Problem statement formation

The challenge is to build a machine learning model that can predict unit sales more accurately than current subjective methods. The model should consider varying factors like store information, item specifics, promotions, and historical sales data. The aim is to balance inventory levels - avoiding both overstocking (leading to waste) and understocking (resulting in lost sales and customer dissatisfaction).

Context

Forecasting is an essential process in various fields, including government, science, and business. In the retail sector, particularly in grocery stores, accurate forecasting is crucial for inventory management. This project will focus on the use of machine learning for time-series forecasting to predict store sales for Corporación Favorita, a prominent grocery retailer in Ecuador. The goal is to develop a model that accurately predicts the unit sales of items across various Favorita stores.

Criteria for success

Success in this project will be measured by the model's accuracy in forecasting unit sales. The model should significantly outperform traditional forecasting methods. Additionally, success will also be gauged by the model's adaptability to different store environments and product types.

Scope of solution space

The solution will involve the development of a time-series forecasting model using machine learning techniques. This may include exploring various algorithms like ARIMA, LSTM (Long Short-Term Memory), and Prophet. The model should be capable of handling large datasets and incorporate factors such as promotions, seasonal trends, and store-specific dynamics. The evaluation metrics include Adjusted R-squared, MAPE/MAE, actual vs. predicted target, residuals distribution. Interpretability tools such as SHAP, InterpretML can also be applied.

Constraints

Data Availability: Dependence on historical sales data, store, and item information.
Computational Resources: Handling and processing large datasets effectively.
Model Generalizability: Ensuring the model is applicable across various stores and products.

Stakeholders

Corporación Favorita: The primary client who will use the forecast for inventory management.
Customers: Beneficiaries of more efficient stocking, leading to better product availability.
Data Science Team: Responsible for model development and data analysis.

Data sources

The primary data source will be Corporación Favorita's historical sales data at different locations, which includes:
Train.csv – time series of features store_nbr, family, and onpromotion, sales

Test.csv – same features as the train.csv
Store.csv – store metadata including city, state, type, and cluster
Oil.csv – daily oil price
Holidays_events.csv

The plan is to go through the classical data science project process – data wrangling, exploratory data analysis (EDA), preprocessing/baseline modeling, extended modeling. The deliverables include the code in Jupyter Notebook, report including details on methodology, analysis, and findings, and presentation that summarize the project, approach, and outcomes.