International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

End-to-End Predictive Analytics Pipeline of Sales Forecasting in Python for Business Decision Support Systems

Adarsh Reddy Bilipelli*

Independent Researcher

Received 01 Dec 2022, Accepted 20 Dec 2022, Available online 21 Dec 2022, **Vol.12, No.6 (Nov/Dec 2022)**

Abstract

In the modern world economy where competition becomes the dominant element, the importance of sales forecasting in business strategy development, inventory management, and resource cannot be overestimated. Appropriate sales forecasting is critical in improving inventory management, the forecast the demands and strategic planning in the retail business. This study aims to attain a higher level of accuracy in sales projections by examining data-driven methods using state-of-the-art machine learning (ML) techniques applied to the Walmart dataset. The methodology is followed as an extensive data preprocessing method such as processing missing data and the deletion of outliers towards the integrity of data. Changes that would be of significant importance, like data normalization, label encoding of categorical values, and feature engineering would be performed to improve the quality of model input. Because of this efficiency and good predictive power on structured data, the XGBoost algorithm is utilized. Model evaluation is carried out using the standard regression metrics, such as the coefficient of determination (R^2) and root mean square error (RMSE), with results of 0.946 and 21.77, respectively, to assess the correctness and reliability. It is seen that a comparative analysis against those traditional models showcases, how the proposed approach has a greater forecasting capability, which is a useful tool to be used in support of data-driven decisions within a retail setting.

Keywords: *Sales Forecasting, Walmart Dataset, Machine Learning (ML), Deep learning (DL), Regression Metrics, Business Decision Support.*

1. Introduction

Sales forecasting today has become an important element of business strategy and operational planning in the current period of data-driven decisions. Proper sales forecasting helps organizations to manage inventory, develop the supply chain process, financial planning and to maximize customer satisfaction [1]. An end-to-end predictive analytics pipeline is provided to sales forecasting using Python, including advanced ML methods in the context of Enterprise Resource Planning (ERP) applications. This is in response to the fact that traditional statistical methods of data forecasting have become insufficient in the face of the exponential growth in the amount of transactional and behavioural data, as well as the use of complex patterns and seasonal trends.

Business analytics is an essential part of a contemporary decision support system with the development of data engineering and analytics [2]. Such analytics are important in demand and sales forecasting, which makes Sales and Operations Planning (S&OP) as efficient as possible. Advancement in e-commerce and logistics industries have largely changed the velocity and sophistication of supply chain demand [3]. By estimating future product demand, precise sales forecasting enables manufacturers and retailers to make the right sets of decisions concerning marketing, procurement, production and supply chain management activities of the business.

Traditional forecasting techniques, founded on extrapolations of historic records on sales and simple mathematical models [4], commonly have significant error rates because they fail to consider organizational complexities and external factors [5]. The new frontiers in the field of ML have initiated the replacement of the static models by dynamic models that are data-driven and learn based on the history of previous sales data [6]. The methods that have proven to be good at making sales predictions are techniques like regression, decision trees, neural networks and the ensemble models. Moreover, ML models can incorporate external factors like economic fluctuations, the competition in the market, the dynamics of

*Corresponding author's ORCID ID: 0000-0000-0000-0000
DOI: https://doi.org/10.14741/ijcet/v.12.6.17

customers, and the social media sentiment analysis, making the models more robust and giving high accuracy to the predictions.

Machine learning as one of the constituents of Business Intelligence (BI) helps to achieve the tasks of determining future trends of the market, as well as, to make strategic decisions. Demand forecasting which is conducted by determining the future need of a product or service by taking past and present data has also become an essential business feature in the competitive business world [7]. Production, distribution, or retailing organizations stand to gain a lot by proper sales forecasting in a timely and an accurate manner which helps them to maximize the use of resources in the planning, marketing, and managing aspects of an organization [8]. In that connection (as shown in Figure 1), machine learning allows computational systems to react to managerial decisions in order to analyze past statistics and forecast future performance [9].

Fig.1 Word Cloud for Sales Forecasting

The objective of the proposed approach is to design an end-to-end predictive analytics pipeline using Python for accurate sales forecasting within a business decision support system. Sales data, such as that from retail or grocery environments, is analyzed to identify patterns and forecast future quantities based on key predictive features[10]. Accurate forecasting enhances an enterprise's ability to make informed, probabilistic decisions throughout various stages of strategic and operational planning. ML thus serves as a pivotal tool for advancing sales forecasting capabilities in modern data-driven enterprises.

Motivation and Contribution of Paper

The motivation behind this study stems from the growing need for accurate and data-driven sales forecasting models in the retail sector to support informed business decision-making. The conventional forecasting models have been inadequate in processing huge and sophisticated data, thus narrowing down the scope of their predictability. As the semi-structured retail data has increased, as with the Walmart dataset, there is a subsequent requirement to maximize the potential of the already defined machine learning procedures, particularly XGBoost, that have a higher degree of accuracy and overall scale. The goal of this study is to create an effective forecasting system that should be able to handle the quality problem of data, uncover complex patterns by utilizing feature engineering, and deliver relevant predictions that can inform inventory control, supply chain improvement, and trade strategy in retailing business.

The paper suggests a clear and straightforward list of steps to develop the sales forecasting pipeline, including raw data gathering through interpretation of the results of sales just to the Walmart dataset.

The selected methodology includes the vital pre-processing stages: missing values treatment, outliers removal, data normalization, label encoding, and feature engineering to provide data sanity and model preparation.

The application of the XGBoost algorithm provides a scalable and high-performance approach to forecasting, demonstrating its suitability for large-scale retail data analysis.

The 80:20 training: testing ratio ensures that the model is rigorously tested and trained without any bias.

The application of the XGBoost algorithm takes into consideration the knack of the algorithm to handle structured data effectively and provide high predictive precision during sales forecasting applications.

The R^2 and RMSE are used to evaluate the effectiveness of the model and provide a rather quantitative determination of the accuracy and generalizability

A comparative analysis is carried out to benchmark the proposed model with conventional approaches with the point highlighting its superiority in terms of the accuracy of forecasting.

Justification and Novelty

The uniqueness of this research is associated with the combination of Convolutional Neural Network (CNN) similar to the intrusion detection of network with the Self-Adaptive Differential Evolution compatible with hyperparameter optimization. Such CNNs have shown great potential in processing of images but the use of CNNs in network intrusion detection is less studied. The proposed research solves the problem of model optimization by applying nature-inspired optimization algorithm, namely SADE to tune CNN hyperparameters, increasing its performance compared to classical machine learning algorithms. The justification for this approach is grounded in the increasing complexity of modern cyberattacks, where traditional signature-based detection methods often fail to recognize new and evolving threats. The proposed CNN-based method, combined with SADE optimization, offers a more adaptive and precise solution for identifying intrusions, with the potential for improved detection accuracy, reduced false positives, and greater scalability in real-world applications.

Structure of paper

The following is the paper's outline An analysis of the literature on sales prediction in Python as a business intelligence tool is presented in Section II. Section III gives the suggested method, which comprises data, model execution, experimental outcomes, and conclusions following Section IV's extensive description. As a final section, Section V discusses the study's limitations and suggests avenues for further research.

Literature Review

The most recent study on lightweight data in sales forecasting using Python for business support systems is compiled in this survey of the literature. Table I provides an overview of current studies, emphasising methods, results, main conclusions, difficulties and constraints, and further research.

Chen et al. (2021) provided Walmart with a neural network sales forecast algorithm. In addition, it tests NN models using datasets made available by the Kaggle site. When tested against competing ML models, NN model consistently outperforms them. They fare better than the Linear Regression and SVM algorithms in terms of RMSE, which are 2.92 and 2.58 points lower, respectively. In addition, NN model is interpreted using SHAP to provide accurate attribute mining across all dimensions for effective prediction. Successful sales in today's customer-focused business environment necessitate a delicate balancing act between meeting customer demand and cutting inventory expenses. Businesses can gain a lot from having reliable sales forecasts because they help them increase [11].

Zamil et al. (2021) The previous observations-based transaction data set file is what it reads and writes, so you don't need to worry about how it works. One feature variable that alters the linear regression model's sales forecasts is television media. An output for the OLS model type's regression and a set of coefficients for use in linear regression are provided. You can make charts and plots that you can see with Python's Seaborn tool. A big part of running a shop is figuring out what the sales. Using ML to make intelligent predictions about the future can help find the set of feature factors that affect guesses about sales growth [12].

Calixto and Ferreira (2020) A multinational products moving organisation provided the authors with 594 salespeople, and they utilised a Naive Bayes model to sort them into pre-established categories. The dataset includes sales data from nearly three years. Poor, Good, and Outstanding are the three categories used for the classification. Classify ourselves according to key performance indicators (KPIs) such as growth rate and percentage, annual sales variability, opportunities created, customer base line, target attainment, and many more. The authors used a confusion matrix in conjunction with other measures, such as True Positives, True Negatives, and the F1 score, to evaluate the model's performance. A total of 92.50% accuracy was attained by the model [13].

Niu (2020) suggests a model for predicting Walmart sales that makes use of the XGBoost algorithm and carefully processes features. These methods efficiently mine attributes across several dimensions, enabling the creation of more accurate predictions. The XGBoost sales forecast model is tested with Walmart sales data using datasets given by the Kaggle competition. The study concludes that the model performs adequate. As compared to the other ML techniques, this one performs better in the real world. When compared to the Logistic regression algorithm and the Ridge algorithm, the study's RMSSE metric is 0.141 and 0.113 times lower, respectively. In addition, the study delves into the importance rating of features and garners some helpful recommendations [14].

Khan et al. (2020) a vital part in decision support systems that provide effective data analysis across different business processes for organizations. Demand forecasting is an important decision-making process that begins with gathering raw sales data from the market and continues with the prediction of future sales and product demand using sophisticated analytical approaches. An ML engine compiles data from many sources and uses it to make weekly, monthly, and quarterly demand predictions for goods and commodities. The efficiency of operations is closely related to the accuracy of the predictions, hence it is critical to have reliable demand forecasts. Applying the suggested method to real-time organizational data can lead to intelligent demand forecasting for retail establishments with an accuracy of up to 92.38%, according to simulation results [15].

Swami, Shah and Ray (2020) The largest software company in Russia, 1C Company, generously supplied us with a challenging time-series dataset consisting of daily sales data. By analysing historical data, they want to predict how much each product and store will sell next month. The network was trained using Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) to predict what the week ahead will bring. When comparing the methods that were utilised, their success is measured by the root mean squared error (RMSE) of the actual and expected target values. As was previously noted, XGBoost outperformed LSTM on this dataset, which was accounted for by its somewhat higher sparsity [16].

Table 1 Summary of Recent Studies on AI Integration in Agile Cycle in Continuous Integration

Author	Methodology	Datasets	Key Contributions	Challenges & Limitation	Future Scope
Chen et al. (2021)	NN model with SHAP for interpretation; evaluated using	Walmart sales dataset from Kaggle	SHAP-interpretable NN model beats Linear Regression and Support Vector Machines, with RMSE improvements of 2.92	Lack of detail on overfitting control; no mention of hyperparameter tuning	Extend to hybrid models; consider seasonal/holiday effects for

	RMSE		and 2.58, respectively		generalizability
Zamil et al. (2021)	Linear Regression using OLS; Seaborn for visualization	Past transactional sales dataset	Identified TV media as a major influence on sales; demonstrates feature selection using regression	Limited to linear relationship assumptions; lacks performance metrics like RMSE	Expand to nonlinear ML models; include more diverse media variables
Calixto et.al. (2020)	Naive Bayes Classification based on KPIs	Sales data from 594 salespeople over 3 years	Classified salespeople into Not Performing, Good, and Outstanding with 92.5% accuracy; KPI-based modeling	Focuses on classification, not forecasting; domain-specific model	Incorporate time-series analysis; test on diverse industries
Niu et.al. (2020)	XGBoost model with engineered features and RMSSE metric	Walmart dataset from Kaggle	Reduced RMSSE by 0.141 and 0.113 points, respectively, and outperformed Logistic and Ridge regression; prioritised features according to their interpretability	Details of feature engineering steps are limited; lacks real-time adaptability	Implement real-time feature engineering; ensemble with deep models
Khan et al. (2020)	ML-based intelligent demand forecasting integrated in DSS	Real-time organizational sales data	Achieved 92.38% accuracy in demand forecasting; multi-source integration; weekly, monthly, quarterly predictions	Lack of algorithm-specific insights; simulation setup not well explained	Develop domain-specific models; integrate reinforcement learning
Swami, et.al. (2020)	XGBoost and LSTM time-series prediction; RMSE-based evaluation	Daily sales data from a Russian software firm (1C Company)	Demonstrated that XGBoost outperforms LSTM on a sparse dataset	LSTM underperformance due to sparsity; dataset specifics not disclosed	Improve LSTM with data augmentation; hybrid LSTM-XGBoost models

Methodology

A sales forecasting approach using the Walmart data was chosen and presented in Figure 2.

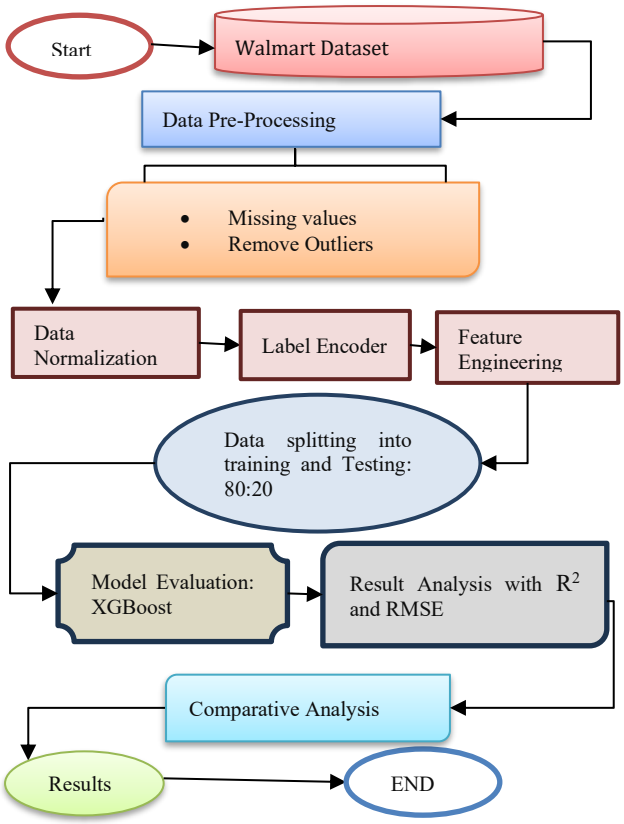

Fig.2 Flowchart Diagram of the Detection using the Walmart Dataset for Sales Prediction

The pre-treatment of the Walmart data acquisition would be previous by means of an intensive data pre-processing phase. At this step, one can fill in the missing values and delete outliers to make the data valuable and consistent. Subsequently, data normalization is undertaken to equalize the features within a normalized frame, label coding is performed to convert categorical variables into numerical representations, and feature engineering is carried out to derive and create meaningful features on which the model can be trained. After pre-processing the data, it would be divided into training and test components in proportion to 80:20 to make model training and unbiased test possible. Model development is performed using the XGBoost algorithm, which is known to be highly accurate and computationally efficient in processing structured data. R^2 and RMSE, two common regression metrics, are utilised in the model. The XGBoost model is benchmarked and compared to others. Finally, the working process concludes with the interpretation and presentation of results. Each step and process of a flowchart is explained below:

Data Collection

This research uses an extensive data set that is unstructured and found on Kaggle thus being referred to as the Walmart dataset. This dataset has the departmental store sales data of 45 particular Walmart stores. Our primary objective is to predict the weekly sales for every category throughout the year. Data is divided into two categories validation and training. Training data includes the time period from February 5, 2010, to November 1, 2012, and a goal variable called the holiday indicator and Weekly Sales. Except

missing weekly sales numbers, the format of the test dataset is same to that of the training dataset. In addition, an additional file called stores is required for correct data formatting; this file contains metadata about the type and physical size of each store. Of the dataset. Another goal of the given study is to consider the issue of the existence of factors like weather conditions, prices of fuel, holidays, markdowns, unemployment rates, consumer price indices, and their significant impact on weekly sales of the store.

Data Visualization

In data visualization, this segment shows results of ML models used to predict Walmart sales in an ERP system. It also contains a detailed study of the dataset through the numerous statistical and graphical methods. The weekly sales, temperature and unemployment trends were analyzed using visualization techniques, in the form of box plots, histograms and correlation heat maps. Such visualizations gave information about the outliers, distributions of data, and relationships between features. Full plots were generated with libraries (Matplotlib, Seaborn, and Plotly). Important trends like average monthly and annual sales trend, distribution of sales store-wise and department wise as well as effect of temperature on sales were well extrapolated. Other diagrammatic illustrations, such as pie charts and correlation tables, helped to elaborate the analysis. The accuracy of the forecasting models was analyzed based on the metrics that provide some insight into the level of model accuracy and predictivity, as follows:

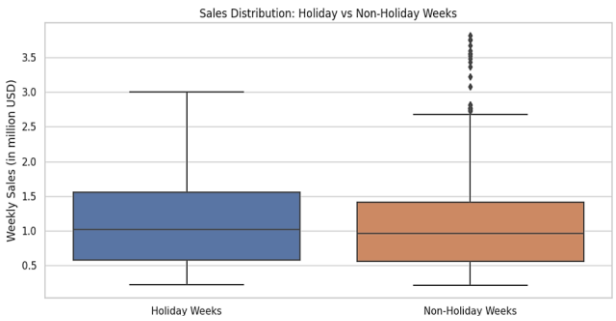

Fig.3 Average weekly Sales vs. non-holiday Weeks for Sales Prediction

Figure. 3 analysis highlights a significant difference in weekly sales between holiday and non-holiday periods, with holiday weeks averaging approximately \$1.12 million compared to \$1.04 million for non-holiday weeks, indicating a weekly advantage of \$82,000. Holiday weeks not only exhibit higher median sales but also demonstrate greater variability, as shown by a wider interquartile range and longer whiskers. This suggests that while holiday periods generally drive higher sales, they also experience greater fluctuations, likely due to the varying impact of different holidays and major events, such as Christmas or Black Friday, which yield peak sales. In contrast, minor holidays result in more moderate increases. Outliers in the non-holiday category indicate occasional high-performing weeks, yet the overall distribution clearly favors holiday weeks in terms of median, upper quartile, and maximum values. These patterns reflect typical consumer behavior during holiday seasons, where increased expenditure on gifts, celebrations, and seasonal goods contributes to elevated sales performance.

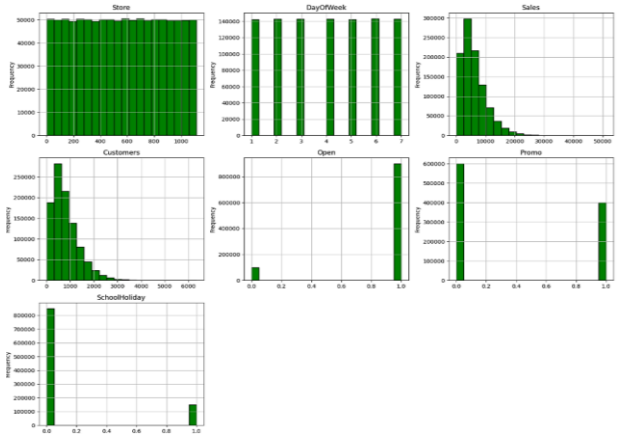

Fig.4 EDA of Walmart Dataset for Sales Forecasting

Figure. 4 presents the distribution of various features in the sales dataset. The 'Store' feature exhibits a uniform distribution across all stores, indicating a balanced representation of the dataset. The 'DayOfWeek' feature is uniformly distributed from 1 to 7, showing that each day is equally represented. Both 'Sales' and 'Customers' demonstrate a right-skewed distribution, where a majority of the observations are concentrated at lower values. The 'Open' variable reveals that most stores are open, with only a small portion being closed. Similarly, the 'Promo' feature indicates that a significant number of days involve active promotional campaigns. The 'SchoolHoliday' attribute shows that the majority of days are not school holidays. These distributions provide insight into the data composition, supporting informed preprocessing and feature engineering for sales forecasting models.

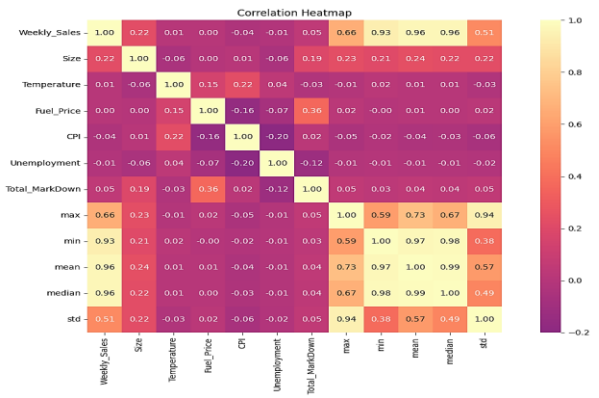

Fig.5 Correlation Matrix Heatmap of Walmart Dataset

Figure. 5 illustrates the correlation heatmap, which highlights the interdependencies among the key variables within the sales dataset. It is evident that features such as `mean`, `median`, and `std` exhibit a strong positive correlation with each other, particularly with `Weekly_Sales`, demonstrating coefficients of 0.93, 0.96, and 0.91 respectively. Conversely, attributes like `Fuel_Price`, `CPI`, and `Unemployment` show minimal to negligible correlation with `Weekly_Sales`, with correlation values approximating zero. Notably, `Total_MarkDown` displays a moderate correlation (0.66) with `Weekly_Sales`, suggesting its potential relevance in sales prediction. This heatmap serves as a visual diagnostic tool for identifying highly correlated features, which can be instrumental in dimensionality reduction and enhancing model interpretability in forecasting tasks.

Data Preparation

Processing data is an important part of making accurate detection models, especially when you want to compare them. It is very important to provide the models with uniform data, so when testing how well they work, you should only consider the models themselves and not the way the data was presented. It shows the steps that are needed to get rid of noise and outliers from the data that was picked. There is a lot of extra meaning in the facts that doesn't need to be there, so they need to be cleared out. The following list shows in detail the different steps of pre-processing:

Missing values: The purpose of the markdown columns was to help Walmart analyse how markdowns impacted sales. Specifically, each markdown column has more than 250000 NaN values, indicating that there are several missing values in the markdown data. In their case, substituting 0 for the missing value works. It shows that there is no markdown for that week.

Remove Outliers: Outliers in sales data were detected using statistical methods and removed based on business logic. This step ensures data consistency and improves the accuracy of the sales prediction model.

Data Normalization

The data is scaled in this study using min-max normalisation. In order to keep the original data inside a limited interval while maintaining their relationships, this method linearly transforms them into a specified range, usually [0, 1]. A linear function converts the dataset's minimum and maximum values to 0 and 1, respectively, to achieve the transformation. The following linear Equation (1) is used to calculate the normalised value:

$$x'_{i,n} = \frac{x_{i,n} - \min(x_i)}{\max(x_i) - \min(x_i)}(nMax - nMin) + nMin \quad (1)$$

where the values of the i —The characteristic are represented by max and min, respectively.

Categorial Encoding

Categorical label encoding can be described as pre-processing method in ML and applied to convert categorical variables into numerical values. Under the binary categorical features, all the categories are often represented as 0 and 1. Such an approach can be applied particularly to the feature, which may assume only two discrete values, e.g. yes/no or true/false. The conversion of the category values to 0 and 1 allows using them in ML algorithms, where input needs to be numerical.

Feature Engineering

The process of feature engineering is aimed at increasing the accuracy of predictions and lowering the complexity of computing with the data (Walmart). Firstly, data types can be reduced to lower-precision floating-point formats to conserve memory. The timestamp is subsequently decomposed into other dates, including month, day of the week, day of the month, weekend indication, and year, in order to analyse time-based seasonal variations and dynamics. Statistical characteristics are calculated within specified periods, and captured as lag variables, statistics based on prices and the aggregate measures, i.e., maximum, minimum and median sales. Last, but not least, recursive feature elimination with cross-validation method (RFECV) is deployed to iteratively eliminate useless or low-utilization features and keep such features that may make considerable contribution to the work of the model.

Train-Test Split

A dataset can be partitioned into a training set and a testing set using the 80:20 split. The training set is utilised to train the model, whereas the testing set is utilised for evaluation.

Model classifcation of XGBoost

A regression tree model, XGBoost adheres to the same decision-making principles as the traditional decision tree. One attribute-based condition is represented by each internal node in a regression tree, and a decision outcome is represented by each score in each leaf node [17] and the final forecast by adding up the scores projected by K separate trees. The formulation of Equation (2) is provided below:

$$\hat{y} = \Sigma_{k=1}^{k} f_K(x_i),\ f_K \epsilon F \quad (2)$$

The variables x_i, f_K, and FRepresent the i-th sales training sample, the $K - th$Tree's score, and the space of functions that contains all the regression trees, respectively.

In order to penalize the model's complexity, XGBoost uses the same gradient boosting as GBM, but it improves upon the regularized objective slightly. Equation (3) is defined as follows:

$L = \Sigma_i\, l\,(\hat{y}_{i'} y_i), \sum_k \Omega\ (f_K)$ (3)

The distance between the prediction $\hat{y}_i$ and ground-truth y_i is measured by l, a differentiable convex loss function, in this case where LIs the total objective function. In the following, the regularisation term (4) is defined as Ω.

The difference between the actual value y_i and the anticipated value $\hat{y}_i$ is measured by l, a differentiable convex loss function, in this case, where LStands for the overall objective. A description of the regularisation term Ω Equation (4) is provided here.

$\Omega(f) = \gamma^T + \frac{1}{2}\ \lambda \|\omega\|$ (4)

The quantity of tree leaves is controlled by constants γ and λ, while the score of each leaf is encoded by w. The simplest objective is achieved by XGBoost, as opposed to GBM, by expanding the loss function and removing the constant term using second-order approximation. This allows for faster optimization of the objective in a broader context. A wide range of problems can be addressed with XGBoost.

Besides the regularized objective that was already stated, XGBoost uses two more methods to cut down on overfitting even more [18]. The first is shrinking, which uses a factor coefficient to increase the weights of newly added trees after each step of tree boosting, making each tree less essential and allowing new trees to enhance the model. Second, column sub-sampling is an improvement over row sub-sampling when it comes to preventing over-fitting.

Further features that contribute to XGBoost's efficient model running include a cache-aware prefetching approach, out-of-core computation, an approximation technique for exact greedy, data storage in in-memory units for parallel learning, and more. Larger datasets can be processed quickly by XGBoost because of the aforementioned strategies.

Performance Evaluation Matrices

Accuracy, precision, recall, and F-score were the conventional measures used to assess the suggested model. However, as the model takes on a multiclass classification problem, the process of calculating these metrics differs slightly from the binary classification scenario. It is common practice to provide precision, recall, and F-score using either a macro-average or a weighted-average strategy in order to give a fairer picture of the model's performance.

Root Mean Squared Error (RMSE)

The strong relationship between RMSE and MSE makes it a popular tool to employ alongside MSE. Squaring the mean squared error (RMSE) puts the error values back on the same scale as the original data. Compared to the MSE classification problem, RMSE is simpler to understand since the error numbers are in the same units as the original data. The following describes what RMSE Equation (5) means:

$RMSE = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2$ (5)

Y_i , $\hat{y}_i$, and nAre defined in the same way as previously. Reduced Mean Squared Error (RMSE) gives a more straightforward picture of the typical magnitude of the model's prediction errors by squaring it.

R-Squared (R2)

In regression analysis, R^2 is a crucial statistic since it determines how much the independent variables account for the dependent variable's variance. Equation (6) can be used to determine it:

$R^2 = 1 - \frac{\Sigma_{i=1}^{N} (y_i - \hat{y}_1)^2}{\Sigma_{i=1}^{N} (y_i - \bar{y})}$ (6)

where $y_i$and $\hat{y}_i$ are defined as before, and y is the mean of the actual sales values. Assuming a perfect fit ($R^2 = 1$), The anticipated and actual answers' values are in perfect agreement.

Results Analysis and Discussions

The experimental setup was performed on a local system equipped with an Intel Core i7-12700H Processor running at 2.70 GHz, supported by 32.0 GB of RAM. This configuration provided sufficient computational capability to execute model training and evaluation efficiently. The proposed model achieved an R-squared value of 0.946 for sales forecasting using Python within a business decision support framework. The XGBoost model was evaluated on the Walmart dataset using standard regression metrics. The performance results are summarized in Table II, as presented below:

Table 2 Results Of Xgboost Model: Regression Performance on Walmart Dataset for Sales Prediction

Parameter	XGBoost
RMSE	21.77
R2	0.946

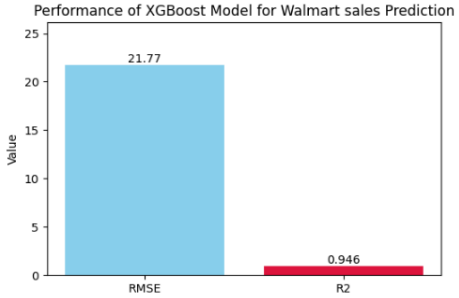

Fig.6 Bar Graph of XGBoost Model Performance for Sales Prediction of Walmart

The XGBoost model's regression performance was shown in Table II and the associated bar plot in Figure 6, which show the sales prediction results for the Walmart dataset. As it turned out, the model's RMSE was just 21.77, which is a modest average prediction error. In addition, the model's coefficient of determination (R^2) It was 0.946, which means that it accurately accounts for approximately 94.6% of the variation in the sales data. The XGBoost algorithm accurately predicted sales and uncovered hidden patterns in the dataset, as demonstrated by these results.

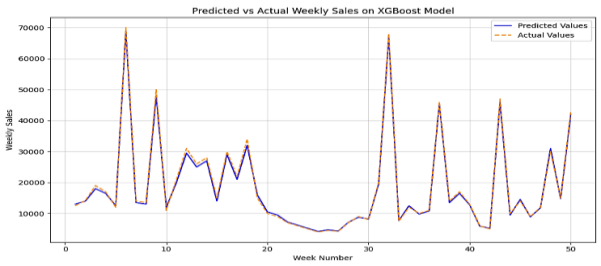

Fig.7 Predicted and Actual Values Graph for Sales Prediction Based on the XGBoost Model

Figure 7 displays the predicted vs real Weekly Sales on XGBoost Model, a time-series that illustrates the discrepancy between the predicted and real sales for fifty weeks using an XGBoost model. Sales values at the week level are displayed on the Y-Axis, while the X-axis denotes the week number. The blue solid line indicates the estimated sales yielded by the XGBoost model, while the orange dashed line portrays the actual sales data. Both curves have a close follow-up, depicting that the model has a high level of prediction accuracy, with minimal deviation. The model seems to capture the rises and dips in the sales patterns well indicating that the model must be picking up the temporal trends and seasonality.

Comparative Analysis

Following Random Forest (RF), Support Vector Regression (SVR), and ARIMA, this section compares and contrasts the proposed XGBoost model with several prominent ML models in the field. The experimental conditions were kept the same in all the models both in training and testing to create a fair comparison. The comparison of the outcomes is presented in Table III to highlight the higher performance of the developed XGBoost model in the most important classification measures.

Table 3 ML And DL Models' Comparison for Sales Forecasting on the Walmart Dataset

Models	R2
XGBoost	0.946
RF[19]	87
SVR[20]	0.7102
ARIMA[21]	74

Table III shows the results of a comparison study of various ML and DL models that were used to forecast sales using the Walmart dataset! The coefficient of determination (R^2) is the appropriate performance metric since it indicates the extent to which the independent variables may explain or anticipate the dependent variable's behaviour. The XGBoost algorithm among the models reviewed has the highest predictive capability with an R2 score of 0.946 implying that it has captured the intricacies of the patterns better using the sales data. The value of R 2 in the Random Forest (RG) model is 0.87 whereas R2 in the Support Vector Regression (SVR) is 0.7102. Another conventional time series forecasting model, the ARIMA model, yields relatively poor results with an R-squared value of 0.74. Such findings highlight the efficiency of ML ensemble proponents, especially XGBoost, at performing sales forecasting challenges compared to traditional statistical measure and other ML models.

Results from a comparison of ML and DL models trained on the Walmart dataset show that ensemble-based models outperform more conventional statistical and regression models when it comes to sales prediction job. The prediction capabilities of random forest and gradient boosting methods outperform those of support vector regression and autoregressive integrated moving average models. In order to identify intricacies in massive retail sales data, these findings demonstrate the efficacy and strength of advanced ML ability.

Conclusion And Future Work

Business organizations need a sales prediction system to manage a wide range of information in a large volume. The speed and accuracy of data processing techniques are the business decisions. The ML methods emphasized in this study paper are capable of offering a good mechanism for data tuning as well as decision-making. In this paper, effective sales forecasting framework is proposed, which is achieved using Walmart data with comprehensive preprocessing data pipeline and the XGBoost algorithm to maximize prediction accuracy. The data preparation procedure that involves outlier checking, normalization, transformation of labels, and feature processing has played a significant role in the model to define the sales patterns correctly. Standard regression measures to evaluate the reliability and robustness of the proposed method in producing accurate forecasts demonstrate its stability and soundness. The results confirm the role that ML could have in enabling data-driven decision-making in the retail process.

To improve in the future, it is possible to expand the framework by including DL models to LSTM or hybrid architecture to incorporate temporal relationships in sales. Additionally, the ability to incorporate external sources of information, such as economic indicators, promotional activities, and

seasonal trends, would enhance accuracy and enable more dynamic forecasting systems that operate in real-time.

References

[1] N. Liu, S. Ren, T.-M. Choi, C.-L. Hui, and S.-F. Ng, "Sales Forecasting for Fashion Retailing Service Industry: A Review," Math. Probl. Eng., vol. 2013, pp. 1–9, 2013, doi: 10.1155/2013/738675.

[2] S. S. S. Neeli, "Key Challenges and Strategies in Managing Databases for Data Science and Machine Learning," Int. J. Lead. Res. Publ., vol. 2, no. 3, p. 9, 2021, doi: 10.5281/zenodo.14672937.

[3] T. Boone, R. Ganeshan, A. Jain, and N. R. Sanders, "Forecasting sales in the supply chain: Consumer analytics in the big data era," Int. J. Forecast., vol. 35, no. 1, pp. 170–180, Jan. 2019, doi: 10.1016/j.ijforecast.2018.09.003.

[4] A. Hicham and B. Mohammed, "Hybrid intelligent system for Sale Forecasting using Delphi and adaptive Fuzzy Back-Propagation Neural Networks," Int. J. Adv. Comput. Sci. Appl., vol. 3, no. 11, pp. 122–130, 2012, doi: 10.14569/IJACSA.2012.031120.

[5] B. S. S. Ramya and K. Vedavathi, "An Advanced Sales Forecasting Using Machine Learning Algorithm," Int. J. Innov. Sci. Res. Technol., vol. 5, no. 5, 2020.

[6] S. Londhe and S. Palwe, "Hybrid Customer-Centric Sales Forecasting Model Using AI ML Approaches," in Recent Trends in Intensive Computing, 2021, pp. 314–321. doi: 10.3233/APC210210.

[7] S. S. S. Neeli, "Ensuring Data Quality: A Critical Aspect of Business Intelligence Success," Int. J. Lead. Res. Publ., vol. 2, no. 9, p. 7, 2021.

[8] Z. Huo, "Sales Prediction based on Machine Learning," in 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), IEEE, Mar. 2021, pp. 410–415. doi: 10.1109/ECIT52743.2021.00093.

[9] S. Morsi, "A Predictive Analytics Model for E-commerce Sales Transactions to Support Decision Making: A Case Study," Int. J. Comput. Inf. Technol., vol. 9, no. 1, Jan. 2020, doi: 10.24203/ijcit.v9i1.3.

[10] S. Patangia, "Sales Prediction of Market using Machine Learning," Int. J. Eng. Res., vol. V9, no. 09, Sep. 2020, doi: 10.17577/IJERTV9IS090345.

[11] J. Chen, W. Koju, S. Xu, and Z. Liu, "Sales Forecasting Using Deep Neural Network and SHAP techniques," in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), IEEE, Mar. 2021, pp. 135–138. doi: 10.1109/ICBAIE52039.2021.9389930.

[12] A. M. A. Zamil, N. M. Nusairat, T. G. Vasista, M. M. Shammot, and A. Yousef, "Prediction of Sales Based on an Effective Advertising Media Sale Data: a Python Implementation Approach," Acad. Strateg. Manag. J., vol. 20, no. Special Issue 2, pp. 1–9, 2021.

[13] N. Calixto and J. Ferreira, "Salespeople Performance Evaluation with Predictive Analytics in B2B," Appl. Sci., vol. 10, no. 11, p. 4036, Jun. 2020, doi: 10.3390/app10114036.

[14] Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," in 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), IEEE, Oct. 2020, pp. 458–461. doi: 10.1109/ICBASE51474.2020.00103.

[15] M. A. Khan et al., "Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3003790.

[16] D. Swami, A. D. Shah, and S. K. B. Ray, "Predicting Future Sales of Retail Products using Machine Learning," pp. 1–6, Aug. 2020.

[17] X. dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," in 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, Jan. 2021, pp. 480–483. doi: 10.1109/ICCECE51280.2021.9342304.

[18] L. Zhang, W. Bian, W. Qu, L. Tuo, and Y. Wang, "Time series forecast of sales volume based on XGBoost," J. Phys. Conf. Ser., vol. 1873, no. 1, p. 012067, Apr. 2021, doi: 10.1088/1742-6596/1873/1/012067.

[19] M. Sarisa, V. N. Boddapati, G. K. Patra, C. Kuraku, S. Konkimalla, and S. K. Rajaram, "An Effective Predicting E-Commerce Sales & Management System Based on Machine Learning Methods," J. Artif. Intell. Big Data, vol. 1, no. 1, May 2020, doi: 10.31586/jaibd.2020.1110.

[20] M. I. Abdullahi, G. I. O. Aimufua, and U. A. Muhammad, "Application of Sales Forecasting Model Based on Machine Learning Algorithms.," in Proceedings of the 28th iSTEAMS Multidisciplinary & Inter-tertiary Research Conference, Society for Multidisciplinary and Advanced Research Techniques - Creative Research Publishers, Oct. 2021, pp. 205–216. doi: 10.22624/AIMS/iSTEAMS-2021/V28P17.

[21] D. K. Deepa and G. Raghuram, "Sales Forecasting Using Machine Learning Models," Int. Sci. J. Eng. Manag., vol. 25, no. 05, pp. 3928–3936, May 2021.