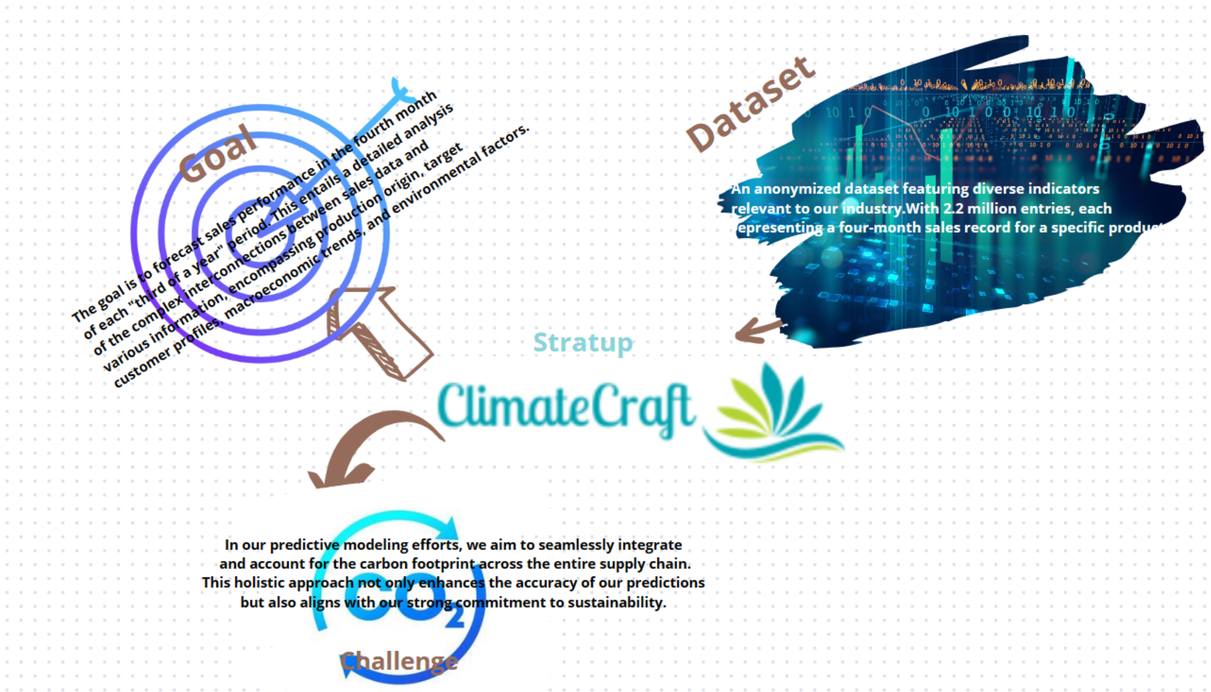
CLIMATECRAFT

Groupe 22

Project Background and Description



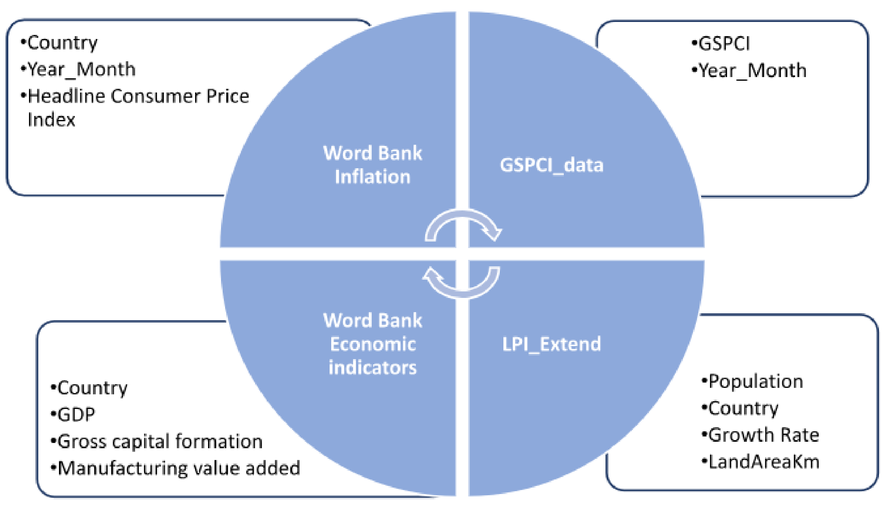
Model explanation

This research paper outlines the application of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to a large-scale supply chain data project. Initially comprising (3,173,346, 19) data points, the dataset presented a significant challenge due to its extensive information on product sales, regions, operations, and other critical features in terms of both size and complexity.

Data Understanding: The dataset underwent exploration to discern its structure and identify key features for the predictive model. Essential attributes included *id_product*,*Region*,*Country*,*Site*, *Operations*,*Zone*, *Cluster*,*Reference proxy*,*Product Line proxy*,*Division proxy*, *Customer Persona proxy*,*Strategic Product Family proxy*,*Product Life Cycle Status*,*Date*,and monthly sales data for four consecutive months (*Month 1* to *Month 4*). Notably, *Month 1* contained null values for May to July 2023 due to unavailable August data, a consideration for subsequent interpretation.

Data Preprocessing: To improve dataset quality and handle missing values, duplicates were removed, resulting in a refined dataset of shape (1,762,970, 19). Numeric features related to monthly sales were converted from strings to float data types. Illogical values, where monthly sales exceeded 50,000 in one month but were zero in another, were treated as outliers and eliminated to ensure dataset consistency.

Post data exploration, the data preprocessing phase involved merging the training data with GSPCI, LPIextend, and Worldbank_economics datasets. Relevant features were selected, and merging was performed using country and/or Year-Month. Some Nan values appeared after merging, and the 'mean imputation' technique was chosen to address this issue.



Model selection: Initially, we considered Xgboost as a potential model for our regression task. However, this choice necessitated encoding, and most of the encoders we were familiar with seemed inappropriate for our specific case. Even One-hot encoding posed challenges due to the potential generation of a large number of features. After conducting research, we opted for the Catboost Regressor, which adeptly handles categorical values, considers their order, and is generally resilient to overfitting. Upon implementing the Catboost Regressor for the first time, we observed improved performance compared to our initial Xgboost model. Subsequently, we dedicated the remaining time to fine-tuning hyperparameters and adding a regularization parameter to achieve optimal results. While we initially attempted GridSearch, its computational demands proved extensive, leading us to opt for manual fine-tuning.

From the outset, our aim was to bring creativity to the preprocessing phase, operating under the belief that even a basic model could yield satisfactory results. This perspective guided our decision not to explore alternative models, as we focused on refining the chosen Catboost Regressor for superior predictive accuracy.

Energy/CO2 consumption

With the objective of mitigating energy consumption and carbon emissions within the production chain, our strategy revolves around the utilization of our meticulously developed predictive modeling. The core concept hinges on the precision of our sales prediction model, wherein the more accurate the forecast, the more effectively diverse enterprises can tailor their production outputs to align with anticipated demand. This, in turn, translates to the production of quantities that precisely match market needs, thereby diminishing energy consumption associated with the production of raw materials and transportation. The crux of the approach lies in the symbiotic relationship between the accuracy of sales predictions and the subsequent alignment of production processes. A heightened level of precision enables businesses not only to streamline their manufacturing processes but also to curtail the ecological impact linked to surplus or deficit production. In practical terms, our predictive model is adept at forecasting zero values, allowing for the strategic decision to either produce minimal quantities or, in certain cases, abstain from production entirely, particularly for products with low or no market demand. This strategic utilization of our model contributes not only to energy savings but also to efficient inventory management. Furthermore, the visionary trajectory for our model involves expanding its predictive capabilities to encompass multiple months. This foresight aims to provide enterprises with a comprehensive and highly detailed outlook, enabling them to plan and execute production strategies over an extended time horizon. By extending the predictive horizon, businesses can cultivate a clearer understanding of market trends and demands, thereby enhancing the overall efficiency of their supply chain operations. In essence, our advanced predictive modeling not only aids in optimizing energy consumption and reducing carbon footprints but also empowers businesses with a strategic tool for agile and forward-thinking production planning. This aligns seamlessly with the principles of sustainable business practices and positions enterprises at the forefront of environmentally conscious and resource-efficient operations.

Conclusion

In conclusion, our analysis and predictive modeling have yielded valuable insights for optimizing the supply chain of the specified product during the fourth month of each quarter. Leveraging a comprehensive dataset we successfully developed a robust predictive model. The integration of additional datasets has significantly enriched our understanding and improved the accuracy of our predictions. This holistic approach enables us to consider various factors influencing sales, facilitating informed decision-making and the anticipation of market trends. Our predictions, validated against the interim benchmark, affirm the reliability of our model. However, recognizing the dynamic nature of the market, we emphasize the importance of continuous refinement and updates to the model as new data becomes available. This adaptive approach ensures the sustained accuracy of our predictions amidst changing market dynamics. Aligned with our startup's commitment to environmental consciousness, our focus extends beyond sales predictions. We have proposed measures to reduce the overall carbon footprint, aligning our supply chain model with sustainable business practices. Looking ahead, we recommend ongoing collaboration with stakeholders, vigilant monitoring of key variables, and regular model updates to maintain predictive efficacy. By implementing these measures, our start-up can not only optimize production and minimize overproduction but also contribute to a more sustainable and ecofriendly supply chain. As we anticipate the final leaderboard on Sunday, we are confident that our comprehensive approach to supply chain optimization and sustainability positions our start-up as an industry leader. Our journey showcases the impactful role of datadriven decision-making in shaping an environmentally responsible and efficient supply chain management system.

Name of participants

- **Zakaria El Kassimi**
- **Yosr ben-jemaa**
- **Asma boukhdhir**
- **MOHAMMED KAFILE**
- **Ghaith makhlouf**
- **NAJLAA SRIFI**