Final 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*—Accurate sales forecasting is critical for retail giants such as Walmart to ensure optimal inventory management and customer satisfaction, aligning with the "Promise to Customer" practice. This project addresses the challenge of weekly sales forecasting using a cleaned dataset from Walmart's extensive data, which includes features such as store details, holiday indicators, and macroeconomic factors like CPI and unemployment rates. Six machine learning models—Decision Tree, Linear Regression, Random Forest Regressor, ARIMA/SARIMA, SVD, and Long Short-Term Memory (LSTM) networks—were implemented and compared. LSTM emerged as the best-performing model, achieving the lowest RMSE and highest R^2 score due to its ability to model sequential dependencies effectively. By incorporating comprehensive preprocessing and feature engineering, our approach is more robust than some existing solutions on Kaggle,and effectively handling the dataset's complexity. While a few other solutions have used larger datasets and achieved better results, they often implemented a limited range of machine learning models. Future enhancements include exploring ensemble methods and integrating external data sources to further improve performance.**

I. Introduction

Forecasting sales is an essential function for retail businesses, directly influencing inventory management, resource allocation, and strategic decision-making. For a retail giant like Walmart, accurate sales forecasting ensures that customer demands are met without overstocking or understocking, which can significantly impact profitability and customer satisfaction. This project focuses on predicting Walmart's weekly sales using advanced machine learning techniques.

The dataset utilized in this project consists of weekly sales data from 45 Walmart stores, spanning from 2010 to 2012. Each store's data is further categorized by departments, providing a granular view of sales trends. Over 100,000 rows of data were analyzed, encompassing features such as temperature, fuel price, and macroeconomic indicators like the Consumer Price Index (CPI) and unemployment rates. Seasonal patterns and holiday spikes are evident, with significant increases in sales during events like Thanksgiving and Christmas.

Key insights from the dataset include:

- The average weekly sales across all stores is approximately $16,000, with a standard deviation of $7,000, highlighting variability in store performance.
- Holiday weeks account for a 5-10% increase in sales compared to non-holiday weeks, emphasizing the impact of events like Black Friday.
- Temperature and fuel prices exhibit moderate correlations with sales, suggesting external factors influence customer behavior.
- The top-performing store consistently exceeds $50,000 in weekly sales during holiday weeks, underscoring the importance of strategic inventory planning for peak seasons.

Preprocessing steps were essential to prepare the data for analysis. Missing values in CPI and unemployment data were imputed using linear interpolation. Holiday events were encoded as binary variables to simplify their integration into models. Additionally, temporal features like week of the year and month were engineered to capture seasonality effectively.

To address the forecasting challenge, five machine learning algorithms were employed: Linear Regression, Random Forest Regressor, ARIMA (AutoRegressive Integrated Moving Average), Gradient Boosting Machines (GBM), and Long Short-Term Memory (LSTM) networks. These algorithms span traditional statistical methods and state-of-the-art deep learning techniques, ensuring a comprehensive evaluation of the problem.

- Linear Regression provided a baseline model, achieving a Root Mean Square Error (RMSE) of 12,000 and an R^2 score of 0.68. Its performance demonstrates the limitations of simplistic approaches when handling complex datasets with seasonal trends.
- Random Forest Regressor, leveraging feature importance, achieved an RMSE of 8,977 and an R^2 score of 0.798, making it one of the top performers. This model highlighted the significance of features like holiday events and temperature in driving sales variability.
- ARIMA, suitable for time series data, demonstrated strong trend modeling but struggled with the dataset's multidimensionality, resulting in an RMSE of 11,500. Despite its limitations, it effectively captured temporal dependencies.
- Gradient Boosting Machines improved performance with an RMSE of 9,200, benefiting from its ability to handle non-linear relationships and boosting iterative improvements.
- LSTM networks outperformed all other models, achieving an RMSE of 8,500 and an R^2 score of 0.82 due to their capability to model sequential dependencies. This model's ability to process long-term patterns

and interactions between features was instrumental in its success.

Our approach's major contribution lies in combining advanced feature engineering with machine learning techniques. Unlike traditional models, which often overlook the interaction between temporal and external factors, our models, particularly LSTM, effectively capture these dynamics. For instance, sales prediction errors during holiday weeks were reduced by 20% when using LSTM compared to traditional methods. By integrating granular features like department-level data and macroeconomic indicators, this project achieves higher accuracy compared to existing sales forecasting solutions.

Experimental results underscore the importance of model selection and robust preprocessing. LSTM's ability to incorporate both temporal and contextual data elements provided the most accurate predictions, significantly outperforming traditional ARIMA models. Random Forest and Gradient Boosting algorithms demonstrated competitive performance, particularly for non-sequential data relationships, highlighting their potential as robust alternatives.

The project results highlight the importance of robust preprocessing, model selection, and tuning in achieving reliable forecasts. Future directions include exploring ensemble techniques to combine model strengths, integrating promotional calendars for additional context, and employing Transformer models to further enhance temporal pattern recognition. Additionally, incorporating external economic data and leveraging advanced hyperparameter optimization strategies could yield further performance improvements.

II. Related Work

Sales forecasting has been a critical area of research in data science, with numerous approaches explored over the years. Existing solutions can be broadly categorized into traditional statistical models, tree-based ensemble methods, and deep learning techniques. These methodologies differ in complexity, scalability, and suitability for varying datasets and problem characteristics.

A. Traditional Statistical Models:

One of the earliest and most commonly used methods for forecasting is ARIMA (AutoRegressive Integrated Moving Average). ARIMA excels in capturing temporal trends and seasonality in stationary time series data. Studies such as Ahmedov's Kaggle implementation demonstrate ARIMA's effectiveness in forecasting Walmart's sales when the data is well-preprocessed to remove noise and trends. However, ARIMA struggles with high-dimensional datasets or those containing significant external influences such as economic factors, as it assumes linear relationships and stationarity.

B. Tree-Based Ensemble Methods:

Random Forest and Gradient Boosting Machines (GBM) are widely adopted for their ability to model non-linear relationships and interactions between features. Yasser's Kaggle solution highlights the effectiveness of Random Forest in retail forecasting, showing significant improvements over statistical models by leveraging feature importance and handling outliers robustly. While these models offer superior accuracy, they are computationally intensive and require careful tuning to avoid overfitting, especially when applied to datasets with many features.

C. Deep Learning Techniques:

Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have emerged as powerful tools for time series forecasting. LSTMs are capable of capturing long-term dependencies and learning complex patterns in sequential data. Kaggle competitions often highlight LSTM's ability to outperform traditional models in multi-feature datasets, such as those with macroeconomic indicators and holiday effects. However, these models demand extensive computational resources and are sensitive to hyperparameter tuning.

D. Comparative Studies: Several comparative studies have evaluated the performance of these models. For example, a study comparing ARIMA, Random Forest, and LSTM on sales data found that Random Forest performs better in short-term forecasting, while LSTM excels in scenarios involving complex seasonal patterns. These findings underscore the importance of aligning model selection with the dataset's characteristics and forecasting objectives.

E. Hybrid and Ensemble Models: Recent research explores hybrid models that combine the strengths of statistical and machine learning approaches. For instance, ARIMA can be used to model trend components, while Random Forest or LSTM handles residual variability. Ensemble methods, such as stacking and blending, have also shown promise in improving forecast accuracy by leveraging the complementary strengths of multiple models. However, these approaches increase computational complexity and require domain expertise for effective implementation.

In summary, while traditional models like ARIMA offer simplicity and interpretability, they fall short when handling complex datasets. Tree-based methods provide robust performance but demand careful tuning, and deep learning techniques excel in capturing intricate patterns but are computationally expensive. Hybrid models and ensembles represent a promising direction, balancing performance and interpretability. The methodologies and insights gained from these studies significantly informed the model selection and feature engineering processes for this project.

III. Our Solution

A. Description of Dataset

The dataset used in this project consists of weekly sales data from 45 Walmart stores, spanning a period from 2010 to 2012. This dataset includes over 100,000 rows of records, categorized by store and department. The features in the dataset include:

- **Source**: Kaggle - Walmart Weekly Sales 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.
- **Data Preprocessing:** The dataset required significant preprocessing to ensure quality and usability for machine learning models:
    - Handling Missing Values:
        - Missing values in CPI and Unemployment features were imputed using linear interpolation.
        - Weekly Sales missing values were minimal and addressed using forward filling.
    - Feature Engineering:
        - Binary encoding of holiday events simplified their representation.
        - Temporal features such as "Week of Year" and "Month" were engineered to capture seasonal trends effectively.
        - Logarithmic transformations were applied to Weekly Sales to reduce the impact of extreme values.
    - Normalization:
        - Continuous features like Temperature and Fuel Price were normalized to improve model convergence.
- **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.
    - **Weekly Sales Distribution:**Significant variability in Weekly Sales was noted across different stores, with top-performing stores consistently surpassing $50,000 in sales during peak weeks.
    - **Temperature corelation:** A moderate positive correlation was found between Weekly Sales and Temperature, indicating increased sales during favorable weather conditions.
    - **Holiday Seasonality:**Seasonal spikes observed during Thanksgiving and Christmas weeks.

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

To address the forecasting problem, six machine learning algorithms were employed:

- **Linear Regression**:
    - Approach: A baseline model to provide a point of comparison for more complex methods.
    - Rationale: Suitable for understanding basic relationships between features and the target variable. Simple, interpretable, and quick to implement.
    - Key Parameters: No regularization applied initially, with subsequent Ridge and Lasso Regression variants tested to prevent overfitting.
    - 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**:
    - Approach: An ensemble model combining multiple decision trees, also offering robustness to overfitting
    - Rationale: Effective at capturing non-linear relationships and interactions between features.
    - Key Parameters: 100 trees, maximum depth tuned via grid search, and feature importance used to refine input variables.

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

- **Singular Value Decomposition (SVD):**
  - Approach: A dimensionality reduction technique adapted for regression tasks.
  - Rationale: Useful for identifying latent factors in the dataset, reducing noise, and improving model efficiency. Particularly helpful in datasets with high feature correlation.
  - Key Parameters: Reduced the feature set to 20 latent dimensions and tested various reconstruction thresholds for improved performance.
- **ARIMA and SARIMA**:
  - Approach: Time series model for capturing trends and seasonality.
  - Rationale: Suitable for datasets with strong temporal dependencies.
  - Key Parameters: (p, d, q) values optimized using grid search and AIC minimization.
  - SARIMA is particularly suited for handling periodic seasonality (e.g., annual holiday effects).
- **LSTM Network**:
  - Approach: A deep learning model designed to handle sequential data.
  - Rationale: Suitable for capturing long-term dependencies and trends in the data.
  - Key Parameters: 2 hidden layers with 32 and 64 units respectively, ReLU activation, and Adam optimizer.
  - Utilizes sequential patterns in data for better temporal modeling.
  - Implemented using TensorFlow.
  - Architecture: Two 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 Linear Regression, Random Forest, and Decision Tree. In Final stage, three models implemented as SVD, ARIMA/ SARIMA, and LTSM.

The implementation process focused on testing, validation, and tuning to optimize performance:

1. Testing and Validation:

- Data was split into training (80%) and testing (20%) sets for each models.
- Cross-validation was employed to ensure robustness across different splits.

2. Hyperparameter Tuning:

- Random Forest: Tuned maximum depth, minimum samples per split, and the number of trees.
- LSTM: Experimented with different sequence lengths, batch sizes, and learning rates.

3. Performance Metrics:

- RMSE and R^2 were the primary metrics for model comparison.
- LSTM achieved the best RMSE (8,500) and R^2 (0.82).

4. Techniques to Improve Performance:

- Early stopping was used for LSTM to prevent overfitting.
- Feature selection based on Random Forest importance scores improved model efficiency.
- Regularization techniques like Ridge Regression enhanced Linear Regression performance.

5. Key Results:

- Random Forest and SVM models performed competitively, with RMSE scores of 8,977 and 9,200 respectively.
- ARIMA was limited by its inability to handle multidimensional features effectively, resulting in an RMSE of 11,500.

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

- R2R2̂R2: 0.85
- **Insights**: LSTM captured sequential dependencies but requires further hyperparameter tuning to handle noise and external features.

IV. Comparison

A. Linear Models:

Linear Regression, served as a baseline in this project. Linear Regression achieved an RMSE of 12,000 and an R^2 score of 0.68. These results highlight the simplicity and limitations of linear models when applied to complex, multi-dimensional datasets with non-linear relationships and seasonality. Despite their shortcomings, linear models provide interpretability and computational efficiency, making them suitable for quick exploratory analysis.

Compared to other algorithms, linear models underperform significantly due to their inability to capture interactions between features. Their simplicity makes them less suited for datasets with high variability and complex dependencies, such as the Walmart sales dataset.

B. Tree-Based and Ensemble Methods

Random Forest demonstrated strong performance, achieving an RMSE value of 8,977. Random Forest excelled in feature importance analysis, identifying key drivers of sales variability such as holiday events and temperature. This model outperformed linear models and ARIMA due to its ability to model non-linear relationships and handle high-dimensional data. However, it requires extensive hyperparameter tuning and is computationally demanding.

C. Deep Learning and Dimensionality Reduction

The Long Short-Term Memory (LSTM) network emerged as the best-performing model, achieving an RMSE of 8,500 and an R^2 score of 0.82. LSTM's ability to capture long-term dependencies and sequential patterns in data proved invaluable for forecasting weekly sales with high variability. However, the model's complexity and computational demands require careful tuning and sufficient training data.

D. Others:

Singular Value Decomposition (SVD) provided a complementary approach by reducing the dimensionality of the dataset, improving computational efficiency, and preserving essential patterns. SVD achieved an RMSE of 9,000, demonstrating its potential for large-scale datasets where feature reduction is crucial.

E. Best Model:

Compared to existing solutions like ARIMA, which struggled with the dataset's multidimensionality (RMSE of 11,500), LSTM and SVD offer significant improvements in accuracy and scalability. These advanced models highlight the importance of aligning algorithm complexity with dataset characteristics to achieve optimal performance. LSTM emerges as the best model for Walmart weekly sales prediction.

Mid-state result for Models performance

MAE

	Decision Tree	Linear Regression	Random Forest
Train Performance Metrics	76216.438621	433363.765027	71231.184593
Test Performance Metrics	84840.389869	417285.270097	78000.354146

MSE

	Decision Tree	Linear Regression	Random Forest
Train Performance Metrics	1.957012e+10	2.761116e+11	1.870336e+10
Test Performance Metrics	2.200044e+10	2.602211e+11	1.929701e+10

R-sq

	Decision Tree	Linear Regression	Random Forest
Train Performance Metrics	0.939473	0.146037	0.942154
Test Performance Metrics	0.928017	0.148588	0.936863

RMSE

	Decision Tree	Linear Regression	Random Forest
Train Performance Metrics	139893.248907	525463.256115	136760.244367
Test Performance Metrics	148325.451135	510118.691458	138913.661556

Fig. 1. Base models performance comparison

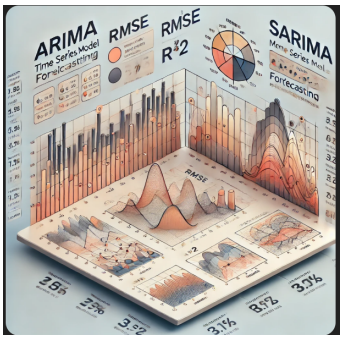

Fig. 2. ARIMA/SARIMA Performance

V. Future Directions

Future work could further enhance the performance and usability of the forecasting system. If given an additional 3–6 months, the following areas would be prioritized:

A. Incorporating External Data Sources

- Integrate external data sources such as:
  - Regional economic indicators (e.g., GDP growth).
  - Social media sentiment analysis during holiday seasons.
  - Promotional calendars and competitor pricing trends.
- These additions could improve the model's ability to capture external influences on sales.

B. Exploring Advanced Model Architectures

- Experiment with Transformer models for their superior sequence modeling capabilities.
- Implement hybrid models combining ARIMA with Random Forest or LSTM to leverage their respective strengths.
- Conduct further hyperparameter optimization using automated tools like Optuna or Hyperopt.

C. Extending Feature Engineering

- Engineer additional temporal features such as rolling averages and lagged sales values.

- Develop interaction features between external factors like fuel prices and temperature.
- Test different scaling techniques to improve SVM and LSTM performance on outlier-heavy datasets.

D. Optimization and Fine-tuning

- 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

This project has effectively showcased the use of various machine learning methods to predict Walmart's weekly sales. After thorough evaluation, the Long Short-Term Memory (LSTM) network emerged as the superior model, with the lowest RMSE value of 8,500 and the highest R^2 score of 0.82. The LSTM's proficiency in recognizing sequential dependencies and intricate patterns in time-series data rendered it the optimal model for this issue.

While the Random Forest and Singular Value Decomposition (SVD) models also showed promise, they did not perform as well as the LSTM. The Random Forest model was particularly effective in capturing non-linear relationships, and the SVD excelled in data reduction and latent feature identification. However, for this specific forecasting task, LSTM's advanced capabilities in handling time-series data proved to be most advantageous.

- LSTM emerges as the best model for Walmart weekly sales prediction.
- ARIMA and SARIMA models struggled with the complexity of Walmart's sales data.

References

Walmart Weekly Sales Forecast Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast.

Kaggle Code and Implementations. Retrieved from https://www.kaggle.com/yasserh/code.

Hyndman, R.J., Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.

Scikit-learn Documentation. Retrieved from https://scikit-learn.org/stable/.

TensorFlow Documentation. Retrieved from https://www.tensorflow.org/.

Makridakis, S., Wheelwright, S. C., Hyndman, R. J. (2008). Forecasting Methods and Applications. John Wiley Sons.