WCS Project

Store Sales – Time Series Forecasting

Group 11 – Adithya, Konstantinos, Sudarsan

Introduction

- We picked the **Store Data Forecasting** using Machine Learning.
- We will try to answer
  - If Holiday Events influence Sales Prices?
  - If Oil Prices influence Sales Prices?
- Our objective is to develop a machine learning model for accurate store sales predictions. It would minimize stockouts, reduce waste, and optimize pricing strategies
- Similar Time-Forecasting machine learning algorithms can be applied to other problems.
- Create generalized pipeline in BRANE

Dataset

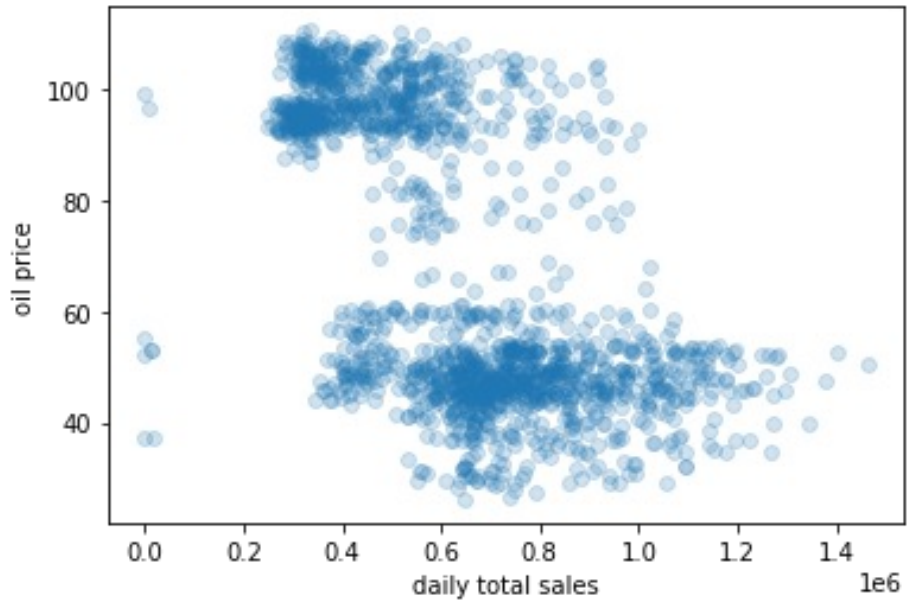
- The dataset for this competition consists of time series data from Favorita stores in Ecuador, including features such as store number, product family, promotion status, and sales.
- The training data in **'train.csv'** contains information on store numbers, product families, promotion status, and the corresponding sales.
- The test data in **'test.csv'** has the same features, and the objective is to predict the sales for the 15 days following the last date in the training data.
- Supplementary files include store metadata, daily oil prices, and information on holidays and events.

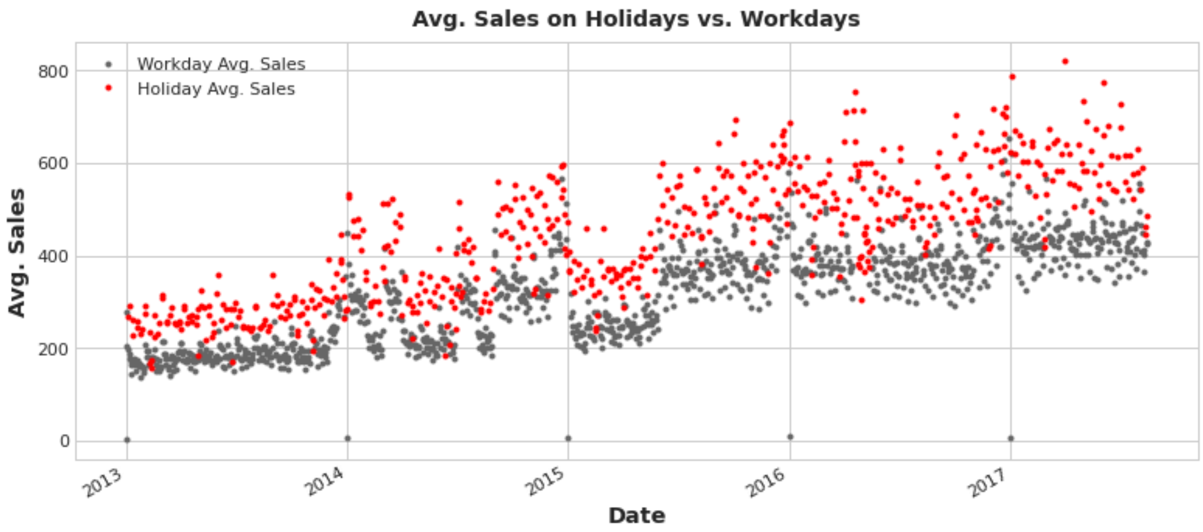
Data cleaning

- Feature Selection and Transformation: Remove unnecessary columns and convert selected columns to the appropriate data types.
- Aggregation:
  - Group the data by store numbers and sum the sales for each date.
  - Merge the processed data with store location information using store numbers.
  - Merge transaction data with the inputs and add lag features.
  - Adjust holiday data to include additional holidays and events.
- Scale the total sales data using MinMaxScaler and create lag features for time series analysis.
- Modify the test data to include a "sales" column filled with zeros and select relevant columns.
- Data Validation and Cleanup: Remove any remaining missing values and set the "date" column as the index.

Analysis

- Sales vs Oil Prices:
  - No obvious linear correlation between both of them. Observed Trend – Higher the Oil Price, Lower the Sales.
- Sales vs Holiday/Events:
  - Generally sales are higher on the day of holidays and events.
  - Fine grained analysis on impact of different holiday on sales using AB test.

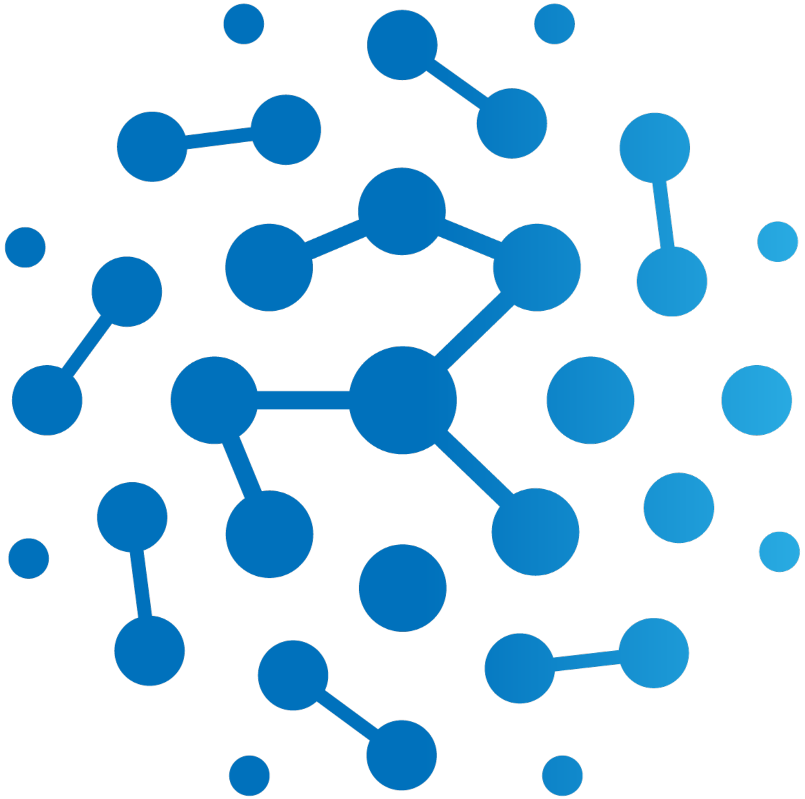




Training and Prediction

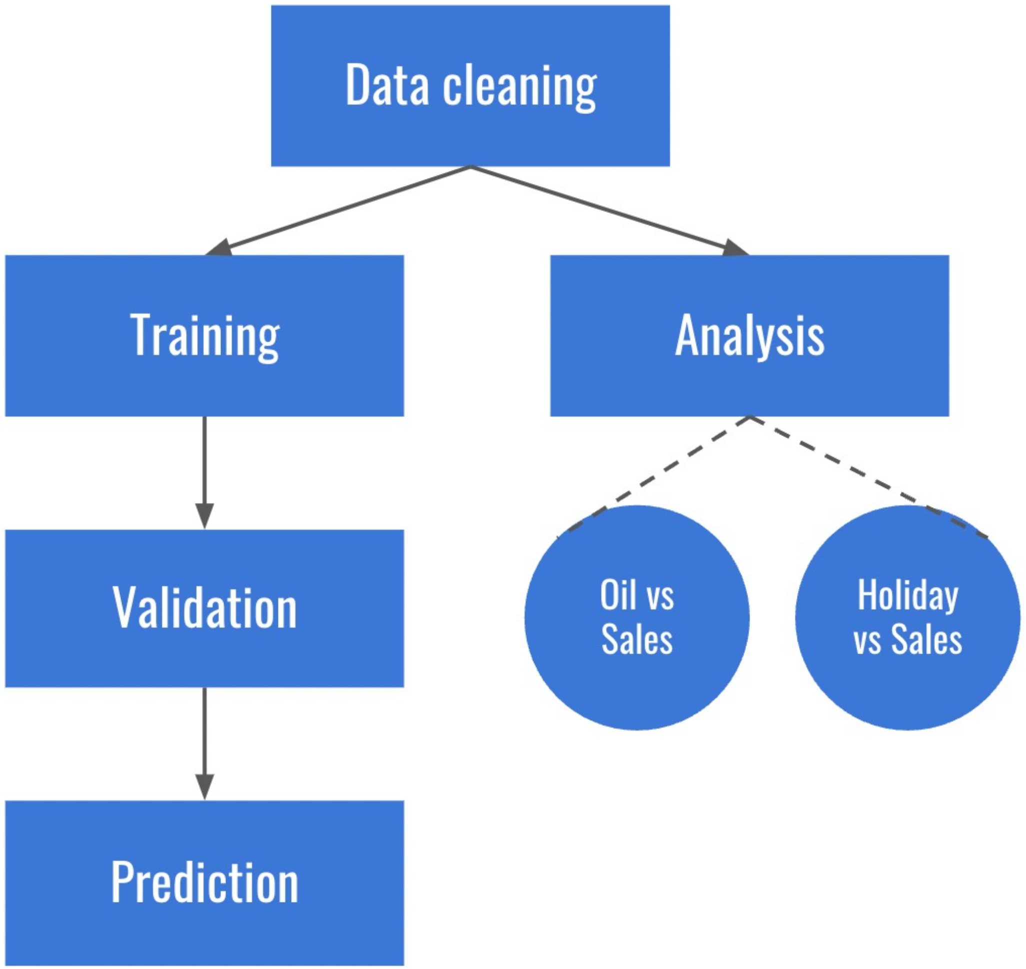
- In the training phase, the linear regression model learns the relationship between the input features and the target variable
- In the prediction phase, the model takes the input features of the testing data (such as date, store number, product family, and holiday information) and calculates the predicted sales values.
- It is a simple and interpretable model that works well when the relationship between the features and the target variable is approximately linear.

Brane



- Orchestrates the jobs on the compute cluster.
- Like SLURM job manager in DAS
- Defines and executes the pipeline jobs
- Easy to use interface

Pipeline



- Initial pipeline design for the problem
- The arrow represent data dependency
- The Oil v Sales and Holiday v Sales jobs can be done parallelly.
- Pipeline can be more fine-grained as we explore.

Discussion

- The work is still in progress.
- We need to fine tune the functions to be more reusable.
- Looking to parallelize the operations as planned and identify data dependencies across jobs.

Questions?