Journal of Theoretical and Applied Information Technology
28th February 2025. Vol.103. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFICIENT APPROACH FOR PREDICTING SALES USING SUPERVISED MACHINE LEARNING ALGORITHMS

ANUSHA CHINTAPANTI[1], SANDIPAN MAITI[2]

[1]Research scholar, VIT-AP University, School of Computer Science and Engineering, India.
[2]Associate Professor, VIT-AP University, School of Computer Science and Engineering, India.
E-mail: [1]anusha.21phd7145@vitap.ac.in, [2]sandipan.maiti@vitap.ac.in

ABSTRACT

The significant impact in businesses is generally affected by manufacturing, planning, supply chain, marketing, warehousing, logistics, and resource management, usually managed by sales forecasting. Casual forecasting techniques and the correlations between factors are used to anticipate future sales behaviour without relying on historical data and trends. Despite the wide usage in research and application, there are severe drawbacks regarding the forecasting techniques related to classic time series. The sales related to supermarkets, along with association rules, regression techniques, time series algorithms, etc., are estimated by numerous available methods. This paper explains constructing a prediction model based on a supervised machine learning algorithm known as Ada Boost to estimate possible sales for 45 Walmart stores in various locations. It is a great opportunity for researchers to predict sales for Walmart, as it is the largest store existing in the world. The sales will be affected on a periodic basis during an event or holidays. This affect might also extend on a daily basis.

Keywords: *Forecasting, Supervised Machine Learning Algorithms, Unsupervised Machine Learning Algorithms, Time Series, Adaboost Algorithm.*

1. INTRODUCTION

Forecasting sales is crucial for managing manufacturing, supply chain, and impacting marketing, logistics, warehousing, and resource management. Competitive advantages are essential, requiring business plans obtained through sales forecasting [1]. In the data-driven retail sector, facing optimization challenges, predicting commodity sales is challenging in a dynamic business environment. Retailers must enhance sales predictions to reduce operating costs, increase sales, and improve customer satisfaction [2]. Success in retail demands accurate sales predictions for inventory management, optimal distribution, and customer satisfaction across outlets [3]. Retailers find it challenging to predict sales due to numerous factors, including external influences like weather conditions and competition from other stores [4]. Supermarket sales are forecasted using various Machine Learning techniques, particularly ensemble methods like random forest regression. This method constructs decision trees and averages their predictions for higher accuracy. Additionally, a feature similarity-based K-NN Regression predicts data points by comparing them to the training set. Support Vector Machines, part of the SVR supervised learning algorithm, determine the best line of fit by maximizing the number of points. In the random forest approach, an optimal data split is employed, and input data subsamples replicate the bootstrap [5].

Boosters aim to strengthen a weak classifier by combining it with a stronger one. The process involves sequentially using weak models to build a model, starting with the initial model using training data. Errors in the first model are addressed by constructing a second model, and predictions are combined to achieve maximum accuracy, as illustrated in Figure 1. AdaBoost, an early boosting algorithm for binary classification, falls under supervised learning in machine learning. It requires a target variable in the training data and can solve both classification and regression problems. AdaBoost, short for Adaptive Boosting, merges multiple weak classifiers into a effective single one [6].

A Sales Forecasting methodology is established through literature review, involving training the model with machine learning algorithms like Random Forest Regression, Linear Regression, SVR, KNN Regression, and Ada Boost. The

obtained results are analysed to draw conclusions. The paper is organized as follows: Section 2 presents the literature review on algorithm approaches. Section 3 details the methodology with various algorithm applications for sales prediction. Section 4 presents result for the approaches, followed by the conclusion in Section 5. And Declaration is provided in Section 6.

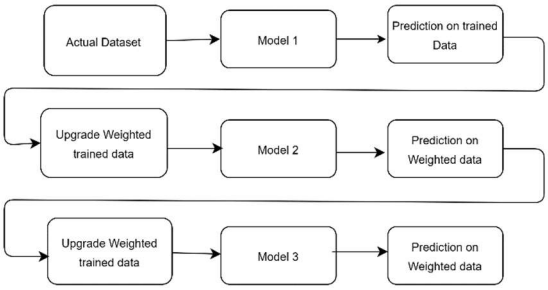

Figure 1:Ada boost classifier

2. LITERATURE REVIEW

Microsoft utilizes time series and regression algorithms for forecasting continuous values such as product sales or demand [7]. The Microsoft Time Series Algorithm proves helpful for predicting continuous variables without requiring additional columns, making it advantageous for trend forecasting. This algorithm uses data to predict anomalies in sales and demand, incorporating new data for ongoing forecasts. Microsoft Time Series stands out for its unique ability to make cross-predictions, understanding the relationship between two series to develop models. For instance, it can address situations where the sales of one car impact forecasts for another. To enhance accuracy, a combination of autoregressive tree model with cross prediction, autoregressive integrated moving average (ARIMA), and both algorithms together is employed. Short-term predictions utilize the ARTXP algorithm, while long-term predictions rely on ARIMA [7].

Analysing linear regression determines the relationship between two variables through a linear equation fitted to observed data. The explanatory variable, explaining something (predictor), and the dependent variable, targeting something, are termed explanatory and dependent variables. The goal is to find the best fit line in training and testing data, applying this method to predict commodity demand based on store sales. Sales forecasting aids in better decision-making for Production and Supply Chain Management [8]. KNN Regression, utilizing data from past cases, predicts new values by measuring similarity. 'Feature similarity' extracts features from data, predicting values for new data points based on their similarity to the nearest neighbours. The K-nearest neighbour method calculates inverse distance weighted averages using distance functions like Euclidean, Minkowski, and Manhattan, similar to those used for classification.

Due to association rule discovery [9] and its diverse applications, Data Mining has gained popularity. The aim of data analysis is to unveil common patterns in datasets, with the Association rule mining process searching for correlations and relationships between items. A recent study revealed that 80% of Indian mobile phone buyers also purchase headphones for improved audio quality. Shelke et al. [10] discuss a variety of Machine Learning algorithms, including those applied in e-commerce, e-marketing, and logistics. Rule induction, among several ML techniques, is widely used in data mining [11],[12]. Previous studies on sales prediction [13] employed both regression and boosting algorithms, with boosting algorithms outperforming regression algorithms. Zhao-Li Sun et al. [14] conducted real-time analysis using an Extreme Learning Machine (ELM) neural network to understand the relation between sales amount and key demand-influencing factors, employing a neural network-based backpropagation model for sales forecasting.

In sales forecasting, neural networks prove more effective with de-seasonalized data compared to linear models, and their superiority is notable over regression models with seasonal indicator variables [15]. In the dynamic fashion retail industry, predicting sales is challenging due to changing customer tastes and short product life cycles. Sales prediction methods include Bayesian analysis, exponential smoothing, Bayesian estimation, evolutionary neural networks, and Artificial neural networks [16]. Thomassey et al. [17] proposed a hybrid forecasting model combining fuzzy logic, neural networks, and evolutionary algorithms. Manpreet et al. [18] take a big data perspective in predicting Walmart's sales, leveraging Spark framework technologies like Python API and Scala.

Collecting data is pointless if it cannot be effectively analysed, understood, and applied [19]. Harsoor et al. [20] achieved highly accurate predictions of Walmart store sales by utilizing Hadoop, MapReduce, and Hive. Data analysis and

visualization employ tools like Apache Spark, Hadoop MapReduce Framework [21], Hadoop Distributed File Systems (HDFS) [22], and high-level languages like Python and Scala. Wazid, Katal, and Goudar et al. [23] recommend several parallel programming tools for handling Big Data. Sharma, Chauhan, and Kishore's study reveals Spark's superiority in analysing Big Data compared to Hadoop, MapReduce, or MapReduce [24]. Spark is noted for being 100 times faster than other data analysis techniques [25].

Various algorithms are employed in Sales Forecasting across industries, and their selection significantly impacts forecast outcomes. MapReduce or Hadoop Distributed File System is supplanting Big Data in short-term statistical model applications. Retailers typically grapple with the complexities of algorithm selection and implementation. This paper aids retailers in choosing the right Machine Learning Algorithm for their specific purposes.

2.1. Comparison

In Comparison with all other algorithms in the literature, the model we used getting good performance score then other models. In our model got better result than previous model. Here in existing model experiments done for individual year. But in our model, we combined all three years and did experiments.

2.2. Machine Learning Models

A machine learning model predicts competition data for the M3 time series using 1000 time series, and Gaussian Process Regression and Multilayer Perceptions emerge as the most efficient methods [26]. If other analysis methods lack machine learning techniques, there's potential for reduced accuracy in results. Machine learning, as suggested by the authors, can bridge the gap between planned and actual performance, offering greater flexibility than traditional statistical models [27]. Time series models using traditional methods are less robust than machine learning models, impacting repeatability, and require manual repetition instead of retraining, affecting generalization [28].

A brief study examined the textile market's versatility and distributor needs. Basic statistical analysis was substituted with deep learning in data sets using an alternative to traditional machine learning methods. To ensure completeness, a machine learning model needs training and testing at various stages, with periodic retraining to prevent deficiencies. The feature-based approach considers diverse variables as potential features, particularly beneficial for models with numerous parameters [29].

2.3. About Dataset

Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	05-02-2010	1643690.9	0	42.31	2.572	211.0964	8.106
1	12-02-2010	1641957.44	1	38.51	2.548	211.2422	8.106
1	19-02-2010	1611968.17	0	39.93	2.514	211.2891	8.106
1	26-02-2010	1409727.59	0	46.63	2.561	211.3196	8.106
1	05-03-2010	1554806.68	0	46.5	2.625	211.3501	8.106
1	12-03-2010	1439541.59	0	57.79	2.667	211.3806	8.106
1	19-03-2010	1472515.79	0	54.58	2.72	211.2156	8.106
1	26-03-2010	1404429.92	0	51.45	2.732	211.018	8.106
1	02-04-2010	1594968.28	0	62.27	2.719	210.8204	7.808
1	09-04-2010	1545418.53	0	65.86	2.77	210.6229	7.808
1	16-04-2010	1466058.28	0	66.32	2.808	210.4887	7.808
1	23-04-2010	1391256.12	0	64.84	2.795	210.4391	7.808
1	30-04-2010	1425100.71	0	67.41	2.78	210.3895	7.808
1	07-05-2010	1603955.12	0	72.55	2.835	210.34	7.808
1	14-05-2010	1494251.5	0	74.78	2.854	210.3374	7.808
1	21-05-2010	1399662.07	0	76.44	2.826	210.6171	7.808
1	28-05-2010	1432069.95	0	80.44	2.759	210.8968	7.808
1	04-06-2010	1615524.71	0	80.69	2.705	211.1764	7.808
1	11-06-2010	1542561.09	0	80.43	2.668	211.4561	7.808
1	18-06-2010	1503284.06	0	84.11	2.637	211.4538	7.808
1	25-06-2010	1422711.6	0	84.34	2.653	211.3387	7.808
1	02-07-2010	1492418.14	0	80.91	2.669	211.2235	7.787

Walmart +

Figure 2:Walmart Dataset of 45 stores

Every day, retailers like Walmart handle vast transactions, requiring a delicate balance between inventory management and meeting customer demands. Accurate sales predictions are crucial for maximizing revenue and profit. Current forecasting methods often rely solely on extending statistical trends, requiring additional data for customer and product analysis. There is a need for a simpler model using only previous sales data to forecast product sales. Leveraging the latest machine learning techniques allows for more precise event forecasting, demonstrated with the use of a Walmart sales dataset from Kaggle in our experiment [30].

3. METHODOLOGY

This section discusses the implementation flow using orange tool. The flow begins with load the dataset into Orange tool. This will be followed by imputation of data, making the data as timeseries, selection of sample data in which the selection will be based on fixed proportion, number of instances, and cross validation, selection of columns & rows as per our requirement, then have to check outliner data, from that we can get the Inlier data also. Then the inlier data is connecting to unsupervised machine learning algorithms and finally checking the sale prediction. Furthermore, the next section will discuss about a brief overview of Orange tool in Section 3.1. This description will be followed by the implementation of machine learning algorithms to the sample data for sales prediction in Section 3.2.

and benefit of Connecting inlier data to supervised machine learning models discussed in section 3.3.

3.1. About Orange Tool

Numerous software options are available for machine learning, data mining, and visualization, with this paper specifically exploring into the open-source Orange software. Orange stands out for supporting interactive data visualization and exploratory qualitative data analysis through a visual programming front-end. It excels in tasks like data interpretation, displaying data tables, feature selection, training predictive models, comparing learning algorithms, and creating visual representations using a canvas interface. Users can interactively explore visualizations or feed them into other widgets. The versatility of Orange extends to diverse fields like genomic research, biomedicine, bioinformatics, and education. It proves instrumental in the development and testing of new machine learning algorithms for implementing bioinformatic and genetic techniques. Furthermore, Orange serves as an effective teaching platform for machine learning and data mining methods in education [31].

3.2. Implementation of Machine Learning Algorithms

In this study, a variety of supervised machine learning algorithms, including K-NN regression, random forest regression, linear regression, SVM, and Ada Boost algorithms, are employed. The prediction of several commodities at Walmart involves the utilization of linear regression, considering factors like fuel prices, previous sales, holidays, and unemployment rates. The dataset encompasses 45 Walmart stores, and data cleaning procedures were applied using the Orange tool to ensure the accuracy and reliability of the analysis. This rigorous approach enhances the robustness of the study's findings and contributes to the precision of the predictive models utilized in forecasting commodity sales at Walmart.

In our approach, the first step involves selecting the Walmart dataset in the Orange tool, designating weekly sales as the targeted variable. Following this, data imputation is performed to address missing data, and the dataset is configured as time series data. Orange tool provides default impute methods to ensure data integrity. In the context of data imputation, replacing missing data with a suitable value is essential to preserve the dataset's integrity, as removing data can significantly reduce dataset size, potentially impacting bias and impairing analysis. In our dataset, we observe no visible difference between imputed and non-imputed datasets, supporting the robustness of our approach. This particular process ensures the quality and reliability of our dataset for subsequent analysis and modelling.

We specifically choose our data as time series data, aligning with the nature of our task to predict sales based on temporal patterns. Time series forecasting involves making scientifically informed predictions grounded in historical data. Through the analysis of this historical data, models are constructed, serving as guides for future strategic decisions. This analytical process goes beyond simple prediction, as it requires understanding and interpreting past trends to project future outcomes. Consequently, our approach encompasses both the analytical examination of historical data and the forecasting of future trends, ensuring a comprehensive and evidence-based foundation for our predictions.

To ensure the efficiency and representativeness of our analysis, we opt for a 70% sample data selection from the entire dataset. The Orange tool's sample widget offers a range of methods for data sampling, facilitating the creation of a representative and complementary dataset that encompasses instances related to the input set. Execution of this algorithm involves providing the input dataset, followed by selecting the sample data option. In the context of Cross Validation, all data instances undergo division into complementary subsets. The subsets, excluding the one selected by the user, are output as Data Sample, while the chosen subset is output as the remaining Data [32]. This particular sampling process enhances the robustness and reliability of our analysis, contributing to the accuracy of our subsequent modelling and predictions.

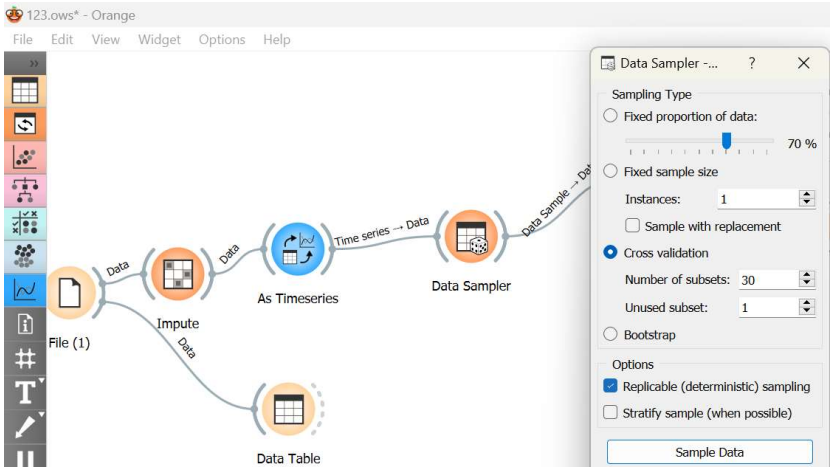

Figure 3:Sample data

After data sampling, we are selecting columns & rows as per our requirement. The Select Columns widget serves as a powerful tool, empowering users to exactly define the architecture of their data domain through the manual selection of attributes and the explicit assignment of their roles. In the robust framework of Orange, attributes are systematically categorized into three classes: ordinary attributes, optional class attributes, and meta-attributes. This categorical breakdown proves especially valuable when constructing sophisticated models such as classification models. For instance, in the process of creating a classification model, the data domain encompasses a set of attributes, each contributing to the overall understanding, alongside a distinct class attribute that defines the target variable.

It's worth noting that while meta-attributes themselves do not actively contribute to the modelling process, their unique role comes into play as certain widgets within Orange have the capacity to harness them as instance labels. This complex flexibility highlights the adaptability of the Orange data analysis framework, allowing users to derive meaningful insights from diverse data structures while tailoring the modelling process to meet specific analytical objectives.

This tool chooses a subset from a given dataset according to conditions defined by the user. Instances meeting the specified criteria are then directed to the output Matching Data channel. The conditions for data selection are expressed as a set of conjunctive terms, meaning that selected items must satisfy all the terms within the 'Conditions.' Condition terms are established by selecting an attribute, choosing an operator from a provided list based on whether the attribute is discrete, continuous, or a string, and optionally specifying the value for the condition term. The operators vary depending on the nature of the attribute, with different options available for discrete, continuous, and string attributes. Here we are selecting data as year wise for our analysis.

In our pursuit of predicting sales, our initial step involves scrutinizing the dataset for any potential outliers. Detecting and addressing outliers is crucial to ensure the accuracy and reliability of our predictive models. For inlier data, we apply various models and systematically compare their performance to identify the most effective one. In the case of outliers, we employ the Local Outlier Factor algorithm, utilizing Euclidean distance as the metric for discerning and isolating these anomalies. This comprehensive approach, addressing both inliers and outliers, contributes to the robustness of our analysis and ensures that our sales predictions are based on a dataset free from any potential distortions.

An outlier refers to a data point that resides outside the vicinity of other data points or exhibits a distinct nature. The identification of outliers within a dataset is achieved through the implementation of an algorithm called the Local Outlier Factor (LOF). This algorithm designates a data point as a local outlier if it stands apart within its local neighbourhood. The key criterion for identifying an outlier using LOF is the density of this local neighbourhoods. If the density of data points across the entire dataset is not uniform, indicating varying

degrees of proximity, then the performance of LOF is deemed effective in discerning outliers. This refinement approach enhances the algorithm's ability to accurately pinpoint and characterize outliers within the dataset, contributing to a more precise analysis and prediction process.

In the list of distance measure techniques, Euclidean distance stands out as the most widely employed method. This technique serves as a fundamental metric for measuring the similarity between objects, particularly within the context of cluster analysis. Euclidean distance essentially quantifies the geometric distance between points in a multidimensional space, facilitating the assessment of their closeness or dissimilarity. Cluster analysis, leveraging Euclidean distance, is instrumental in grouping similar objects or entities based on their spatial relationships. While Euclidean distance is a versatile and widely applicable measure of similarity, it's important to note that certain situations may necessitate exceptions. In such cases, the measurement of similarity must adhere to specific constraints or requirements, emphasizing the importance of tailoring the approach to the unique characteristics of the data or context [33].

Let's consider two objects, denoted as x = (a_{x1}, a_{x2}, …, a_{xn}) and y = (a_{y1}, a_{y2}, …, a_{yn}), each described by n numeric attributes. The distance between these two objects is precisely defined as the Euclidean distance, which is calculated using the following formula:

$$Dist(x,y) = \sqrt{(a_{x1} - a_{y1})^2 + (a_{x2} - a_{y2})^2 + \cdots + (a_{xn} - a_{yn})^2}$$

In this equation, Distance(x,y) represents the Euclidean distance between objects x and y. The summation is carried out over each corresponding attribute of the objects, with $(a_{xi},a_{yi})^2$ quantifying the squared difference between the respective attribute values. The square root of the sum provides the overall Euclidean distance, serving as a robust measure to measure the dissimilarity or proximity of the objects based on their numeric attributes. This mathematical representation offers a precise and widely accepted methodology for quantifying the distance between two objects in a multi-dimensional space defined by their numeric attributes.

In the illustrative context of Figures 4, our analysis focuses on the utilization of inlier data connected to several supervised machine learning algorithms. These algorithms encompass a diverse set, including the gradient boost algorithm, random forest, KNN, Ada boosting, and SVM algorithms, all employed to predict sales. Notably, our rigorous evaluation indicates that the Ada Boosting algorithm consistently delivers the best performance score among these methodologies.

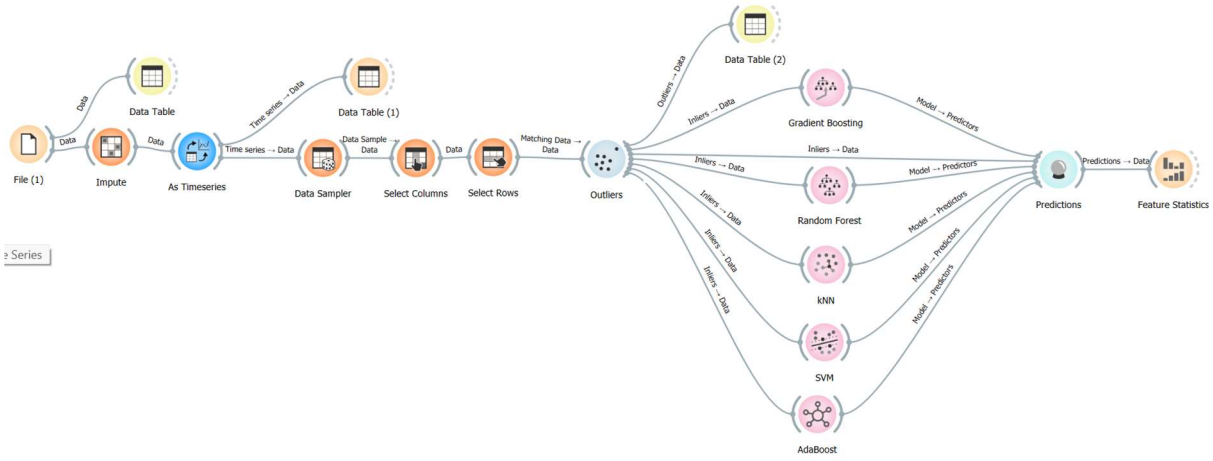

Figure 4:Inlier Data To Machine Learning Algorithms

3.3. Connecting inlier data to supervised machine learning models:

Inlier data serves as a crucial foundation for separating and predicting intricate patterns inherent in the dataset. By seamlessly integrating with advanced machine learning models like Gradient Boosting, Random Forest, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and AdaBoost, inliers facilitate the extraction of

meaningful insights and foster a comprehensive understanding of the underlying structure. These models leverage the inherent regularities identified in inlier data to make accurate predictions, enabling a more robust and effective analysis.

In contrast, outliers, characterized as data points exhibiting unusual or anomalous behaviour, stand out as exceptions within the dataset. While inliers represent the normative or typical elements, outliers deviate significantly from the expected patterns. The identification and treatment of outliers are crucial steps in refining the accuracy and reliability of machine learning models, ensuring that the predictive power of the algorithms is not compromised by unusual data points. Thus, the distinction between inliers and outliers becomes pivotal in the pursuit of extracting meaningful insights and making informed decisions from complex datasets.

Gradient Boosting, as a powerful mechanism, excels in combining predictions from numerous weak learners, typically decision trees, to create a difficult predictive model. Through the application of Gradient Boosting, one can separate difficult patterns and establish relationships within inlier data, contributing to a comprehensive understanding of the underlying data dynamics. This method's strength lies in its ability to capture complex relationships and refine predictions on inlier patterns by iteratively fitting weak learners to the residuals of preceding models. Random Forest stands out as another noteworthy ensemble learning approach. This method constructs multiple decision trees and combines their predictions to craft a more robust and accurate model. By aggregating the outcomes of diverse decision trees, Random Forest adeptly captures the inherent structure of inlier data. This holistic approach, focusing on various sides of the data, enables a thorough comprehension of inlier patterns, offering a refinement perspective on the intricacies within the dataset. The interaction between Gradient Boosting and Random Forest underscores their collective capability in extracting meaningful insights from inlier data and enhancing the predictive accuracy of machine learning models.

The k-nearest neighbour algorithm, a versatile classification method, operates by assigning data points in feature space to the majority class of their k-nearest neighbours. This approach provides a dynamic means of understanding the inherent patterns within a dataset. Exploring deeper into its capabilities, KNN offers a valuable tool for identifying inlier data points through a meticulous comparison of similarities between instances. The concept of inlier identification becomes particularly insightful within the KNN framework, as close proximity among points in the feature space is a key factor in capturing the structural intricacies of inliers, often indicative of shared class membership.

By analysing the close relationships among neighbouring points, KNN not only classifies data points effectively but also focus on the cohesive structures within the dataset. Inliers, characterized by their proximity and similarity, tend to form clusters in feature space, a phenomenon that the k-nearest neighbour algorithm adeptly exploits. This method's sensitivity to local patterns allows for a significance exploration of the relationships between inliers, unveiling valuable insights into the cohesive groupings that might signify shared characteristics. In essence, the k-nearest neighbour algorithm not only excels in classification tasks but also serves as a discerning tool for uncovering the underlying structures of inliers. Its support on proximity and similarity metrics enables a detailed exploration of the feature space, offering good understanding of how inliers cluster and contribute to the extensive patterns within a dataset.

In Support Vector Machines (SVM), the quest for the hyperplane that maximally segregates different classes of data points is undertaken through the utilization of a sophisticated supervised learning algorithm. This process involves meticulous optimization to identify the optimal hyperplane that distinctly separates diverse classes, contributing to SVM's ability in classification tasks. Notably, the versatility of SVM extends to training the algorithm specifically to separate hyperplanes that effectively distinguish inlier data from other data points, marking a key application in uncovering distinctive patterns within the dataset.

The effectiveness of SVM in recognizing patterns within inlier data is further highlighted by the thoughtful consideration of the kernel function and hyperparameters. The kernel function plays a crucial role in transforming the input space, allowing SVM to operate in higher-dimensional feature spaces and revealing intricate relationships that may be obscured in the original data. By selecting an appropriate kernel function and fine-tuning hyperparameters, practitioners can tailor SVM to the specific characteristics of the dataset, enhancing its ability to discern and understand the nuanced patterns embedded in inlier data. In essence, SVM

not only excels in creating hyperplanes that maximize class separation but also proves to be a versatile tool for distinguishing inlier data amidst the broader dataset. The careful selection of the kernel function and hyperparameters highlights the adaptability of SVM, allowing it to capture and influence the inherent patterns within inliers, contributing to a more refined and accurate analysis of complex datasets.

AdaBoost, a significant ensemble learning technique, arranges weak learners in a sequential manner, with each subsequent learner focusing on emphasizing the previously misclassified points, thereby refining the model's predictive capabilities. What sets AdaBoost apart is its adaptive ability to concentrate on inlier data points, particularly those that the ensemble's weak learners might inaccurately classify. This adaptability enhances the model's overall performance on inlier patterns, making it a versatile and robust approach in handling complex datasets. The strength of AdaBoost lies in its iterative nature, where emphasis is dynamically placed on correcting errors made by the ensemble, leading to a continuous improvement in its accuracy, especially in the identification and understanding of inlier data. By adaptively adjusting its focus, AdaBoost becomes expert at capturing the subtle distinctions present in inliers, contributing to a more refined pattern recognition.

To optimize the performance of a machine learning model that incorporates inlier data, a comprehensive approach is taken. This involves training the model on the inlier data, fine-tuning hyperparameters to enhance its precision, and validating its generalization on unobserved inliers. This meticulous process ensures that the model not only performs well on the training data but also demonstrates robustness when exposed to new instances of inlier patterns. As practitioners engage in model evaluation, it becomes imperative to consider the nature of the data and the specific problem at hand. The effectiveness of a model is inherently tied to the context in which it operates, requiring a thoughtful examination of its performance with regard to the unique characteristics of the dataset and the distinctions of the problem domain. This contextual awareness is pivotal in determining the true efficacy of a model in addressing the complexities associated with inlier patterns and, consequently, in making informed decisions within the machine learning.

To quantify the performance of our predictive models, we employ various performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics provide a comprehensive and understanding of the predictive accuracy and effectiveness of each algorithm. The obtained results, particularly the superior performance of the Ada Boosting algorithm, serve as valuable insights for refining and optimizing our predictive models in the pursuit of accurate sales forecasting.

4. RESULTS

This section depicts the comparison of various factors such as performance scores like MSE, RMSE, MAE, and MAPE, by using inlier data.

4.1. Comparison of Performance Scores

This section discusses the performance score values that are obtained using orange tool for the various factors mentioned above. The performance score values are compared by applying the various algorithm techniques such as Inlier data. The comparison of various algorithms based on the factors MSE, RMSE, MAE and MAPE by applying Inlier data and unsupervised algorithms to compare the performance scores of all three years sales data. As seen in Fig.10, when Inlier data is applied, AdaBoost algorithm has the best performance score for the factors mentioned when compared with other algorithms.

Show perfomance scores

Model	MSE	RMSE	MAE	MAPE
Gradient Boosting	30455369221.085	174514.668	128176.995	0.167
Random Forest	5714435701.207	75593.887	39859.090	0.036
kNN	227350787891.153	476813.158	402608.089	0.554
AdaBoost	137527423.253	11727.209	5378.482	0.009
SVM	344473771954.990	586918.880	501800.403	0.813

Figure 5: Performance Score using inlier data for all three years data

Model	MSE	RMSE	MAE	MAPE
Gradient Boosting	29288042799.072	171137.497	124369.837	0.163
Random Forest	6099306866.186	78098.059	41712.981	0.038
AdaBoost	186643797.951	13661.764	6621.666	0.011
SVM	335996377157.028	579651.945	495197.042	0.801
kNN	226517692837.223	475938.749	401370.724	0.551

Figure 6: Performance Score Using Inlier data for 2010 data

Model	MSE	RMSE	MAE	MAPE
Gradient Boosting	30977446666.110	176004.110	127811.817	0.170
Random Forest	6845861839.147	82739.723	43841.246	0.039
AdaBoost	194200458.335	13935.582	6721.122	0.010
SVM	330504528799.510	574895.233	487701.005	0.779
kNN	224110900868.569	473403.529	399205.109	0.545

Figure 7: Performance Score Using Inlier data for 2011 data

Model	MSE	RMSE	MAE	MAPE
Gradient Boosting	29288042799.072	171137.497	124369.837	0.163
Random Forest	6316824081.920	79478.450	41803.737	0.038
AdaBoost	186643797.951	13661.764	6621.666	0.011
SVM	335996377157.028	579651.945	495197.042	0.801
kNN	226517692837.223	475938.749	401370.724	0.551

Figure 8: Performance Score using Inlier data for 2012 data

By applying inlier data to unsupervised algorithms to 2010, 2011, 2012 sales data, figure 6, figure 7, figure 8 illustrates the comparison between various algorithms based on MSE, RMSE, MAE and MAPE. When evaluated based on Inlier data in Figs.6, 7, 8 compared to other algorithms for the factors mentioned, it is shown that AdaBoost algorithm has the highest performance score.

4.2. Results Analysis

The analysis based on difference between the proposed method and the existing method [5]. In the table it is shown that the inlier data, connected to Random Forest, SVM, KNN and Ada Boost for 2010-year data is considered.

Table 1:Analysis for 2010 data using Inlier data

Models	MSE Diff
Random Forest	7.75%
SVM	-4.93%
KNN	-71.62%
Extra tree regression Vs Ada Boost	96.62%

Table 2:Analysis for 2011 data using Inlier data

Models	MSE Diff
Random Forest	-68.8%
SVM	-23.67%
KNN	-81.32%
Extra tree regression Vs Ada Boost	94.94%

Table 3:Analysis for 2012 data using Inlier data

Models	MSE Diff
Random Forest	-68.8%
SVM	-23.67%
KNN	-81.32%
Extra tree regression Vs Ada Boost	94.94%

Here we can observe that, for random forest has positive difference of MSE value between existing method and proposed method. Means that we got lesser values of MSE, RMSE and MAE in proposed method. But in SVM and KNN has huge negative difference of MSE, RMSE and MAE between existing method and proposed method. Means that we got greater values of MSE, RMSE and MAE in proposed method. Similarly, differences between Extra tree regression and Ada boost were in existing model and proposed model respectively, has huge positive difference of MSE, RMSE and MAE between existing method and proposed method. Means that we got lesser values of MSE, RMSE and MAE in proposed method. Hence that proves Ada Boost is the best algorithm for sales prediction.

Similarly, shown in tables 2 and 3, where the inlier data for 2011, 2012 years is analysed using Random Forest, SVM, KNN, and Ada Boost. Compared with existing model and proposed model, Extra tree regression and Ada boost had huge positive differences in MSE, RMSE, and MAE. Therefore, the proposed method has lower MSE, RMSE, and MAE values. Hence that proves Ada Boost is the best algorithm for sales prediction. In below figures 9, 10, 11 are showing Performance

score without using Inlier data for 2010, 2011, 2012 respectively.

Show perfomance scores

Model	MSE	RMSE	MAE	MAPE
Random Forest	7228032965.694	85017.839	41538.124	0.037
SVM	337909498919.047	581299.836	492877.278	0.794
kNN	225644455946.677	475020.480	397144.738	0.548
Gradient Boosting	33146983043.688	182063.129	132859.124	0.182
AdaBoost	195881336.871	13995.761	6812.298	0.010

Figure 9:Performance score without using Inlier data for 2010

Model	MSE	RMSE	MAE	MAPE
Random Forest	7448170627.271	86302.785	42665.071	0.038
SVM	337463006814.375	580915.662	492459.183	0.794
kNN	226254924943.989	475662.617	397884.488	0.547
Gradient Boosting	32957082195.146	181540.855	131090.228	0.176
AdaBoost	235205027.294	15336.396	7453.882	0.011

Figure 10:Performance score without using Inlier data for 2011

Model	MSE	RMSE	MAE	MAPE
Random Forest	7519156352.702	86713.069	42568.665	0.038
SVM	337463006814.375	580915.662	492459.183	0.794
kNN	226254924943.989	475662.617	397884.488	0.547
Gradient Boosting	32957082195.146	181540.855	131090.228	0.176
AdaBoost	235205027.294	15336.396	7453.882	0.011

Figure 11:Performance score without using Inlier data for 2012

The analysis from the tables 4, 5, 6 compares the proposed method (without using inlier) with the existing method [5]. The results shows that the Random Forest algorithm has a positive difference in MSE. By this value we are calculated RMSE, and MAE, this is also indicating that the proposed method has lower values for these metrics compared to the existing method. On the other hand, SVM and KNN algorithms have a large negative difference in MSE, RMSE, and MAE, indicating that the proposed method has higher values for these metrics. Similarly, the Extra Tree Regression and Ada Boost algorithms have a large positive difference, indicating that the proposed method has lower values for MSE, RMSE, and MAE. Based on these findings, it can be concluded that Ada Boost is the best algorithm for sales prediction.

Table 4:Analysis for 2010 data without using Inlier data

Models	MSE Diff
Random Forest	71.23%
SVM	-5.5%
KNN	-70.96%
Extra tree regression Vs Ada Boost	96.46%

Table 5:Analysis for 2011 data without using Inlier data

Models	MSE Diff
Random Forest	-84.74%
SVM	-26.27%
KNN	-83.05%
Extra tree regression Vs Ada Boost	93.87%

Table 6:Analysis for 2012 data without using Inlier data

Models	MSE Diff
Random Forest	17.53%
SVM	-8.43%
KNN	-84.07%
Extra tree regression Vs Ada Boost	94.04%

5. CONCLUSION

Compared with Random Forest Regression, Gradient Boost, KNN and SVM Techniques, Ada Boost Technique is the more suitable technique to predict Walmart Store sales in the future based on the dataset used. This result could also be used by other retail store owners to determine their sales figures, and instead of spending time in doing analysis with other Supervised ML Algorithms, they could opt directly for Sales Prediction using Ada Boost or Random Forest Approach. It would also be beneficial for other retailers to perform demand analyses on the same basis. According to the study, we were able to understand that external factors, such as holidays, unemployment rate, CPI, etc. are also important to predict the performance of a retail store.

6. DECLARATION

Funding Declaration: There has been no significant financial support for research.

Competing Interests:

There is no Conflict of Interest.

Author's Contribution:

Anusha Chintapanti wrote the manuscripts done the experiments and Dr. Sandipan Maiti guided for identifying the problem.

Preprint Information:

Anusha Chintapanti, Sandipan Maiti. Efficient approach for Predicting Sales using Supervised Machine Learning Algorithms, 23 August 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3255369/v1]

REFERENCES:

[1] Omar, H.A., Liu, D.-R.: Enhancing sales forecasting by using neuro networks and the popularity of magazine article titles. In: 2012 Sixth International Conference on Genetic and Evolutionary Computing, pp. 577–580 (2012). IEEE.

[2] Jain, A., Menon, M.N., Chandra, S.: Sales forecasting for retail chains. San Diego, California: UC San Diego Jacobs School of Engineering (2015).

[3] Berry, M.J., Linoff, G.S.: Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management. John Wiley & Sons, (2004)

[4] Ferreira, K.J., Lee, B.H.A., Simchi-Levi, D.: Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & service operations management 18(1), 69–88 (2016)

[5] Raizada, S., Saini, J.R.: Comparative analysis of supervised machine learning techniques for sales forecasting. International Journal of Advanced Computer Science and Applications 12(11) (2021)

[6] https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost

[7] Mekala, P., Srinivasan, B.: Time series data prediction on shopping mall. Int. J. Res. Comput. Appl. Robot 2(8), 92–97 (2014)

[8] Sohrabpour, V., Oghazi, P., Toorajipour, R., Nazarpour, A.: Export sales forecasting using artificial intelligence. Technological Forecasting and Social Change 163, 120480 (2021)

[9] Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender systems: An introduction–Cambridge university press. New York 352 (2010)

[10] Shelke, R., Dharaskar, R., Thakare, V.: Data mining for supermarket sale analysis using association rule. Int. J. Trend Sci. Res. Dev 1(4) (2017)

[11] Bose, I., Mahapatra, R.K.: Business data mining—a machine learning perspective. Information & management 39(3), 211–225 (2001)

[12] Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp. 617–620 (2018). IEEE

[13] Krishna, A., Akhilesh, V., Aich, A., Hegde, C.: Sales-forecasting of retail stores using machine learning techniques. In: 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), pp. 160–166 (2018). IEEE

[14] Sun, Z.-L., Choi, T.-M., Au, K.-F., Yu, Y.: Sales forecasting using extreme learning machine with applications in fashion retailing. Decision support systems 46(1), 411–419 (2008)

[15] Chu, C.-W., Zhang, G.P.: A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of production economics 86(3), 217–231 (2003)

[16] Liu, N., Ren, S., Choi, T.-M., Hui, C.-L., Ng, S.-F., et al.: Sales forecasting for fashion retailing service industry: a review. Mathematical Problems in Engineering 2013 (2013)

[17] Thomassey, S., Happiette, M., Castelain, J.-M.: A global forecasting support system adapted to textile distribution. International Journal of Production Economics 96(1), 81–95 (2005)

[18] Singh, M., Ghutla, B., Jnr, R.L., Mohammed, A.F., Rashid, M.A.: Walmart's sales data analysis-a big data analytics perspective. In: 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 114–119 (2017). IEEE

[19] Sullivan, G.: Interpreting qualitative data: methods for analysing talk, text and interaction. JSTOR (2003)

[20] Harsoor, A.S., Patil, A.: Forecast of sales of Walmart store using big data applications. International Journal of Research in Engineering and Technology 4(6), 51–59 (2015)

[21] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M., Shenker, S., Stoica, I.: Fast and interactive analytics over

Hadoop data with spark. Usenix Login 37(4), 45–51 (2012)

[22] Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM 51(1), 107–113 (2008)

[23] Katal, A., Wazid, M., Goudar, R.H.: Big data: issues, challenges, tools and good practices. In: 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 404–409 (2013). IEEE

[24] Sharma, M., Chauhan, V., Kishore, K.: A review: MapReduce and spark for big data analysis. In: 5th International Conference on Recent Innovations in Science, vol. 5 (2016)

[25] H. Pandey, Is Spark really 100 times faster on stream or its hype? vol. 2, Sept 2016

[26] Ahmed, N.K., Atiya, A.F., Gayar, N.E., El-Shishiny, H.: An empirical comparison of machine learning models for time series forecasting. Econometric reviews 29(5-6), 594–621 (2010)

[27] Bohanec, M., Borstnar, M.K., Robnik-Sikonja, M.: Integration of machine learning insights into organizational learning: A case of b2b sales forecasting. In: Blurring the Boundaries Through Digital Innovation: Individual, Organizational, and Societal Challenges, pp. 71–85 (2016). Springer.

[28] Maciel, L.S., Ballini, R.: Design a neural network for time series financial forecasting: Accuracy and robustness analysis. Anales do 9º Encontro Brasileiro deFinan ̧cas, Sao Pablo, Brazil (2008)

[29] Thomassey, S., Fiordaliso, A.: A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems 42(1), 408–421 (2006)

[30] https://www.kaggle.com/input/retail-analysis-with-walmart-data-Dataset used for modelling

[31] ttps://en.wikipedia.org/wiki/Orange(software) Features

[32] https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest

[33] Data Mining Concepts and Techniques, Third Edition, Jiawei Han University of Illinois at Urbana–Champaign, Micheline Kamber, Jian Pei, Simon Fraser University.