Sales Prediction using Machine Learning Algorithm

Dr. V. Latha Jothi
Professor
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode – 638012

Aditya B., Arthika S., Jayashree S.
Final Year Student
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode – 638012

Abstract-Businesses must perform time-series forecasting of seasonal item sales because it enables them to predict future demand and modify their inventory accordingly. This investigation compares the performance of three well-known machine learning methods for time-series forecasting of seasonal item sales: Support Vector Machine (SVM), Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), and Multi-layer Perceptron (MLP). A dataset of historical sales data is used to evaluate the algorithms. The data is split into training and testing sets, and measures like Mean Absolute Error (MAE), Relative Absolute Error (RAE), and Root Mean Squared Error (RMSE) are used to assess each algorithm's performance. The analysis's findings support the assertion that Multilayer Perceptron (MLP) provides greater accuracy than other methods incalculating the seasonal sales of the historical data.

Keywords—Multi-Layer Perceptron, Time-series forecasting, SARIMAX, SVM, Machine Learning

I. INTRODUCTION

In the past, super markets produced goods without taking demand or sales volume into account. Data about the demand for items on the market is needed for any manufacturer to decide whether to expand or decrease the production of multiple units. Businesses that compete in the market without taking these values into account risk losing out. To gauge demand and sales, many businesses use different factors [15].

Accurate and timely revenue forecasting, also known as sales forecasting or revenue forecasting, can give businesses involved in the production, distribution, or retail of goods important insight in today's fiercely competitive climate and rapidly changing consumer landscape [14]. While long-term forecasts can address a variety of issues, short-term forecasts mostly assist with production planning and stock management.[15]

Due to the short shelf lives of many products in these businesses, which result in income losses in both shortage and surplus conditions, sales forecasting is particularly crucial. A lack of products results from too many orders, while a lack of opportunities results from too few orders. As a result, the rivalry in the food market is always changing as a result of elements including pricing, advertising, and rising consumer demand.

Machine learning techniques can be used to automatically create precise sales forecasting models using the wealth of sales data and related data. This strategy is significantly easier. It is adaptable, so it can adjust to data changes, and it is not biased by the quirks of a single sales manager Yet, it runs the risk of overestimating the human expert's prediction's accuracy, which is typically faulty. For instance, businesses once produced goods without taking demand or sales volume into account, which resulted in a number of issues. Data on consumer demand for items is crucial for any producer to decide whether to increase or decrease the number of units produced because they don't know how much to sell. Companies will suffer losses if they don't take these guidelines into account when they compete in the market. Various businesses use various metrics to estimate their market and sales.

Companies have historically relied on several statistical models such time series and linear regression, feature engineering, and random forest models to acquire future sales and demand prediction. There are multiple

methods for forecasting sales. Time series are collections of data points that are kept over a predetermined amount of time and are used to predict the future. The term "time series" refers to a group of data points that are gathered over time at consecutive, evenly spaced points.

Time-series forecasting is an essential task in business analytics, where the goal is to predict future values of a time-dependent variable based on historical data. In the context of seasonal item sales, accurate forecasting is crucial for businesses to plan production, inventory management, and marketing activities effectively. Machine learning algorithms have emerged as popular techniques for time-series forecasting due to their ability to learn from past data and identify complex patterns. In this comparative analysis, we focus on three machine learning algorithms like Support Vector Machine (SVM), Seasonal Autoregressive Integrated Moving Average with exogenous variables (SARIMAX), and Multi-layer Perceptron (MLP). SVM is a supervised learning algorithm that can be used for regression analysis, while SARIMAX is a statistical model that incorporates seasonality into a time-series forecast. MLP is a neural network-based algorithm that can learn nonlinear relationships between inputs and outputs.

II. LITERATURE REVIEW

1. Cluster-based hierarchical demand forecasting for perishable goods.

For retailers, demand forecasting is especially crucial when it comes to supply chains for perishable goods like fresh food. Due to their necessity to be offered as fresh as possible and speedy deterioration, such commodities are created and delivered daily. Underestimation and overestimation of demand have a negative impact on the retailer's profits. Consumers dislike stock-outs, and at the end of the day, unsold goods must be thrown away. To aid with operational choices, multivariate ARIMA models are used to forecast daily demand. Using point-of-sale data from an industrialized bakery chain, we assess the strategy and demonstrate that it is feasible to simultaneously enhance availability and reduce loss. Top-down forecasting is equally accurate as direct forecasting, which enables lower computing costs.

2.*A network traffic forecasting method based on SA optimized ARIMA–BP neural network.*

The traditional linear and non-linear network traffic forecasting methods are unable to anticipate traffic in the future with sufficient accuracy. Based on SA (Simulated Annealing) improved ARIMA (Autoregressive Integrated Moving Average model)-BPNN, this study proposes a method to predict network traffic (Back Propagation Neural Network). The non-linear model BPNN, the linear model ARIMA, and the optimization algorithm SA are all heavily utilized in this strategy. Since it is impossible to precisely specify the patterns in network traffic, the results of the traditional single model and hybrid models' predictions of network traffic are very wrong. A novel ARIMA-BPNN-based method is therefore proposed for forecasting network traffic.

3.*Sales prediction using machine learning algorithms.*

Machine learning is transforming every part of life, and this has had a big impact on actual-world events as well. Education, healthcare, engineering, commerce, entertainment, and transportation are just a handful of the many industries where machine learning has made a significant impact. Because it ignores how consumers actually make purchases, the conventional approach to attaining sales and marketing goals no longer benefits firms. This study's goal is to propose a factor for predicting Big Retail Companies' expected sales while accounting for their historical sales. A thorough analysis of sales prediction is conducted using machine learning techniques like linear regression, K-neighbors regression, and XG are used.

4. An advanced sales forecasting using machine learning algorithm.

A retailer can estimate its anticipated future revenues for income earned over a specific period of time with the aid of sales forecasting. As a result, time is crucial in sales forecasting. For sales forecasting in this study, we used data mining approaches including ARIMA models and XG Boost algorithms, which are more effective at manipulating trending sales analyses. The material was once subjected to extensive review in order to identify patterns and outliers that can hinder the prediction system. As it provides crucial information on the visitors a store

might anticipate on a particular day, sales forecast plays a crucial role in increasing how successfully stores can operate.

5. *Intelligent sales prediction using machine learning techniques*.

Sales forecasting is becoming more and more crucial as e-commerce grows, as accurate and prompt forecasting can assist e-commerce businesses in resolving all the supply and demand-related uncertainty and lowering inventory costs. The majority of commercial businesses rely largely on data sources and demand forecasts of sales patterns. The accuracy of sales projections has a significant effect on business. Data mining techniques are very practical tools for unlocking buried information from a sizable dataset to improve predicting precision and effectiveness. Dealing with huge data and accurate sales forecasting is challenging, though. The use of various data mining techniques could solve these problems. We provided a quick analysis of sales data and sales projection in this report. The use of various data mining techniques could solve these problems. We provided a quick analysis of sales data and sales projection in this report. Every person and every business must be aware of customer demand far in advance of any season to prevent product shortages. As time passes by, the desire of the businesses to be more accurate regarding the predictions will expand tremendously. Improved projections are directly correlated to the company's earnings.

6. A study of demand and sales forecasting model using machine learning algorithm.

A basic task in commerce is sales forecasting. In this way, the application of machine-learning models for sales prediction analysis was researched. Reviewing scientific literature will help determine whether there are any advantages over traditional statistical methods. This study has finished its in-depth analysis of predictive models in an effort to improve future sales forecasts. Millions of reviews are written every day, which makes it difficult for a customer to decide whether to purchase the goods or not. It is challenging and time-consuming for product manufacturers to research this many different viewpoints. the classification of reviews using Logistic Regression and the prediction of sales using the three machine learning methods, Linear Regression, Decision Tree (DT), and Random Forest (RF).

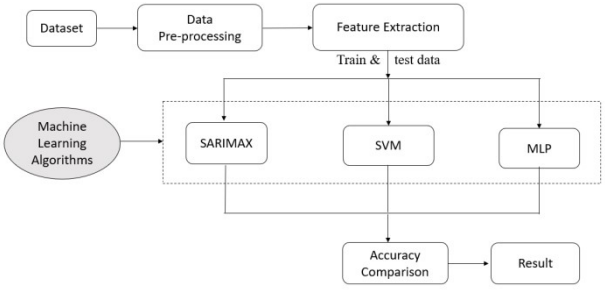
7. *Machine learning based sales forecasting system*

Sales forecasting tries to anticipate future demand for sales numbers, reserve the quantity of products, and implement marketing strategies based on the outcomes of the forecasting. In addition to avoiding wasteful overstocking, maintenance expenses, and sales demand patterns, an accurate and reliable forecasting system plays a significant role in decision-making activities in the departments relating to sales, production, purchasing, finance, and accounting. The results of the sales forecasting might be influenced by numerous things. However, during their research, researchers only collected a small number of information. In this project, both internal and external factors are examined. These include the weather, fuel costs, holidays, CPI, unemployment rate, and discounting tactics, all of which are thought to have a direct impact on consumer demand for sales in supermarkets and their departments. This analysis also takes into account the opinions of subject-matter experts. An experiment process is started to assess the performances of machine learning algorithms after past research papers and publications are used to determine appropriate approaches and algorithms. Additionally, based on those key variables, a novel forecasting approach is suggested. It was created as an accurate machine learning-based sales forecasting system for local supermarkets and the departments in Walmart USA supermarkets to meet the gaps in existing solutions.

III. METHODS

Machine Learning algorithms are used to predict the accuracy of the sales which will be used to determine whether to increase or decrease the production. The performance of these algorithms is compared using different accuracy measurement methods.

Row ID	Order ID	Order Date	Ship Date	Ship
1	CA-2017-152156	08/11/2017	11/11/2017	Second
2	CA-2017-152156	08/11/2017	11/11/2017	Second
3	CA-2017-138688	12/06/2017	16/06/2017	Second
4	US-2016-108966	11/10/2016	18/10/2016	Standa



A. Data Collection

Identify the relevant sources of historical sales data, such as point-of-sale systems, transaction databases, or online sales platforms. Collect data from as many sources as possible to ensure a comprehensive dataset. Ensure that the collected data is complete and includes all relevant information, such as sales date, item sold, quantity sold, price, and any other relevant variables. Check the quality of the collected data, including the accuracy of the data entry, any missing values, and any anomalies or outliers in the data. Remove any irrelevant data and clean the data before further analysis. Ensure that the data is in a format that is suitable for machine learning algorithms. This may involve converting the data into a structured format, such as a .csv or Excel file, and labelling the data appropriately.

B. Data Preprocessing

The process of Eliminating any duplicates, missing data, or outliers from the gathered data. This can be accomplished through data cleaning methods including interpolation, imputation, or the complete removal of the problematic data points. To guarantee that all variables have the same range of values, normalize or scale the data. Techniques like min-max scaling, normalization, or log transformation can be used for this. With the gathered data, find and extract pertinent features that can be utilized to train machine learning algorithms. To enhance model performance, this can entail adding new features or combining current features. Dividing the pre-processed data into training and testing sets will allow you to test the machine learning models' accuracy with a subset of the data.

C. Feature Extraction

Extract time-based features such as day of the week, month, season, or year to capture any temporal patterns in the sales data. Create lagged features by shifting the sales data by a certain number of time periods. Create rolling window features by calculating statistics such as mean, median, or standard deviation over a rolling window of time periods. This can capture any trends or patterns in the sales data over time. Incorporate any relevant exogenous variables, such as weather data or holiday dates that may impact sales patterns. If the sales data includes text-based features such as item descriptions or customer feedback, extract relevant features using techniques such as Natural Language Processing (NLP) or sentiment analysis.

D. Prediction

The proposed system for forecasting seasonal item sales using machine learning algorithms. SVM is a popular algorithm used for time-series forecasting. It works by finding a hyper plane that separates the data into different classes, with the aim of minimizing the prediction error. Seasonal Autoregressive Integrated Moving Average with exogenous variables (SARIMAX) is a statistical algorithm used for time-series forecasting. It models the time-series data as a combination of seasonal, autoregressive, and moving average components, and incorporates

exogenous variables to improve forecast accuracy. Multi-layer perceptron (MLP) is a type of neural network used for time-series forecasting. It works by using multiple layers of interconnected neurons to model complex relationships between the input data and the output forecasts. Ensemble methods, Ensemble methods such as random forests or gradient boosting can be used to combine the predictions of multiple machine learning algorithms. This can improve the accuracy and stability of the generated forecasts. Evaluation metrics, once the machine learning algorithms have been trained and tested on the pre-processed data, evaluate the accuracy of the generated forecasts using metrics such as Mean Absolute Error, Relative Absolute Error, or Root Mean Squared Error. He suggested method can be utilized to give accurate and dependable forecasts of seasonal item sales by utilizing various prediction approaches and assessing the correctness of the forecasts provided.

E. PERFORMANCE METRICS

i)MEAN ABSOLUTE ERROR

Regression models use the process performance metric known as Mean Absolute Error (MAE). The mean absolute error for a model with respect to a set of test data is the average of the actual values for each instance of the specific prediction errors in the test set. For instance, every forecast error is represented by the discrepancy between the anticipated value and the actual value. One metric for calculating and evaluating the performance of the machine learning model is mean absolute error.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Where:
n = the number of errors,
Σ = summation symbol (which means “add them all up”),
|xi – x| = the absolute errors.

ii)ROOT MEAN SQUARE ERROR

Root Mean Square Error, sometimes referred to as root mean square deviation, is one of the techniques most frequently used to evaluate the accuracy of forecasts. It demonstrates the Euclidean separation between forecasts and observed true values.

Calculate the residual (difference between prediction and truth) for each data point, along with its norm, mean, and square root in order to determine the root-mean-square error (RMSE). Due to the fact that it requires and uses real measurements at each projected data point, RMSE is frequently utilised in supervised learning applications.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

iii) RELATIVE ABSOLUTE ERROR

The average of the actual values (RAE) and the relative squared error serve as a straightforward predictor for both the relative absolute error and the absolute error (RSE). Hence, the error is just the entire absolute error, not the total squared error. In order to normalise the total absolute error, the relative absolute error divides the total absolute error by the total absolute error of the simple predictor.

Relative Absolute Error Ei of an individual model i.is evaluated by the equation:

$$E_i = \frac{\sum_{j=1}^{n} \left| P_{(ij)} - T_j \right|}{\sum_{j=1}^{n} \left| T_j - \overline{T} \right|}$$

IV. MACHINE LEARNING ALGORITHM

Forecasting means predicting events of the future, typically based on previous records. For a long time, statistical models were commonly used for the conducting of predictions. The role of generalization in Machine Learning has been considered. Because there isn't much previous data available for a given time series, this impact could be utilized to anticipate sales when a new store or product is introduced [4]. The seasonal autoregressive integrated moving average (SARIMA), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) are supervised learning techniques that are used to predict sales.

A. Support Vector Machine

Support Vector Machine or SVM is the common Supervised Learning algorithms used for both Classification and Regression issues. The SVM algorithm aims to build the best line or decision boundary that can divide n-dimensional space into conveniently place the new data point in the right category. The optimal choice boundary is called a hyper plane. SVM chooses extreme points vectors that help to create a hyper plane. Such extreme cases are called help vectors [9]. The equation for Support Vector Regression is:

$$f(x) = x' \beta + b$$

Multilayer Perceptron

The Multi-layer Perceptron is used in addition to the feed forward neural network (MLP). The hidden layer, the output layer, and the input layer are the three different types of layers. The input layer is where the signal is received for processing at the input layer. The required tasks, such as classification and prediction, are finished by the output layer. The input and output layers are sandwiched between an arbitrary numbers of hidden layers that make up the MLP's true computational engine. Similar to a feed forward network, data flows from the input to the output layer of an MLP in the forward direction. The MLP's neurons are trained with the use of the back propagation learning technique. Because MLPs are designed to mimic any continuous function, they are able to tackle problems that cannot be separated linearly. The primary use cases for MLP include pattern classification, recognition, prediction, and approximation. Multilayer Perceptron's equation is

$e=target-f(m2)$ or e=target−m2

where f(·) is the activation function of the output layer.

C. SARIMA

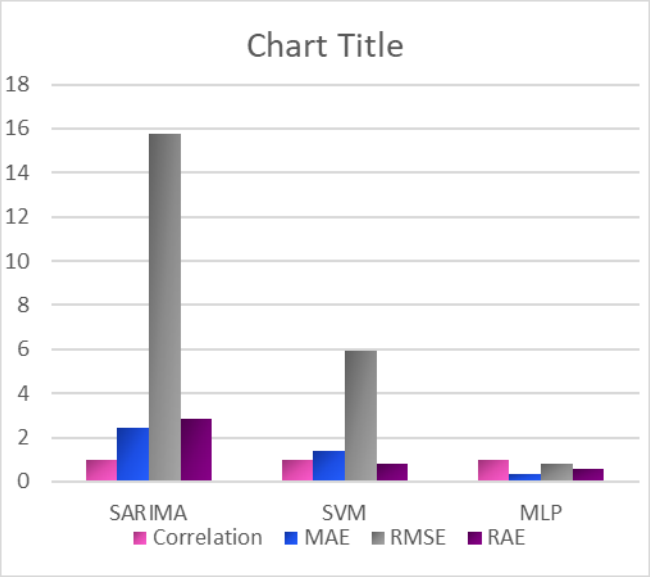
The seasonal component of unilabiate time series data is specifically supported by the ARIMA extension known as Seasonal Autoregressive Integrated Moving Average, or Seasonal ARIMA. Together with three new hyper parameters to calculate the auto regression (AR), differencing (I), and moving average (MA) for the seasonal component of the series, it also contains an additional parameter for the seasonality period. A seasonal ARIMA model is produced by adding additional seasonal components to the ARIMA. Although they also incorporate backshifts of the seasonal period, the words that make up the model's seasonal section are pretty comparable to those that make up its non-seasonal portions.

SARIMA (p, d, q) x (P, D, Q, s)

the parameters for these types of models are as follows: p and seasonal P: indicate number of autoregressive terms

V. RESULT

By utilizing evaluation criteria like mean absolute error, mean squared error, or root mean square error, you may compare the predicting accuracy of the various machine learning methods. This will make it easier to determine which algorithm is best for the available dataset. To assess the projections' accuracy, visualize the generated forecasts and contrast them with the actual sales data. Line plots, scatter plots, and time series decomposition plots are some of the methods that can be used to do this. By contrasting the anticipated values with the actual values, analyses the forecasting inaccuracies. Recognize any systematic biases or trends in the errors that can point to forecasting model flaws.



VI. CONCLUSION

Sales forecasting is essential to the business sector in all industries. Sales revenue analysis will assist in obtaining the information required to estimate both the revenue and the income with the aid of the sales predictions. The proposed system for forecasting seasonal item sales using machine learning algorithms can provide accurate and reliable forecasts that can help businesses optimize their inventory management and increase sales revenue. Data collection, data pre-processing, feature extraction, prediction, and outcome analysis are only a few of the system's crucial phases. Methods like the multi-layer perceptron, the seasonal autoregressive integrated moving average with exogenous variables, and the support vector machine can be used for forecasting. Performance indicators including correlation, mean absolute error, root mean square error, and relative absolute error are calculated for each of the three machine learning algorithms. According to research, the Multilayer Perceptron is the most effective technique for forecasting sales.

ACKNOWLEDGMENT

We would acknowledge our guide for development of the Sales Prediction using Machine Learning Algorithm with full support and guidance. We would like thank the clinic who contribute their data to study.

REFERENCES

[1] Ali Heydarzadegan, Yaser Nemati, Mohsen Moradi. Evaluation of Machine Learning Algorithms in Artificial Intelligence. International Journal of Computer Science and Mobile Computing

[2] Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: A comparison of Artificial Neural Networks and traditional methods. Journal of Retailing and Consumer Services, 8(3), 147–156. Alpaydin, E. (2009).

[3] Anjali Jagwani. A review of machine learning in education. Journal of Emerging Technologies and Innovative Research.

[4] Ayushi Chahal, Preeti Gulia. Machine Learning and Deep Learning. International Journal of Innovative Technology and Exploring Engineering (IJITEE).
[5] Bhuvaneswaria. A, Venetiaa. T.A. "Predicting periodical sales of products using a machine learning algorithm". Department of Computer Applications PSG College of Technology Coimbatore, India.
[6] Bohdan M Pavlyshenko. Machine-learning models for sales time series forecasting. Data, 4(1):15, 2019
[7] Chakraborty, A., & Kar, A. K. (2017). Swarm intelligence: A review of algorithms. Nature-Inspired Computing and Optimization, 475–494.
[8] Gensler, A., Henze, J., Sick, B., & Raabe, N. (2016). Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM neural networks. In Proceeding of the IEEE international conference on systems, man, and cybernetics (SMC) (pp. 002858–002865). IEEE.
[9] Giering, M. (2008). Retail sales prediction and item recommendations using customer demographics at store level. ACM SIGKDD Explorations Newsletter, 10(2), 84–89.
[10] Grigorios Tsoumakas. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1):441–447, 2019
[11] Indrastanti R. Widiasari, Lukito Edi Nugroho, Widyawan. Deep learning multilayer perceptron (MLP) for flood prediction model International Conference on Innovative and Creative Information Technology
[12] JN Hu, JJ Hu, HB Lin, XP Li, CL Jiang, XH Qiu, and WS Li. State-of-charge estimation for battery management system using optimized support vector machine for regression. Journal of Power Sources.
[13] John. D. Kelleher, Brian Mac Namee, Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies
[14] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1):69–88, 2016.
[15] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. American Economic Review, 105(5):481–85, 2015.
[16] Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor. "Sales prediction using machine learning algorithms". International Research Journal of Engineering and Technology. June 2020.
[17] Robert Aberg, Christopher Dahlen. "Predicting sales in a food store department using machine learning".
[18] Sekban, Judi. "Applying machine learning algorithms in sales prediction." (2019).
[19] Stephen Marsland. Machine learning: an algorithmic perspective. CRC press,2015.
[20] Yasaman Ensafia, Saman Hassanzadeh Amin, Guoqing Zhang, Bharat Shah.Time-series forecasting of seasonal items sales using machine learning –A comparative analysis. International Journal of Information Management Data Insights 2 (2022)