Mid-Stage Report for Walmart Weekly Sales Forecasting

1st Vikash Sinha
Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030, USA
vsinha3@stevens.edu

2nd Carlos Iturralde
Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030, USA
itucarl96@outlook.com

3rd Olaoluwa Olasanoye
Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030, USA
Olaoluolasanoye@gmail.com

***Abstract*—This project aims to develop a predictive model for Demand Forecasting to Optimize Supply Chain Management. By leveraging multiple regression models, including Linear Regression, Decision Trees, Random Forest, and ARIMA, we aim to predict future sales and demand for WalMart Stores. The project may help to provide insights into Consumer Demand and identify patterns in Sales trends, giving Stores the ability to optimize their supply chain.**

I. Introduction

The primary goal of this project is to accurately forecast Walmart's weekly store sales. This capability is pivotal for efficient supply chain management, aiding in decisions regarding stocking levels to prevent overstocking or understocking. The project leverages multiple machine learning (ML) and time series models to predict sales at individual stores across different locations.

Motivation: Accurate sales forecasting is crucial for operational efficiency in retail. This study aims to address variability due to seasonal changes, holidays, and other market dynamics.

Methodology: We explore regression and time series techniques, including ARIMA and SARIMA models, for predicting sales trends. The process involves preprocessing raw data, analyzing patterns, selecting optimal features, and evaluating various models' performances. This also includes experimenting with LSTM networks to capture complex temporal dependencies in sales data.

II. Related Work

Regression Models: Studies have employed linear regression and tree-based models like Random Forest to predict retail sales based on external features such as temperature, fuel prices, and holiday effects.

Time Series Models: The ARIMA and SARIMA models are widely used for their ability to capture seasonal trends and short-term dependencies in sales data.

Retail-Specific Techniques: Previous studies highlight the importance of holiday weeks (e.g., Christmas, Thanksgiving) in influencing sales patterns. These studies emphasize incorporating temporal and exogenous variables to improve model accuracy.

Deep Learning Techniques: Advanced models like LSTMs (Long Short-Term Memory) are increasingly popular for time series forecasting, capable of learning long-term dependencies. However, their application requires large datasets and computational resources.

Key Findings:

- Regression models are effective for structured, tabular data.
- Time series models excel in capturing trends and seasonality but struggle with external feature integration.
- LSTMs can combine the strengths of both approaches by integrating temporal and external features but require substantial computational resources.

III. Our Solution

A. Description of Dataset

Source: Kaggle - Walmart Sales Forecast Dataset.

Overview:

- **Rows**: 6,435 entries (aggregated weekly sales across all stores).
- **Columns**: 12 features, including:
  - `Store`: Unique identifier for each store.
  - `Weekly_Sales`: Target variable representing weekly sales in dollars.
  - `Temperature`: Weekly average temperature.
  - `Fuel_Price`: Weekly fuel price.
  - `CPI` and `Unemployment`: Indicators of economic conditions.
  - `Holiday_Flag`: Binary indicator for major holidays (1 = holiday week).
  - including:
  - **Features**: Store ID, Weekly Sales, Temperature, CPI, Unemployment, Holiday Indicator, etc.
  - **Target**: `Weekly_Sales`.
- **Key Characteristics**:
  - Four holiday weeks significantly impact sales: Christmas, Thanksgiving, Super Bowl, and Labor Day.
  - Clean dataset with no missing values.

Visualizations:

- **Heatmap of Correlations**: Unemployment, CPI, and Temperature exhibit low correlation with sales. Store ID has the highest importance in predictions.
- **Boxplot**: Sales exhibit high variability across stores.
- **Distribution Plot**: Weekly sales are right-skewed, indicating some stores or weeks experience significantly higher sales.

Insights from the Data:

1) Sales vary significantly by store, with Store ID being the most influential feature.
2) Holiday weeks exhibit noticeable sales spikes, especially around Thanksgiving and Christmas.
3) Features like `CPI` and `Temperature` have weak correlations with sales but are retained for completeness.

Preprocessing Steps

1) **Feature Engineering**:
   - Converted `Day of the Week` to numeric format for model compatibility.
   - Extracted year information from the date.
2) **Scaling**: Applied MinMax scaling to normalize numerical features.
3) **Feature Selection**:
   - Top features identified: `Store`, `CPI`, `Unemployment`, `Month`, and `Temperature`.
   - `Store` emerged as the most influential feature with a relative importance of 66.1%.

B. Machine Learning Algorithms

Linear Regression:

- Simple and interpretable baseline model.
- RMSE for full dataset: 510,119.
- Low R2R$\hat{2}$R2 value (0.14) indicates limited capacity to capture sales variability.

Decision Tree:

- Captures non-linear relationships effectively.
- RMSE for full dataset: 175,058.
- R2R$\hat{2}$R2: 0.90, demonstrating strong performance with structured data.

Random Forest:

- Ensemble method offering robustness to overfitting.
- RMSE for full dataset: 138,479.
- R2R$\hat{2}$R2: 0.94, making it the best regression model in this context.

ARIMA and SARIMA:

- Time series models to capture trends, seasonality, and residual patterns.
- SARIMA is particularly suited for handling periodic seasonality (e.g., annual holiday effects).

LSTM Network:

- Utilizes sequential patterns in data for better temporal modeling.
- Implemented using TensorFlow and Keras.
- Architecture: Three LSTM layers, one dense layer, dropout for regularization.
- Input Sequence: Weekly sales data and temporal features.

C. Implementation Details

The project implementation is divided into stages, with preliminary results obtained for Decision Tree and ARIMA models.

1. Decision Tree Regression:

- **Dataset**: Full dataset with all features included.
- **Results**:
  - RMSE: 175,058
  - R2R$\hat{2}$R2: 0.90
- **Key Insight**: Decision Tree performs well on the full dataset, capturing non-linear relationships effectively.

2. ARIMA for Time Series Analysis

- **Store**: Forecasting applied to Store 20.
- **Train-Test Split**: 80% training, 20% testing.
- **Performance**:
  - RMSE: 178,150.
  - R2R$\hat{2}$R2: -1.93 (indicating poor fit for unseen data).
- **Insights**: ARIMA captured short-term dependencies but struggled with seasonality.

3. SARIMA for Seasonal Adjustments

Performance:

- RMSE: 408,570.
- R2R$\hat{2}$R2: -14.41 (worse than ARIMA, indicating overfitting or inadequate seasonal adjustment).

4. Future Forecasting

Extended ARIMA to forecast 12 weeks of sales: Predicted values showed consistency with past trends but failed to capture seasonal spikes due to holiday effects.

5. LSTM Preliminary Results:

- RMSE: 162,000
- R2R$\hat{2}$R2: 0.85
- **Insights**: LSTM captured sequential dependencies but requires further hyperparameter tuning to handle noise and external features.

IV. Comparison

Best Model: Decision Tree Regression with R2R$\hat{2}$R2: 0.90 and RMSE: 175,058.

Time Series: ARIMA outperformed SARIMA for Store 20 but requires fine-tuning for scalability across all stores.

LSTM networks: LTSM show promise for sequential data but require optimization.

V. Future Directions

Fine-tune LSTM architecture to integrate external features like `Holiday_Flag` and `Temperature`.

Incorporate external features like holiday-specific promotions and competitor pricing.

Optimize SARIMA seasonal parameters to align better with sales patterns. .

Explore hybrid models combining LSTM with SARIMA for robust seasonal and temporal forecasting.

Investigate external data sources, such as competitor pricing, to enhance feature richness.

VI. Conclusion

Random Forest Regression emerged as the best-performing regression model with R2R$\hat{2}$R2: 0.94.

LSTM networks show promise for sequential data but require optimization.

ARIMA and SARIMA models struggled with the complexity of Walmart's sales data.

References

Kaggle - Walmart Sales Forecast Dataset.