**Store sales forecasting**

Nishita Matlani (002866323)

Dhir Thacker (002819144)

**Team 07**

**Introduction**

**Background**

The Store Sales Forecasting project is driven by the necessity to navigate the challenges of retail management in an economically volatile environment, which heavily relies on oil revenues. This dependency on oil prices introduces significant fluctuations in the economy, directly impacting consumer purchasing power and, consequently, retail sales. Retailers face the ongoing challenge of matching supply with variable demand, necessitating accurate sales forecasts to minimize costs associated with overstocking or understocking and optimize staffing levels. The project addresses these needs by developing predictive models that can forecast sales with high accuracy, using a dataset that includes sales history, promotional activities, and external economic factors like oil prices and holidays.

**Motivation**

In the heart of Ecuador's fluctuating economy, retailers face the challenge of staying ahead in a rapidly changing consumer landscape. This project is born from the necessity to empower businesses with the foresight to navigate these shifts. By integrating data science and machine learning, our aim is to transform raw sales data into actionable insights. This not only enables informed decision-making on inventory and promotions but also paves the way for a retail environment that adapts in real-time to the ebb and flow of market demands.

**Goal**

Our mission is to harness cutting-edge technology to forge a predictive model that doesn’t just respond to market changes but anticipates them. We envision a tool that offers crystal-clear daily sales predictions across diverse store fronts and product lines, optimizing operations and boosting profitability. Beyond numbers, we aim to decode the complex dance between sales trends and influencing factors like promotions and holidays, offering a deeper understanding of consumer behavior. This is not just about predicting the future; it's about creating it.

**Methodology**

**Data Cleaning and Preprocessing:** Load and merge different datasets (sales, stores, transactions, oil prices, holidays/events) to create a comprehensive dataset for analysis. Address missing values, duplicate rows, and inconsistent data. For instance, missing values in oil data are filled using backfill, and categorical columns with missing values are filled with the most common value (mode) in each respective column.

**Exploratory Data Analysis (EDA):** An extensive EDA will be conducted to understand the underlying patterns and distributions in the data. This includes analyzing the distribution of sales, transactions, and the impact of promotions and holidays on sales and visualizations such as histograms, scatter plots, and bar charts. Examining relationships between variables (e.g., sales vs. oil prices, sales vs. promotions). Identify highly correlated features that may be important for modeling.

**Feature Engineering:** Create new features like 'Year', 'Month', 'Weekday Name', lag features, rolling averages, and more to capture temporal trends and patterns in the data that are useful for forecasting.

**Modeling:** Apply statistical and machine learning models to forecast sales. Selecting appropriate models like Linear regression, random forest, decision trees and XGBoost to capture the nature of sales data.

**Model Evaluation:** Assess the model's accuracy and reliability in forecasting sales using metrics like RMSE (Root Mean Square Error) or MAE (Mean Absolute Error). Compare different models to select the best performer. Analyze where the model's predictions deviate from actual sales to understand potential areas of improvement in feature engineering or model selection.

**Dataset Description**

The dataset contains a diverse set of variables, including information about store operations, product characteristics, external factors, and temporal dynamics. This variety allows for a comprehensive analysis that considers multiple facets of the retail environment. By examining the relationships and interactions between these variables, analysts can gain a deeper understanding of the factors driving sales.

There are total 5 csv files and 14 columns divided among those csv files.

5 CSV files:

1. train.csv
2. test.csv
3. sample\_submission.csv
4. stores.csv
5. oil.csv
6. holidays\_events.csv

Features Included in the Dataset

1. store\_nbr: Identifies the store.
2. family: Identifies the type of product sold.
3. onpromotion: Indicates whether the product was being promoted at the store on a given date.
4. sales: Total sales for a product family at a particular store on a given date.
5. city: City where the store is located (from stores.csv).
6. state: State where the store is located (from stores.csv).
7. type: Type of store (from stores.csv).
8. cluster: Grouping of similar stores (from stores.csv).
9. dcoilwtico: Daily oil price (from oil.csv).
10. type: Type of holiday or event (from holidays\_events.csv).
11. locale: Locale where the holiday is observed (from holidays\_events.csv).
12. locale\_name: Name of the locale (from holidays\_events.csv).
13. description: Description of the holiday or event (from holidays\_events.csv).
14. transferred: Indicates if the holiday was transferred to another date (from holidays\_events.csv).

**Dataset Link:** https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data