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

Cyril Neba C. [a*], Gerard Shu F. [b], Gillian Nsuh [c], Philip Amouda A. [a], Adrian Neba F. [a], F. Webnda [a], Victory Ikpe [d], Adeyinka Orelaja [a] and Nabintou Anissia Sylla [e]

[a] *Department of Mathematics and Computer Science, Austin Peay State University, Clarksville, Tennessee, USA.*
[b] *Gianforte School of Computing, Montana State University, Bozeman, Monatana, USA.*
[c] *School of Business Analytics, University of Quinnipiac, Hamden, Connecticut, USA.*
[d] *Department of Economics and Decision Sciences, Western Illinois University, Macomb, Illinois, USA.*
[e] *University of Arkansas, Graduate and Interdisciplinary Studies, Statistics and Analytics, Fayetteville, Arkansas, USA.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

Original Research Article

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

_____

*Corresponding author: Email: cneba@my.apsu.edu;

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. Our analysis revealed varying levels of bias across different models. Linear Regression, Multiple Regression, and GLM exhibited moderate bias, suggesting some systematic errors in predictions. Decision Tree showed slightly higher bias, while Random Forest demonstrated a unique scenario of negative bias, implying systematic underestimation of predictions. However, models like GBM, XGBoost, and LGB displayed biases closer to zero, indicating more accurate predictions with minimal systematic errors. 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.

Insights from the fairness measurements revealed that Linear Regression, Multiple Regression, and GLM showed consistent predictive performance across both subgroups. Meanwhile, Decision Tree performed similarly for holiday predictions but exhibited better accuracy for non-holiday sales, whereas, Random Forest, XGBoost, GBM, and LGB models displayed lower MAE values for the non-holiday subgroup, indicating potential fairness issues in predicting holiday sales.

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 [1-3].

The dataset utilized in this study was sourced from Kaggle [4,1,2,3] and comprises Walmart sales data collected from February 5, 2010, to October 26, 2012. This dataset encompasses a wide range of features, including holiday flags, temperature, fuel prices, the Consumer Price Index (CPI), unemployment rates, and weekly sales data. The

sales data were collected from various Walmart stores located across the United States, providing a diverse representation of retail operations within the country. The temporal scope of the dataset spans approximately two and a half years, offering insights into sales patterns and trends over this period. By capturing data from multiple Walmart stores across different regions of the United States, the dataset reflects the geographical diversity of retail markets, thereby enhancing its applicability to broader retail analytics. 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 [5], 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 [6,7]. 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 [8].

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 [9].

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 [10]. Similarly, our analysis of environmental data demonstrated the utility of regression models in predicting deaths caused by ambient ozone pollution [11]. 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 [12]. Furthermore, our comparative analysis of stock price prediction models shed light on the performance of different machine learning algorithms in financial forecasting tasks [13].

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 [14]

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 [14]

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:

4.2 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 (Fig. 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.

4.3 Outliers in temperature

Temperature is a significant environmental factor that can influence consumer behavior and purchasing patterns. Outliers in the Temperature variable (Fig. 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.

4.4 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 (Fig. 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.

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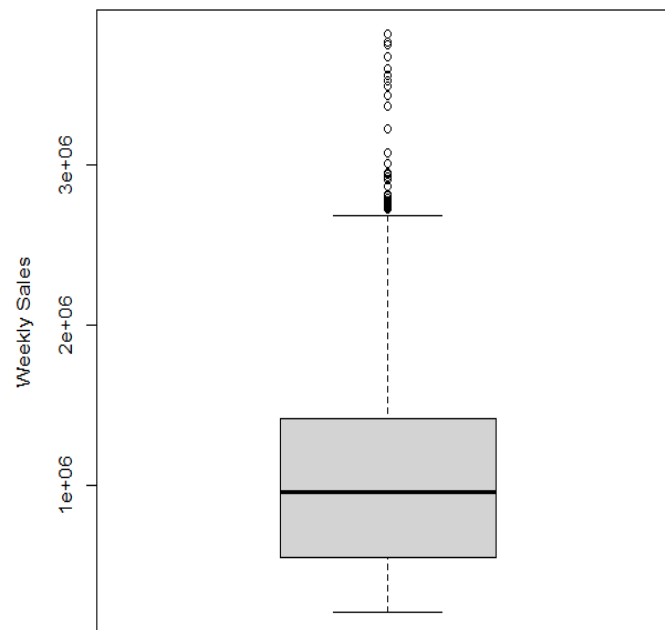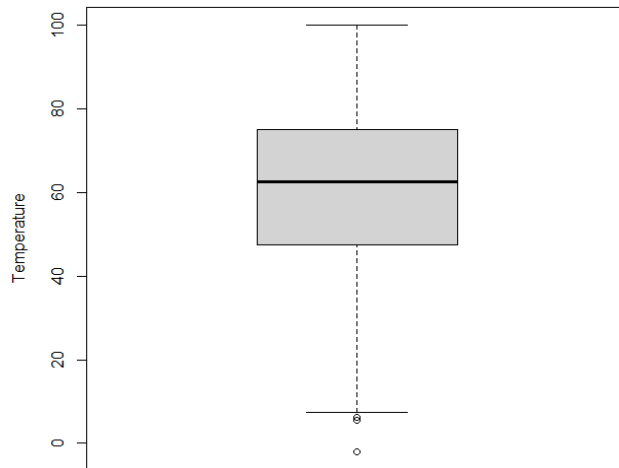

Fig. 1. Boxplot of weekly sales
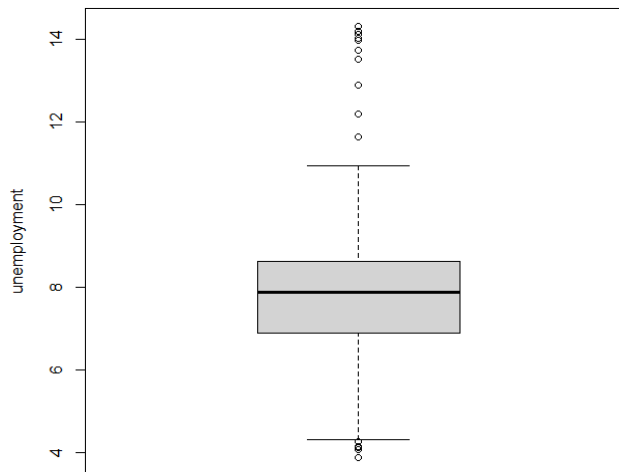
Fig. 2. Boxplot of temperature
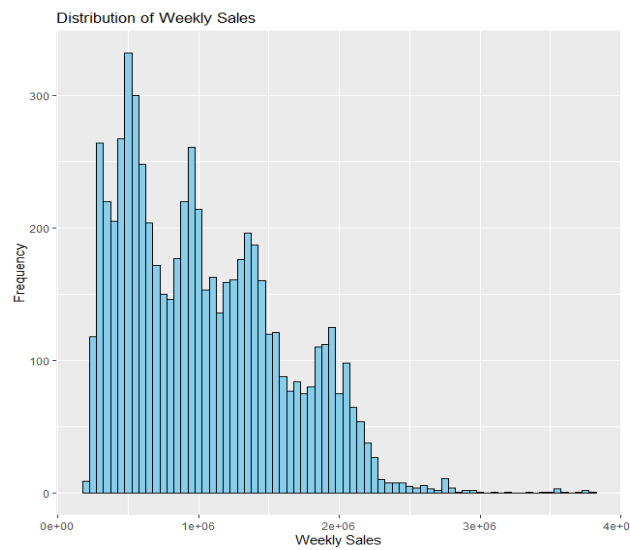

Fig. 3. Boxplot of Unemployment

Fig. 4. Histogram of walmart weekly sales

A correlation plot (Fig. 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.

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.
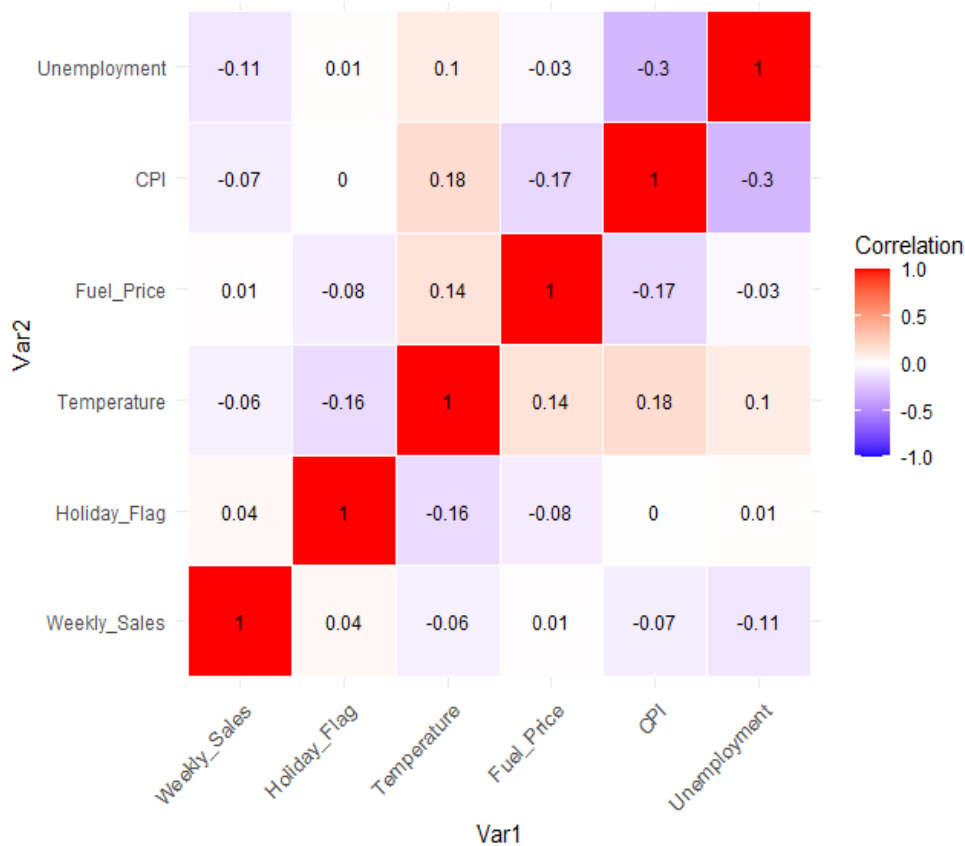
Fig. 5. Correlation plot between variables

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

Fig. 6. Average weekly sales vs non-holiday weeks

4.5 Summary statistics

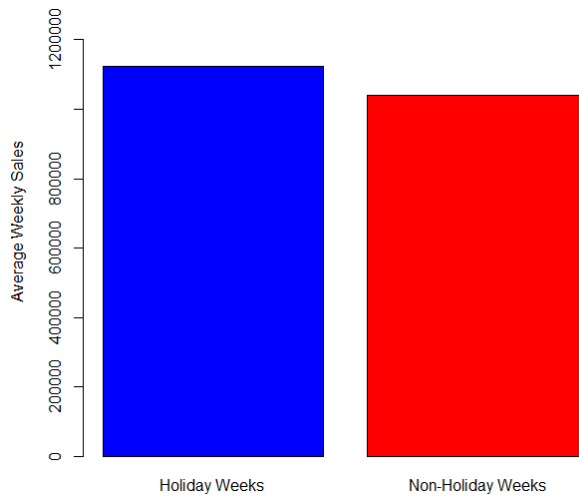A summary statistic was carried out and the results (Table 2) provided valuable insights into the central tendency, variability, and distribution of the variables relevant to our predictive modeling efforts.

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%

4.6 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).

4.7 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° F and 74.94° F

4.8 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.

4.9 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.

4.10 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 Modeling of the walmart sales dataset

5.2.1 Interpretability of machine learning models

Interpretability in machine learning is pivotal, encompassing the capacity to elucidate and rationalize the mechanisms through which a model generates predictions [15]. This facet assumes paramount importance across diverse domains, notably in retail predictive modeling, where stakeholders necessitate discerning insights into the underlying determinants steering model decisions.

5.2.2 Linear regression, multiple regression, and GLM

Traditional regression models such as Linear Regression, Multiple Regression, and Generalized Linear Models (GLM) inherently offer high interpretability [16]. These models delineate the association between input features and the target variable by fitting a linear equation. The coefficients assigned to each input feature explicate both the magnitude and direction of its impact on the target variable. For instance, a positive coefficient signifies a positive correlation, while a negative coefficient denotes a negative relationship.

5.2.3 Decision tree

Decision trees stand out for their intuitive and transparent interpretability [6]. They segment the feature space into distinct partitions based on conditional if-else statements, with each node representing a decision based on a specific feature. Traversing the tree enables an explicit understanding of the sequential decisions leading to a prediction. This inherent transparency renders decision trees invaluable for comprehending the rationale behind model predictions.

5.2.4 Random Forest and Gradient Boosting Machines (GBM)

Random Forest and Gradient Boosting Machines (GBM) represent ensemble techniques, comprising multiple decision trees. Although individual trees within these ensembles may lack interpretability, the amalgamation of predictions from numerous trees enhances accuracy [7]. Techniques such as feature importance elucidate the relative significance of each feature in predicting outcomes, thus augmenting interpretability to a certain extent.

5.2.5 XGBoost and LightGBM

XGBoost and LightGBM, advanced gradient boosting frameworks renowned for their predictive prowess, introduce a layer of complexity compared to traditional models like linear regression. Nonetheless, they offer features such as SHAP (SHapley Additive exPlanations) values, which quantify the contribution of each feature to model predictions [5]. SHAP values furnish invaluable insights into feature importance, facilitating a deeper comprehension of model behavior.

5.3 Understanding predictions

Comprehending how models derive predictions necessitates the application of various techniques such as feature importance, partial dependence plots, and SHAP values [15]. These methodologies empower analysts to discern which features wield substantial influence on predictions and elucidate how alterations in feature values impact the predicted outcome. Furthermore, visual aids like decision tree diagrams and SHAP summary plots play a pivotal role in interpreting model behavior and effectively communicating insights to stakeholders.

In summation, while certain models offer inherent interpretability, others necessitate the adoption of more sophisticated techniques for a comprehensive understanding of their predictions. By leveraging appropriate interpretability tools and methodologies, analysts can glean invaluable insights into model behavior, instill trust in predictions, and facilitate informed decision-making processes in the realm of retail analytics.

5.3.1 Model building

To prevent the effect of overfitting, we opted to use default hyperparameter tuning for the different models especially because the dataset was not quite large enough and therefore was not too complex for the models to effectively capture the patterns in the data.

For future projects on large datasets with a lot more complexities, non-default hyperparameter tuning will be employed.

Linear Regression

- Utilizing the lm() function to fit a linear regression model to the training data.
- **Hyperparameter tuning:** In R, when using the lm() function for linear regression, there is no family parameter to specify because lm() is specifically designed for fitting linear models with Gaussian errors by default. So, in the case of lm(), Gaussian (normal) errors are assumed by default.
- **Prediction**: Generating predictions on the testing data using the predict() function.

Multiple Regression

- Utilizing the lm() function to fit a linear regression model to the training data.
- **Hyperparameter tuning:** In R, when using the lm() function fits the model using ordinary least squares (OLS) regression, which estimates the coefficients for each predictor variable in the model. Since OLS regression does not involve regularization, there are no hyperparameters like regularization strength (e.g., lambda in Ridge regression or alpha in LASSO regression) to tune. Therefore, there is no hyperparameter tuning process for multiple linear regression using the lm() function in R.

Generalized Linear Model (GLM)

- Fitting a GLM using the glm() function with the appropriate family parameter (e.g., gaussian for continuous outcomes).
- **Hyperparameter tuning:** In R, when using the glm() function for linear regression, the default family parameter is gaussian, which assumes a Gaussian (normal) distribution of errors, suitable for continuous outcome variables. R will therefore assume family = gaussian by default.
- **Prediction:** After building the model, we make predictions on the testing data.

Decision Tree

- Building a decision tree model using a suitable package (e.g., rpart or party).
- **Hyperparameter tuning:** The decision tree model is built using the rpart package in R. The rpart package does not involve hyperparameter tuning in the traditional sense like other machine learning algorithms. Instead, it builds a decision tree based on recursive partitioning of the data.
- **Prediction:** After building the model, we make predictions on the testing data.

Random Forest

- Training a Random Forest model with the randomForest() function.
- Hyperparameter tuning: When using R, the default values in the randomForest() function in R are typically:
  n_estimators = 500
  max_features = sqrt(number of features)
  max_depth = NULL (i.e., nodes are expanded until all leaves are pure or until all leaves contain less than min_split samples)
  min_samples_split = 2
  min_samples_leaf = 1
  bootstrap = TRUE
- **Prediction:** After building the model, we make predictions on the testing data.

Gradient Boosting Machine (GBM)

- Building a GBM model with the gbm() function from the gbm package.
- **Hyperparameter tuning:** When using the gbm package in R, the default hyperparameters for the gbm() function are as follows:

    Number of Trees (n.trees): Typically set to 100.
    Shrinkage/Learning Rate (shrinkage): Default value is 0.001.
    Tree Depth (interaction.depth): Default value is 1.
    Minimum Observations in a Node (n.minobsinnode): Default value is 10.
    Sampling Fraction (bag.fraction): Default value is 0.5.
- **Prediction:** After building the model, we make predictions on the testing data.

XGBoost

- Training XGBoost model using the function xgboost()
- **Hyperparameter Tuning :** In R, the default hyperparameter tunings for xgboost () are as follows:

        nrounds = 100,
        max_depth = 6,
        eta = 0.3,
        gamma = 0,
        subsample = 1,
        colsample_bytree = 1,
        min_child_weight = 1,
        lambda = 1,
        alpha = 0,
        objective = "reg:squarederror",
        booster = "gbtree")

- **Prediction:** After building the model, we make predictions on the testing data.

LightGBM

- **Training** LightGBM model using the function lgb.train().
- **Hyperparameter Tuning :** In R, the default hyperparameter tunings for lgb.train().is as follows:
    objective = "regression",
    num_iterations = 100,
    learning_rate = 0.1,
    max_depth = -1,
    num_leaves = 31,
    min_data_in_leaf = 20,
    bagging_fraction = 1.0,
    feature_fraction = 1.0,
    lambda_l2 = 0.0,
    lambda_l1 = 0.0

- **Prediction:** After building the model, we make predictions on the testing data.

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.

6 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.

Assessing the Bias and Fairness of the Machine Learning Models

i.   Method of Bias Assessment:

We began by evaluating biases across different models, specifically examining how model predictions deviated from actual sales. This process involved the following steps:

Calculation of Metrics:

Functions were defined to calculate Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as metrics for assessing biases.

Assessment of Biases:

Using the defined functions, biases were assessed for each model by subgrouping the data based on the "Holiday" variable. For each subgroup, MAE and RMSE were calculated to quantify the deviation between predicted and actual sales during holidays and non-holidays.

Results Summary:

The bias results were compiled into a dataframe, providing insights into biases across holiday and non-holiday periods for each model.

ii.   Method of Fairness Evaluation:

To further evaluate fairness and bias concerning holiday and non-holiday attributes, we conducted a subgroup analysis as follows:

6.1 Model prediction separation

Model predictions were separated into holiday and non-holiday subgroups for each model, including Linear Regression, Multiple Regression, GLM, Decision Tree, Random Forest, GBM, XGBoost, and LightGBM.

6.2 Performance metric calculation

MAE was calculated separately for holiday and non-holiday subgroups for each model, providing a comparative analysis of prediction accuracy across different time periods.

6.3 Results compilation

The results were compiled into a dataframe, summarizing MAE values for holiday and non-holiday subgroups across all models.

By conducting detailed bias assessments and fairness evaluations, we aimed to provide a comprehensive understanding of model performance across different temporal attributes, particularly focusing on holidays. These analyses contribute to a more robust assessment of predictive model fairness and performance in real-world retail scenarios.

In our study, we meticulously examined bias and fairness within our machine learning models to ensure the ethical and equitable treatment of temporal segments, specifically holiday and non-holiday periods.

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

6.4 Bias assessment

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

6.5 Fairness assessment

To evaluate fairness, we analyzed 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.

6.6 Practical implications of study findings for retail businesses

The findings of this study hold significant practical implications for retail businesses, particularly in the realms of inventory management and marketing strategies. By leveraging the insights derived from various machine learning models, retailers can optimize their operations and drive more effective decision-making processes.

6.7 Optimizing inventory management

Effective inventory management is crucial for retail businesses to minimize costs and meet customer demand. Machine learning models, such as Linear Regression, Multiple Regression, and Generalized Linear Models (GLM), offer high interpretability, allowing retailers to understand the relationships between different variables and inventory levels [16]. For instance, by analyzing historical sales data and identifying key factors influencing demand, retailers can forecast future inventory requirements more accurately. This enables them to reduce excess stock, avoid stockouts, and improve overall inventory turnover.

Decision trees, known for their intuitive interpretability, can be used to segment inventory based on various features such as seasonality, product category, and sales trends [6]. By traversing the decision tree, retailers can develop specific strategies for different segments, ensuring that inventory levels are optimized for each category. This targeted approach enhances inventory efficiency and responsiveness to market changes.

Ensemble methods like Random Forest and Gradient Boosting Machines (GBM) further enhance inventory management by providing robust predictions through the aggregation of multiple decision trees [7]. Techniques such as feature importance help retailers identify the most influential factors affecting inventory levels, enabling them to prioritize key drivers and make data-driven decisions to optimize stock levels.

6.8 Enhancing marketing strategies

Retailers can also utilize machine learning insights to refine their marketing strategies. Advanced models like XGBoost and LightGBM, while complex, offer tools such as SHAP (SHapley Additive exPlanations) values to explain model predictions [5]. By understanding the contribution of each feature to the model's output, retailers can identify the most effective marketing channels, customer segments, and promotional tactics.

For example, SHAP values can reveal which customer attributes (e.g., demographics, purchase history) significantly impact the likelihood of responding to a marketing campaign. This enables retailers to tailor their marketing efforts to target high-value customers with personalized promotions, increasing the effectiveness of their campaigns and boosting customer engagement.

Additionally, partial dependence plots and feature importance analyses can help retailers understand the relationship between marketing spend and sales outcomes [5]. By identifying the optimal allocation of marketing resources across different channels, retailers can maximize their return on investment and enhance overall marketing efficiency.

The practical implications of the study's findings are therefore profound for retail businesses. By harnessing the interpretability and predictive power of machine learning models, retailers can optimize inventory management and enhance marketing strategies. These insights enable retailers to make informed decisions, reduce operational costs, and improve customer satisfaction, ultimately driving business growth and competitiveness in the dynamic retail landscape.

6.9 Potential ethical considerations in data collection, model deployment, and decision-making

The implementation of machine learning models in retail businesses brings numerous benefits, but it also raises several ethical considerations. These considerations span data collection, model deployment, and decision-making processes, and addressing them is crucial to maintain trust and ensure fair practices.

6.10 Data collection

6.10.1 Privacy concerns

Retailers often collect vast amounts of customer data, including purchase histories, browsing behaviors, and personal information. Ensuring customer privacy and adhering to data protection regulations such as GDPR and CCPA is essential. Unauthorized use or inadequate protection of personal data can lead to breaches of privacy and loss of customer trust [17].

6.10.2 Bias in data

Collected data may contain biases that, if not addressed, can lead to unfair outcomes. For instance, historical sales data might reflect existing prejudices or inequalities, which, when used to train models, can perpetuate these biases. Ensuring that data collection methods are unbiased and representative is critical to developing fair models [18].

6.11 Model deployment

6.11.1 Algorithmic bias

Machine learning models can inadvertently perpetuate or exacerbate biases present in the training data. For example, a model predicting customer preferences might favor certain demographic groups over others, leading to discriminatory practices. Regularly auditing models for bias and implementing fairness constraints can mitigate this risk [19].

6.12 Transparency and accountability

Retailers must ensure that their models are transparent and interpretable. Stakeholders should be able to understand how decisions are made and who is accountable for these decisions. Black-box models, which are difficult to interpret, can lead to ethical dilemmas, especially when decisions have significant impacts on individuals [20].

6.13 Impact on employment

Automation through machine learning can lead to workforce reductions, affecting employees' livelihoods. Retailers should consider the societal impact of deploying such technologies and explore ways to retrain and redeploy affected employees to mitigate negative consequences [21].

6.14 Decision-making based on model predictions

6.14.1 Fairness in decision-making

Decisions based on model predictions should be fair and equitable. For instance, in marketing, ensuring that all customer segments are treated fairly and not excluded based on biased model outputs is crucial. Establishing fairness criteria and regularly reviewing decision-making processes can help achieve this [22].

6.14.2 Informed and ethical use of predictions

Retailers should use model predictions ethically, ensuring that decisions made do not harm customers or exploit their data unfairly. For example, dynamic pricing models should not unfairly target vulnerable customers with higher prices. Ethical guidelines and oversight can help prevent such practices [23]

6.14.3 Customer impact and autonomy

Decisions that significantly impact customers, such as credit approvals or personalized pricing, should be made with consideration of their potential effects. Providing customers with the ability to understand and challenge decisions made by models respects their autonomy and promotes ethical standards [24]

Addressing ethical considerations in data collection, model deployment, and decision-making is vital for maintaining trust and fairness in retail businesses. By implementing transparent, accountable, and fair practices, retailers can leverage the benefits of machine learning while upholding ethical standards and ensuring positive outcomes for all stakeholders.

7 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.

Our exploration into predictive modeling for Walmart sales therefore 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.

To summarize, sales prediction modeling encompasses several technical aspects crucial for accurately forecasting future sales based on historical data and relevant factors. First, data collection and preprocessing are essential steps in gathering and cleaning historical sales data, ensuring its quality and reliability [25]. Feature engineering plays a vital role in creating predictive variables that capture factors influencing sales, such as product attributes, pricing dynamics, seasonality, and external factors like economic indicators and promotional activities [26]. Model selection involves choosing appropriate techniques based on the data characteristics and business requirements, ranging from traditional time series forecasting methods like ARIMA to machine learning algorithms such as random forests or gradient boosting machines [27]. During model training, historical data is split into training and validation sets, and models are trained to minimize prediction errors using optimization techniques [28]. Model evaluation relies on metrics like mean absolute error and root mean squared error to assess predictive accuracy and compare model performance [29]. Hyperparameter tuning further refines models for optimal performance, while model deployment involves integrating them into production environments for real-time predictions and continuous monitoring and maintenance [30]. By leveraging these technical aspects, organizations can develop robust sales prediction models that provide valuable insights for decision-making and strategic planning in retail and other industries [31].

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.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that generative AI technology like ChatGPT has been used during writing or editing of this manuscript. The ChatGPT, version GPT-4 was used, and it is based on the GPT (Generative Pre-trained Transformer) model developed by OpenAI. The source of this technology is OpenAI, a research organization focused on artificial intelligence. This generative AI technology was utilized to assist in drafting and refining the text. The use of this AI tool included providing language suggestions and ensuring coherence and clarity throughout the manuscript.

Prompts used included the following.

1. Revise the passage to enhance clarity and coherence.
2. Identify areas where the language can be polished for better readability.
3. Refine the wording to ensure precision and accuracy.
4. Strengthen transitions between ideas for smoother flow.
5. Check for any ambiguities or inconsistencies and clarify them.
6. Consider alternative phrasing to convey the same meaning more effectively.
7. Eliminate unnecessary repetition and redundant phrases.
8. Ensure the tone remains consistent throughout the text.
9. Look for opportunities to vary sentence structure to maintain engagement.
10. Proofread for grammatical errors and typos to enhance overall quality.

Consent

Customers should be fully informed about the types of data being collected and how it will be used. Transparent data collection practices, including obtaining explicit consent from customers, are necessary to ensure ethical standards. Failure to do so can lead to ethical breaches and potential legal consequences.

Competing Interests

Authors have declared that they have no known competing financial interests or non-financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Iwok UU, Udofia KM, Obot AB, Udofia KM, Michael UA, Kingsley AI. Evaluation of Machine Learning Algorithms using Combined Feature Extraction Techniques for Speaker Identification. J. Eng. Res. Rep; 2023.
Accessed on 2024 May 25];25(8):197-216.
Available:https://journaljerr.com/index.php/JERR/article/view/973

[2] J. Pierre SM, Luz MN, Camile LB, Andre BN, Telesphore NL, Cedric KT, Boniface EW. Model of a Neural Network for Solving Systems of Inequalities with Three Real Unknowns . J. Adv. Math. Com. Sci; 2023.

Accessed on 2024 May 25];38(9):51-64.
Available:https://journaljamcs.com/index.php/JAMCS/article/view/1804

[3]   Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas MJ. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews. 2015;71:804-18.

[4]   Kaggle. Walmart sales data; 2024.
Available:https://www.kaggle.com/code/yasserh/walmart-sales-prediction-best-ml-algorithms/input

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

[6]   Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

[7]   Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29(5):1189-1232.

[8]   Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu TY. LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems. 2017;3146-3154.

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

[10]   Cyril Neba C, Gerard Shu F, Adrian Neba F, Aderonke Adebisi P, Kibet F, Webnda; Philip Amouda A. 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). 2023;8:9.
Available:www.ijisrt.com. ISSN - 2456-2165 :1841-1866.
Available:https://doi.org/10.5281/zenodo.8413849

[11]   Cyril Neba C, Gerard Shu F, Adrian Neba F, Aderonke Adebisi P, Kibet F, Webnda; Philip Amouda A. 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). 2023;8(9):1867-1884.
Available:www.ijisrt.com. ISSN - 2456-2165 ,
Available:https://doi.org/10.5281/zenodo.8414044.

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

[13]   Cyril Neba C, Gillian Nsuh, Gerard Shu F, Philip Amouda A, Adrian Neba F, Aderonke Adebisi P, Kibet, F.Webnda. 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). 2023;8(10):636-647.
Available:www.ijisrt.com. ISSN - 2456-2165,
Available:https://doi.org/10.5281/zenodo.10040460.

[14]   Chen J, Kaur R, Li C. A seasonal decomposition approach to sales forecasting with application in; 2018.

[15]   Lundberg,SM, Lee SI. A unified approach to interpreting model predictions. In Advances in neural information processing systems. 2017;4765-4774.

[16]   Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.

[17]   DalleMule L, Davenport TH. What's your data strategy? Harvard Business Review. 2017;95(3):112-121.

[18]   Solove DJ. Introduction: Privacy self-management and the consent dilemma. Harvard Law Review. 2013;126:1880-1903.

[19]   Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning. fairmlbook.org; 2019.

[20]   Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys. 2021;54(6):1-35.

[21]   Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning; 2017. arXiv preprint arXiv:1702.08608.

[22]   Verma S, Rubin J. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). 2018;1-7.

[23]   Martin KE. Ethical implications and accountability of algorithms. Journal of Business Ethics. 2019;160(4):835-850.

[24]   Citron DK, Pasquale FA. The scored society: Due process for automated predictions. Washington Law Review. 2014;89(1):1-33.

[25]   Delen D, Kuzey C, Uyar A. A comprehensive review of data mining-based fraud detection research. Expert Systems with Applications. 2019;109:1-20.

[26]   Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2018.

[27]   Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS One. 2018;13(3):e0194889.

[28]   Brownlee J. Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future. Machine Learning Mastery; 2021.

[29]   Montgomery DC, Jennings CL, Kulahci M. Introduction to time series analysis and forecasting. John Wiley & Sons; 2015.

[30]   Provost F, Fawcett T. Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media; 2013.

[31]   Bessen J. AI and Jobs: The role of demand. NBER Working. 2019;24235.

_____