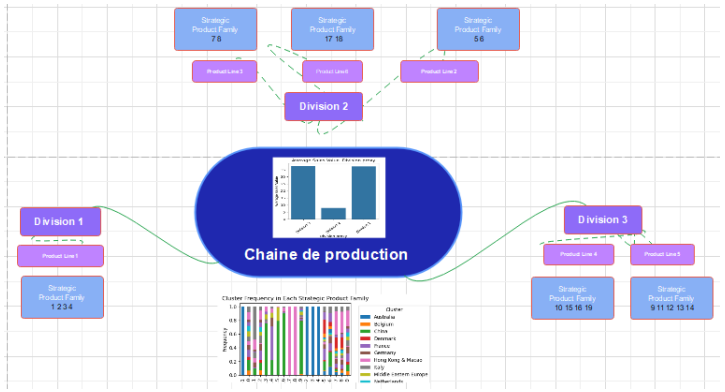

Figure 1: Study of production chain and sales by division

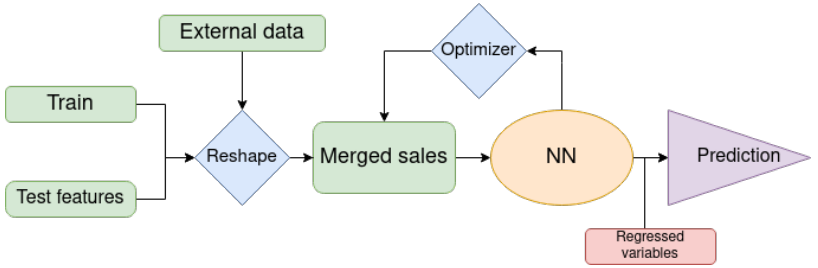

Figure 2: Data pre-processing and model pipeline

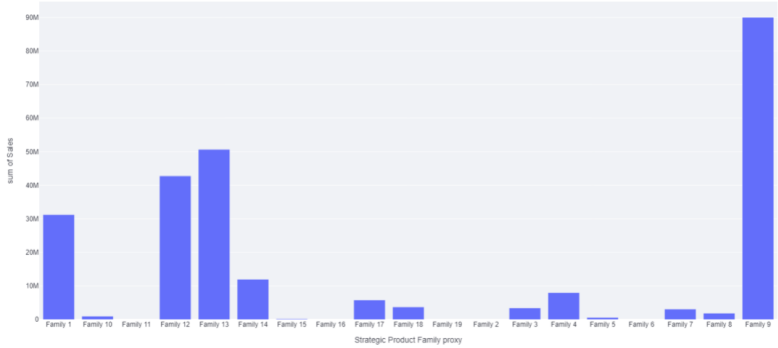

Figure 3: Sales volume by Strategy Product Family

values. We decided to go one step further and learn one specific linear model for each of the 9 date values, which ended up showing high variance across dates, not only because of the model intrinsic variance but possibly also due to the variance of sales values and data quality across dates. This approach allows a better understanding of the impact of the chosen period on the regression quality, and could be interesting for further pursuit in combination with more powerful models.

Although linear models yield good results, achieving a score of nearly 40%, we found that we could improve our predictions using more complex methods such as Gradient Boost and Deep Learning. Our most accurate solution consisting on a Multi-layer Perceptron (MLP) with two hidden layers, each with 100 nodes, and one hot encoding for categorical inputs. We have employed L2 regularization on the model weights, with a coefficient tuned by grid search, and optimized it using the Adam algorithm with early stopping. In addition to that, we performed cross-validation for overfitting control.

5 Deployment Strategy

Our prediction model is useful for filling in missing data, and reconstituting the dataset has enabled us to obtain sales trends for each product over 36 months. This has enabled us to create an interactive site that can be used to monitor and, above all, anticipate sales by country and by production line. Thanks to this tool, storage throughout the supply chain can be optimised.

6 Sales Forecasting and Sustainability

To make the supply chain as sustainable as possible, we need to anticipate customer needs as effectively as possible, but also reduce C02 emissions throughout the supply chain. The graph shows that it is production family number 9 where green investments should be made. Finally, we can see that China represents the largest sales volume, so we need to reduce the carbon impact of the supply chain in this territory.

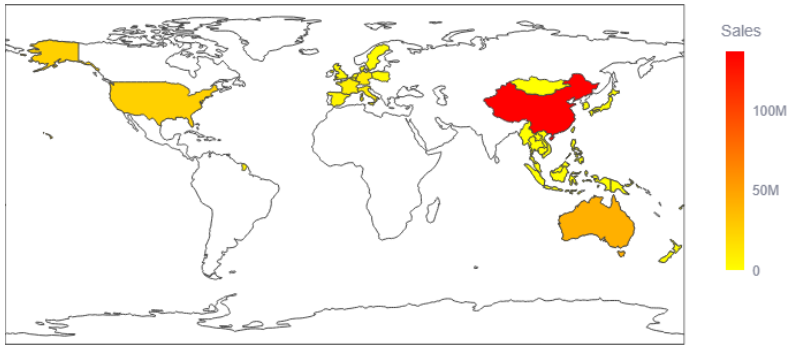

Figure 4: Map of countries by sales volume