

WALMART SALES PREDICTION

Team Members:

A Hari priya-20BDS0352

Amogh A M -20BDS0172

K Bindhu sree- 20BDS0356

Vennela G-20BDS0146

1. Introduction

Sales forecasting is a critical process for businesses to make informed decisions and anticipate their short-term and long-term performance. This project aims to develop a predictive model for estimating future sales in Walmart stores and analyzing the impact of holidays on sales. We aim to develop a predictive model using machine learning algorithms such as regression, Random Forest, and Decision Trees. We try to create a strong model that predicts sales for Walmart locations by training and evaluating the data using several techniques. In addition, we examine the impact of vacations to offer insightful planning and optimization advice for promotional markdown events. Walmart may make data-driven decisions to improve operational efficiency and maximize profitability by comprehending the effects of these vacations. We incorporate the Flask web framework to guarantee usability and accessibility, allowing us to create a user-friendly interface for accessing and using the prediction model. We seek to offer useful insights that direct strategic decision-making and contribute to Walmart's performance in the fiercely competitive retail sector by utilizing past sales data and taking holiday impacts into account.

2. Literature Survey

Every retailing organization understands the need of using time series data to forecast future product sales fluctuations in terms of resource management and planning. This article examines the viability of traditional time series models, hybrid models based on time series model and machine learning model, and machine learning model in predicting Walmart sales to find an effective method to improve the accuracy of sales forecasting of retail goods which are strongly influenced by season and holiday. Walmart grocery sales data from 2011-01-29 to 2016-06-19 are used to train and test the Prophet model. The findings reveal that the Prophet model and the LightGBM model's Root Mean Square Errors (RMSE) are, respectively, 0.694 and 0.617, demonstrating that the machine learning model performs well in the sales forecasting of retail stores[1].

In the current digital age, anticipating Walmart's gross sales utilizing machine learning analysis and making sales predictions is a breath-taking undertaking. One option to handle the situation is to use machine learning to analyze Walmart's dataset and forecast future sales, which is the most crucial component of strategic planning. Machine learning techniques are employed in this situation to predict new results by using historical data as input. Therefore, in our projects, future sales are forecasted using these algorithms by selecting the method that provides the best accuracy and then analyzing by adding new variables that forecast future sales [2].

A key component of contemporary commercial sales issues is sales projection. Dynamic sales challenges can be solved using machine learning, particularly supervised machine learning algorithms, which can identify complicated and unpredictable trends as well as a variety of potentially significant variables. This study uses three alternative regression models (Multiple Linear Regression, Elastic-Net Regression, and Polynomial Regression) to forecast Walmart's future weekly sales. Additionally, statistical measurements (such R^2 and RMSE) are used to assess the model's quality. Holiday, Date, Type, and Stores are pertinent variables that clearly had an

impact on weekly sales, according to the data. The simplest Multiple Linear Regression Model, which has a moderate R²-Score of around 0.933 and the RMSE with the smallest change between the training and test sets, generates the best sales predictions in terms of the prediction model. Overall, these findings provide guidance for selecting acceptable models for sales prediction and offer advice to retail businesses for developing sales strategies [3].

These days, a key technology in the retail sector is sales forecasting. Business owners can properly forecast the sales of thousands of products and base their decisions on them thanks to sophisticated machine learning and deep learning algorithms. In this study, it is suggested to make use of a CatBoosting-based sales forecasting system. The Walmart sales dataset, by far the largest dataset in this industry, is used to train the algorithm. We successfully engineered features to increase prediction speed and accuracy. Our model achieves an RMSE of 0.605 in the experiments, outperforming established machine learning techniques like Linear Regression and SVM. Our method requires less fine-tuning than existing methods, which enhances its capacity to generalize to other custom datasets and broadens its potential applications [4].

In the modern world, machine learning is revolutionizing every industry. One can increase the revenues of business-to-consumer (B2C) models incorporating retail chains by projecting or forecasting sales. This study looks at several areas of data-driven marketing trade prediction. The findings of this study will be helpful to store managers by giving a thorough comparison of several forecasting approaches for estimating retail chain sales. In this study, forecasting models that can analyze the seasonality of sales are being compared. It is done and analyzed to compare several regression models, including linear regression, k-nearest neighbor, decision tree regression, random forest regression, and XGB regression [5].

The retail sector is essential both during and after an epidemic. Management teams must grasp precise sales predictions and comprehend the more important aspects if firms and outlets are to grow. The Random Forest Regressor model is used in this study to forecast the weekly sales with a high degree of accuracy using data from Walmart's weekly sales[6].

Businesses can create acceptable sales strategies with the aid of accurate sales projections. I'm going to develop a sales forecasting model in this essay. I must first preprocess the historical data, identify any feature structures that are obvious, and remove any data that is irrational. In contrast to SVM and linear regression models, the LightGBM model I present has a lower Root Mean Square Error of 2.09 and stronger predictive power[7].

Accurate sales forecasting can boost manufacturing productivity, increase production efficiency, and boost industry competitiveness for businesses. The goal of this essay is to realize the long-term sales projection for WalMart using the LSTM and LightGBM models. We also performed feature engineering processing on the data, combining abnormal data and extracting data features to obtain processed data that is convenient for modeling, and then used the LSTM model to learn and forecast sales because the amount of data provided by the materials is enormous and inconsistent. The proposed model in this research has a lower RMSSE than conventional linear regression models and SVM models, according to experiments, and is more predictive[8].

Exact sales forecasting is essential to modern retail firms that run a massive chain of businesses since it determines the growth and success of the company.

Businesses may efficiently allocate resources, such as cash flow and production, and create more informed business plans, thanks to sales forecasting. In this research, we provide a machine learning-based methodology for accurate and efficient sales forecasting.

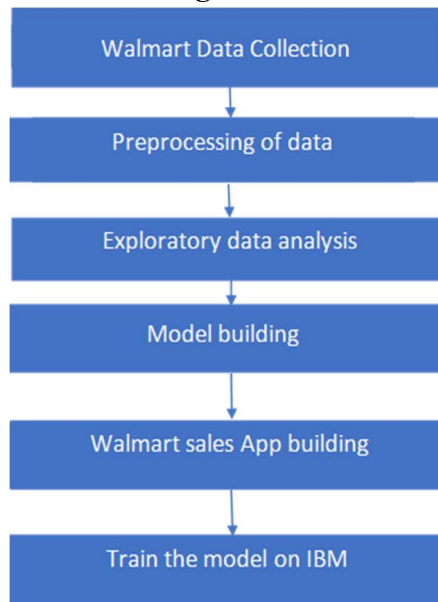
For extracting features from historical sales data, feature engineering is first carried out. Furthermore, we made use of these attributes to predict future sales volume using eXtreme Gradient Boosting (XGBoost). The experiment's findings on a publicly available Walmart retail products dataset provided by the Kaggle competition show that our suggested model performs incredibly well for sales prediction while utilizing less memory and processing power[9].

The application of sequence modeling to problems like speech recognition, time series forecasting, and context identification has demonstrated great promise.

The Walmart dataset has been used to train a variety of sequence models, including vanilla LSTM, stacked LSTM, bidirectional LSTM, and convolution neural network-based-LSTM. A comparison of the models' performance using mean squared error (MSE) and weighted mean absolute error (WMAE) metrics is reported. Exploratory data analytics has been used to identify pertinent aspects for the multivariate training of the learners. The Local Interpretable Model Agnostic Explanation (LIME) model is used to explain away the crucial variables involved in the prediction job, which further makes these sequence models interpretable. The performance of the stacked LSTM model is better than other learners, according to empirical findings using the Walmart sales dataset. Additionally, the LIME module complements the stacking model, which is the most generalizable, by providing an explanation for its predictions based on pertinent variables[10].

3. Theoretical Analysis

3.1. Block diagram



The entire process involves collecting Walmart sales data and pertinent external data like economic indicators, weather data, and competitor information, removing missing numbers, outliers, and discrepancies from obtained data, transforming, normalizing, and combining data sources, understanding data patterns, trends, and seasonality through statistical analysis and visualization, discovering sales drivers, choosing a forecasting model (time series, machine learning) based on data properties and project needs. Historical data training is to be done for accurate sales forecasting and a sales prediction tool or platform is created using the trained model. This provides real-time insights, visualizations, and reports for decision-making and company planning. The model is trained with IBM's Watson Studio or Watson Machine Learning to improve training and model performance.

3.2. Software Design

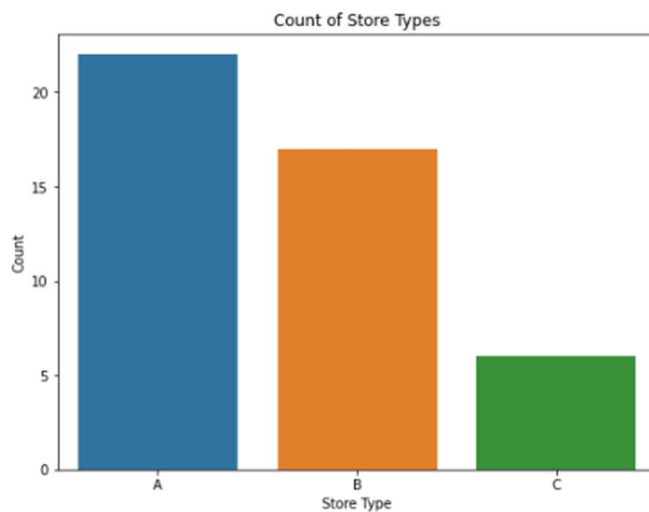
Some of the software requirements for the Walmart sales project includes Anaconda navigator software with python packages numpy, pandas, scikit-learn, matplotlib, seaborn, pmdarima and Flask micro web framework to develop web application.

4. Experimental Investigations

These are some of the analysis of data during experimentation:

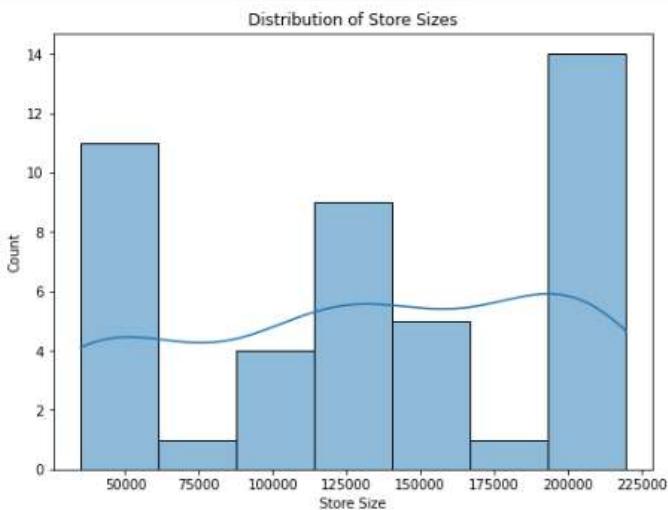
Countplot of store types:

```
# Countplot of store types
plt.figure(figsize=(8, 6))
sns.countplot(data=stores, x='Type')
plt.title('Count of Store Types')
plt.xlabel('Store Type')
plt.ylabel('Count')
plt.show()
```



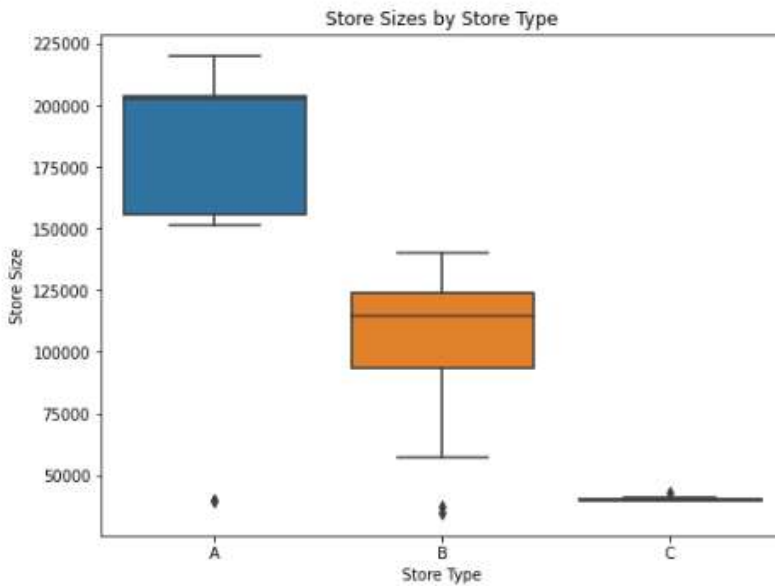
Distribution of store sizes:

```
# Distribution of store sizes
plt.figure(figsize=(8, 6))
sns.histplot(data=stores, x='Size', kde=True)
plt.title('Distribution of Store Sizes')
plt.xlabel('Store Size')
plt.ylabel('Count')
plt.show()
```



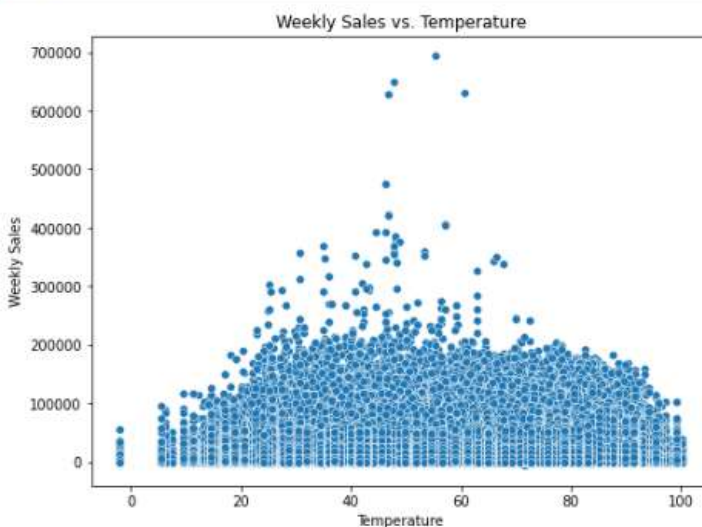
Boxplot of store sizes by store type:

```
# Boxplot of store sizes by store type
plt.figure(figsize=(8, 6))
sns.boxplot(data=stores, x='Type', y='Size')
plt.title('Store Sizes by Store Type')
plt.xlabel('Store Type')
plt.ylabel('Store Size')
plt.show()
```



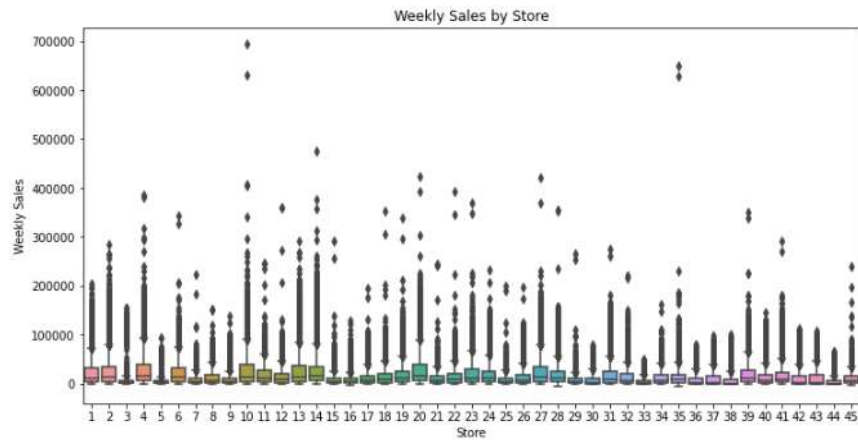
Scatter plot of weekly sales vs temperature:

```
# Scatter plot of weekly sales vs. temperature
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='Temperature', y='Weekly_Sales')
plt.title('Weekly Sales vs. Temperature')
plt.xlabel('Temperature')
plt.ylabel('Weekly Sales')
plt.show()
```



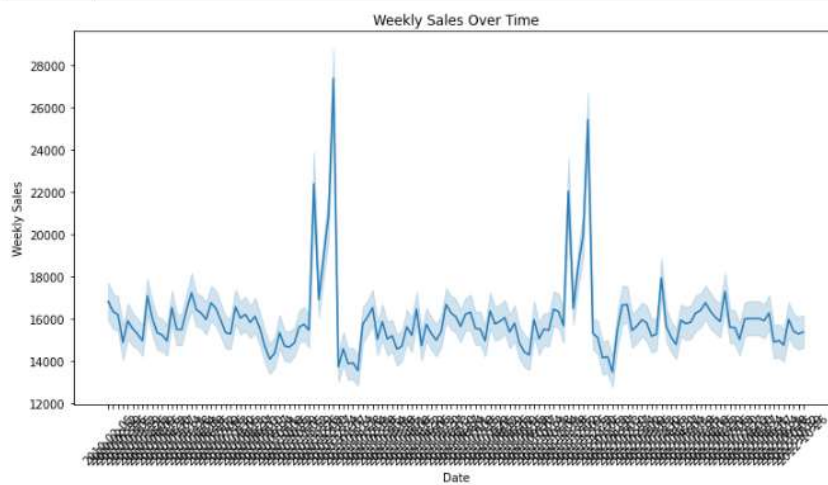
Boxplot of weekly sales by store:

```
# Boxplot of weekly sales by store
plt.figure(figsize=(12, 6))
sns.boxplot(data=train, x='Store', y='Weekly_Sales')
plt.title('Weekly Sales by Store')
plt.xlabel('Store')
plt.ylabel('Weekly_Sales')
plt.show()
```



Line plot of weekly sales over time:

```
# Line plot of weekly sales over time
plt.figure(figsize=(12, 6))
sns.lineplot(data=df, x='Date', y='Weekly_Sales')
plt.title('Weekly Sales Over Time')
plt.xlabel('Date')
plt.ylabel('Weekly_Sales')
plt.xticks(rotation=45)
plt.show()
```



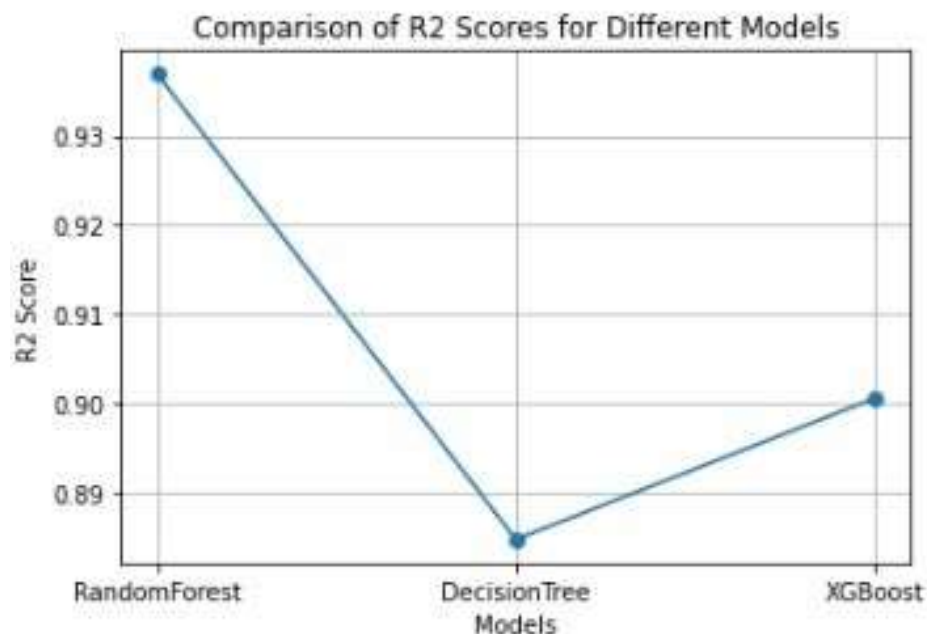
Numerous techniques have been shown to perform well when evaluating machine learning models for Walmart datasets. In the Walmart datasets, common machine learning techniques including random forests, decision trees, linear regression, and XGBoost have been employed to forecast sales.

On comparing all these four models we observe that the random forest is the optimal model for the Walmart project, since it has 0.93 as the R2 score.

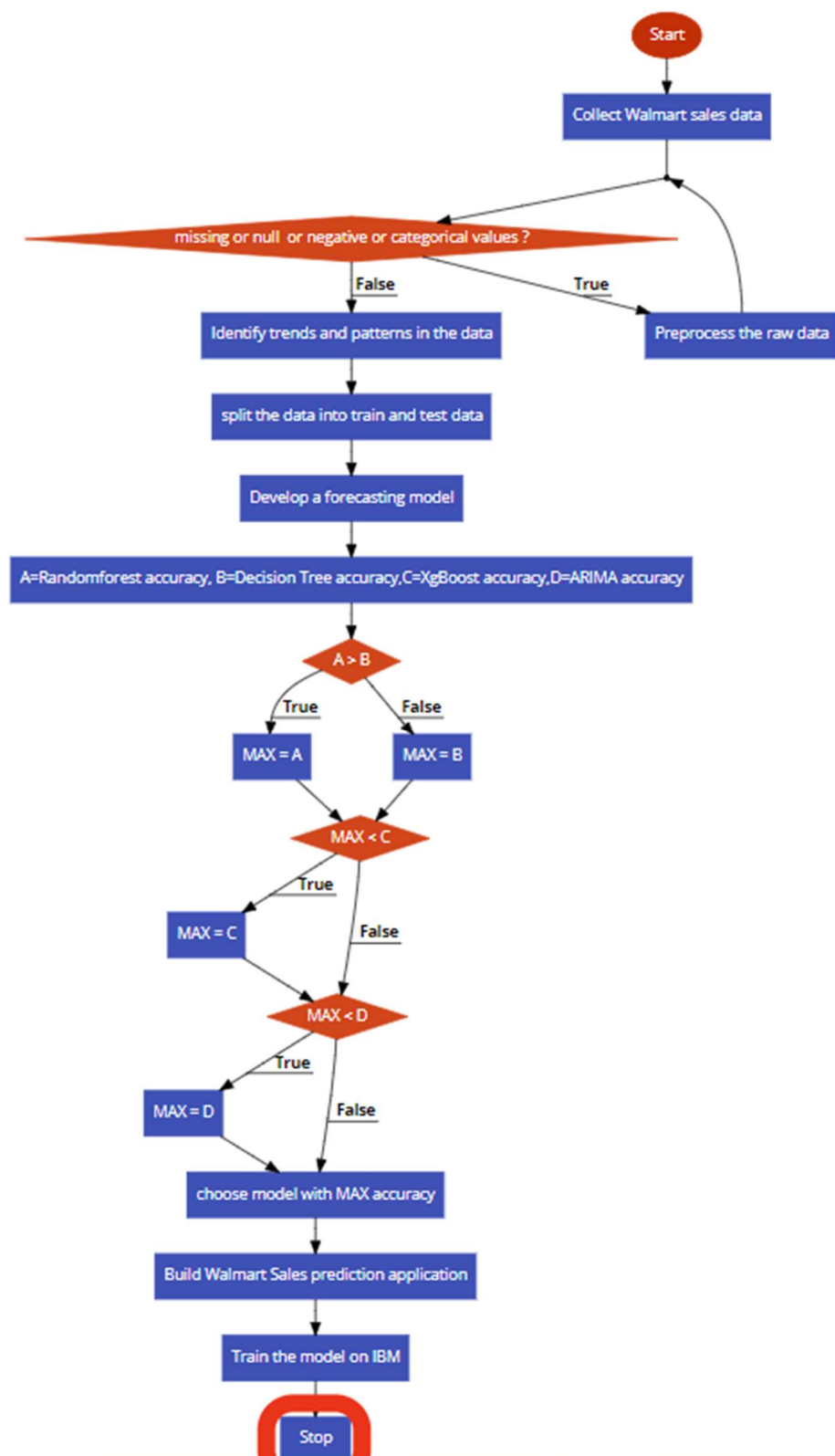
Comparison table:

Name of the model used	R2 score	RMSE(Root mean square error)
Linear Regression	-14.75985023617903	21977.570095745654
Decision Tree	0.8861117397424264	7778.514498701807
Random Forest	0.9373547251134687	5565.621005092228
XGBoost	0.9006349040489972	6752.763224191785

Comparing Random Forest, Decision Tree and XGBoost using Line Plot:



5. Flowchart



6. Result

Comprehensive data analysis was performed and thereby the more important features were identified and various machine learning models were trained upon the 'Walmart Sales' dataset. The following results were obtained:

Linear Regression:

```
In [86]: lr.score(x_val, y_val)
Out[86]: 0.06149084741656563

In [87]: from sklearn.metrics import mean_squared_error, r2_score
mse = mean_squared_error(y_pred, y_val)
r2 = r2_score(y_pred, y_val)
print('Root Mean Square Error = ', np.sqrt(mse))
print('R2 Score = ', r2)

Root Mean Square Error = 21977.570095745654
R2 Score = -14.75985023617903
```

Decision Tree Regression:

```
In [69]: rms_dt = np.sqrt(mean_squared_error(y_pred_dt, y_val))
r2_dt = r2_score(y_pred_dt, y_val)
print('RMSE of DT = ', rms_dt)
print('R2 Score of DT = ', r2_dt)

RMSE of DT = 7778.514498701807
R2 Score of DT = 0.8861117397424264
```

Random Forest Regression:

```
In [72]: rms_rf = np.sqrt(mean_squared_error(y_pred_rf, y_val))
r2_rf = r2_score(y_pred_rf, y_val)
print('RMSE of RF = ', rms_rf)
print('R2 Score of RF = ', r2_rf)

RMSE of RF = 5565.621005092228
R2 Score of RF = 0.9373547251134687
```

XGBoost :

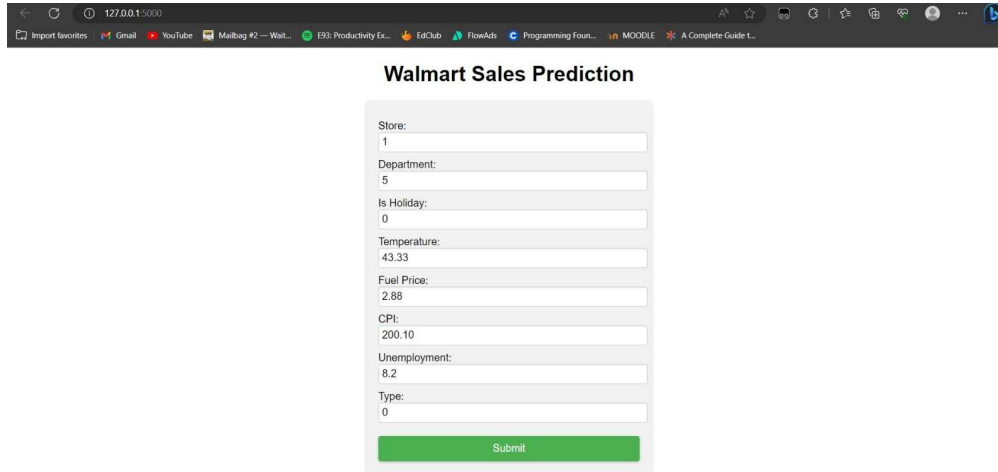
```
In [83]: print('Accuracy:', xg_reg.score(x_val, y_val)*100, '%')

rms = mean_squared_error(y_val, y_pred, squared=False)
print('RMSE:', rms)

r2_xg = r2_score(y_pred, y_val)
print('R2 score of XGBoost:', r2_xg)

Accuracy: 91.139826858775 %
RMSE: 6752.763224191785
R2 score of XGBoost: 0.9006349040489972
```

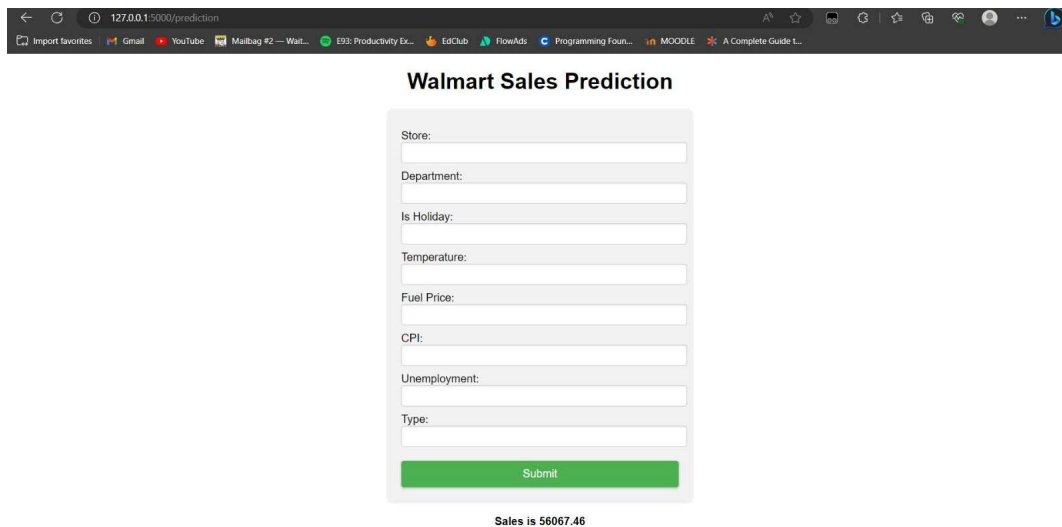
Website Overlay:



The screenshot shows a web browser window with the address bar displaying "127.0.0.1:5000". The page title is "Walmart Sales Prediction". The form contains the following fields and values:

Field	Value
Store:	1
Department:	5
Is Holiday:	0
Temperature:	43.33
Fuel Price:	2.88
CPI:	200.10
Unemployment:	8.2
Type:	0

A green "Submit" button is located at the bottom of the form.



The screenshot shows the same web browser window after the form has been submitted. The page title remains "Walmart Sales Prediction". The input fields are now empty. Below the form, the output is displayed:

Sales is 56067.46

7. Advantages and Disadvantages

Forecasting Walmart sales has several benefits.

1. **Effective Resource Allocation:** By ensuring that the appropriate amount of inventory is accessible at the appropriate moment, accurate sales forecasting aids Walmart in making efficient resource allocation decisions. As a result, operating expenses are optimized, lost sales opportunities are reduced, and overstocking or understocking is avoided.
2. **Better customer service:** Walmart can better satisfy consumer demand and guarantee product availability by precisely estimating sales. Since customers can find the things they require when they visit Walmart stores or shop online, this increases customer happiness and loyalty.

3. **Promotional Planning:** Walmart can better plan and carry out promotional efforts with the help of sales forecasts. Walmart can time promotions to coincide with times of strong demand, boosting sales and maximizing the effects of marketing initiatives by comprehending sales trends and forecasting future demand.
4. **Supply Chain Optimisation:** Walmart is able to optimize its supply chain operations because of accurate sales forecasts. In order to ensure smooth operations, cut costs, and minimize supply chain disruptions, it aids in managing logistics, production, and procurement processes.
5. **Making Strategic Decisions:** Sales forecasting offers insightful information for making strategic decisions. The projections can be used by Walmart to pinpoint expansion prospects, develop expansion plans, decide on prices wisely, and improve store layouts to boost sales.

Forecasting Walmart sales has certain drawbacks.

1. **Data Restrictions:** Since sales forecasting mainly relies on previous data, data restrictions may have an impact on forecast accuracy. Forecasts may be less accurate as a result of elements like missing data, outliers, or modifications in market conditions that are not reflected in the past data.
2. **Uncertainty and Variability:** Due to a variety of factors, including changes in customer behavior, the economy, competition, and outside events, sales forecasting is subject to uncertainty and variability. These uncertainties have the potential to reduce forecast accuracy and increase risk in the decision-making process.
3. **complicated and Dynamic Environment:** With changing consumer preferences, market trends, and competitive landscapes, the retail industry is both highly dynamic and complicated. Accurate predicting in such a setting can be difficult and necessitate constant monitoring, adjusting, and adapting of forecasting models.
4. **Reliance on Historical Patterns** Too much sales forecasting may fail to account for new trends, consumer preferences, or quick changes in the market. To improve the precision of projections, it is crucial to supplement analysis of past data with market information, customer insights, and other outside elements.
5. **Interpretability Issues:** Complex machine learning algorithms, particularly those used in forecasting models, may be difficult to understand. Stakeholders may find it challenging to accept the predictions or base their actions on the predicted results if they are unable to comprehend the underlying causes affecting the projections.

Overall, while Walmart sales forecasting has many benefits for allocating resources, providing for customers, and making strategic decisions, it also has issues with data limits, unpredictability, and interpretability. Robust data analysis, sophisticated modeling methods, domain knowledge, continuous monitoring, and predictive process adaptability are all necessary to meet these challenges.

8. Applications

Optimisation of Promotional Markdown Events: The conclusions drawn from the investigation can help Walmart make the most of their promotional markdown activities. The business may strategically plan and allocate resources for these occasions, assuring optimal efficacy and return on investment, by analyzing the influence of holidays on sales.

Inventory Management: Walmart may streamline their inventory management procedures with the help of accurate sales forecasting. The organization can avoid stockouts and excess inventory by anticipating sales trends and knowing the influence of holidays on inventory levels. This results in increased productivity, lower expenses, and more customer satisfaction.

Resource Allocation: Walmart can allocate resources wisely by following the advice provided by the sales forecasting model. The business can deploy resources like personnel, marketing initiatives, and operational assistance based on realistic sales forecasts.

9. Conclusion

In conclusion, this study examines how several regression models might be used to forecast Walmart sales. First off, the Kaggle platform's initial data set has been cleaned. Consequently, it is discovered through the comparison of individual attributes that Size, Date, Store, and other features have various impacts on weekly sales. Then, three different regression models are used to predict sales, and it was discovered that the Multiple Linear Regression Model, which has the moderate R^2 values and the RMSE with the smallest difference between the training and test sets, performed better than the more complicated Elastic-Net Regression Model and Polynomial Regression Model. More predictive models will be tested in the future, and the new models will be used to solve additional sales issues. This study has important ramifications for how merchants organize their inventories and sales. It is advised that businesses and managers completely comprehend the findings made by specialists in recognising potential sales issues, educate their staff on the needs of the market, and incorporate their expertise into the decision-making process. Overall, these results provide a framework for selecting relevant models for sales prediction problems and give Walmart and similar organizations advice on how to create sales strategies based on influencing factors.

10. Future Scope

There are a several ways to improve Walmart's analysis .This involves including extra variables like weather conditions and competitor data, investigating cutting-edge machine learning techniques like neural networks, customizing holiday weighting based on specific holidays, taking into account dynamic effects and lead-up time, integrating real-time data for up-to-date forecasting, conducting a comparative analysis of algorithms, deploying the model in a live environment for evaluation, and continuously improving and expanding the p These initiatives will help Walmart further improve its decision-making, marketing tactics, and overall performance.

11. Bibliography

- [1] Jiang, H., Ruan, J., & Sun, J. (2021, March). Application of machine learning model and hybrid model in retail sales forecast. In *2021 IEEE 6th international conference on big data analytics (ICBDA)* (pp. 69-75). IEEE.
- [2] Mounika, S., Sahithi, Y., Grishmi, D., Sindhu, M., & Ganesh, P. (2021). Walmart Gross Sales Forecasting Using Machine Learning. *Journal of Advanced Research in Technology and Management Sciences (JARTMS)*, 3(4), 22-27.
- [3] Chen, Z. (2023, April). Sales Forecast of Walmart on Account of Multivariate Regression and Machine Learning Methods. In *Proceedings of the International Conference on Financial Innovation, FinTech and Information Technology, FFIT 2022, October 28-30, 2022, Shenzhen, China*.
- [4] Ding, J., Chen, Z., Xiaolong, L., & Lai, B. (2020, December). Sales forecasting based on catboost. In *2020 2nd international conference on information technology and computer application (ITCA)* (pp. 636-639). IEEE.
- [5] Baby, A., & Newniz, A. K. (2023, May). Performance Analysis and Evaluation of Regression Models for Sales Forecasting. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 227-232). IEEE.
- [6] Pang, S. (2022, October). Retail Sales Forecast Based on Machine Learning Methods. In *2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA)* (pp. 357-361). IEEE.
- [7] Qiao, Z. (2020, October). Walmart Sale Forecasting Model Based On LightGBM. In *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 76-79). IEEE.
- [8] Xie, H. H., Li, C., Ding, N., & Gong, C. (2021, January). Walmart Sale Forecasting Model Based On LSTM And LightGBM. In *2021 2nd International Conference on Education, Knowledge and Information Management (ICEKIM)* (pp. 366-369). IEEE.
- [9] Shilong, Z. (2021, January). Machine learning model for sales forecasting by using XGBoost. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)* (pp. 480-483). IEEE.
- [10] Narang, R., & Singh, U. P. (2023, May). Interpretable Sequence Models for the Sales Forecasting Task: A Review. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 858-864). IEEE.

APPENDIX

Source code Link: <https://github.com/Bindhu2708/Walmart-sales-prediction>