Hi!ckathon 2023 4th edition: AI & Supply Chain

Team 34

Nardi Xhepi, Trung Dan Phan, Lucas Selini,
Mathilde Heibig, Théo Moret, Jérémy Pinault

1. Project Background and Description

Considering the highly legitimate concerns about the environment's current state and knowing the potential impact of AI on environment related KPIs, one of the main points of our project is to propose an impactful and sustainable solution. Furthermore, we emphasize the importance of simultaneously monitoring and optimizing sustainability and performance as critical metrics. To address these considerations, we have developed an inventive solution grounded in supply chain principles, detailed in this paper. Our solution entails a predictive algorithm utilizing the XGBoost model. The algorithm was trained to forecast the sales value for various products during the last month of a given quarter.

2. Model Explanation

2.1 Preprocessing

In order to prepare the dataset for the computation we followed multiple steps:

Categorical Variables: Our main training dataset contains many categorical variables with a major part of them being proxy variables. As a result, we must encode these variables. We used One-Hot Encoding to deal with these features.

Sales Value Imputation: We possess four columns related to sales values (Month 1 to 4), with Month 4 serving as our target. We confronted two primary challenges concerning sales values. Initially, the columns were in 'str' format, necessitating parsing. Secondly, Month 1 sales data exhibited missing values. To address this second issue, we replaced the missing values by the mean sales of each product.

Geographical Data:

The original dataset contains six columns related to geographical details tied to the product: Region, Country, Site, Operation, Zone, and Cluster. In an effort to streamline the dataset's features, we opted to focus solely on the most detailed geographical information available, encapsulated within the Site variable. We computed geographical distance between the starting location and destination of the shipments using (latitude, longitude) information.

<u>**Data Aggregation:**</u>

We had access to additional datasets to potentially help increase our predictive power. While time did not allow us to experiment with all, we believe each dataset had its merit to be included for further analysis.

- <u>World bank economic data:</u> The economic well-being of a country or region is closely intertwined with its purchasing power, which, in turn, impacts on the sales of specific products. Consequently, we could have decided to incorporate key economic indicators, including GDP, Final Consumption Expenditure Growth, Imports, Exports, Industry variables among others.
- <u>World bank inflation data:</u> Inflation-related indices appear pertinent for predicting sales, as elevated inflation may signify a reduction in consumption and, consequently, lower sales.
- <u>GSCPI data:</u> The Global Supply Chain Pressure Index is another extremely powerful KPI to get information on the market state.
- <u>LPI data:</u> Population and growth rates, as well as other LPI data related to the quality of infrastructure and logistics resources in the countries.

2.2 Model Implementation

To design our model, our starting point was the fact that we mainly had to deal with tabular data. Indeed, even though the outcome to predict (sales of Month 4) can be thought of as the next time step of a time series (through the sales value from Months 1, 2, and 3), the vast majority of the available features consist in tabular variables that have no temporal structure.

After testing several algorithms like linear regression, random forest, and support vector machine, we realized our best performance using gradient boosting regression model. Here are the features that are considered by our model:

- Sales that have been made during the first three months (quantitative).
- Distances (km) between the shipper’s warehouse and the country where the sale was made (quantitative).
- LPI Grouped rank for the arrival and departure country (quantitative).
- Product Line category (one-hot encoded)
- Customer Persona category (one-hot encoded)
- Strategic Product Family category (one-hot encoded)
- Product Life Cycle Status category (one-hot encoded)
- Whether the first month is January, May, or September (one-hot encoded)

- The matching year

We think that replacing the Sites and Country by the distances, one-hot encoding the remaining features and adding the Logistics grouped index positively impacts our model, while keeping it relatively simple and frugal.

3. Carbon Footprint

Considering the carbon emission in the design of our product, the utilized process includes two main features described below.

3.1. Efficient design:

As explained previously, we use a smart and light algorithm in order to limit the environmental impact of our product. In addition, we could consider the use of pipeline to optimize pre, and post-processing of the data to further improve its efficiency.

3.2 Monitor and Ensure quality:

After a thorough analysis of the macro-economic environment in which the company operates, we identified key strategies whose implementation would have a significant impact on reducing the environmental impact and CO_2 emissions of the company.

By leveraging machine learning and data-driven insights to streamline supply chain processes, our model enables us to predict sales accurately, leading to more efficient inventory management and a significant reduction in overall waste. We recommend using more sustainable transportation methods, such as consolidating shipments, optimizing routes, and choosing low-emission modes of transport. Moreover, we recommend transitioning to renewable energy sources for manufacturing facilities, warehouses, and distribution centers. Moreover, the supply chain involves many different parties. A collaborative approach to reduce carbon emissions throughout the supply chain is essential.

4. Conclusion

Considering both performance and environmental aspects, we obtain of final score of **0.4357** calculated with the bellow formula:

$RMSE\_hfct = (R_0 - \ 0.8 * RMSE_{pred}) \ / \ R\_0$

Where $R\_0 = RMSE(y_{true} \, , [0] * \text{len} \ (y_{true} \ ))$

At this point, other perspectives for enhancing our model can be considered. One potential improvement involves adjusting our target metric to incorporate a new feature penalized by a production factor. This modification aims to mitigate the risk of over-producing goods.

As for tackling carbon emission reduction, we understand that reducing CO_2 emission is a long process. It is important to set target goals and continuously monitor energy usage and carbon emissions. Through a commitment to continuous evaluation and refinement, the pursuit of sustainability in the context of CO_2 reduction remains an enduring and evolving commitment.