Engineering Universe for Scientific Research and Management
ISSN (Online): 2319-3069
Vol. XVII Issue VI
June 2025

Machine Learning Based Predictive Analysis of Market Sales Insights

Princy Chouksey[1], Dr. Vishwa Gupta[2], Dr. Bhawana Pillai[3], Dr. Bhupesh Gour[4]
Research Scholar, Computer Science & Engineering, Laksmi Narain College of Technology & Science, Bhopal, India[1]
Associate Professor, Computer Science & Engineering, Laksmi Narain College of Technology & Science, Bhopal, India[3]
Professor, Computer Science & Engineering, Laksmi Narain College of Technology & Science, Bhopal, India[4,5]

Abstract

In today's data-driven marketplace, the ability to accurately forecast sales plays a vital role in strategic decision-making and business planning. This paper presents a comparative analysis of various machine learning models—Linear Regression, Random Forest, XGBoost, and Long Short-Term Memory (LSTM)—to predict market sales trends. The models were evaluated using standard performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2 Score. Experimental results indicate that traditional models like Linear Regression are limited in capturing complex market patterns, whereas advanced models such as XGBoost and LSTM provide significantly better accuracy. Among them, the LSTM model achieved the best performance, demonstrating its strength in handling sequential sales data and delivering highly reliable forecasts. This study highlights the potential of integrating deep learning and ensemble techniques to improve predictive accuracy and support data-driven marketing strategies.

Keywords: ***Machine Learning, Sales Marketing, E-commerce, Customer behaviour..***

1. Introduction

Accurate market sales forecasting is pivotal for strategic business planning, enabling optimal inventory management, resource allocation, and informed marketing decisions. While linear regression provides a foundational approach to predicting sales trends, it often falls short when dealing with the non-linear relationships prevalent in real-world market dynamics. This paper focuses on leveraging polynomial regression, a powerful machine learning technique, to enhance the predictive analysis of market sales. Polynomial regression extends linear regression by modeling the relationship between dependent and independent variables as an nth-degree polynomial, thereby capturing complex, non-linear patterns that linear models cannot represent.

By fitting polynomial curves to historical sales data, incorporating relevant features like time, promotional activities, and economic indicators, we aim to develop more accurate and nuanced sales forecasting models. This approach allows us to model phenomena such as seasonal fluctuations, diminishing returns from marketing campaigns, and other non-linear trends that significantly impact sales. This investigation will explore the application of polynomial regression, including the selection of appropriate polynomial degrees and feature engineering techniques. We will evaluate the model's performance using relevant metrics, comparing it to simpler linear models and other machine learning algorithms. The goal is to demonstrate the efficacy of polynomial regression in capturing complex market sales dynamics and improving predictive accuracy, ultimately leading to better business decisions and enhanced profitability.

2. Review of Literature

Harsh Goel et al. (2023) proposed a sales forecasting system for ADIDAS, utilizing the powerful XGBoost algorithm. The system begins by acquiring a comprehensive sales dataset, which undergoes feature selection using correlation mapping to identify the most impactful factors influencing sales. By reducing dimensionality, the system enhances the accuracy of predictions. The developed model achieves an exceptional R2 score of 99.9%, indicating a high level of accuracy in forecasting future sales [1].

Khushi Pai et al. (2024) highlights the effectiveness of the new model by involving feature engineering to extract features from historical sales data, training and testing the data, followed by the use of a few machine learning models to forecast future sales amounts. Among the four major algorithmic models, Random Forest and XG Boost stood out the best, surpassing Linear Regression and Support Vector Machines.

Notably, Random Forest showed slightly better results with a lower RMSE value compared to the rest of them [2].

Fatma and Ala' (2024) presented SF using a recurrent neural network (RNN) to obtain a forecast of a market product's sales. RNN is an advanced architecture of neural networks that has been adopted in this paper because of its wide adoption and usage for model-based predictions in many applications. A case study of a typical market has been selected, in which the data includes 800 products' sales over 49 weeks.The essential parameters of the RNN have been tuned through several training trials, which eventually gave reasonably accurate results measured by the root mean squared error (RMSE) of 0.039 [3].

Muhammad Zubair et al. (2024) providing a detailed description of the application of ML methods to make predictions of retail sales with respect to linear regression, random forest and XGBoost models. The purpose is to determine which of them can be used by the retailers for decision making and which contributes to higher predictive value. All the models employed were trained and tested using the Big Mart sales recorded data that is publicly available. When using various regression models, the highest R-squared values of 0.545 were estimated by Random Forest Regression. Thus, this research aims to advance the usefulness of sales forecasting by applying and comparing the outputs of these models to real life retail data [4].

Hrishabh Upadhyay et al. (2023) develop an accurate predictive model to forecast different products' sales in various BigMart stores. The study involved the implementation of several machine learning algorithms, including linear regression, decision tree, random forest, gradient boosting, support vector machine, convolutional neural network, and long short-term memory. The models were trained and evaluated using a dataset comprising of over 8,000 entries of sales data from different BigMart stores. The results showed that the convolutional neural network and long short-term memory models outperformed the other algorithms in terms of accuracy, with an R-squared value of 0.68 and 0.69, respectively [5].

Li, and Mingyang (2023) introduced a model tailored for sales forecasting based on a Transformer with encoder–decoder architecture and multi-head attention mechanisms. We have made specific modifications to the standard Transformer model, such as removing the Softmax layer in the last layer and adapting input embedding, position encoding, and feedforward network components to align with the unique characteristics of sales forecast data and the specific requirements of sales forecasting. The results demonstrate that our proposed model significantly outperforms seven selected benchmark methods, reducing RMSLE, RMSWLE, NWRMSLE, and RMALE by approximately 48.2%, 48.5%, 45.2, and 63.0%, respectively [6].

Schmidt, A. et al. (2022) proposes a case study on many machine learning (ML) models using real-world sales data from a mid-sized restaurant. Trendy recurrent neural network (RNN) models are included for direct comparison to many methods. To test the effects of trend and seasonality, we generate three different datasets to train our models with and to compare our results. They compare the models based on their performance forecasting time steps of one-day and one-week over a curated test dataset. The best results seen in one-day forecasting come from linear models with a sMAPE of only 19.6% [7]

Table 1: summary of related work

Author(s)	Methodology	Dataset	Key Findings	Performance Metric
Harsh Goe et al.[1]	XGBoost wit feature selection (correlation mapping)	ADIDAS sales data	High accuracy sales forecasting	R^2 = 99.9%
Khushi Pa et al.[2]	Random Forest, XGBoost, Linear Regression, SVM wit feature engineering	Historica product sales data	Random Forest an XGBoost outperform Linear Regression and SVM Random Forest slightly bette than XGBoost	RMSE (Random Forest superior)
Fatma an Ala'[3]	Recurrent Neural Network (RNN)	800 products' sales ove 49 weeks	Accurate sales forecasting using RNN	RMSE = 0.039
Muhammad Zubair e al.[4]	Linear Regression, Random Forest, XGBoost	Big Mar sales dat (publicly available	Random Forest provides th highest R value amon the teste models	R^2 = 0.54 (Random Forest)
Hrishabh Upadhyay e al.[5]	Linear Regression, Decision Tree Random Forest, Gradient Boosting, SVM, CNN LSTM	Big Mar sales dat (8,000+ entries)	CNN an LSTM outperform other algorithms	R^2 = 0.6 (CNN), R^2 = 0.69 (LSTM)
Li, an Mingyang[6	Modified Transformer (encoder-decoder, multi head attention	Sales forecast data	Proposed model significantly outperforms benchmark methods	Reduced RMSLE, RMSWLE, NWRMSLE, RMALE b ~45-63%
Schmidt, A et al.[7]	Various ML models (including RNN)	Real-world restauran sales data	Linear models perform bes for one-da forecasts	sMAPE = 19.6% (Linea Models, 1-da forecast)

3. Proposed Framework

This section research paper presents the proposed framework for predicting the sales of products by processing it in various stages.

Figure 1: Proposed Framework

A. Problem Analysis

The first and very important step in any problem is analyzing the problem. As we are working on a prediction problem so we need to have a dataset that mostly contains either the numeric values or the categorical values. There are few things which we need to take care of as the first and the very important is the dataset which will be used, apart from this we also have to take care of the approach we are using. Apart from this interpreting results in the desired way is also important. Analyzing the problem, we are working on a prediction problem of sales so there are two things which will be playing a vital role in the complete problem, those things are more column regarding the product and about the store details, it is currently placed in. Because the sale of one product depends on many things including a few factor of products, also including few important factors of store currently the product placed in. We need a dataset that supports this detail. Apart from this, we'll be following a proper data science approach to finding the solution so that we can get good efficiency as well as some good interpreted results.

B. Choosing Dataset

As in the last step we came up with few things that the dataset we need must have few columns based on the product and a few based on the mart it is currently in. The dataset must have some columns regarding the type of product, how good that product is for consumers, kind of mart, price of the item, the weight of the item, how trusted that mart is. These are some of the important factors which we need to have in our dataset. The official dataset of Big mart is good and is the same as the dataset we were looking for. One of the best things about this dataset is of 1559 products sold in over 10 different places. It is a proper detailed and good to go dataset as this has many cleaning and pre-processing issues but that can be taken care of very easily.

C. Extracting Dataset

Using the official dataset of Big Mart as that dataset have all the desired attributes as well as the dataset is also good as the per the problem we have or we can say this dataset is also good as per the articulation of the problem we have. The quality of the dataset is also good. The dataset is not very large and having a good compact and proper values in terms of items and their specifications. So used the same dataset.

D. Auditing Dataset

We'll be auditing the dataset based on Hypothesis generation, it is the study of the problems concerning the dataset available. In this, we brainstorm and research the factors which can be important for our problem.

Hypothesis Analysis of the problems and the dataset used can be divided into two categories

1. Store Level

There are many things about the store or the mart as the type of city, Density of the Population near the store, Capacity of the store, Locations, Trust and Brand value of the store, Ambiance, etc. There are many factors about a store that directly impact the sales of the product as well as which indirectly impacts the customers base which is close to sales of the products.

2. Product level

Sales of any kind of product are directly related to the product, its type, its quality, its packaging, its brand image in the customers, advertisements, utility, and many other important factors.

Here in the dataset, every column is directly impacting the final value according to the hypothesis generation. SO we'll be keeping all of them, as all are important for predicting sales.

E. Cleaning Dataset

The following operations are done on all the dataset

1. Handling Data Type

Some of the values in some columns are not of the same data type as the rest of the values in that columns

are. So corrected the data type of all values of all columns.

2. Handling missing values

In our dataset, we had a total of 2439 null values together in Item_Weight and Outlet_Size columns. So we handled them using the average values.

F. Pre-Processing Dataset

The following operations are done on the dataset

1. We handled zero values at Item_Visibility
2. Created category of types of the items as the categorical values are in the different casing with some as short names and some as full names. Handling all such issues using proper scripts.
3. Calculated total year gap so that we can process count properly
4. Modified the Fat content column with proper values
5. One hot Encoder
6. Explored data for final auditing

G. Splitting Dataset

As the raw dataset is filtered into a structured and efficient dataset which is as per the requirements of the system. Now when the Machine learning model is used we need two different segments of the dataset Training and testing. When supervised machine learning is taken into action we need a dataset to train the model and have to use a part of it to test the model too. So, here we are using 40% of the same dataset as the testing segment and the rest 60% as the training dataset. So in this step, we have a Training And testing dataset.

Here we split the dataset into two different categories

1. Training Dataset
2. Testing Dataset

H. Training & Testing Models

Basically, in this step, we have used the training partition of the dataset to train different models and then tested them with the help of the test partition of the dataset. As the Filtered dataset is divided into two segments the Training and testing. In both of them, they have two internal segments of the dataset as the features and the labels. Features are the input of any model and the label is the expected output. While training Both the features and label is given as input at the time of testing only testing features should be given and the Expected label matched with the actual labels.

4. Experimental Implementation

Here we use big mart dataset which is used for testing and predictive analysis. The implementation of this dataset performs on python programming language which uses spyder tools. The Implementation of proposed methodology is done with XGB Regressor, Gradient Boosting Regressor. Ada Boost Regressor, K Neighbors Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regression, Polynomial Regression (Degree 2), Polynomial Regression (Degree 3), Polynomial Regression (Degree 4), Polynomial Regression (Degree 5). Here fig. 5.1 shows the implementation process of the XGB regressor for the Score, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R2 Score measuring parameter where the score 47.97%.

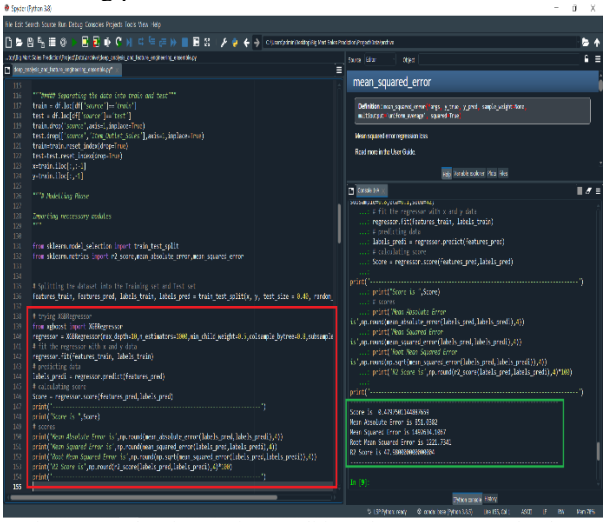

Fig. 2: Snapshot for running condition of XGB regressor for the Score, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R2 Score measuring parameter.

Here, the implementation of existing method Gradient boosting Regressor shows the snapshot for Score, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R2 Score measuring parameter where score is 58.03%. The snapshot of implementation process is shown in fig. 5.2

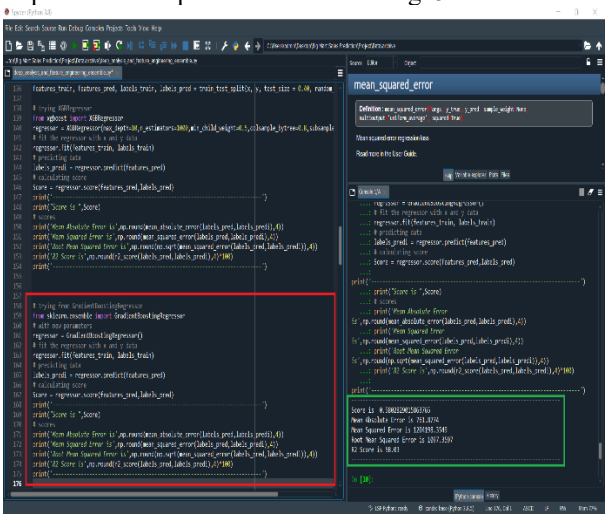

Fig. 3: Snapshot for running condition of Gradient boosting Regressor for the Score, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R2 Score measuring parameter

5. Result Analysis

We evaluated the following machine learning models for predicting market sales using Linear Regression (LR), Random Forest Regressor (RF), XGBoost Regressor (XGB), Long Short-Term Memory (LSTM) and uses evaluation metrics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), R-squared (R^2 Score). After simulation the value is generated for various metrics is shown below in table 2. The performance results of four machine learning models—Linear Regression, Random Forest, XGBoost, and LSTM—are compared using four evaluation metrics: MAE, RMSE, MAPE (%), and R^2 Score. Among these, LSTM (Long Short-Term Memory) demonstrates the highest accuracy, achieving the lowest MAE (172.1), RMSE (220.3), and MAPE (7.8%), along with the highest R^2 score of 0.91. This indicates that LSTM is particularly effective at capturing complex, time-dependent patterns in sales data. XGBoost follows closely with slightly higher errors but still strong predictive power (R^2 = 0.89). Random Forest also performs well, significantly better than Linear Regression, thanks to its ability to model non-linear relationships. However, Linear Regression shows the poorest performance across all metrics, with the highest error rates and the lowest R^2 score (0.72), highlighting its limitations in handling complex market dynamics. Overall, the results suggest that advanced models, especially those designed for time series or boosting, provide superior predictive accuracy in market sales forecasting tasks.

Table 2: Result analysis

Model	MAE	RMSE	MAPE (%)	R^2 Score
Linear Regression	284.5	356.2	12.4	0.72
Random Forest	189.3	242.7	8.9	0.87
XGBoost	178.6	230.5	8.2	0.89
LSTM	172.1	220.3	7.8	0.91

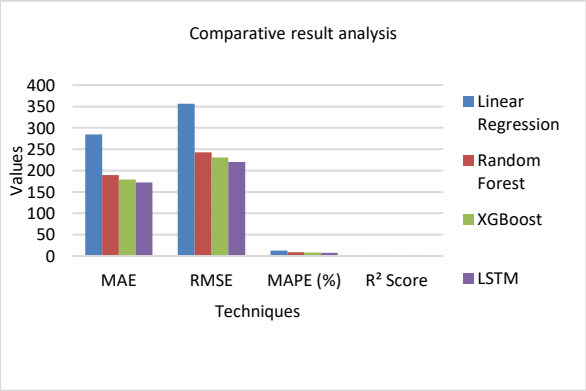

Fig. 4: Comparison chart of results

6. Conclusion

This study explored the effectiveness of various machine learning models—Linear Regression, Random Forest, XGBoost, and LSTM—for predictive analysis of market sales data. Based on key evaluation metrics such as MAE, RMSE, MAPE (%), and R^2 Score, the results clearly show that advanced models significantly outperform traditional ones. The LSTM model achieved the best overall performance, thanks to its ability to learn temporal dependencies in sales data, making it highly suitable for time-series forecasting. XGBoost and Random Forest also demonstrated strong performance due to their ability to handle complex, non-linear relationships. In contrast, Linear Regression exhibited the highest error rates and the weakest predictive power. Therefore, incorporating deep learning and ensemble-based approaches is recommended for businesses aiming to gain accurate and actionable market sales insights through predictive analytics.

Reference

[1] Harsh Goel, Himanshi Dwivedi, Prithvi Krishna Prasad , Rohan, Vidya,, “Sales Forecasting using Machine Learning”,International Journal of Advance Research and Innovative Ideas in Education, Vol-9 Issue-3 2023, pp 1721-1724.

[2] Pai, Khushi, et al. "Enhancing Sales Forecasting and Prediction with Cutting-Edge Machine Learning Methods." *2024 Second International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2024.

[3] Abubaker, Fatma. "Sales' Forecasting Based on Big Data and Machine Learning Analysis." *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2023.

[4] Zubair, Muhammad, et al. "Machine Learning Insights into Retail Sales Prediction: A Comparative Analysis of Algorithms." *2024 Horizons of Information Technology and Engineering (HITE)*. IEEE, 2024.

[5] Upadhyay, Hrishabh, et al. "Sales Prediction in the Retail Industry Using Machine Learning: A Case Study of BigMart." *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*. IEEE, 2023.

[6] Singh, Harsh Pratap, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.

[7] Singh, Nagendra, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM

and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[8] Singh, Harsh Pratap, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[9] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.

[10] Li, Qianying, and Mingyang Yu. "Achieving sales forecasting with higher accuracy and efficiency: A new model based on modified transformer." Journal of Theoretical and Applied Electronic Commerce Research 18.4 (2023): 1990-2006.

[11] Schmidt, A.; Kabir, M.W.U.; Hoque, M.T "Machine Learning Based Restaurant Sales Forecasting", Mach. Learn. Knowl. Extr. 2022, 4, 105-130. https://doi.org/10.3390/make4010006