Project Documentation: Walmart Sales Analysis

Table of Contents

1. Introduction

Walmart is one of the largest retail corporations in the world. Analyzing its sales data can provide insights into sales patterns, product performance, and factors influencing sales. This project focuses on analyzing Walmart's sales data to understand these aspects and make data-driven decisions.

2. Objective

The main objectives of this project are:
- To analyze the sales data and identify key trends and patterns.
- To forecast future sales using machine learning models.
- To provide actionable insights and recommendations for improving sales performance.

3. Data Description

The dataset used in this project includes Walmart sales data from various stores across different regions. The primary variables in the dataset are:
- Store: The store number.
- Date: The date of the sales record.
- Weekly_Sales: The sales amount for the given week.
- Holiday_Flag: Indicates whether the week is a special holiday week (1) or not (0).
- Temperature: The average temperature in the region.
- Fuel_Price: The cost of fuel in the region.
- CPI: The consumer price index.
- Unemployment: The unemployment rate.

4. Data Preprocessing

Data preprocessing steps included:
- Handling missing values.
- Converting data types as necessary.
- Feature engineering, such as extracting month and year from the date.
- Normalizing or scaling numerical features.
- Encoding categorical variables.

5. Exploratory Data Analysis (EDA)

EDA involves visualizing and summarizing the data to uncover patterns and insights:
- *Sales Trends*: Analyzing weekly, monthly, and yearly sales trends.
- *Holiday Impact*: Assessing the impact of holidays on sales.
- *Store Performance*: Comparing sales performance across different stores.
- *Correlation Analysis*: Evaluating the relationship between sales and other variables such as temperature, fuel price, CPI, and unemployment.

6. Model Development

Several machine learning models were developed to forecast sales:
- *Linear Regression*: A baseline model to predict weekly sales.
- *Random Forest*: An ensemble method to capture non-linear relationships.
- *XGBoost*: A boosting algorithm to improve prediction accuracy.
- *ARIMA*: A time-series model to forecast future sales based on historical data.

Model Evaluation

The models were evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Cross-validation was used to ensure the robustness of the models.

7. Results and Discussion

The results showed that:
- Sales typically increase during holiday weeks.
- Certain stores consistently outperform others.
- Temperature and fuel prices have a noticeable impact on sales.
- The XGBoost model provided the most accurate sales forecasts, followed by Random Forest.

8. Conclusion

This project successfully analyzed Walmart's sales data and provided valuable insights. The XGBoost model was identified as the best predictor for future sales. Recommendations for improving sales include focusing on high-performing stores and

optimizing stock levels during holidays.

9. References

- Kaggle Dataset: [Walmart Sales Data](https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data)
- Scikit-learn Documentation: [Scikit-learn](https://scikit-learn.org/stable/)
- XGBoost Documentation: [XGBoost](https://xgboost.readthedocs.io/en/latest/)
- ARIMA Model: [Statsmodels](https://www.statsmodels.org/stable/tsa.html)

This documentation provides a comprehensive overview of the Walmart sales analysis project, detailing each step from data collection and preprocessing to model development and results interpretation.