Walmart Sales Forecasting

Final Report: Walmart Sales Forecasting

Business Problem

Walmart faces challenges in accurately forecasting sales, which impacts inventory management, staffing, and promotional strategies. Overstocking leads to increased costs while understocking results in missed revenue opportunities. Additionally, external factors like holidays, inflation, and unemployment further complicate predictions. The objective of this project is to develop models that enhance forecasting accuracy, enabling Walmart to make informed decisions.

Data Analysis

The dataset includes weekly sales data from Walmart stores, with features such as:

- **Weekly Sales** (target variable),
- **CPI**, **Unemployment**, and **Fuel Price** (economic predictors),
- **Holiday Indicator** (to distinguish holiday weeks).

Preprocessing Steps:

- Converted the date column into **Year**, **Month**, and **Week** to incorporate temporal trends.
- Created lagged features (**Lag_1**, **Lag_2**) and rolling averages (**Rolling_4**) to capture past sales behavior.
- Verified no missing values, ensuring dataset completeness.

Exploratory Data Analysis:

- Sales exhibit strong seasonality, with spikes during holidays.
- A boxplot revealed higher sales during holiday weeks compared to non-holiday weeks.
- Correlation analysis showed that **CPI** and **Unemployment** negatively correlate with sales, highlighting the economic impact on consumer behaviour.

Modelling Approach

Four models were implemented to address the forecasting problem:

1. **Linear Regression**:
    - Baseline model used to understand relationships between predictors and sales.
    - Holidays increase sales by approximately 80,000 units, while CPI and unemployment negatively impact sales.
    - Performance: RMSE = 557,239, with a low R-squared value of 0.025, indicating limited predictive power.

2. **Decision Tree**:
    - Captures non-linear interactions between predictors.
    - Key splits include **Unemployment ≥ 9** and **CPI ≥ 189**, which reflect the impact of economic conditions on sales.
    - Performance: RMSE = 537,942, with improved interpretability but limited performance at extreme values.
3. **ARIMA**:
    - Time-series model that accounts for seasonality and trends.
    - Best performance among all models: RMSE = 51,245, MAPE = 2.02%.
    - Ideal for short-term operational planning, especially during holidays.
4. **XGBoost**:
    - Advanced machine learning model leveraging lagged and rolling features.
    - Identified holidays and CPI as significant predictors.
    - Performance: RMSE = 98,151, slightly higher than ARIMA but offering valuable feature insights.

Challenges Encountered

1. **High RMSE Values**:
    - Due to the scale of weekly sales (often exceeding millions), RMSE values appeared large. This was mitigated by also using MAPE for relative error measurement.
2. **Trade-off Between Interpretability and Accuracy**:
    - While Decision Tree and ARIMA provided clearer insights, XGBoost offered superior flexibility but required more feature engineering.
3. **Handling Seasonality**:
    - Capturing holiday effects required creating new features like lagged sales and rolling averages to account for temporal trends.

Findings

- ARIMA provided the best short-term forecasts, particularly for holiday-driven sales spikes.
- XGBoost identified key drivers like holidays and CPI, aiding in long-term planning.
- Linear Regression served as a baseline but struggled with complex relationships.
- Decision Tree highlighted actionable thresholds for CPI and Unemployment, guiding business strategies.

Conclusion and Impact

This project demonstrates that a hybrid approach combining ARIMA and XGBoost can significantly improve Walmart's forecasting accuracy. Key takeaways include:

- **Short-Term Planning**: ARIMA effectively handles seasonal trends for operational adjustments.
- **Strategic Insights**: XGBoost identifies critical drivers, enabling targeted promotional strategies.
- **Operational Efficiency**: These models can reduce costs by optimizing inventory and staffing while maximizing revenue during peak periods.

By implementing these forecasting techniques, Walmart can adopt a more dynamic and data-driven approach to decision-making, ensuring resilience in a competitive retail environment.