Predicting Retail Sales Success: A Comparative Analysis of Machine Learning, Neural Networks, and Time Series Models.

Submitting to: Liverpool John Moores University

Sai Sharan Paspunuri • Student ID: 1096356 •29th July, 2024

Overview

- Introduction/Background
- Literature Review
- Problem Statement
- Methodology
- Results & Discussion
- Conclusion and future works

The Retail Landscape-A World of Uncertainty

- The retail industry is a dynamic and unpredictable landscape. From seasonal shifts to surprise promotions, retailers face a constant challenge in predicting future demand.

Forecasting: Accurate sales forecasting is essential for navigating this uncertainty and achieving success in retail. Reliable predictions empower retailers to make critical decisions about:

- Inventory Management: Ensuring the right products are available at the right time.
- Pricing Strategies: Optimizing prices to maximize revenue.
- Marketing Campaigns: Targeting the right customers
- Staffing Levels: Adjusting workforce to meet fluctuating demand.

Literature Review: A Foundation of Knowledge

A Historical Perspective:

- Traditional Methods: Retail sales forecasting has evolved from traditional statistical methods like moving averages and exponential smoothing, which are well-suited for stable data patterns.
- Machine Learning Enters the Stage: The emergence of machine learning algorithms, including decision trees, support vector machines, and gradient boosting, has brought more sophisticated methods for handling non-linear relationships in sales data.
- Deep Learning's Promise: Deep learning models, such as LSTMs, CNNs, and recurrent neural networks, are revolutionizing forecasting, particularly for complex time series data and long-term predictions.

Literature Review: A Foundation of Knowledge

Key Research Insights & Unmet Needs:

- Strengths and Limitations: Existing research highlights the strengths and limitations of various methods, including the challenge of interpreting deep learning models and the difficulty of traditional models in capturing complex non-linear patterns.

- Comprehensive Model Comparison: Several studies highlight the need for more comprehensive research on comparing and evaluating different models (statistical, machine learning, and deep learning) in real-world retail contexts.

Problem Statement

Beyond Traditional Approaches:

- Traditional forecasting methods, often relying on simple statistical models, struggle to keep up with the complexities of the modern retail landscape. Retailers are increasingly turning to advanced techniques. They are often clueless to decide on the model.
- Machine Learning: Leveraging algorithms to identify complex patterns in data.
- Deep Learning: Using powerful neural networks to capture intricate relationships and improve accuracy.
- Time Series Models: Analyzing historical sales patterns to predict future trends.
- Several studies highlight the need for more comprehensive research on comparing and evaluating different models (statistical, machine learning, and deep learning) in real-world retail contexts.

Research Goals and Objectives

- Research Aim: This thesis aims to evaluate and select the most effective model for predicting retail sales, considering the complexities of the retail environment.

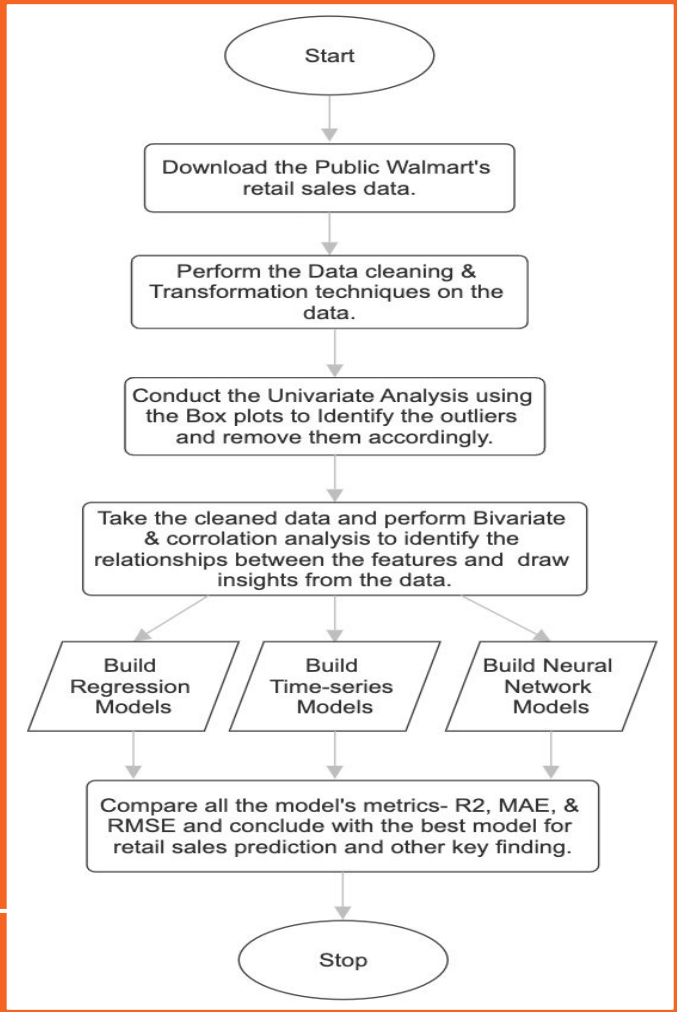
Key Objectives:

- Build and implement various machine learning models (XGBoost, Random Forest, SVM, Linear Regression), time series models (ARIMA, SARIMAX), and deep learning models (ANN, LSTM) using the Walmart retail sales dataset.
- Evaluate and compare the performance of different models using relevant metrics (R-squared, MAE, RMSE), highlighting their strengths and limitations in a retail context.
- Based on the evaluation, provide practical recommendations for retailers on choosing the most effective forecasting models for their forecasting needs.

Research Methodology

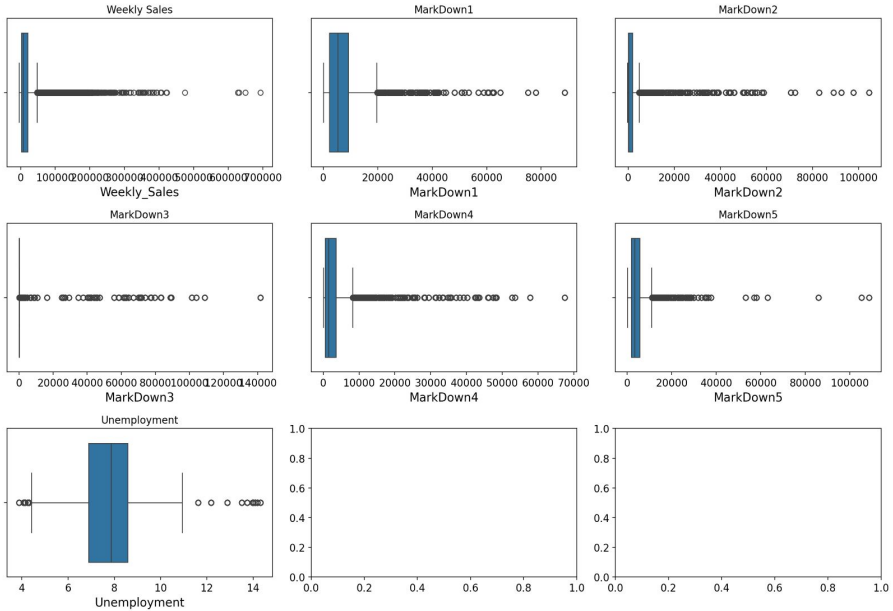
Dataset Overview:

- The dataset, which includes data from 45 stores spanning 65 departments, covering three years (2010-2012).
- Store information (Store ID, Size, Type, Sales)
- Time information (Date, Year, Month, Day of Week, Holiday)
- External factors (CPI, Unemployment, Fuel Price)
- Promotional data (MarkDown1 to 5)

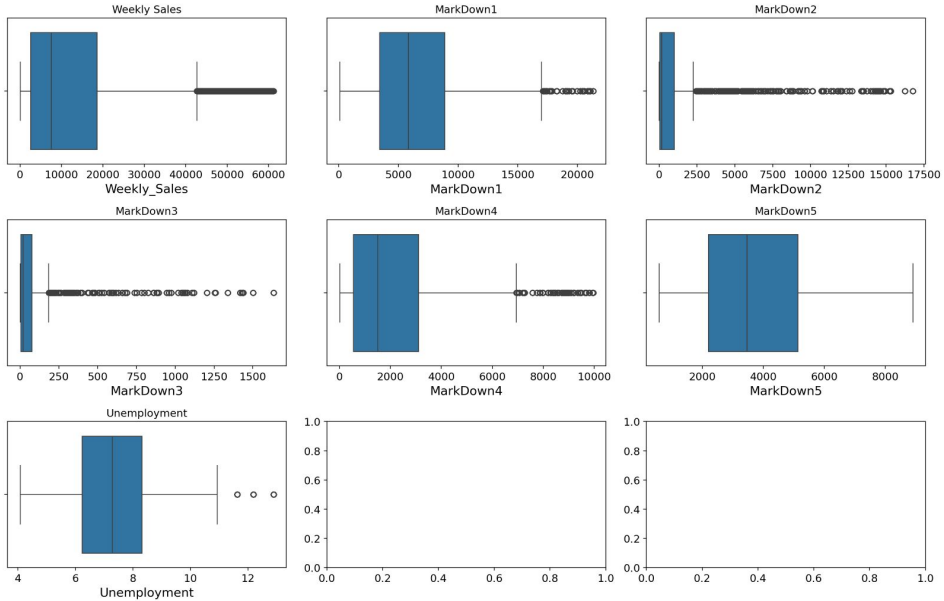


Univariate Analysis Using Boxplots

Before Outlier Treatment:



After Outlier Treatment:



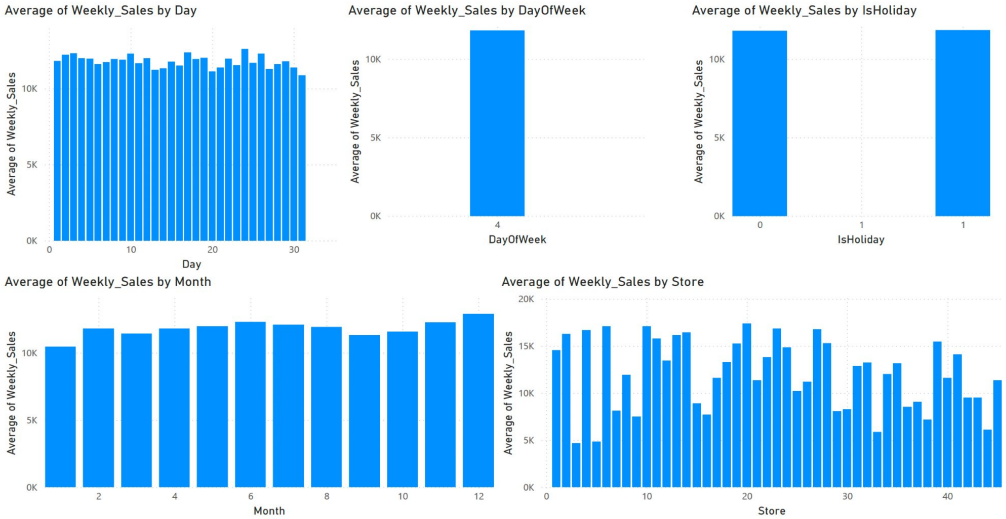
Data Visualization: Bivariate Analysis

Insights:

Day of Month Impact: The relationship between the day of the month and weekly sales revealed no discernible trend. This suggests that sales performance does not appear to be significantly influenced by the specific day of the month.

Holiday Effect: The visualization clearly indicates that holidays do not have a significant impact on sales. This finding suggests that promotional periods and other marketing initiatives might be more influential in driving sales than holidays alone.

December Demand: Sales data indicates a clear peak in sales during December, highlighting the strong impact of holiday shopping on overall retail performance.



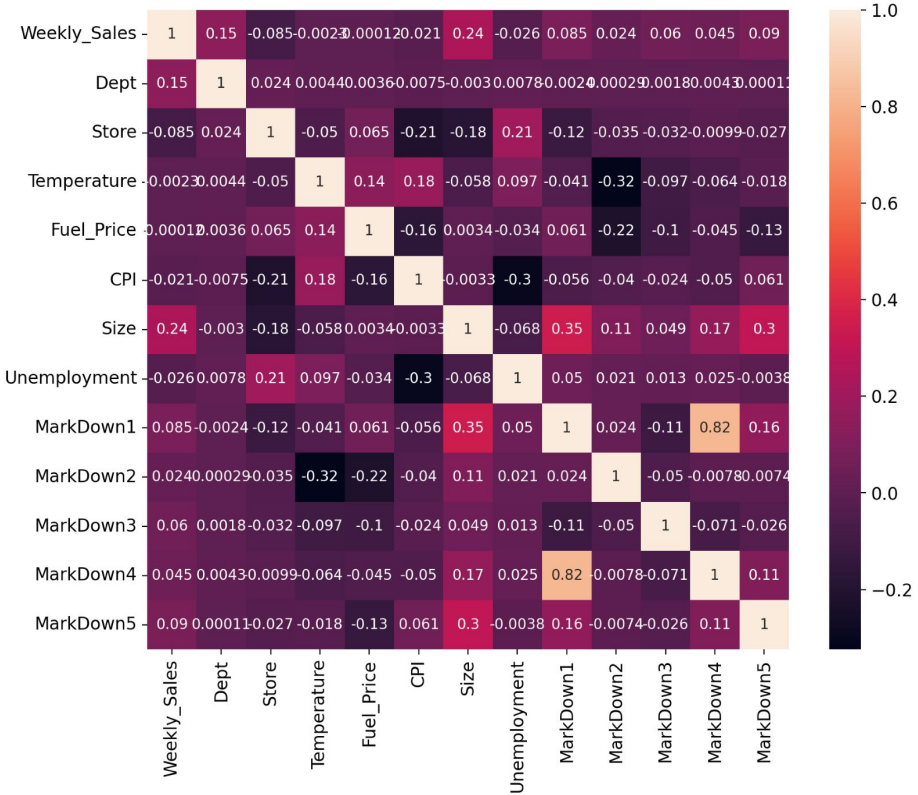
Store-Specific Performance: There is a significant differences in sales performance across different stores. Stores 3, 5, 33, 38, and 44 consistently had the lowest sales, while stores 20 and 10 showed the highest sales. This suggests that store-specific factors, significantly influence sales performance.

Correlation Map

Insights:

Our correlation map analysis revealed a surprising finding:

There was no strong or statistically significant relationship between the target variable (Weekly_Sales) and most of the independent variables examined. While a few variables showed a very slight negative correlation, their influence was negligible. The highest correlation observed was with "Size," indicating a moderately weak positive correlation of 0.24. Based on this analysis, we can conclude that the selected independent variables do not appear to have a direct or inverse causal relationship with weekly sales.



Regression Models

Model Selection Overview:

- Linear Regression: This simple model served as a baseline to understand the basic linear relationships within the data.
- Regularized Linear Models (Lasso, Elastic Net, Ridge): These models were included to address overfitting and reduce the impact of irrelevant features
- Non-linear Models (SVM, KNN): We included models like SVM and KNN to investigate non-linear relationships within the data, as retail sales are often influenced by complex factors .
- Ensemble Methods : These models were chosen because they often achieve higher accuracy by combining multiple decision trees. They are particularly well-suited for handling large, complex datasets and reducing overfitting.

Regression Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
XGBoost Regression	0.95165	3014.199	1812.385
Random Forest Regression	0.94765	3137.518	1640.912
LG Boost Regression	0.82263	5842.45	4014
AdaBoost Regression	0.7286	7226.53	5264.93
KNN Regression	0.17608	12442.98	8806.89
LASSO Regression	0.06354	13265.54	9994.91
Elastic Net	0.06354	13265.54	9994.91
RIDGE Regression	0.06352	13265.73	3187.56
SVM Regression	0.00959	13642.32	8587.45

Time Series Models

Model Selection Overview:

- ARIMA: This widely used model captures trends and seasonality, providing a baseline for comparison.
- SARIMAX: By allowing for the inclusion of external factors, SARIMAX addresses the need to incorporate variables like promotions and economic indicators that influence retail sales.
- Auto ARIMA: Automating parameter selection through Auto ARIMA streamlined the process and potentially identified more optimal settings for the ARIMA model.
- These models were chosen to explore both traditional statistical methods and more advanced approaches capable of incorporating external factors, offering a comprehensive evaluation of time series techniques for retail sales forecasting.

Time Series Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
ARIMA	0.000282	13942.26	10789.3
SARIMAX	-0.0007	13945.18	10958.075
Auto ARIMA	-0.00076	13945.18	10958.07

Neural Network Models

Model Selection Overview:

- ANN (Artificial Neural Network): This basic architecture served as a foundation, providing a baseline for comparison with more sophisticated models. It helped demonstrate the potential of neural networks for time series forecasting.
- Tuned ANN: By hyperparameter tuning the ANN, we sought to optimize its performance for this specific dataset, potentially improving its accuracy and addressing challenges like overfitting.
- LSTM Bidirectional: Bidirectional LSTM was chosen to explore its ability to capture more intricate temporal dependencies in sales data by processing the data in both forward and backward directions. This approach is particularly valuable for time series with complex patterns.

Neural Network Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
ANN	0.36308	10940.14	7548.15
Tuned ANN	0.29192	11535.09	7951.03
LSTM Bidirectional	-0.35194	16129.86	10250.99

Key Findings on Model Performance

- **Gradient Boosting Dominance:** The results suggest that gradient boosting models (XGBoost and LightGBM) consistently outperformed other models, achieving high R-squared values and lower error metrics (RMSE and MAE).
- **Time Series Challenges:** Traditional time series models (ARIMA, SARIMAX, and Auto ARIMA) struggled to capture the complex dynamics of the data, resulting in lower accuracy and higher error metrics.
- **Deep Learning Potential:** Neural networks, including ANN and LSTM, showed promise but required careful tuning to achieve satisfactory results.

Experiments to improve the efficiency

Design and Execution:

Experiment 1: Removing Promotional Columns: We evaluated XGBoost performance without using the "MarkDown" columns (representing promotional discounts). This experiment sought to understand the influence of promotional activities on sales predictions.

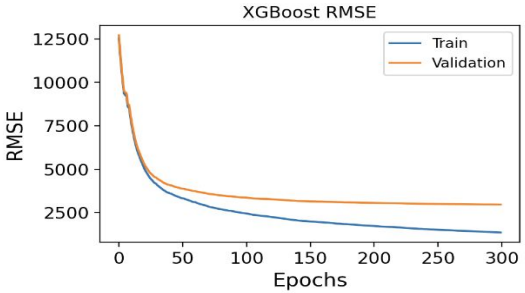
Experiment 2: Removing Store Size and Type: We tested the model's performance excluding the "Size" and "Type" columns to determine their impact on the model's accuracy.

Experiment 3: Using Only Promotional Columns: We trained the model solely on the promotional features ("MarkDown 1 to 5") to assess their predictive power.

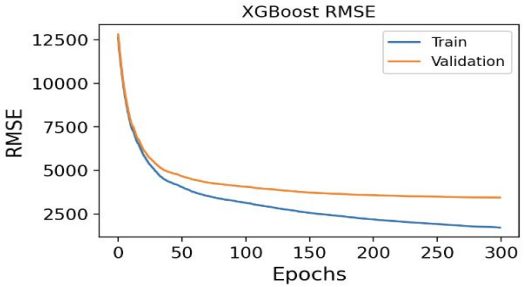
Experiment 4: Combining Promotional Columns: We combined all five promotional columns into a single feature, aiming to capture a more comprehensive view of promotional activity and its impact on sales.

Experiments on XGBoost Model			
Experiments	**R Squared**	**RMSE**	**MAE**
XGBoost Model Experiment 1	0.95165	3014.199	1812.385
XGBoost Model Experiment 2	0.93661	3451.309	2176.194
XGBoost Model Experiment 3	0.00699	13664.701	10408.86
XGBoost Model Experiment 4	0.95235	2992.263	1760.227

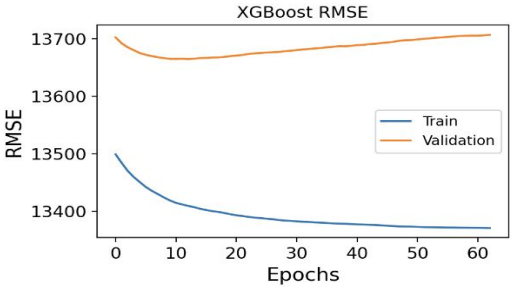
RMSE Plots of the Experiments



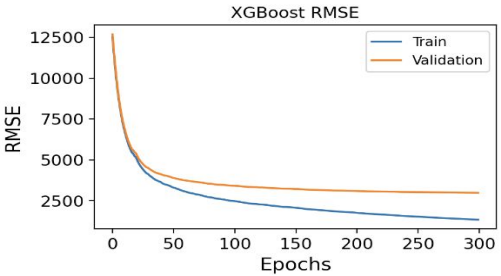
RMSE plot of Experiment 1



RMSE plot of Experiment 2



RMSE plot of Experiment 3



RMSE plot of Experiment 4

Results & Discussion

Gradient Boosting Excellence: Gradient boosting models (XGBoost and LightGBM) consistently outperformed other models, including linear regression, support vector machines, traditional time series models (ARIMA, SARIMAX, Auto ARIMA), and deep learning models (ANN, LSTM). This highlights their superior accuracy and ability to capture complex patterns in retail sales data.

Data Limitations: The analysis revealed potential limitations in the dataset, such as the lack of certain features, potential data bias, and the reliance on weekly data. Further research could explore data augmentation techniques and incorporate additional data sources.

The Post-Promotional Challenge: The research underscores the significance of accurately forecasting sales during the post-promotional period, which is often overlooked in traditional forecasting. The inclusion of "promotional period" as a feature proved essential for capturing post-promotion effects.

Conclusion & Future Work

Conclusion:

This research provides guidance on selecting the most appropriate forecasting model to forecast the sales and data characteristics & feature importance in the dataset of a retail business.

Future Work:

Hybrid Model Architectures: Exploring the combination of different deep learning architectures, such as NCDEs and ANCDEs, is promising for handling the complexities of retail sales data.

Data Augmentation and Feature Engineering: Researching data augmentation techniques, such as synthetic data generation and feature transformations, can enhance model performance, especially when dealing with limited data. Further exploration of feature engineering techniques is needed to improve model accuracy.

Conclusion & Future Work

Future Work:

Multi-Step Forecasting: While the research focused on single-step forecasting, exploring multi-step forecasting models that predict sales over multiple future time periods can provide more comprehensive insights and support long-term planning.

Integration with Business Processes: Investigating how these advanced forecasting models can be seamlessly integrated into existing retail operations and decision-making processes is crucial for practical implementation and business impact.

These future research directions offer promising avenues for enhancing retail sales forecasting, providing retailers with more accurate predictions and valuable insights to optimize their operations and improve customer experiences.

Sincere Thanks to Everyone for the Opportunity