République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Constantine 2 - Abdelhamid Mehri -

Faculté des Nouvelles Technologies de l'Informatique et de la Communication Département des Technologies des Logiciels et Systèmes d'Information

Mini project SAD

Option : ILSI

Analyse de Walmart Sales

Réalisé par : KAHOUL Abd El Madjid
BOURAOUI Zakaria

- Session 2023/2024 -

Contents

I Introduction

This project report constitutes a thorough analysis of historical sales data derived from 45 distinct Walmart stores. The principal objective centers around predicting department-wide sales, necessitating a meticulous examination of various variables and contextual elements. As we meticulously navigate through the intricacies of the dataset, implement rigorous statistical methodologies, and formulate a predictive model, this report articulates a discerning exploration within the dynamic milieu of retail analytics. The ensuing revelations and implications encapsulated herein are poised to significantly augment our comprehension of sales prediction dynamics within the retail industry.

II Description of Data

The dataset utilized in this analysis comprises historical sales data for 45 Walmart stores situated in diverse regions. Each store is segmented into various departments, forming a comprehensive dataset crucial for predicting department-wide sales.

II.1 Dataset Components

- `stores.csv`: An anonymized file containing information about the 45 stores, including type and size.
  - Type: Categorical data indicating the type of store.
  - Size: Numerical data representing the size of the store.
- `train.csv`: Historical training data spanning from 2010-02-05 to 2012-11-01.
  - Store: Store number.
  - Dept: Department number.
  - Date: Date of the week.
  - Weekly Sales: Sales for the given department in the given store.
  - IsHoliday: Binary indicator of whether the week is a special holiday week.
- `test.csv`: Similar to `train.csv`, but with weekly sales withheld for prediction.
- `features.csv`: Additional data related to store, department, and regional activity.
  - Store: Store number.
  - Date: Date of the week.
  - Temperature: Average temperature in the region.
  - Fuel Price: Cost of fuel in the region.
  - MarkDown1-5: Anonymized data related to promotional markdowns (available after Nov 2011).
  - CPI: Consumer price index.
  - Unemployment: Unemployment rate.
  - IsHoliday: Binary indicator of whether the week is a special holiday week.

  `Holiday Weeks`: The dataset incorporates several promotional markdown events, which precede major holidays such as the Super Bowl, Labor Day, Thanksgiving, and Christmas. Evaluation weights holiday weeks five times higher than non-holiday weeks, emphasizing the impact of these events on sales.

III Statistical analysis

III.1 Data Exploration and Time Series Analysis

III.1.1 Time-Related Features

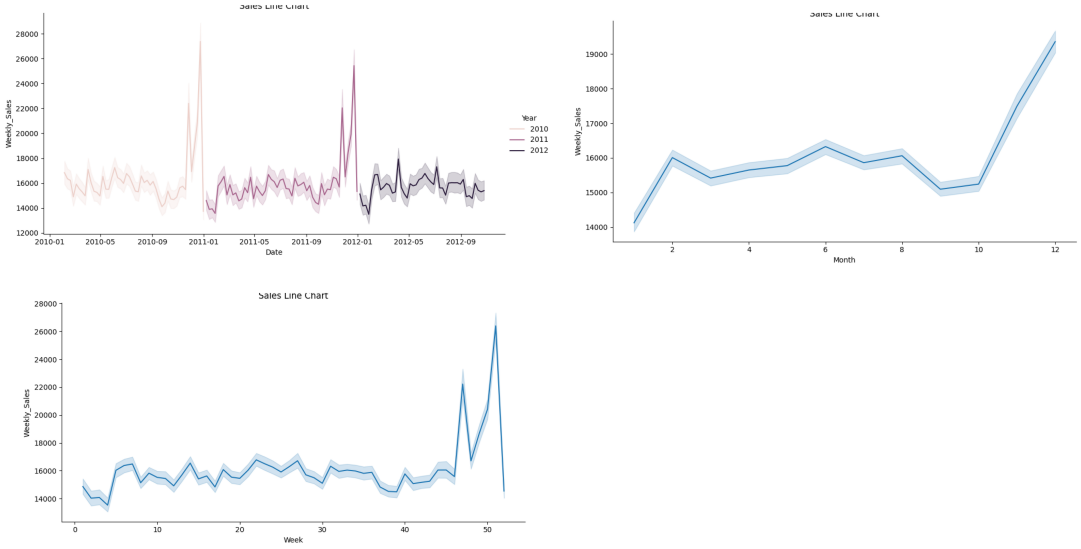

Figure 1: Weekly Sales Over Time (year, month ,week)

Discussion: While no significant year-wise trend is observed, weeks 45 to 50 (November and December) exhibit a notable increase in sales, driven by Christmas and Thanksgiving. Acknowledging and utilizing this seasonality is essential for optimizing strategies during these high-sales periods.

III.2 Feature Engineering and Data Merging

Data Merging: We enriched the feature sets for each data point by merging information from `train.csv`, `stores.csv`, and `features.csv`. This comprehensive approach provides a more holistic view, incorporating store details, sales data, and additional features for a more robust analysis.

III.3 Correlation Analysis

III.3.1 Categorical Features

Discussion: Notable correlations exist between categorical features and `Weekly_Sales`. `Department` exhibits a strong positive correlation (0.537), emphasizing its significant association with weekly sales. In contrast, `Type` has a weaker correlation (0.036). Considering these correlations is crucial for predicting and optimizing weekly sales.

III.3.2 Numerical Features

Discussion: `Size` demonstrated the highest correlation with `Weekly_Sales`.

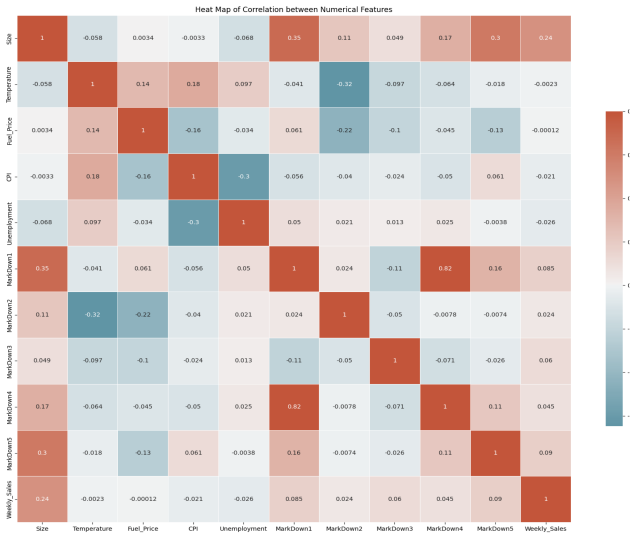

Figure 2: Heat Map - Correlation Between Numerical Features

III.4 Features Importances

Discussion: Key features influencing `Weekly_Sales` were identified as `Dept`, `Size`, `Store`, `Week`, and `CPI`. Recognizing their substantial impact is crucial for accurate sales predictions and strategic decision-making to optimize business outcomes.

III.5 Exploratory Data Analysis (EDA)

Discussion: The analysis emphasizes the significance of Department and Store features for accurate sales forecasting. Notably, Type A stores show the highest sales, and understanding department-specific variations during holidays is essential for optimizing strategies and promotions.

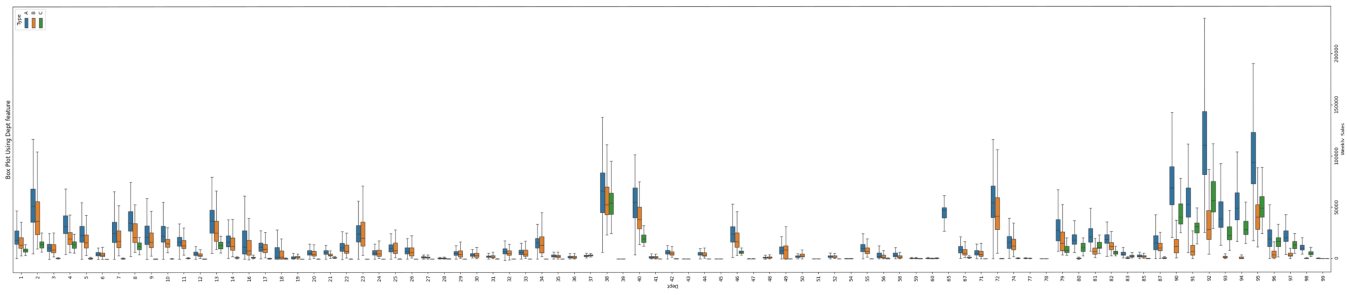

Figure 3: Box Plot - Department Feature

III.6 Inefficacious Feature Extraction

Discussion: Month-end and Month-start features exhibit very low correlation with weekly sales, indicating their limited predictive value. The data shows that these features may not be useful in understanding sales variations. It is suggested to explore alternative temporal features for a more comprehensive analysis.

III.7 Feature Impact Analysis

III.7.1 Economic Indicators Analysis

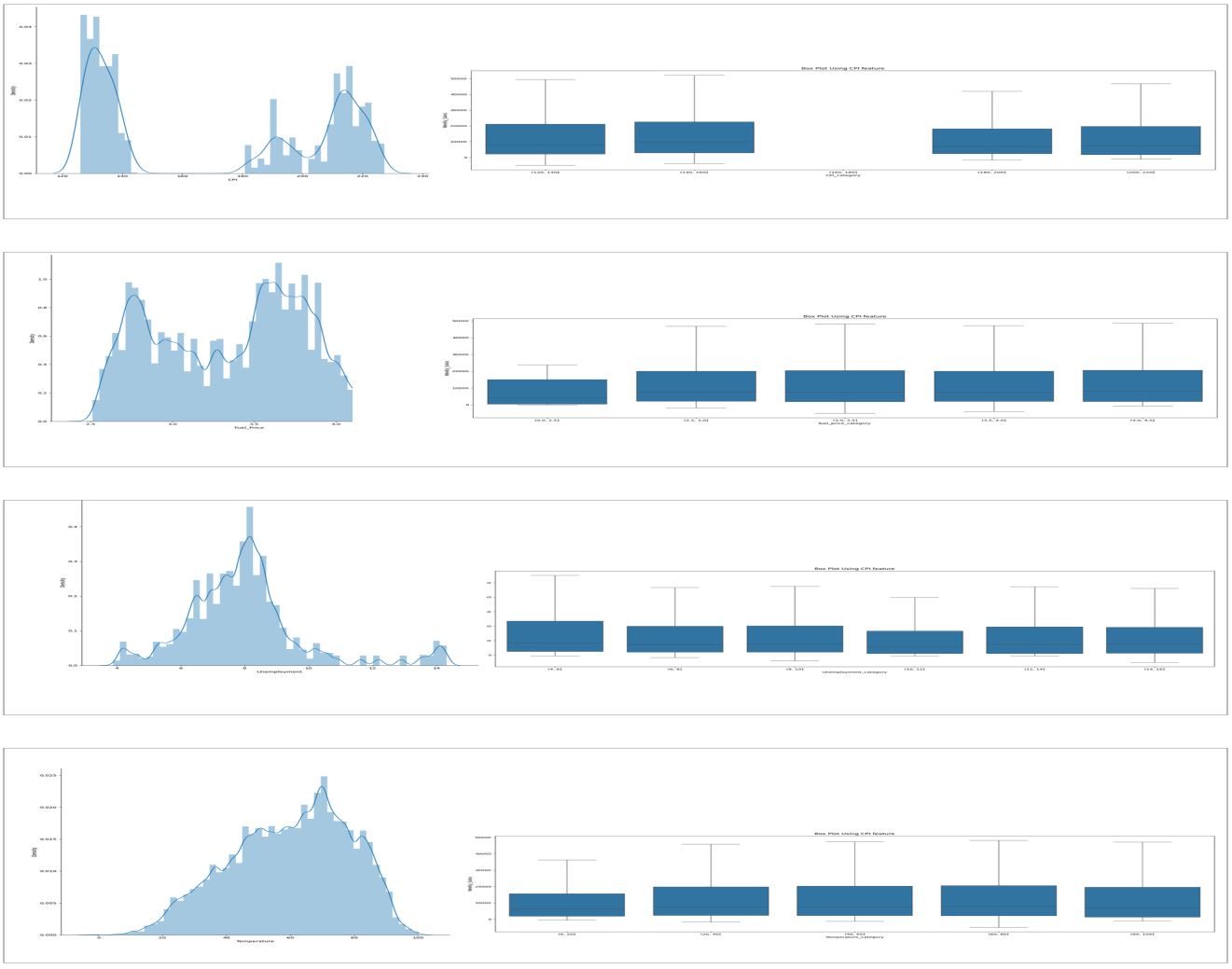

Figure 4: Economic Indicators vs. Weekly Sales

Discussion: Economic indicators, such as CPI, Fuel Price, Unemployment, and Temperature, show minimal impact on weekly sales. Visualizations using box plots and histograms indicate consistent sales patterns across different indicator ranges. Considering their limited influence, it is advisable to remove these features from the model, aligning with the exclusion of Markdowns for similar reasons.

III.7.2 holidays Analysis

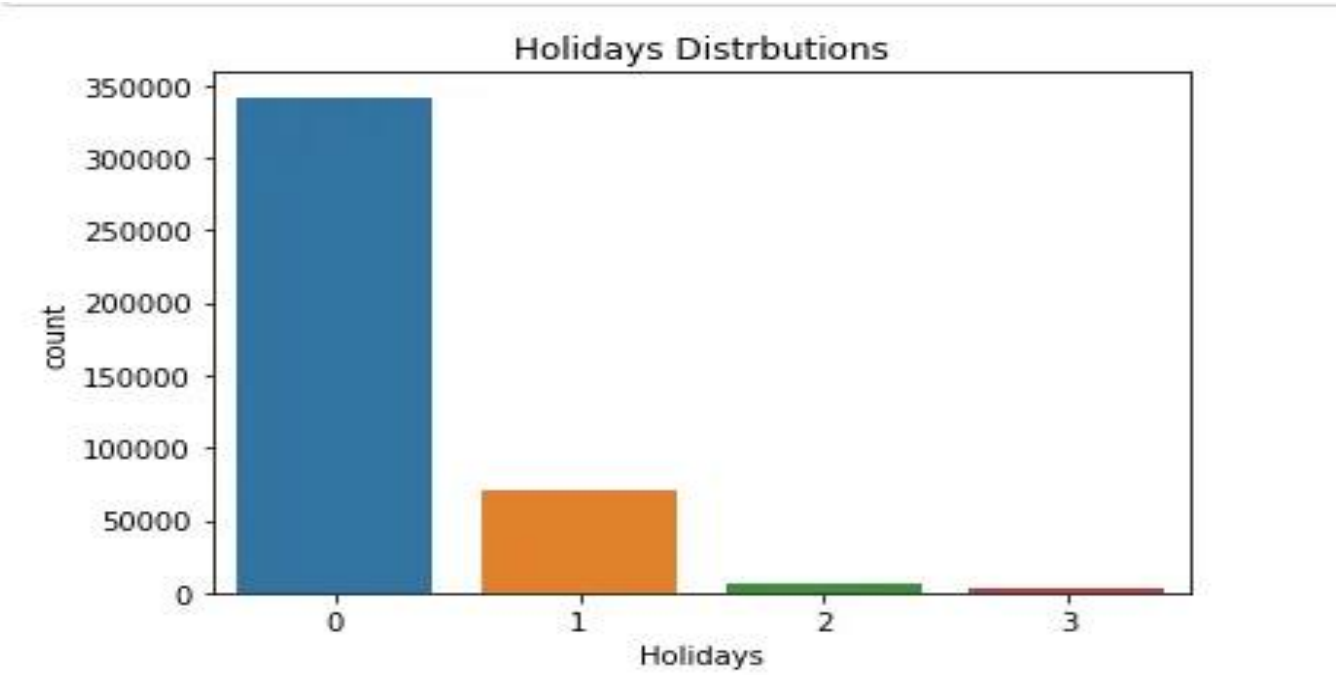

Figure 5: Histogram - Holidays

Discussion: Sales showed changes with the number of holidays in a week, indicating the feature's relevance.

III.8 Handling Missing Values

Discussion: Markdown values were less correlated with `Weekly_Sales`, and certain columns were dropped.

IV Model Evaluation

IV.1 Model Used

Discussion: Various models were evaluated, including linear regression, SVR, RidgeCV, ElasticNet, RandomForest, and XGBoost. The Random Forest model, with hyperparameters {n_estimators: 100, max_depth: 27}, outperformed others in terms of the Weighted Mean Absolute Error (WMAE) metric.

IV.2 Results Obtained

Discussion: The Random Forest model achieved the lowest WMAE score of 1419.06, demonstrating its effectiveness in predicting weekly sales. Feature importance analysis highlighted the significance of `Dept`, `Size`, `Store`, `Week`, and `CPI`.

V Conclusion:

The analysis provides valuable insights into weekly sales influencers, with considerations for seasonal trends, department/store variations, and the impact of holidays. Despite the comprehensive

approach, certain limitations, such as the exclusion of external economic factors and the need for more granular data, are acknowledged. Future directions involve exploring advanced statistical techniques and integrating additional data sources for a more robust predictive model. Lessons learned emphasize the importance of continuous refinement and adaptability in data-driven decision-making for optimizing business outcomes.

VI User Interface (UI) Build

To enhance data analysis and forecasting capabilities, we developed a comprehensive user interface (UI).exemple of what The UI includes :

- Sales Analytics
- Markdown Impact
- Prediction of Weekly Sales ..

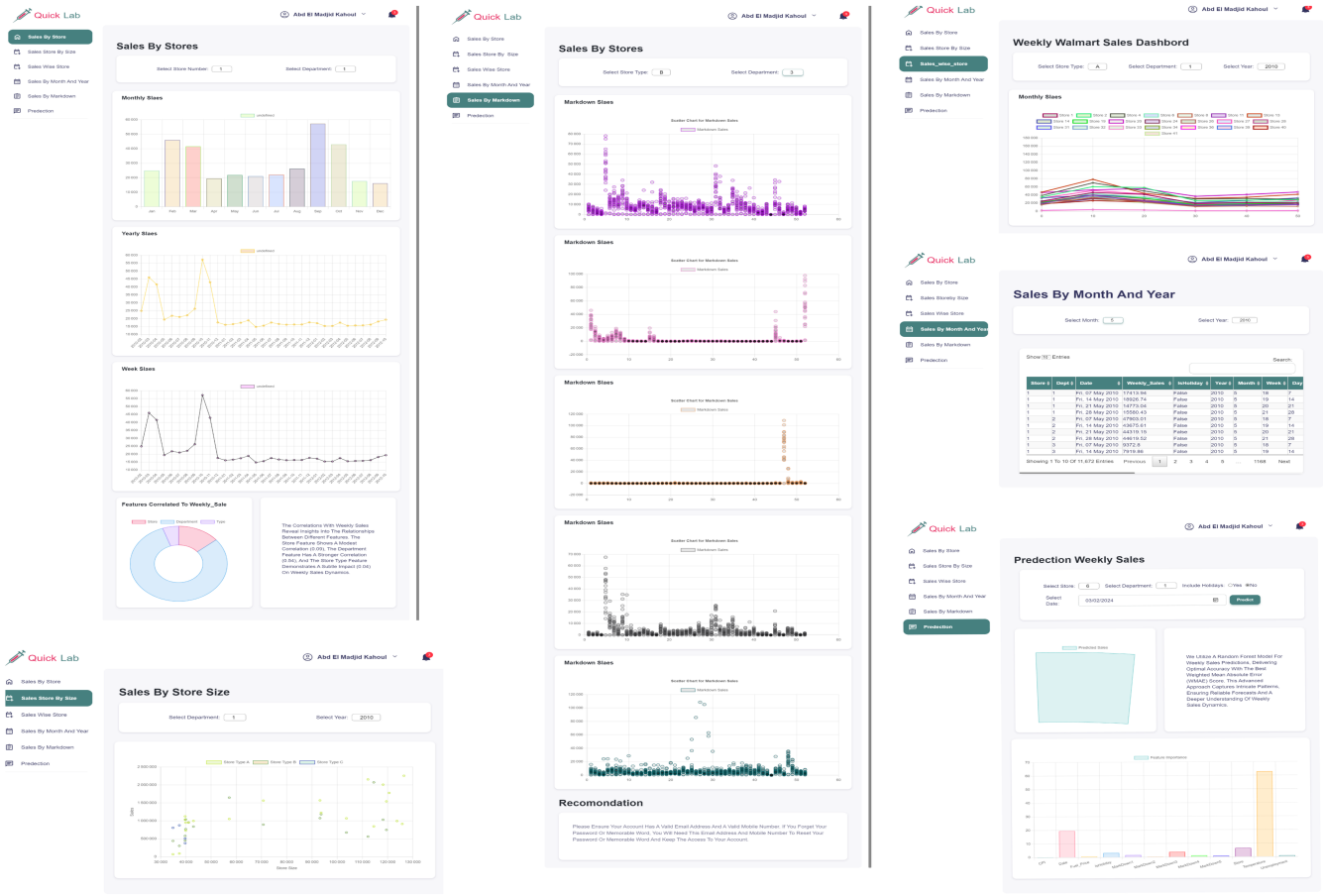

Figure 6: UI interfaces