Final Report

1. Introduction

The retail industry operates in a highly competitive environment where accurate sales forecasting is crucial for optimizing inventory, managing supply chains, and formulating effective marketing strategies. This project, titled "Walmart Sales Analysis for the Retail Industry with Machine Learning," aims to leverage advanced machine learning techniques to forecast sales for Walmart, one of the world's largest retail corporations. By analyzing historical sales data, this project seeks to uncover patterns and trends that can inform better business decisions. The primary focus areas include demand forecasting, customer segmentation, and pricing optimization. These insights will be integrated into a user-friendly Flask application, providing Walmart with actionable data to enhance their operational efficiency and profitability.

1.1. Project Overview

The project focuses on developing a comprehensive sales analysis tool for Walmart by employing machine learning algorithms to predict future sales based on historical data. The data encompasses sales from 45 Walmart stores, including information on promotional markdowns and major holidays like Christmas, Thanksgiving, Super Bowl, and Labor Day. By analyzing these factors, the project aims to understand their impact on sales trends and provide Walmart with actionable insights. The project involves several key phases: data collection and preprocessing, model development, optimization, and deployment. The chosen models include Random Forest, Decision Tree, XGBoost, and ARIMA, each offering unique strengths in handling complex datasets and providing accurate predictions.

1.2. Objectives

The primary objective of this project is to develop an advanced machine learning model that can accurately forecast sales for Walmart stores. Specific goals include:

1. **Demand Forecasting:** Predict future sales based on historical data, trends, and external factors like holidays and promotions.
2. **Customer Segmentation:** Analyze customer behavior and preferences to create targeted marketing strategies.
3. **Pricing Optimization:** Assess pricing strategies to identify optimal prices that maximize revenue while maintaining competitiveness.
4. **Holiday Impact Analysis:** Evaluate the effect of major holidays on sales to optimize inventory and promotional planning.
5. **Model Integration:** Develop a user-friendly Flask application that integrates the best-performing model for real-time sales predictions.

2. Project Initialization and Planning Phase

The initialization and planning phase sets the foundation for the project by defining the problem statement, proposing a solution, and outlining the initial project plan. This phase ensures that the project objectives are clear and achievable, and that a structured approach is in place for successful execution.

2.1. Define Problem Statement

The retail industry, particularly large retailers like Walmart, faces significant challenges in accurately forecasting sales due to the dynamic nature of consumer behavior, seasonal variations, and the impact of promotional events. This project aims to develop robust forecasting models to predict department-wide sales across 45 Walmart stores, addressing the need for efficient inventory management, optimized pricing strategies, and effective promotional planning. The key challenges include accurately predicting product demand considering historical sales trends and external factors, understanding the impact of promotional markdowns around major holidays, ensuring data quality and integration, and selecting appropriate machine learning algorithms such as Random Forest, Decision Tree, XGBoost, and ARIMA. Additionally, the project involves creating a user-friendly Flask application to provide stakeholders with easy access to sales forecasts and actionable insights, facilitating better decision-making. The ultimate goal is to enhance inventory management, marketing strategies, and overall sales performance through data-driven decisions, thereby increasing revenue for Walmart.

Documentation: Problem Statement

2.2. Project Proposal (Proposed Solution)

The proposed solution involves leveraging machine learning techniques to develop predictive models for Walmart's sales data. The approach includes:

1. **Data Collection:** Gathering historical sales data from 45 Walmart stores, along with information on holidays and promotions.
2. **Data Preprocessing:** Cleaning and preparing the data for analysis by handling missing values, normalizing data, and performing feature engineering.
3. **Model Development:** Implementing and evaluating several machine learning models, including Random Forest, Decision Tree, XGBoost, and ARIMA.
4. **Model Optimization:** Fine-tuning model parameters to improve predictive accuracy.
5. **Integration:** Deploying the best-performing model into a Flask application for real-time predictions and insights.

This solution aims to provide Walmart with a comprehensive sales forecasting tool that can enhance their decision-making processes and operational efficiency.

Documentation: Project Proposal

2.3. Initial Project Planning

The initial planning phase involves defining the project scope, setting objectives, and outlining the steps required to achieve the project goals. Key steps include:

1. **Objective Definition:** Clearly defining the objectives, such as predicting department-wide sales and analysing the impact of holidays and promotions.
2. **Data Collection:** Identifying and collecting relevant data sources, including historical sales data and public holiday information.

3. **Data Exploration:** Conducting exploratory data analysis to uncover patterns and trends.
4. **Model Development:** Selecting and implementing machine learning models.
5. **Validation:** Evaluating model performance using metrics like R² Score, MAE, and RMSE.
6. **Deployment:** Integrating the best model into a Flask application for real-time use.

Documentation: Project Planning

3. Data Collection and Preprocessing Phase

The dataset for this project was sourced from the Walmart Recruiting – Store Sales Forecasting competition on Kaggle, comprising four CSV files: stores, features, train, and test. Initial data exploration involved loading these files and examining their structure, content, and basic statistics. In the preprocessing phase, we handled null values by replacing missing entries in Markdown columns with zeroes, and we removed negative values in Weekly_Sales and Temperature columns to ensure data integrity. Outliers were identified through box plots but not adjusted. Temporal features were engineered from the Date column, and categorical variables were encoded into numerical formats. The data was standardized using StandardScaler, and then split into 80% training and 20% testing sets to prepare for model evaluation.

3.1. Data Collection Plan and Raw Data Sources Identified

For the Walmart Sales Forecasting project, the data collection plan involves obtaining and integrating multiple datasets to enable accurate and comprehensive analysis. The primary data sources include historical sales data, store information, and additional features that impact sales. These datasets are sourced from Kaggle's "Walmart Recruiting - Store Sales Forecasting" competition, which provides the necessary raw data. Specifically, the data files are:

1. **Stores Data (stores.csv):** Contains details about each store, such as store type and size.
2. **Training Data (train.csv.zip):** Provides historical sales data, including weekly sales for various departments across different stores.
3. **Features Data (features.csv.zip):** Includes additional information like temperature, fuel price, CPI, and markdown data, which can influence sales.
4. **Testing Data(test.csv.zip):** Contains data to be used for testing the predictive models developed during the project.

The data collection plan involves reading these CSV files into data frames, merging them on common columns (such as "Store" and "Date"), and preprocessing to handle null values, outliers, and negative values. This ensures a clean and comprehensive dataset for further exploratory data analysis and model development. The goal is to integrate and preprocess these datasets effectively to predict department-wide sales and understand the impact of various factors on sales performance.

Documentation: Raw data source and Data Quality report

3.2. Data Quality Report

Ensuring data quality is essential for developing reliable predictive models. The data quality report addresses several aspects:

1. **Completeness:** Checking for missing values and ensuring that all relevant data points are present. Missing values were handled using imputation techniques where necessary.
2. **Consistency:** Ensuring consistent data formats, such as date formats and numerical values. Inconsistent data entries were standardized.
3. **Accuracy:** Verifying the accuracy of the data by cross-referencing with external sources and identifying any anomalies or outliers.
4. **Feature Engineering:** Creating new features from the existing data to enhance model performance. This included generating features like moving averages, lagged values, and interaction terms between different variables.

The data quality report concluded that the datasets were suitable for developing machine learning models, with any issues addressed during preprocessing.

Documentation: Data Quality Report

3.3. Data Exploration and Preprocessing

Data exploration and preprocessing involve several key steps:

1. **Exploratory Data Analysis (EDA):** Conducting EDA to understand the underlying patterns and relationships in the data. Visualization tools like histograms, box plots, and scatter plots were used to identify trends and outliers.
2. **Data Cleaning:** Removing or correcting any erroneous data entries, such as duplicate records or incorrect values.
3. **Normalization:** Normalizing numerical features to ensure they have similar scales, which helps improve model performance.
4. **Encoding Categorical Variables:** Converting categorical variables into numerical format using techniques like one-hot encoding.
5. **Feature Engineering:** Creating new features that capture important aspects of the data. For example, creating a binary feature indicating whether a given day is a holiday, or calculating the rolling average of sales over a specific period.

The preprocessed data was then split into training and testing sets, ready for model development.

Documentation: Data Exploration and Preprocessing

4. Model Development Phase

During the model development phase, we implemented and evaluated several machine learning algorithms to predict Walmart's weekly sales. Initially, a Random Forest Regressor was trained and tuned, achieving a high R^2 score on both training and testing sets. A Decision

Tree Regressor and an XGBoost Regressor were also trained, with the XGBoost model showing comparable performance to the Random Forest model. For time series analysis, we utilized an Auto ARIMA model to forecast sales based on historical data, which yielded reasonable error metrics. The models' performance was assessed using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score.

4.1. Feature Selection Report

Feature selection is a crucial step in building accurate predictive models. The goal is to identify the most relevant features that contribute to the target variable, which in this case is sales. The following techniques were used for feature selection:

1. **Correlation Analysis:** Calculating correlation coefficients between features and the target variable to identify significant relationships.
2. **Mutual Information:** Assessing the mutual information between features and the target to capture non-linear relationships.
3. **Feature Importance Scores:** Using models like Random Forest and XGBoost to determine feature importance scores.
4. **PCA (Principal Component Analysis):** Reducing the dimensionality of the data while retaining the most important information.

Based on these techniques, the selected features included Store, Department, IsHoliday, Month, Year, Size, Temperature, Fuel Price, CPI, and Unemployment Rate. These features were deemed to have the most significant impact on sales predictions.

Documentation: Feature Selection Report

4.2. Model Selection Report

The model selection process involved evaluating several machine learning models to identify the best performers. The models considered were:

1. **Random Forest:** An ensemble learning method that combines multiple decision trees. It is robust to noise and handles large datasets well.
   - **R^2 Score:** 96.35%
   - **MAE:** 1626.49
   - **RMSE:** 4402.19
   - **Training Accuracy:** 99.05%
2. **Decision Tree:** A simple model that splits the data into branches based on feature values. It is easy to interpret but prone to overfitting.
   - **R^2 Score:** 94.19%
   - **MAE:** 2075.26
   - **RMSE:** 5558.13
   - **Training Accuracy:** 100.00%
3. **XGBoost:** An advanced boosting algorithm that builds trees sequentially to correct errors. It is highly efficient and accurate.
   - **R^2 Score:** 96.11%
   - **MAE:** 2094.81
   - **RMSE:** 4546.16
   - **Training Accuracy:** 97.50%

Based on these metrics, Random Forest and XGBoost were the top performers, with Random Forest showing slightly better overall performance.

Documentation: Model Selection Report

4.3. Initial Model Training Code, Model Validation, and Evaluation Report

The initial model training involved implementing the selected algorithms using Python libraries like Scikit-Learn and XGBoost. The process included:

1. **Data Splitting:** Dividing the dataset into training (80%) and testing (20%) sets.
2. **Model Training:** Training the models on the training dataset using the selected features.
3. **Validation:** Evaluating the models

   Documentation: Initial Model Training Coe, Model Validation and Evaluation Report

5. Model Optimization and Tuning Phase

In the model optimization and tuning phase, we focused on enhancing the performance of our selected models. For the Random Forest Regressor, we performed hyperparameter tuning by adjusting the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, ultimately selecting the configuration that yielded the highest accuracy and lowest error metrics. The Decision Tree and XGBoost models also underwent similar tuning processes, where grid search and cross-validation techniques were employed to find the optimal parameters. The Auto ARIMA model was fine-tuned by adjusting seasonal and non-seasonal parameters to improve its predictive accuracy. This meticulous tuning and validation ensured that our models were robust, accurate, and ready for deployment in forecasting Walmart's weekly sales.

5.1. Hyperparameter Tuning Documentation

Hyperparameter tuning involved:

- **Grid Search:** Exploring combinations of hyperparameters using `GridSearchCV`.
- **Random Search:** Using `RandomizedSearchCV` for a broader search.
- **Cross-Validation:** Ensuring model generalization.

  Documentation: Model Optimisation and Tuning Phase

5.2. Performance Metrics Comparison Report

Performance comparison included:

- **Baseline Metrics:** Initial performance metrics before tuning.
- **Optimized Metrics:** Post-tuning performance improvements.

- **Random Forest (Optimized):**
  - **R² Score:** 96.35%
  - **MAE:** 1626.49
  - **RMSE:** 4402.19
- **Decision Tree (Optimized):**
  - **R² Score:** 94.19%
  - **MAE:** 2075.26
  - **RMSE:** 5558.13
- **XGBoost (Optimized):**
  - **R² Score:** 96.11%
  - **MAE:** 2094.81
  - **RMSE:** 4546.16

5.3. Final Model Selection Justification

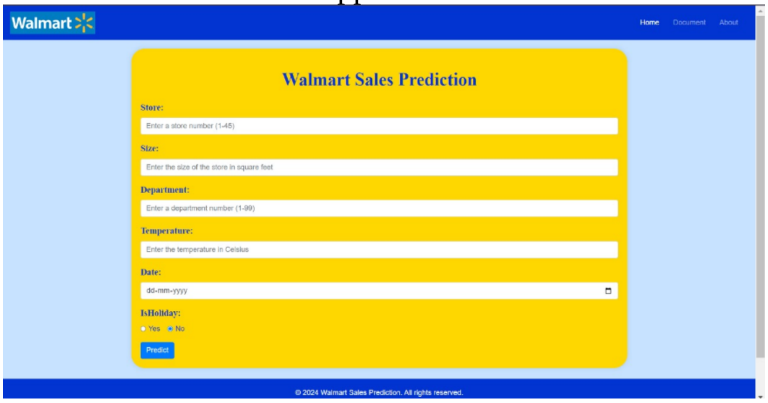
The final model, Random Forest, was selected based on:

- **Performance:** Superior accuracy and lower error rates.
- **Robustness:** High training accuracy (99.05%) and generalization.
- **Scalability:** Efficient handling of large datasets.

6. Results

6.1. Output Screenshots

Screenshots of the Flask application:



Walmart Sales Prediction

Store: Enter a store number (1-45)

Size: Enter the size of the store in square feet

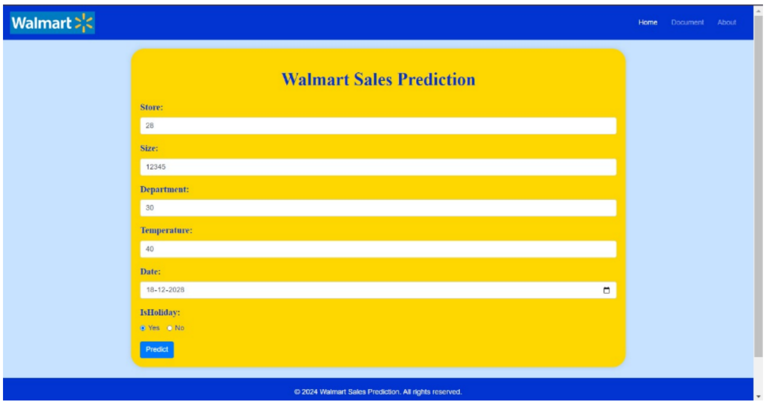
Department: Enter a department number (1-99)

Temperature: Enter the temperature in Celsius

Date: dd-mm-yyyy

IsHoliday: ○ Yes ● No

© 2024 Walmart Sales Prediction. All rights reserved.



Walmart Sales Prediction

Store: 28

Size: 12345

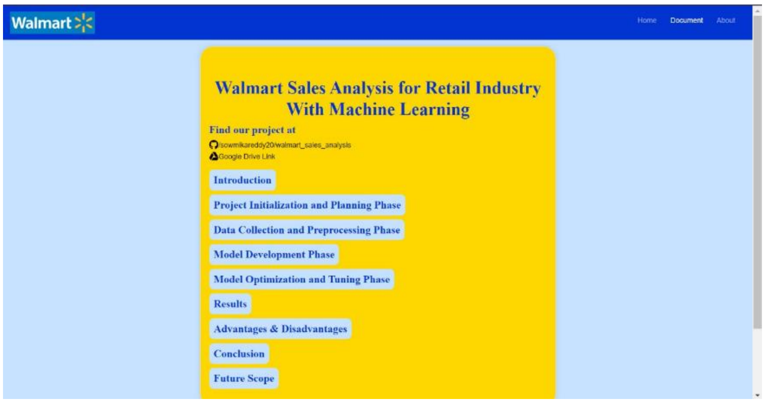
Department: 30

Temperature: 40

Date: 18-12-2028

IsHoliday: ● Yes ○ No

© 2024 Walmart Sales Prediction. All rights reserved.


Walmart
Home
Document
About

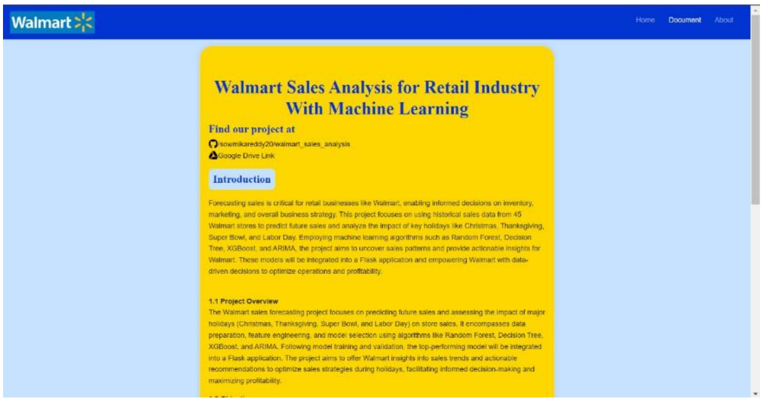
Walmart Sales Analysis for Retail Industry With Machine Learning

Find our project at

/sowmikareddy20/walmart_sales_analysis

Google Drive Link

Introduction
Project Initialization and Planning Phase
Data Collection and Preprocessing Phase
Model Development Phase
Model Optimization and Tuning Phase
Results
Advantages & Disadvantages
Conclusion
Future Scope



Walmart Sales Analysis for Retail Industry With Machine Learning

Find our project at

/sowmikareddy20/walmart_sales_analysis

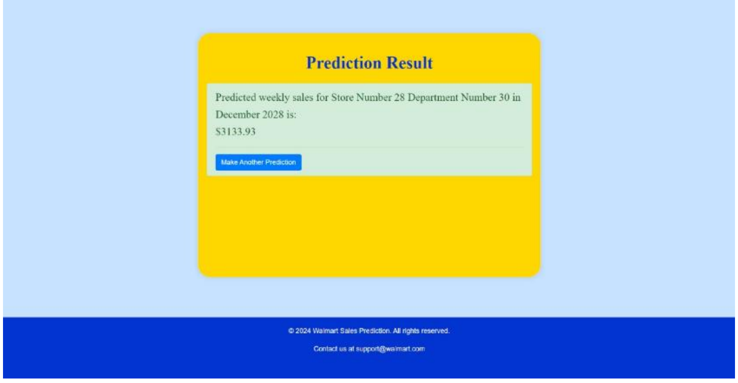
Google Drive Link

Introduction

Forecasting sales is critical for retail businesses like Walmart, enabling informed decisions on inventory, marketing, and overall business strategy. This project focuses on using historical sales data from 45 Walmart stores to predict future sales and analyze the impact of key holidays like Christmas, Thanksgiving, Super Bowl, and Labor Day. Employing machine learning algorithms such as Random Forest, Decision Tree, XGBoost, and ARIMA, the project aims to uncover sales patterns and provide actionable insights for Walmart. These models will be integrated into a Flask application and empowering Walmart with data-driven decisions to optimize operations and profitability.

1.1 Project Overview

The Walmart sales forecasting project focuses on predicting future sales and assessing the impact of major holidays (Christmas, Thanksgiving, Super Bowl, and Labor Day) on store sales. It encompasses data preparation, feature engineering, and model selection using algorithms like Random Forest, Decision Tree, XGBoost, and ARIMA. Following model training and validation, the top-performing model will be integrated into a Flask application. The project aims to offer Walmart insights into sales trends and actionable recommendations to optimize sales strategies during holidays, facilitating informed decision-making and maximizing profitability.

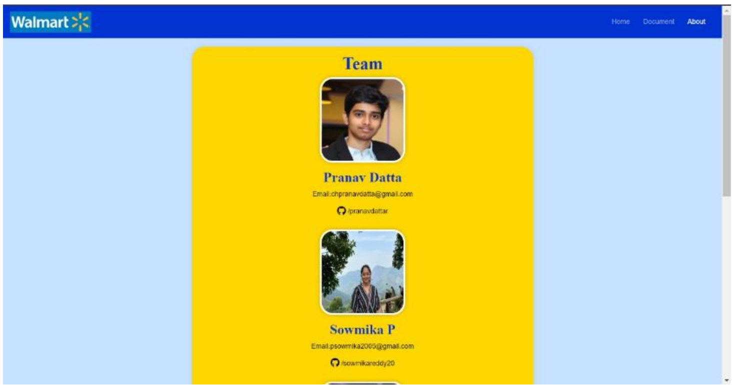

Prediction Result

Predicted weekly sales for Store Number 28 Department Number 30 in December 2028 is:

$3133.93

© 2024 Walmart Sales Prediction. All rights reserved.

Contact us at support@walmart.com



Team

Pranav Datta

Email:chpranavdatta@gmail.com

/pranavdattar

Sowmika P

Email:psowmika2005@gmail.com

/sowmikareddy20

- **Real-Time Sales Forecasts:** Accurate predictions for various departments and stores.
- **Holiday Impact Visualization:** Charts showing sales trends during major holidays.
- **User Interface:** Accessible and intuitive interface for stakeholders.

7. Advantages & Disadvantages

- **Advantages:**
  - **Improved Sales Forecasting:** Accurate predictions help Walmart manage inventory and meet customer demand.
  - **Enhanced Inventory Management:** Reliable forecasts reduce excess inventory and stockouts.
  - **Effective Promotional Strategies:** Analyzes the impact of markdowns and holidays for better promotional planning.
  - **Targeted Customer Segmentation:** Offers insights for personalized marketing efforts.
  - **Optimized Pricing Strategies:** Helps develop strategies to set optimal prices.
  - **Data-Driven Decision Making:** Supports a data-driven approach for retail management.
  - **User-Friendly Application:** Delivers accessible and actionable sales models.
- **Disadvantages:**
  - **Data Limitations:** Historical sales data may not fully capture future trends or unexpected events.
  - **Complex Model Development:** Developing and tuning multiple machine learning models can be complex and time-consuming.
  - **High Weight on Holidays:** Emphasis on holidays might overshadow other important factors influencing sales.
  - **Resource Intensive:** Requires significant computational resources for model training and application development.
  - **Integration Challenges:** Integrating machine learning models into a Flask application may present technical challenges.
  - **Scalability Issues:** Ensuring the application scales effectively can be challenging.
  - **Ongoing Maintenance:** Requires regular updates and maintenance to stay relevant.

8. Conclusion

The conclusion summarizes the project's achievements and highlights the benefits of using machine learning for sales forecasting and holiday impact analysis. Key points include:

1. **Project Success:** The project successfully developed accurate sales forecasting models for Walmart.
2. **Actionable Insights:** Provides Walmart with valuable insights into sales patterns and holiday impacts.
3. **Enhanced Decision-Making:** Supports data-driven decision-making for inventory management, pricing, and promotions.
4. **Future Enhancements:** Outlines potential future enhancements to improve forecasting accuracy and expand the analysis.

9. Future Scope

The future scope outlines potential enhancements and extensions to the project:

1. **Additional Data Sources:** Incorporate economic indicators, weather patterns, and regional events for a comprehensive analysis.
2. **Advanced Models:** Explore advanced machine learning models such as deep learning and ensemble methods.
3. **Real-Time Processing:** Implement real-time data processing and forecasting for quick responses to market changes.
4. **Hyperparameter Tuning:** Further refine hyperparameter tuning to enhance model accuracy and robustness.
5. **Geographical Analysis:** Conduct geographical and store-level analysis for targeted insights.
6. **Seasonal Trends:** Analyze seasonal and long-term trends to guide strategic planning.
7. **Improved User Experience:** Enhance the user experience and visualization capabilities of the Flask application.
8. **Business Integration:** Integrate the forecasting system with existing business processes for efficient operations.
9. **Causal Relationships:** Explore causal relationships between sales, promotions, and holidays for deeper insights.

10. Appendix

10.1. Source Code

Link to source code:
https://drive.google.com/drive/folders/1RGCmtFKDg69nVBMij7Ghimsu3nPUKn8r

10.2. GitHub & Project Demo Link

GitHub link: https://github.com/sowmikareddy20/walmart_sales_analysis/tree/main

Project Demo Link:

https://drive.google.com/drive/folders/1uMy2L4prDtKei4ojC9CExdbu5TOIyqk4?usp=drive_link