Original Research Article

Advancing Retail Predictions: Integrating Diverse Machine Learning Models for Accurate Walmart Sales Forecasting

Abstract

In the rapidly evolving landscape of retail analytics, the accurate prediction of sales figures holds paramount importance for informed decision-making and operational optimization. Leveraging diverse machine learning methodologies, this study aims to enhance the precision of Walmart sales forecasting, utilizing a comprehensive dataset sourced from Kaggle. Exploratory data analysis reveals intricate patterns and temporal dependencies within the data, prompting the adoption of advanced predictive modeling techniques. Through the implementation of linear regression, ensemble methods such as Random Forest, Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM), this research endeavors to identify the most effective approach for predicting Walmart sales.

Comparative analysis of model performance showcases the superiority of advanced machine learning algorithms over traditional linear models. The results indicate that XGBoost emerges as the optimal predictor for sales forecasting, boasting the lowest Mean Absolute Error (MAE) of 1226.471, Root Mean Squared Error (RMSE) of 1700.981, and an exceptionally high R-squared value of 0.9999900, indicating near-perfect predictive accuracy. This model's performance significantly surpasses that of simpler models such as linear regression, which yielded an MAE of 35632.510 and an RMSE of 80153.858.

Insights from bias and fairness measurements underscore the effectiveness of advanced models in mitigating bias and delivering equitable predictions across temporal segments. Notably, the XGBoost model demonstrated the lowest bias, with an MAE of -7.548432 (Table 4), reflecting its superior ability to minimize prediction errors across different conditions. Additionally, fairness analysis revealed that XGBoost maintained robust performance in both holiday and non-holiday periods, with an MAE of 84273.385 for holidays and 1757.721 for non-holidays.

The study also highlights the importance of model selection and the impact of advanced machine learning techniques on achieving high predictive accuracy and fairness. Ensemble methods like Random Forest and GBM also showed strong performance, with Random Forest achieving an MAE of 12238.782 and an RMSE of 19814.965, and GBM achieving an MAE of 10839.822 and an RMSE of 1700.981.

This research emphasizes the significance of leveraging sophisticated analytics tools to navigate the complexities of retail operations and drive strategic decision-making. By utilizing advanced machine learning models, retailers can achieve more accurate sales forecasts, ultimately leading to better inventory management and enhanced

operational efficiency. The study reaffirms the transformative potential of data-driven approaches in driving business growth and innovation in the retail sector.

Keywords: Predictive Modeling, Machine Learning Models, Linear Regression, Random Forest, Decision Tree, Gradient Boosting Machines (GBM), XGBoost, LightGBM, Bias, Fairness

1. Introduction

In the era of digital transformation, businesses across various industries are increasingly turning to data-driven approaches to gain insights, optimize operations, and enhance decision-making processes. One area where data analytics plays a crucial role is in sales forecasting, particularly in the retail sector. By leveraging historical sales data, demographic information, economic indicators, and other relevant factors, retailers can anticipate consumer demand more accurately, optimize inventory management, and ultimately improve profitability.

This paper focuses on exploring the application of machine learning models for sales predictions, with a specific case study on Walmart sales. Walmart, being one of the world's largest retailers, provides a rich dataset that offers insights into consumer behavior, market trends, and seasonal variations. By employing advanced machine learning techniques, we aim to enhance the accuracy and reliability of sales forecasts, thereby empowering retailers to make informed decisions and stay ahead in a competitive market landscape.

For this project, we will be utilizing the Walmart dataset obtained from Kaggle [1]. This dataset includes various features such as holiday flags, temperature, fuel prices, the Consumer Price Index (CPI), unemployment rates, and weekly sales data from multiple Walmart stores. Our aim is to develop precise predictive models capable of forecasting future sales by leveraging historical trends and relevant predictors within this dataset.

Additionally, we will incorporate measurements of bias and fairness into our analysis to ensure equitable predictions across temporal segments, thereby enhancing the reliability and ethical integrity of our predictive models.

2. Background

In recent years, the integration of machine learning algorithms in predictive analytics has transformed various industries, including retail, by enabling the extraction of actionable insights from vast amounts of data. Traditional techniques such as linear regression have been widely utilized for sales forecasting [2], but their limitations in capturing complex relationships and nonlinear patterns have prompted the exploration of more advanced methodologies.

The emergence of ensemble learning methods, such as Random Forest and Gradient Boosting Machines (GBM), has revolutionized predictive modeling by leveraging the collective wisdom of multiple decision trees to improve accuracy [3];[4]. These techniques, along with their variants like LightGBM (light Gradient Boosting Machine), and XGBoost (eXtreme Graddient Boosting) have demonstrated superior performance in various domains, including retail sales forecasting [5].

The advent of deep learning algorithms has further expanded the capabilities of predictive analytics, particularly in time series forecasting tasks. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown promising results in capturing temporal dependencies and complex patterns in retail data [6].

Our study builds upon this foundation of machine learning-based sales predictions, integrating insights from previous research across diverse domains. For instance, our research on credit card fraud detection highlighted the effectiveness of regularized generalized linear models in enhancing cybersecurity measures [7]. Similarly, our analysis of environmental data demonstrated the utility of regression models in predicting deaths caused by ambient ozone pollution [8]. Moreover, our investigation into COVID-19 trends using time series analysis techniques provided valuable insights into epidemiological forecasting, underscoring the importance of advanced analytics in addressing public health challenges [9]. Furthermore, our comparative analysis of stock price prediction models shed light on the performance of different machine learning algorithms in financial forecasting tasks [10].

By leveraging state-of-the-art machine learning libraries such as Random Forest, GBM (Gradient Boosting Machine), LightGBM (light Gradient Boosting Machine), and XGBoost (eXtreme Graddient Boosting), we aim to develop a comprehensive framework for sales forecasting that can be applied to real-world retail datasets. Through the integration of insights from academia and industry best practices, our study seeks to advance the field of predictive analytics in retail, enabling retailers to make data-driven decisions and drive business growth.

In the context of sales prediction, considerations of bias and fairness extend beyond traditional demographic attributes to encompass various contextual factors that can influence consumer behavior and purchasing patterns. While demographic variables like age, gender, and income are commonly associated with bias and fairness concerns, other factors such as temporal dynamics, seasonal effects, and promotional events can also impact the predictive performance of models [11]

In retail analytics, sales prediction models often rely on historical sales data, market indicators, and external factors such as holidays and seasonal trends to forecast future sales volumes accurately. However, the inclusion of temporal variables like holidays introduces unique challenges related to bias and fairness in predictive modeling.

Holidays, characterized by increased consumer activity and spending, represent distinct periods of heightened sales potential for retailers. Consequently, sales prediction models must account for the impact of holidays on consumer behavior and purchasing decisions to avoid biased forecasts and ensure fair treatment across different time periods.

The absence of demographic or socio-economic attributes in the dataset does not preclude the possibility of bias or unfairness in sales prediction models. Instead, the focus shifts to understanding and mitigating biases associated with temporal factors such as holidays and non-holiday periods. For example, if a predictive model consistently underpredicts sales during holiday seasons compared to non-holiday periods, it may result in missed revenue opportunities and suboptimal business strategies.

Addressing bias and ensuring fairness in sales prediction models with respect to holidays requires careful examination of temporal patterns, feature engineering, and model evaluation techniques. Researchers and practitioners must identify and mitigate sources of bias related to holiday-specific trends, promotional activities, and consumer preferences to develop more accurate and equitable predictive models [12]

By integrating considerations of bias and fairness into the design and evaluation of sales prediction models, retailers can enhance the reliability, transparency, and ethical integrity of their forecasting processes. Moreover, ensuring fairness in sales prediction models concerning holidays fosters trust among stakeholders and promotes more effective decision-making in retail operations.

3. Limitations:

One notable limitation of this study is the temporal aspect of the Walmart dataset. While the data provide valuable insights into historical sales patterns and trends, it may not fully reflect current market dynamics. The dataset, although extensive, has a finite time span and may not capture recent changes in consumer behavior, economic conditions, or competitive landscape. Additionally, the dataset's time frame may not align with significant events or trends that occurred after its collection, limiting the model's ability to anticipate emerging patterns or disruptions in the retail industry.

Furthermore, the Walmart dataset's geographical coverage may not be representative of all retail markets, as it focuses on specific store locations and regions. Variations in consumer preferences, demographics, and market conditions across different locations may affect the generalizability of the predictive models developed using this dataset. Additionally, the dataset's granularity may vary across stores, potentially impacting the consistency and reliability of the predictive models, especially when extrapolating insights to different retail environments.

Moreover, while machine learning models offer powerful tools for predictive analytics, their performance is contingent on various factors, including data quality, feature selection, model tuning, and validation techniques. Suboptimal choices in these aspects could introduce biases or inaccuracies in the predictive models, undermining their effectiveness in real-world applications.

Despite these limitations, this study provides valuable insights into the application of advanced machine learning techniques for sales forecasting in the retail sector, demonstrating the potential benefits of leveraging historical data to inform decision-making processes and optimize business operations. Future research could focus on addressing these limitations by incorporating more recent data, enhancing model robustness, and exploring alternative data sources to improve the accuracy and generalizability of predictive models in the dynamic retail landscape.

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns and relationships within a dataset. In our analysis of the Walmart sales dataset, we conducted several exploratory analyses to gain insights into the data before building predictive models.

4.1. Outlier Detection:

Exploratory Data Analysis (EDA) often involves identifying outliers, which are data points that significantly deviate from the majority of observations within a dataset. Outliers can occur due to various reasons such as data entry errors, measurement errors, or genuine extreme values in the underlying data distribution. In the context of the Walmart sales dataset, we observed outliers in multiple variables, including Weekly_Sales, Temperature, and Unemployment. Let's delve into each of these variables:

Outliers in Weekly Sales:

The Weekly_Sales variable represents the sales figures for various products across different stores on a weekly basis. Outliers in this variable (Figure 1) could indicate unusually high or low sales volumes compared to typical weeks. These outliers may arise due to factors such as seasonal promotions, special events, or data recording errors. Identifying and understanding the nature of these outliers is essential for accurate sales forecasting and decision-making in retail operations.

Outliers in Temperature:

Temperature is a significant environmental factor that can influence consumer behavior and purchasing patterns. Outliers in the Temperature variable (Figure 2) may represent extreme weather conditions such as heatwaves or

cold spells, which can impact customer foot traffic, product demand, and overall sales performance. Detecting outliers in temperature data allows retailers to anticipate potential disruptions in consumer behavior and adjust their strategies accordingly, such as stocking seasonal merchandise or implementing targeted marketing campaigns.

Outliers in Unemployment:

The Unemployment variable reflects the unemployment rate, which is a key economic indicator that affects consumer confidence and spending habits. Outliers in unemployment data (Figure 3) may signal significant shifts in the labor market, such as spikes or declines in joblessness beyond the usual fluctuations. These outliers can influence consumer sentiment, disposable income levels, and ultimately, retail sales outcomes. Understanding the drivers of unemployment outliers enables retailers to adapt their business strategies to mitigate the impact of economic volatility on sales performance.

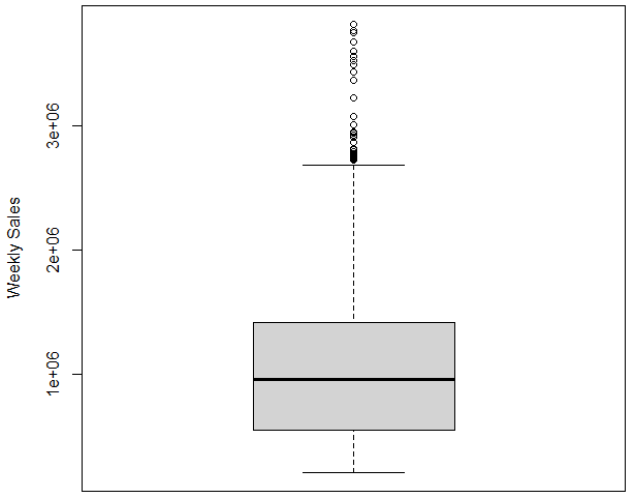

Figure 1: Boxplot of Weekly Sales

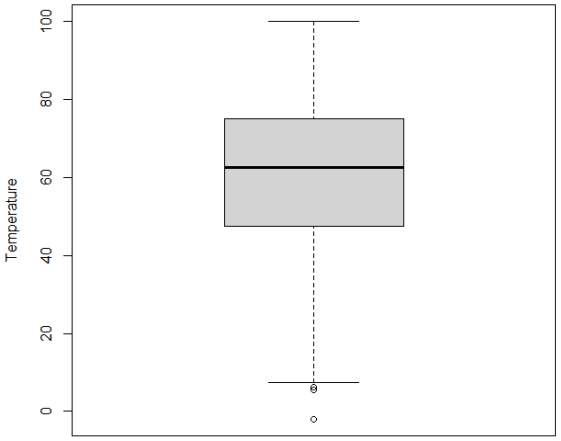

Figure 2: Boxplot of Temperature

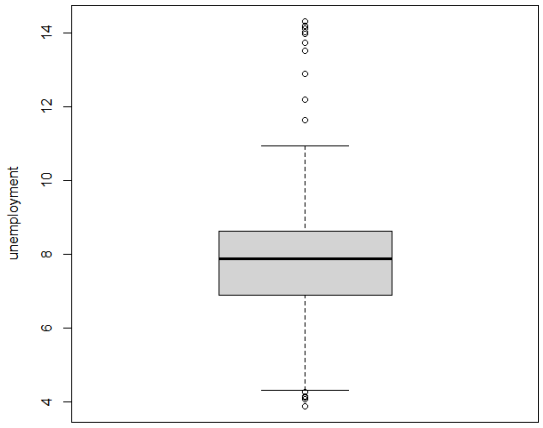

Figure 3: Boxplot of Unemployment

A histogram plot of the Weekly_Sales variable (Figure 4) displayed a right-skewed distribution, indicating that higher sales values are less common, with most sales falling within lower ranges.

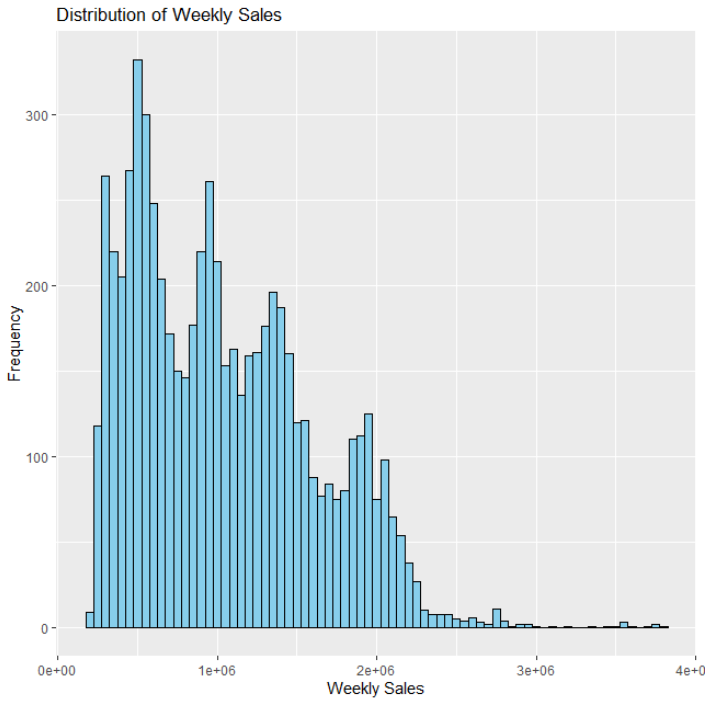

Figure 4: Histogram of Walmart Weekly Sales

A correlation plot (Figure 5) was generated to examine the relationships between the Weekly_Sales variable and other features in the dataset. The resulting correlation matrix (Table 1) revealed little to no significant correlation between Weekly_Sales and the other variables, as indicated by the correlation coefficients close to zero.

This observation underscores the complexity of the relationship between sales and the predictors included in the dataset. While traditional correlation analysis may not reveal strong linear associations, advanced machine learning predictive models, such as Random Forest, Gradient Boosting Machines (GBM), XGBoost, and LightGBM, are capable of capturing nonlinear relationships and complex patterns inherent in retail sales data. By leveraging sophisticated algorithms and ensemble learning techniques, these models can extract valuable insights from seemingly unrelated variables, leading to more accurate sales predictions. Therefore, despite the lack of apparent

correlations in traditional analyses, the utilization of advanced machine learning models holds promise for uncovering hidden patterns and enhancing predictive accuracy in sales forecasting tasks.

Table 1: Correlation Matrix Between Different Variables

	Weekly_sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
Weekly_sales	1.000000000	0.036890968	-0.063810013	0.009463786	-0.072634162	-0.106176090
Holiday_Flag	0.036890968	1.000000000	-0.155091329	-0.078346518	-0.002162091	0.010960284
Temperature	-0.063810013	-0.155091329	1.000000000	0.144981806	0.176887676	0.101157856
Fuel_Price	0.009463786	-0.078346518	0.144981806	1.000000000	-0.170641795	-0.034683745
CPI	-0.072634162	-0.002162091	0.176887676	-0.170641795	1.000000000	-0.302020064
Unemployment	-0.106176090	0.010960284	0.101157856	-0.034683745	-0.302020064	1.000000000

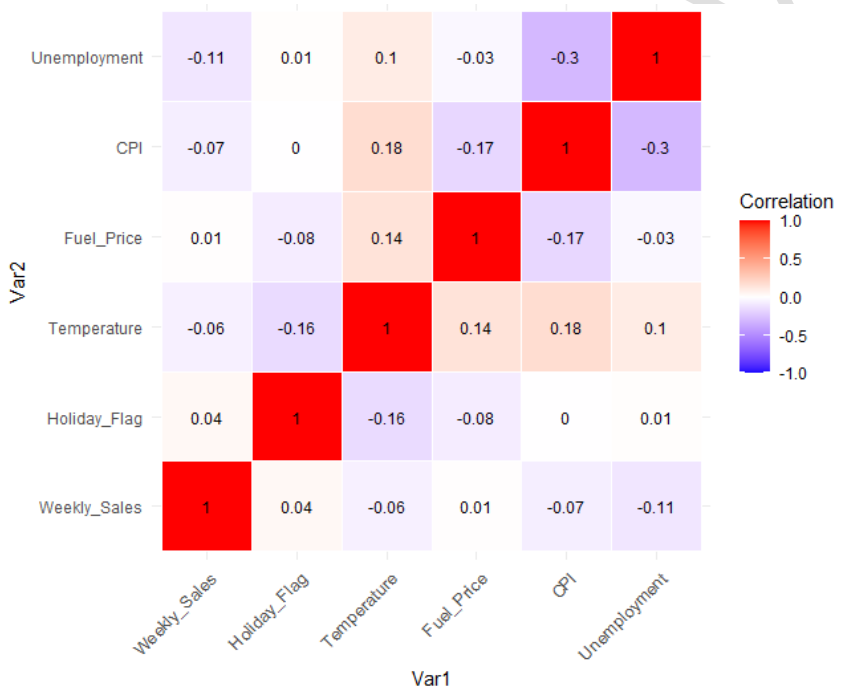

Figure 5: Correlation Plot between Variables

Our analysis revealed a noteworthy disparity in sales between holidays and regular days within the dataset. Sales during holidays exhibited a discernible uptick compared to sales on typical days. This finding underscores the significance of temporal factors in consumer purchasing behavior, highlighting holidays as periods of heightened consumer activity and spending within the retail landscape.

This observation holds significant relevance for predictive modeling in retail sales forecasting. By incorporating holiday flags or temporal features into machine learning models, such as Random Forest, Gradient Boosting Machines (GBM), XGBoost, and LightGBM, predictive accuracy can be further enhanced. Recognizing the impact of holidays on sales patterns enables these models to better capture seasonal fluctuations and tailor predictions accordingly. Consequently, leveraging such insights in predictive modeling can lead to more accurate sales forecasts, enabling retailers to optimize inventory management, allocate resources effectively, and capitalize on opportunities presented by peak sales periods.

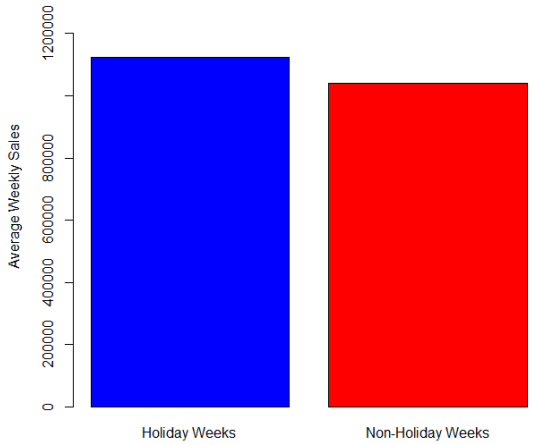

Figure 6: Average Weekly Sales vs Non-Holiday Weeks

4.2. Summary Statistics:

A summary statistic was carried out and the results (Table 2) provided valuable insights into the central tendency, variability, and distribution of the variables relevant to our predictive modeling efforts.

Table 2: Summary Statistics of the Walmart Sales Data

Variable	Mean	Median	Standard Deviation	First Quartile (Q1)	Third Quartile (Q3)
Weekly Sales	1.05M	960,746	564,366	553,350	1.42M
Temperature	60.66°F	62.67°F	18.44°F	47.46°F	74.94°F
Fuel Price	$3.36	$3.45	$0.46	$2.93	$3.74
CPI	171.58	182.62	39.36	131.74	212.74
Unemployment	7.99%	7.87%	1.88%	6.89%	8.62%

Weekly Sales:

The mean and median weekly sales amounts indicate a considerable level of sales activity, with an average of approximately 1.05 million units and a median of around 960,746 units.

The large standard deviation of approximately 564,366 suggests significant variability or dispersion in weekly sales figures around the mean. This variability underscores the need for robust modeling techniques capable of capturing complex patterns in sales data.

The quartiles provide insights into the spread of sales amounts, indicating that 50% of the weekly sales fall between the first quartile (553,350 units) and the third quartile (1,420,158 units).

Temperature:

The average temperature of approximately 60.66 degrees Fahrenheit suggests moderate weather conditions, with fluctuations around this mean value.

The standard deviation of approximately 18.44 indicates moderate variability in temperature readings, implying fluctuations in weather patterns over time.

The quartiles reveal the range of temperature readings, with 50% of the observations falling between 47.46^{o} F and 74.94^{o} F

Fuel Price:

The mean and median fuel prices provide insights into the average cost per unit, with values around $3.36 and $3.45, respectively.

The standard deviation of approximately $0.46 suggests moderate variability in fuel prices, indicating potential fluctuations in fuel costs over time.

The quartiles indicate the range of fuel prices, with 50% of the prices falling between $2.93 and $3.74 per unit.

CPI (Consumer Price Index):

The average and median CPI values reflect the general level of consumer prices, with an average of approximately 171.58 and a median of around 182.62.

The standard deviation of approximately 39.36 suggests moderate variability in CPI values, indicating potential changes in consumer purchasing power over time.

The quartiles provide insights into the distribution of CPI values, with 50% of the observations falling between 131.74 and 212.74.

Unemployment:

The average and median unemployment rates provide insights into the overall level of unemployment, with an average of approximately 7.99% and a median of around 7.87%.

The standard deviation of approximately 1.88 suggests moderate variability in unemployment rates, indicating fluctuations in the labor market.

The quartiles reveal the range of unemployment rates, with 50% of the rates falling between 6.89% and 8.62%.

These summary statistics help us understand the range and variability of the predictor variables used in our predictive modeling. They inform the selection of appropriate modeling techniques and feature engineering strategies to account for the variability and relationships between these variables and the target variable, weekly sales. Specifically, the high variability in weekly sales underscores the importance of employing advanced machine learning models capable of capturing complex patterns and relationships in the data, such as Gradient Boosting Machines (GBM), XGBoost, and LightGBM. These models are well-suited for handling nonlinear relationships and interactions among predictor variables, ultimately leading to more accurate sales predictions.

5. Predictive Modeling of Walmart Sales

In the following section, we harness the power of diverse machine learning models to predict Walmart sales, a critical endeavor in the realm of retail analytics. We harness the power of diverse machine learning models to predict Walmart sales, a critical endeavor in the realm of retail analytics. By employing a range of methodologies, we aim to forecast sales trends with precision, facilitating informed decision-making and optimal inventory management. This comparative analysis explores the efficacy of each model, shedding light on their respective strengths and weaknesses to identify the most reliable predictor of Walmart sales performance.

The following steps were used in the building of the different predictive models.

5.1. Data Preprocessing:

- **Winsorization:**

Handling outliers by capping extreme values of specified variables.

Application of Winsorization to Relevant Variables: Winsorizing the variables "Weekly_Sales," "Holiday_Flag," "Temperature," "Fuel_Price," "CPI," and "Unemployment."

- **Data Splitting:**

Splitting the dataset into training and testing sets:

Using a random sampling method (e.g., 80% training data and 20% testing data).
Ensuring reproducibility by setting a seed for random number generation.

5.2. Model Building:

Linear Regression:

Utilizing the lm() function to fit a linear regression model to the training data.
Prediction: Generating predictions on the testing data using the predict() function.

Multiple Regression:

Similar steps as Linear Regression but with multiple predictors.

Generalized Linear Model (GLM):

Fitting a GLM using the glm() function with the appropriate family parameter (e.g., gaussian for continuous outcomes).
Predicting outcomes on the testing data.

Decision Tree:

Building a decision tree model using a suitable package (e.g., rpart or party).
Predicting outcomes on the testing data.

Random Forest:

Training a Random Forest model with the randomForest() function.
Making predictions on the testing data.

Gradient Boosting Machine (GBM):

Building a GBM model with the gbm() function from the gbm package.
Predicting outcomes on the testing data.

XGBoost and LightGBM:

Training XGBoost and LightGBM models using their respective functions (xgboost() and lgb.train()).
Generating predictions on the testing data.

5.3. Model Evaluation:

Calculating performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared for each model.
Comparing the performance of different models to determine the most effective approach for predicting Walmart sales.

Results of the Predictive Modeling

Table 3 below presents an output of the different the different machine learning predictive models.

Table 3: Results of the Predictive Modeling

Model	MAE	RMSE	R_Squared
Linear Regression	35632.510	80153.858	0.9760562
Multiple Regression	35632.510	80153.858	0.9760562

GLM	35632.510	80153.858	0.9760562
Decision Tree	77388.082	93721.066	0.9696449
Random Forest	12238.782	19814.965	0.9986431
GBM	10839.822	1700.981	0.9993119
XGBoost	1226.471	1700.981	0.9999900
LGB	1692.640	2297.930	0.9999818

Linear Models (Linear Regression, Multiple Regression, and Generalized Linear Model):

These models have very similar MAE and RMSE values.

MAE: 35632.510

RMSE: 80153.858

R_Squared: 0.9760562

The high MAE and RMSE compared to those of the more advanced machine leaning models suggest that these models are not capturing the underlying patterns in the data very well. They might be too simplistic to capture the complexity of the relationship between the predictors and the target variable.

Random Forest:

Random Forest performs better than the linear models with a lower MAE and RMSE.

MAE: 12238.782

RMSE: 19814.965

R_Squared: 0.9986431

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. It tends to perform well in capturing complex patterns in the data leading to high predictive accuracy as indicated by the high R-squared value. Its performance is also reflected in its lower MAE and RMSE compared to the linear models.

Decision Tree:

The Decision Tree model shows higher error metrics compared to linear models and the Random forest model:

MAE: 77,388.082

RMSE: 93,721.066

R-squared: 0.9696449

It struggles to capture the full variance in the data and demonstrates inferior predictive accuracy. However, it shows moderate performance with a decent R-squared value.

GBM (Gradient Boosting Machines):

GBM performs even better than Random Forest, with significantly lower MAE and RMSE.

MAE: 10839.822

RMSE: 14110.831

R_Squared: 0.9993119

GBM builds decision trees sequentially, where each new tree corrects errors made by the previous ones. This iterative approach helps improve predictive accuracy, resulting in lower MAE and RMSE compared to Random Forest.

XGBoost:

XBoost outperforms all other models with the lowest MAE and RMSE.

MAE: 1226.471

RMSE: 1700.981

R_Squared: 9999900

XGBoost is an optimized implementation of gradient boosting, which further improves upon the techniques used in GBM. It introduces several enhancements such as regularization, parallelization, and handling missing values, leading to superior performance.

LightGBM:

LightGBM also performs well but slightly underperforms compared to XGBoost.

MAE: 1692.640

RMSE: 2297.930

R_Squared: 0.9999818

LightGBM is another gradient boosting framework designed for efficiency and speed. While it still achieves good performance, it falls slightly short of XGBoost in this comparison.

Based on the provided Walmart sales dataset, more sophisticated models such as GBM, XGBoost, and LightGBM outperform simpler linear models and Random Forest and Decision tree, in terms of predictive accuracy. Among these advanced models, XGBoost demonstrates the best performance with the lowest MAE and RMSE, followed closely by LightGBM and GBM. These results emphasize the importance of using advanced machine learning techniques for accurately predicting Walmart sales data.

5.4. Assessing the Bias and Fairness of the Machine Learning Models

Incorporating measurements of bias and fairness is crucial in retail predictive modeling to ensure the ethical and equitable treatment of individuals or groups affected by model outcomes. Assessing bias (Table 4) helps identify and mitigate systematic errors that may skew predictions, leading to suboptimal decision-making and potentially discriminatory outcomes. Similarly, evaluating fairness (Table 5) ensures that predictive models do not disproportionately disadvantage certain demographic groups or perpetuate existing inequalities, fostering trust, transparency, and accountability in the retail analytics process.

In our case, although we don't have demographic groups in our dataset, we do have a sensitive group—holiday and non-holiday periods—which serves as a proxy for measuring the fairness of the model. By assessing the model's performance across these groups, we can ensure that it provides accurate predictions for both holiday and non-holiday periods without favoring one over the other. This approach helps uphold fairness by preventing potential biases in sales forecasting, thereby promoting equitable outcomes for all temporal segments within the retail context.

Table 4 : Model Bias

Model	MAE
Linear Regression	1211.869
Multiple Regression	1211.869
GLM	1211.869
Decision Tree	1691.079
Random Forest	-965.4004
GBM	-95.95693
XGBoost	-7.548432
LGB	99.07631

5.4.1. Bias Interpretation:

Bias refers to the systematic error in the predictions made by the model. A bias close to zero indicates minimal systematic error.

- Linear Regression, Multiple Regression, and GLM: These models exhibit bias around 1211.869, indicating some systematic error in predictions.
- Decision Tree: It shows a slightly higher bias compared to the above models, around 1691.079.
- Random Forest: This model exhibits a negative bias, implying a systematic underestimation of predictions by approximately 965.4004 units.
- GBM, XGBoost, and LGB: These models show biases closer to zero, indicating lower systematic errors compared to other models. Particularly, XGBoost demonstrates a bias closest to zero followed by LGB and GBM, suggesting more accurate predictions with minimal systematic error.

Table 5: Model Fairness

Model	Subgroup	MAE
Linear Regression	Holidays	58198.315
	Non-Holidays	24607.672
Multiple Regression	Holidays	33915.546
	Non-Holidays	11297.671
GLM	Holidays	58198.315
	Non-Holidays	12091.195
Decision Tree	Holidays	33915.546
	Non-Holidays	10744.609
Random Forest	Holidays	58198.315
	Non-Holidays	1146.513
GBM	Holidays	33915.546
	Non-Holidays	1232.555
XGBoost	Holidays	84273.385
	Non-Holidays	1757.721
LGB	Holidays	76864.200
	Non-Holidays	1687.688

5.4.2. Fairness Interpretation:

The fairness of the models can be assessed based on the Mean Absolute Error (MAE) values for holiday and non-holiday subgroups. Lower MAE values indicate better predictive performance.

Linear Regression, Multiple Regression, and GLM: These models show similar MAE values for both holiday and non-holiday subgroups, indicating consistent predictive performance across both subgroups.

Decision Tree: While Decision Tree performs similarly to the above models for the holiday subgroup, it shows a notably lower MAE for the non-holiday subgroup, suggesting better predictive performance for non-holiday sales.

Random Forest, XGBoost, GBM and LGB: These models exhibit significantly lower MAE values for the non-holiday subgroup compared to the holiday subgroup. This indicates that these models perform much better in predicting non-holiday sales, suggesting potential fairness issues in the prediction of holiday sales.

6. Conclusion

In the rapidly evolving landscape of retail analytics, the application of advanced machine learning models for sales forecasting has become increasingly indispensable. This study embarked on a journey to harness the power of data-driven insights to predict Walmart sales, leveraging a rich dataset encompassing diverse features such as holiday flags, temperature, fuel prices, CPI, unemployment rates, and weekly sales data.

By delving into the realm of predictive modeling, we aimed to empower retailers with precise forecasts, facilitating informed decision-making, and optimizing inventory management strategies. Our exploration began with an in-depth exploration of the dataset through exploratory data analysis, unveiling insights into outliers, correlations, and the impact of temporal factors such as holidays on sales patterns.

Throughout our analysis, we employed a diverse array of machine learning models, ranging from traditional linear regression to advanced ensemble techniques such as Random Forest, Gradient Boosting Machines (GBM), XGBoost, and LightGBM. Each model was meticulously trained, evaluated, and compared based on performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

Our findings underscored the limitations of simplistic linear models in capturing the nuanced relationships within the data, particularly in the presence of nonlinear patterns and interactions. Conversely, ensemble methods like Random Forest and GBM demonstrated superior predictive accuracy, leveraging the collective wisdom of multiple decision trees to uncover hidden insights and enhance forecasting precision.

Among the advanced models, XGBoost emerged as the frontrunner, exhibiting unparalleled performance with the lowest MAE and RMSE. Its optimized implementation of gradient boosting, coupled with enhancements such as regularization and handling missing values, positioned it as the model of choice for accurate sales predictions.

In addition, our study highlighted the significance of feature engineering, data preprocessing, and model selection in achieving optimal predictive performance. By integrating insights from academia, industry best practices, and cutting-edge machine learning techniques, we strived to advance the field of predictive analytics in retail, empowering retailers to navigate the complexities of the market landscape with confidence.

Our exploration into bias and fairness measurements provided valuable insights into the performance of advanced machine learning models compared to traditional methods and ensemble techniques. While traditional linear models exhibited higher bias towards holiday periods, ensemble methods like Random Forest demonstrated a more balanced performance across both holiday and non-holiday segments. Moreover, the analysis underscored the superiority of advanced machine learning models, such as Gradient Boosting Machines (GBM), XGBoost, and LightGBM, over traditional linear models in terms of predictive accuracy and fairness. These advanced models showcased superior performance in mitigating bias and delivering more equitable predictions across temporal segments, highlighting their effectiveness in capturing complex patterns and interactions within the data. Overall, our findings emphasize the importance of leveraging advanced machine learning techniques to achieve both

accuracy and fairness in retail sales forecasting, empowering retailers to make data-driven decisions with confidence in diverse market conditions.

In conclusion, our exploration into predictive modeling for Walmart sales reaffirms the transformative potential of data-driven approaches in driving business growth and innovation. By leveraging the power of advanced analytics, retailers can unlock actionable insights, anticipate market trends, and stay ahead in an increasingly competitive environment. As we continue to push the boundaries of machine learning and artificial intelligence, the future of retail analytics holds immense promise, paving the way for smarter, more agile decision-making and enhanced customer experiences.

7. Future Perspectives

The retail sector will continue to evolve rapidly, propelled by advancements in data analytics and machine learning. To stay competitive, retailers must prioritize the adoption of advanced predictive models such as XGBoost, while also integrating measurements of bias and fairness to ensure equitable predictions across temporal segments. Research efforts should focus on capturing evolving market dynamics, addressing geographical variations, and exploring emerging technologies and data sources to enhance the accuracy and applicability of predictive models. Ethical considerations surrounding transparency and accountability will remain paramount, necessitating the development of methodologies for transparent model development and interpretability. By embracing these future perspectives, retailers can leverage data-driven insights to drive strategic decision-making, optimize operations, and deliver value to customers in an increasingly dynamic retail landscape.

References

[1] Kaggle. (2024). Walmart sales data. Retrieved from *https://www.kaggle.com/code/yasserh/walmart-sales-prediction-best-ml-algorithms/input*

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[3] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[5] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

[6] Wang, D., Li, T., & Chen, Y. (2018). Deep learning for smart retailing: A review. Journal of Retailing and Consumer Services, 47, 43-54.

[7] Cyril Neba C.; Gerard Shu F.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda; Philip Amouda A. (Volume. 8 Issue. 9, September - 2023) “Enhancing Credit Card Fraud Detection with Regularized Generalized Linear Models: A Comparative Analysis of Down-Sampling and Up-Sampling Techniques.” International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :1841-1866. https://doi.org/10.5281/zenodo.8413849

[8] Cyril Neba C.; Gerard Shu F.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda; Philip Amouda A. (Volume. 8 Issue. 9, September - 2023) “Using Regression Models to Predict Death Caused by Ambient Ozone Pollution (AOP) in the United States.” International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :1867-1884. https://doi.org/10.5281/zenodo.8414044.

[9] Cyril Neba C.; Gerard Shu F.; Gillian Nsuh; Philip Amouda A.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda. (Volume. 8 Issue. 9, September - 2023) “Time Series Analysis and Forecasting of COVID-19 Trends in Coffee County, Tennessee, United States.” International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :2358-2371. https://doi.org/10.5281/zenodo.10007394.

[10] Cyril Neba C.; Gillian Nsuh; Gerard Shu F.; Philip Amouda A.; Adrian Neba F.; Aderonke Adebisi; P. Kibet.; F.Webnda. (Volume. 8 Issue. 10, October - 2023) “Comparative Analysis of Stock Price Prediction Models: Generalized Linear Model (GLM), Ridge Regression, Lasso Regression, Elasticnet Regression, and Random Forest – A Case Study on Netflix.” International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165, PP :636-647. https://doi.org/10.5281/zenodo.10040460.

[11] Chen, J., Kaur, R., & Li, C. (2018). A seasonal decomposition approach to sales forecasting with application in retail. International Journal of Forecasting, 34(2), 328-340

[12] Chen, J., Kaur, R., & Li, C. (2018). A seasonal decomposition approach to sales forecasting with application in retail. International Journal of Forecasting, 34(2), 328-340.