JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

Sales Forecasting Using Machine Learning Algorithm in the Retail Sector

Saira Malik[1*], Muhibullah Khan[1], Muhammad Kamran Abid[1], and Naeem Aslam[1]

[1]Department of Computer Science, NFC IET, Multan, Pakistan.
[*]Corresponding Author: Saira Malik. Email: 2k19mscs116@nfciet.edu.pk

Received: November 21, 2023 **Accepted:** January 29, 2024 **Published:** March 01, 2024

Abstract: The ability to predict future sales is essential for modern firms. The already difficult work of sales forecasting is made much more difficult by a lack of data, missing data values, or outliers. Regression has a stronger relationship with sales forecasting complexity than time series does. Intricate patterns in the dynamics of sales that also involve a range of risk factors can be discovered using machine learning algorithms and supervised machine learning techniques. A company's sales projections need to be correct for it to succeed. By utilizing a reliable sales projection model, businesses may identify potential risks and make smarter decisions. In this study, Rossmann sales data will be analyzed using the Extreme Gradient Boosting (XG-Boost), FB-Prophet, and autoregressive integrated moving average (ARIMA) prediction models. A corporation can reduce costs associated with excess inventory, make future plans, and boost profitability with the use of an accurate sales forecast. Therefore, the model needs to be assessed using statistical techniques like R2, RMSE, and MAE. To determine if models are more accurate at predicting sales, the results are employed.

Keywords: Sales Forecasting; Machine Learning; Time Series; FB Prophet; ARIMA; Extreme Gradient Boosting.

1. Introduction

Since data engineering (DE) & analytics have advanced, business analytics have become a crucial component of all business support systems. In this regard, demand and sales forecasting are key components of business analytics solutions, & businesses need to have a precise sale prediction to employ their S & OP (operations and sales processes). The growth of the e-commerce and logistics sectors in recent years has altered the scope and pace of supply chain demand [1][2]. When it comes to marketing, sales, manufacturing, and procurement planning, manufacturers and retailers may make better decisions with the help of a detailed prediction of a product's prospective sales**.** In the past, Businesses created goods without considering customer demand or sales volume. For each manufacturer to decide whether to increase or decrease the production of numerous units, information about the demand for products on the market is required [3]. Managers at firms and companies frequently forecast sales at random. However, competent managers are become harder to find and less dependable. Sales forecasting can be assisted by computer programmers that can step in for capable managers, when they are absent or provide them with the information, they need to make the best decision by giving potential sales projections. One method to put this idea into effect is to try to create a computer programmer that simulates the skills of professional managers [4]. Without considering these principles, businesses that compete in the market face the threat of failing. Different parameters are applied for the demand and sales analysis for each business. Businesses involved in the production, distribution, or retail of goods can profit enormously from accurate and timely revenue forecasting, also known as revenue forecasting or sales forecasting, in today's fiercely competitive economy and rapidly evolving customer environment [5].

Predicting future sales to aid decision-makers in making better choices regarding planning, supplying, production, and marketing activities is termed sales forecasting. Companies employ a variety of tactics to sustain their sales levels over an economic year. One of these strategies is to organize sales promotions where a variety of goods are provided at lower costs to the merchants in demand to display additional quantities of good for a set period of time. In this project, we will do a predictive study of Rossman's sales using a large data set[6]**.** When models & statistical analysis are employed to create forecasting, sales forecasting is typically thought of as a time-series problem. If machine learning techniques are used to find underlying patterns and trends in historical time series sales data and then use those patterns and trends to predict future transactions, either in the long-term or the short-term, it may also be thought of as a regression problem.

The benefits of sales prediction vary depending on the activity of the supply chain. Sales forecasting may help manufacturing companies with all their planning & decision-making procedures, from inventory management & production planning to marketing initiatives and sales. The lack of an accurate and effective demand prediction solution results in faulty forecasting despite the significance of sales forecasting [7].

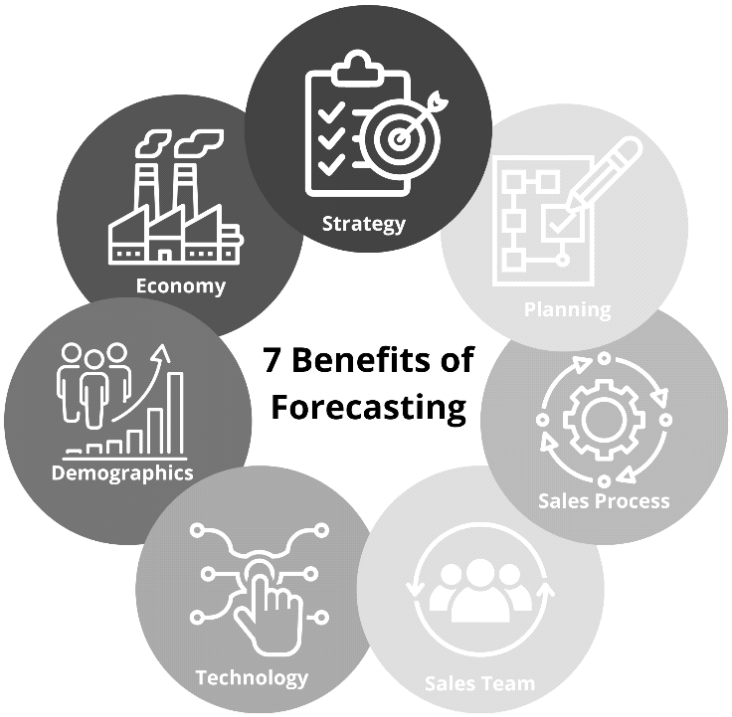

Figure 1. Benefits of Sales Forecasting

It is challenging to estimate the amount of money an organization will save as a result of accurate sales forecasting, even if the foundation of these benefits can be assumed to be. For manufacturers, distributors, and retailers in general, sales forecasting is an important factor to take into account. It is also a key task for many businesses engaged in supply-chain activities. The benefits of sales prediction vary depending on the supply & demand chain activity. Supply management, production planning, sales & marketing initiatives—all of these planning & decision-making processes—may benefit from sales forecasting in manufacturing organizations [8].

A collection of data points that were gathered at regular intervals between related points in time is known as a time series. Using only historical data, time series forecasting generates predictions. There must be enough trustworthy historical data available [6]. The classical time series forecasting techniques are employed in this thesis to establish a baseline against which the machine learning techniques can be measured. A general issue with significant practical value across many fields is time series. Because it makes it possible to infer, with a certain amount of inaccuracy, what a series' future values will be from its past values [9].

The design of the prior system could not produce precise results data that could classify the sales appropriately[10]. Some machine learning regressors have been tested on datasets and can predict the results. We need to develop a system that can predict sales more precisely, and by using machine-learning techniques, we can increase the efficiency and precision of our system.

Two major contributions made by our work are as follows:

1. To determine the crucial characteristics that have the biggest effects on sales.
2. To compare the best algorithms to get the best sales forecasting model.

Section I contains a through introduction to sales forecasting, machine learning, benefits of sales forecasting and objectives. Section II contains a discussion of the literature review. In this part, we examined a few studies that employed machine learning techniques to forecast sales. We discussed the applied approaches in Section III. The suggested algorithms and the outcomes of the experiments are presented in Section IV. Section V presented the Conclusion.

2. Literature Review

In this study, hybrid demand forecasting methods based on machine learning techniques such as ARIMAX and neural networks were developed. The devised technique took time series and explanatory factors into account. The technique was used and assessed in relation to a functioning product. It was revealed that there were statistically significant variations in the improvement of supply chain performance between traditional and ML-based demand forecasting strategies. [11].

Applied several machine learning techniques, such as RFR (Random Forest Regression), SVR (Support Vector Regression), & DTR (Decision Tree Regressor), as well as deep learning techniques—ANN (Artificial Neural Network), & LSTM (Long Short-Term Memory)—to deal with demand prediction based on marketing expenses [12]. As a result, the accuracy of demand forecasting was examined using a real market dataset from a television manufacturer that included advertising costs, sales & demand forecasting via selected machine-learning approaches. Therefore, it was shown that Long Short-Term Memory surpassed other models in delivering extremely accurate predicting results. This study proposed a neural network sale prediction algorithm to forecast Walmart sales. NN models using datasets provided by the Kaggle platform were also evaluated. Studies revealed that the NN model performed better than other machine learning models. In comparison to the SVM and linear regression techniques, the RMSE value was 2.92 and 2.58 smaller, respectively [13]. To forecast product sales, the extreme gradient boosting (XG-Boost) algorithm a supervised learning approach was used in this study [14]. If the provided data are inadequate to attain the desired accuracy, the generated model was trained and evaluated using the AD (Augmented Data) approach. The AD technique offered a fair level of accuracy when compared to the results utilizing the baseline dataset with few records. Using the RMSE and MSE (Mean Square Error) metrics, it was feasible to minimize the prediction error by almost an order of magnitude.

The study [15] used point-of-sale (POS) data gathered over three years to create a sales prediction model based on a day's sales. As a consequence, an L1 regularization-based deep learning model obtained an 86% accuracy rate in forecasting sales. Even when there were hundreds of characteristics for each of the several product categories, the estimated accuracy did not decrease by more than 7%. In contrast, the accuracy decreased by roughly 13% when the logistic regression model was used. These results demonstrate that deep learning was particularly well suited for creating models with various features. The XG-Boost model's implementation and evaluation were described in this study[16]. To enhance data quality, XG-Boost made full use of precisely crafted data filler algorithms for missing values. Based on a thorough investigation of the effects of numerous variables on sales via information collection and data correlation, the most indicative characteristics from the set of features for prediction were selected. Numerous tests have shown that Fore XG-Boost can produce accurate predictions with little overhead.

In this research, several forecasting models were used [17]. Two conventional time-series forecasting techniques, Triple Exponential Smoothing and SARIMA (Seasonal Autoregressive Integral Moving Average), were employed. Then, more advanced methods like the Prophet framework, LSTM (Long Short-Term

Memory), & Convolutional Neural Networks (CNN) were applied. The models' outputs were assessed using a variety of accuracy measurement methodologies, including MAPE (Mean Absolute A percentage Error) & the RMSE (Root Means Squared Error). The outcomes demonstrated that the Stacked LSTM approach beat out the competition. The outcomes also showed that the CNN and Prophet models performed well. In this investigation, The kernel neural network (and linear regression approaches were used to evaluate Rossmann sales data. An organization may increase earnings, save expenditures related to excess inventory, and make wise decisions for the future by using an accurate sales prediction. The model must thus be evaluated using statistical methods like the RMSE and MAPE. The results were used to decide whether a classifier is more suitable for forecasting sales. [10].

The goal of this research was to create a three-step, cluster-based, data-driven demand forecasting method for the retail sector. Customers were first divided into groups according to their recent, frequent, and monetary (RFM) traits. Customers with comparable purchasing habits were identified as a group, establishing an ordered relationship between their transactions. The second step was forecasting demand for each consumer segment using time-series analytic techniques. Last but not least, the predicting results from several time series techniques were combined using Bayesian model averaging (BMA). Analysis of related case studies is used to demonstrate the applicability of the suggested approach and to show how the daily demand prediction accuracy has improved [18].

The CNN-LSTM (Long, short-term Memory-Convolutional Neural Network) model was proposed in research to forecast retail demand. [19]. The Swish Activation Function is present in this particular model. It outperformed ReLU's (Rectified Linear Unit), the most widely used and traditional activation function. Data from 10 retailers, with 50 goods each, were used as the input. In the proposed work, models for forecasting sales included convolutional neural network models, multilayer perceptrons, long short-term memory cells, and and others. The experiment's findings recommend utilizing the CNN-LSTM Model because of its significantly lower RMSE (Root Mean-Squared Error).

Based on a careful examination of the characteristics of a specific algorithm approach, long and short memories neural network models, and in conformity with the information set supplied by a chain of supermarkets in Kaggle, an XG-Boost-LSTM artificial neural network a fusion framework for forecasting sales and a conventional time-series data forecasting framework were built to contrast the experimental results. According to the experimental findings, the XG-Boost-LSTM neural network forecast model surpasses the time series prediction model in terms of accuracy, which can offer a crucial scientific basis for the chain of supermarket sales forecasting [20].

In this study [21] a technology called FB Prophet was suggested for using data from supermarkets to predict sales. A few prediction models, including the ARIMA (Autoregressive Integrated Moving Average), the model additive model and the FB-Prophet model, have been studied in the suggested research effort. FB Prophet was shown to be a better prediction model from the proposed research effort in conditions of reduced error, better forecast & better fitting.

A machine learning-based model for accurate and efficient sales forecasting was put forth in this paper [3]. The initial purpose of feature engineering was to extract features from past sales data. Additionally, the same attributes were utilized for projecting future sales amounts using extreme Gradient Boosting (XG-Boost). The results of the experiment using a publicly accessible dataset of Walmart retail products supplied by the Kaggle database competition revealed that the recommended model performed remarkably well for revenue predictions with fewer computational resources and memory.

In this study, regression techniques were applied to the issue of predicting sales. LASSO, Linear, Ridge as well, Random Forests (RF), a Decision Tree (DT), an Extended Tree, and XG-Boost were the regression techniques employed for these forecasts. Additionally, using a publicly accessible dataset and performance indicators like MAE, WMAE (Weighted MAE) & RMSE, the efficacy of the mentioned models was determined. The findings of the trial show that the RF (random forest) strategy produces superior results to other approaches [1]. Microsoft's Azure ML Studio models were applied to Walmart sales data. The following regression techniques were applied: NN regression, linear regression, Bayesian regression, BDTR (boosted

decision trees regression), and DFR (decision forest regression). The seasonal ETS, non-seasonal ETS, SARIMA, or (seasonal ARIMA), non-seasonal ARIMA, a drift method, mean method, and naive approach of time series analysis were utilized as well as these regression techniques. The outcomes showed that this sales data was best handled using Boosted DT Regression [22].

This study [23] employed multivariate regression models with data from time series models, stock market values, and social media [24]. LSSVR models (least squares support vector regress) were used to handle multivariate data. Three data sources were utilized to forecast monthly sales: stock prices, sentiment ratings from tweets, and hybrid data. Both stock market prices and sentiment ratings derived from tweets are included in the hybrid data. Among the time series models are the naïve approach, the exponentially smoothed (ES) approach, the model of ARIMA, the SARIMA approach, the BPNN, or (backpropagation neural network) structure, & the LSSVRTS model. The findings show that compared to other models employing unique data, LSSVR models can generate more accurate results when using hybrid data and de-seasonalizing strategies.

Table 1. Summary of Existing Work

References	Year	Contribution	Findings
[24]	2021	Three machine learning models, Two deep learning and Two linear models were used	Adding more cost and date information aided with prediction, ML and deep learning models did not appear to have any clear advantages in prediction
[25]	2021	Regression and Classification Model (XG-Boost, Polynomial regression, linear regression, and Ridge regression) were used.	Ridge and XG-Boost gave better prediction than the Linear & polynomial approaches.
[34]	2021	Linear Regression, Ridge, XG-Boost were used	XG-Boost gave better accuracy than traditional machine-learning models
[26]	2020	Combine XG-Boost and feature engineering	XG-Boost Performed better than the Ridge algorithm and logistic regression algorithm
[27]	2020	proposed a GBDT-based Light GBM model	Light GBM had better prediction compared with SVM and linear regression model
[3]	2020	XG-Boost and Feature Engineering were used	XG-Boost showed good results with less processing power and memory

[28]	2020	Data mining, the Best-Worst Method (BWM), and the Random Forest Model)	This technique could be used as a backup plan to enhance customer service, improve inventory management, and foresee future financial gains.
[29]	2019	RNN and ARIMA along with web crawler technology implemented	It was anticipated that the deep learning approach, characterized by a neural network, would overtake other forecasting techniques

3. Materials and Methods

The approach used to complete this thesis work is described in the section that follows, along with a brief overview of the data analytic steps taken to prepare the data for forecasting. The design and structure of the proposed work is presented in Figure 2.

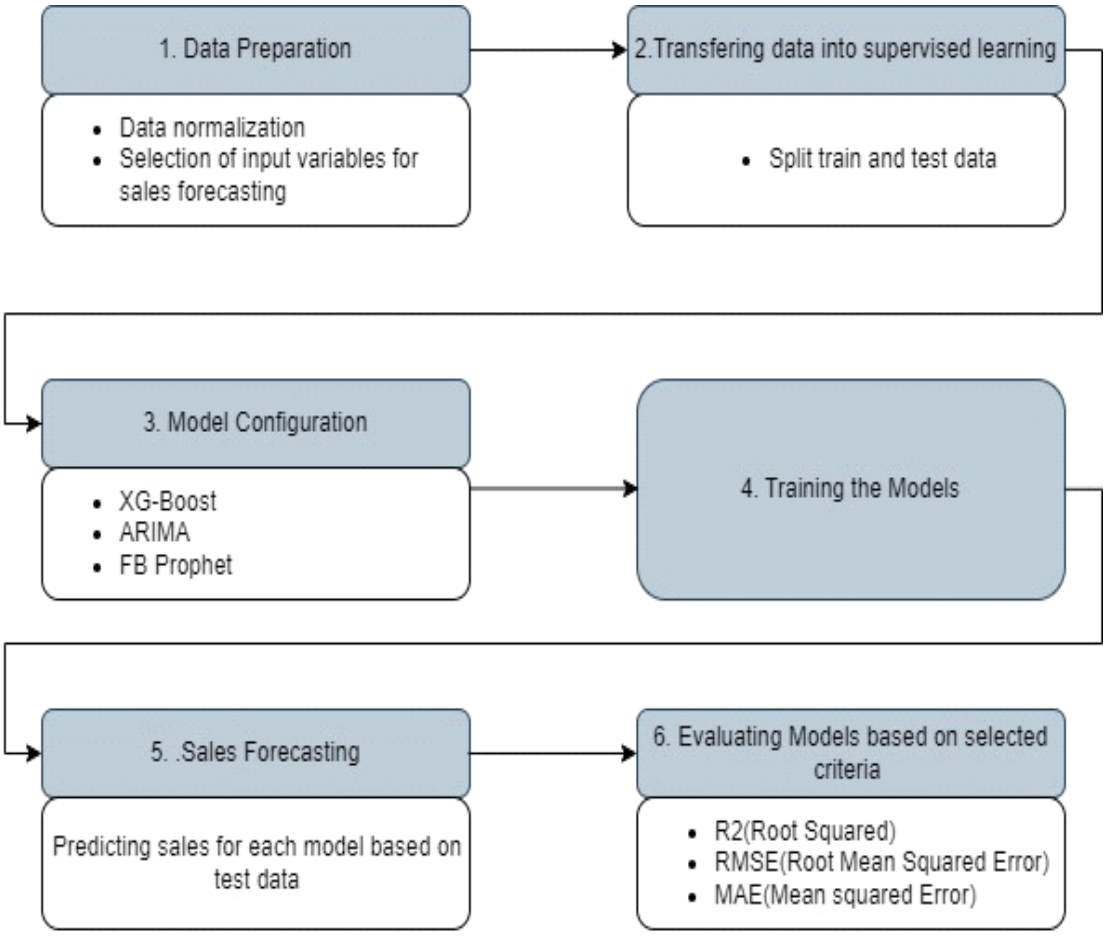

Figure 2. Step by step process of Proposed Methodology

3.1 Overview of the Dataset

Datasets were gathered from the Kaggle website's Rossman store Sales Competition page. Store.csv, Test.csv, Train.csv and Sample Submission are four separate files. In-Store, there are several features.csv

providing store type, selection, competition distance, and price Promo, Promo-Since-Week, Promo-Since-Year, Promo-Interval, Competition-Open-Since-Month, Competition-Open-Since-Year. Features like Id and sales can be found in the sample_submission.csv file. Finally, genuine sales figures and some history data are available in train.csv.

3.2 Performance evaluation measures

Three different algorithms were applied to the dataset, and their accuracy and other statistical metrics were evaluated in order to determine the most effective algorithm. ARIMA, XG-Boost and FB Prophet were the algorithms employed. These algorithms were compared based on the measures chosen to evaluate their performance. The dataset's time series features are taken into consideration while choosing the model for XG-Boost, ARIMA and PROPHET utilizing the splitting approach. The best ratios for validation, training, and test sets were produced by the splitting approach.

Before using time series data in an ML analysis, preprocessing is advised. Data preparation for the forecasting computations involves two critical steps: missing value data normalization and analysis. Additionally, determining if the data is static or not is important since static data contains a time-dependent component that might impair forecast accuracy [17]. The input variables to be utilized for the sales forecast will be chosen after conducting stationary and missing value analyses. A data normalization job will then be employed for additional investigation. A critical stage in forecasting models is selecting the independent factors that will predict a dependent parameter. The fact that there are too many parameters accessible, there is a relationship between the input parameters, or some of the parameters have low predictive ability, among other factors, can make this work challenging. The predictive capacity of the model is improved by selecting the appropriate input parameters, and since there are fewer parameters to calculate with, the required computing time is also reduced.

3.3 Treatment for Missing Values

Handling missing values in the data set is a very fundamental first step because some machine-learning algorithms just remove the rows with missing values, which results in a smaller training data set and ultimately lower predictive accuracy.

3.4 Feature Engineering

The process of selecting the characteristics that are most important for projecting the target variable is known as feature selection. By selecting the elements that are necessary for the predicting of the target variables and deleting the rest, this approach allows for the reduction of the model's dimension. The impact of a variable on prediction accuracy is taken into consideration by feature significance when selecting the top contributing characteristics in the feature space. After the training phase, the model will determine the link between the characteristics or independent variables and the objective or dependent variable. Simply said, the model learns from previous examples that are provided in training data to forecast future situations. The trained model may then be used to predict the values or labels of the test data set.

4. Results

This section serves as as a summary of the research's results and a graphic representation of its progress. First, we present the entire machine-learning pipeline that was created for this project. The outcomes of using this ML pipeline for the data from our case study are then presented. Different machine learning (ML) algorithms have been used to forecast the sales of different retail business datasets, and several evaluation techniques including RMSE (Root-Mean-Squared-Error), MAE (Mean-Absolute-Error), and R2 (Root-Square) are computed and assessed. Different ML algorithms have been used to predict the sales of different retail stores datasets and different evaluation methods like RMSE (Root-Mean-Squared-Error), MAE (Mean-Absolute-Error) and R2 (Root-Square) are calculated and evaluated. examined ML Autoregressive integrated moving averages (ARIMA), extreme gradient boost (XG Boost), and FB Prophet are the models that were employed in the experiment. This study's dataset was obtained via Kaggle.

4.1 Result Analysis

Google Collab, which is suitable for Python programming jobs, served as the simulation tool in this study. Google Collab write and run Python 3 code Without a local setup. Run terminal commands from the Notebook. Imports data from outside sources like Kaggle. Notebooks can be saved to Google Drive.

4.2 Attribute Representation

The Kaggle website provided the dataset. There are 15 variables in the dataset. Included in this list are the following: ID, the store, Sales, Users, Open, State Holiday, School vacation, Store Type, Assortment, Competitive Distance, Tournament Open From Month, Competition Open From Year's Revenue, Promo2, Promo2SinceWeek, Promo2SinceYear, and Promo Interval. The dataset's multivariate values demonstrate that several variables are simultaneously affecting sales.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Id	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance
2	1	1	5	7/31/2015	5263	555	1	1	0	1	c	a	1270
3	2	2	5	7/31/2015	6064	625	1	1	0	1	a	a	570
4	3	3	5	7/31/2015	8314	821	1	1	0	1	a	a	14130
5	4	4	5	7/31/2015	13995	1498	1	1	0	1	c	c	620
6	5	5	5	7/31/2015	4822	559	1	1	0	1	a	a	29910
7	6	6	5	7/31/2015	5651	589	1	1	0	1	a	a	310
8	7	7	5	7/31/2015	15344	1414	1	1	0	1	a	c	24000
9	8	8	5	7/31/2015	8492	833	1	1	0	1	a	a	7520
10	9	9	5	7/31/2015	8565	687	1	1	0	1	a	c	2030
11	10	10	5	7/31/2015	7185	681	1	1	0	1	a	a	3160
12	11	11	5	7/31/2015	10457	1236	1	1	0	1	a	c	960
13	12	12	5	7/31/2015	8959	962	1	1	0	1	a	c	1070
14	13	13	5	7/31/2015	8821	568	1	1	0	0	d	a	310
15	14	14	5	7/31/2015	6544	710	1	1	0	1	a	a	1300
16	15	15	5	7/31/2015	9191	766	1	1	0	1	d	c	4110
17	16	16	5	7/31/2015	10231	979	1	1	0	1	a	c	3270
18	17	17	5	7/31/2015	8430	946	1	1	0	1	a	a	50
19	18	18	5	7/31/2015	10071	936	1	1	0	1	d	c	13840
20	19	19	5	7/31/2015	8234	718	1	1	0	1	a	c	3240
21	20	20	5	7/31/2015	9593	974	1	1	0	0	d	a	2340

Figure 3. Attributes Representation

Thirteen variables from the dataset along with their first five values of the sales dataset are shown in figure 4.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear
0	1	5	2015-07-31	5263	555	1	1	0	1	c	a	1270.0	9.0	2008.0
1	1	4	2015-07-30	5020	546	1	1	0	1	c	a	1270.0	9.0	2008.0
2	1	3	2015-07-29	4782	523	1	1	0	1	c	a	1270.0	9.0	2008.0
3	1	2	2015-07-28	5011	560	1	1	0	1	c	a	1270.0	9.0	2008.0
4	1	1	2015-07-27	6102	612	1	1	0	1	c	a	1270.0	9.0	2008.0

Figure 4. The first five rows of the sales dataset

4.3 Checking Correlation between Columns

The correlation between the variables is calculated and displayed as a heat map after the confirmation that the data is balanced. The positive relationship between the variables is clearly shown by the heat map. Separate training and testing sections of the data set are provided. The algorithms are employed once the data has been prepared. The accuracy of the algorithm was learned from the results. The root mean square error (Root Mean Squared Error), Root-Square(R2), and the use of MAE (Mean Absolute Errors) were used to identify the faults.

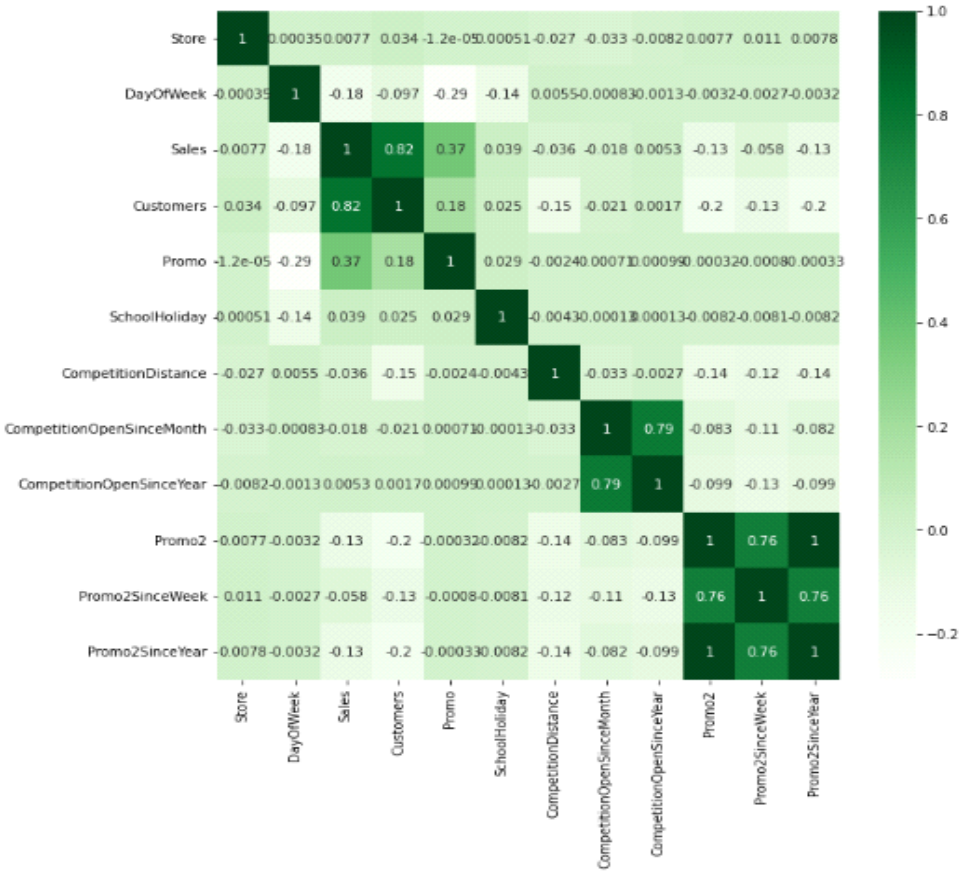

Figure 5. Heat map for correlation of attributes

4.4 Histogram of Attributes

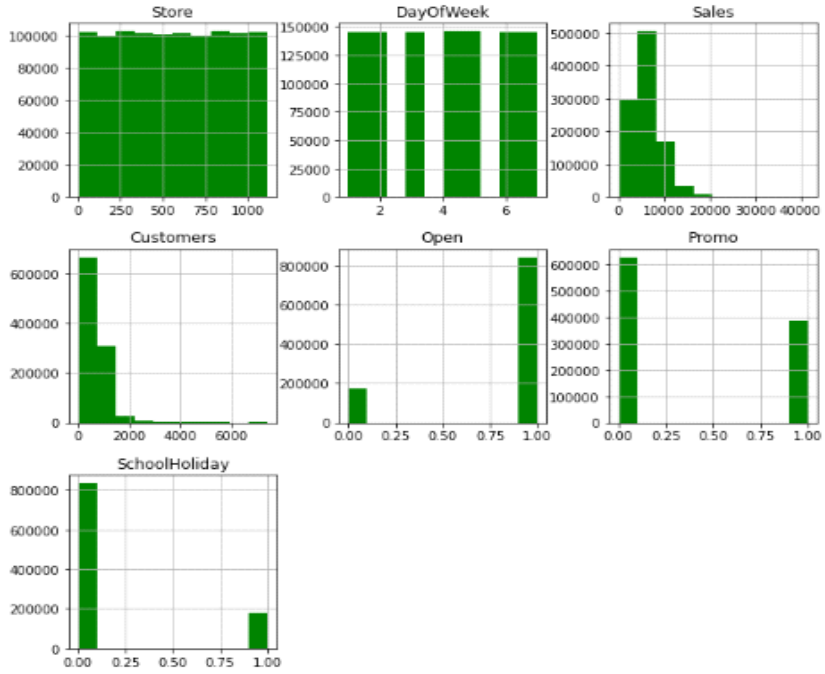

Figure 6. Histogram of Attribute

We can observe from these histograms that each feature has a unique range of distribution. The graph is represented with 10 bins for each variable from the dataset.

4.5 Exploratory Data Analysis (EDA)

Google Collab was used as a tool for exploratory data analysis (EDA) and visualization. Because the efficiency with which the information is prepared and presented determines how successful a machine learning methodology will be. Analyzing the target variable first.

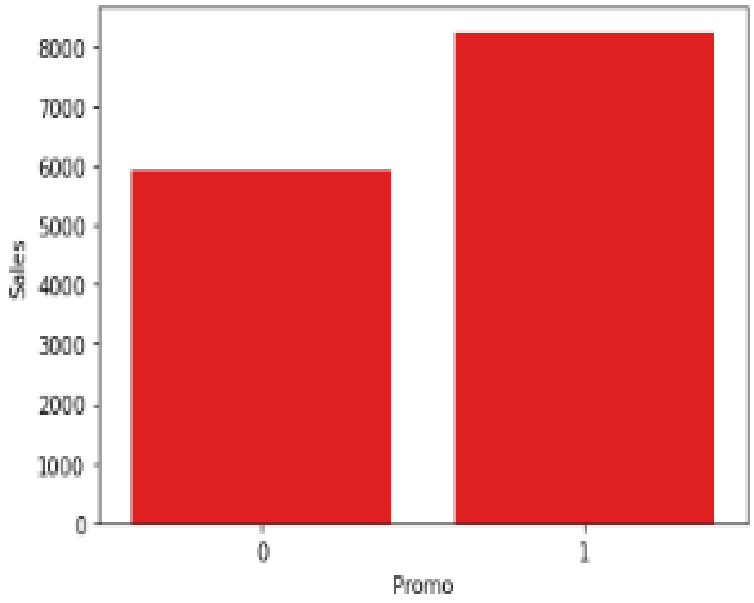

Figure 7. Analyzing the Sales variable

The chart allows one to get the conclusion that the data is fairly balanced. Various data set attributes, such as the Holiday property, which has values of 1 (Open) and 0(Closed)., may also be used with bar plots.

	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday
count	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06
mean	5.584297e+02	3.998341e+00	5.773819e+03	6.331459e+02	8.301067e-01	3.815145e-01	1.786467e-01
std	3.219087e+02	1.997391e+00	3.849926e+03	4.644117e+02	3.755392e-01	4.857586e-01	3.830564e-01
min	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.800000e+02	2.000000e+00	3.727000e+03	4.050000e+02	1.000000e+00	0.000000e+00	0.000000e+00
50%	5.580000e+02	4.000000e+00	5.744000e+03	6.090000e+02	1.000000e+00	0.000000e+00	0.000000e+00
75%	8.380000e+02	6.000000e+00	7.856000e+03	8.370000e+02	1.000000e+00	1.000000e+00	0.000000e+00
max	1.115000e+03	7.000000e+00	4.155100e+04	7.388000e+03	1.000000e+00	1.000000e+00	1.000000e+00

Figure 8. Dataset description

Over two years, stores were halted for 155 days.

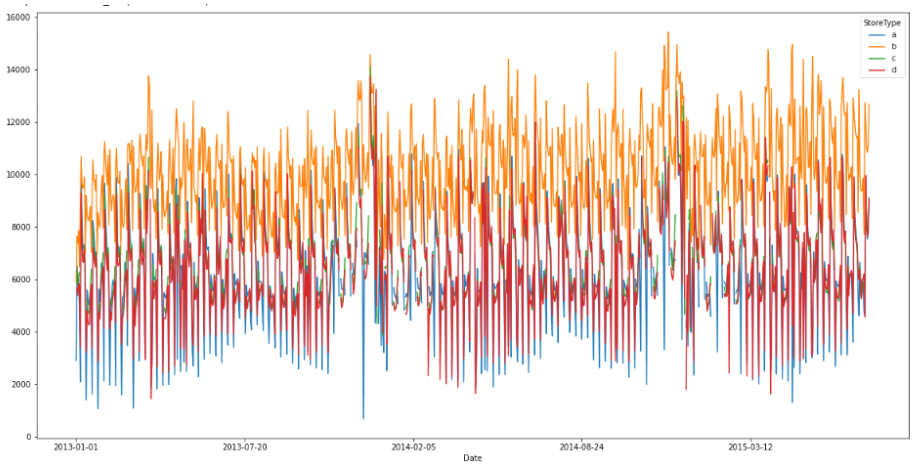

Figure 9. Subplot of four different types of stores

In Figure 9, a Subplot among four different stores data is presented each store is labeled with different color and these are grouped by according to their mean values of total sales of each store.

Table 2. Results of machine learning algorithms applied

Models	R2	RMSE	MAE
XG Boost	0.936	5.2567	4.1429
FB Prophet	0.824	8.9231	7.6714
ARIMA	0.368	1113.4	970.50

Table 2 demonstrates that PROPHET and XG-Boost fit the training and validation sets much better. But ARIMA is worse than XG-Boost PROPHET.

5. Conclusions

The goal of this thesis was to identify which methodology, out of the three under consideration, could anticipate net sales for the shop's dataset the most precisely.

Additionally, one of the goals was to see if machine learning techniques could be utilized to develop sales estimates using the given data. The data set was subjected to three different applications before being compared between them. An overview of the findings from the performance metrics applied to compare the strategies is shown in Table 2 Looking at Table 2 makes it clear that the machine learning techniques work better since they produce superior RMSE, MSE, and R2 outcomes for the testing data set.

The sales data set has been carefully examined to achieve this. In the 60-day projection, models generated values that may be deemed successful. The XG-boost, ARIMA, and PROPHET packages in Google Collab were discovered to be very significant factors in this regard. It is evident that XG-Boost provided better results from Table 2 and ARIMA did not perform well on large-scale dataset, XG-Boost outperformed them all.

5.1 Future Work

In future, we can try to incorporate more accurate data into the ongoing investigation. Machine Learning has the benefit of evaluating data and important variables so that you can use several Machine Learning approaches to construct a systematic approach. Another approach deep learning algorithms can also be studied for the challenges of sales prediction.

References

1. G. Behera, A. Bhoi, and A. K. Bhoi, "A Comparative Analysis of Weekly Sales Forecasting Using Regression Techniques," Lect. Notes Networks Syst., vol. 431, pp. 31–43, 2022, doi: 10.1007/978-981-19-0901-6_4/COVER.
2. T. Weng, W. Liu, and J. Xiao, "Supply chain sales forecasting based on lightGBM and LSTM combination model," Ind. Manag. Data Syst., vol. 120, no. 2, pp. 265–279, 2020, doi: 10.1108/IMDS-03-2019-0170.
3. G. Behera and N. Nain, "A comparative study of big mart sales prediction," Commun. Comput. Inf. Sci., vol. 1147 CCIS, no. October, pp. 421–432, 2020, doi: 10.1007/978-981-15-4015-8_37.
4. G. Tsoumakas, "A survey of machine learning techniques for food sales prediction," Artif. Intell. Rev., vol. 52, no. 1, pp. 441–447, Jun. 2019, doi: 10.1007/S10462-018-9637-Z.
5. Y. jiang Li, Y. Yang, K. Zhu, and J. Zhang, "Clothing Sale Forecasting by a Composite GRU-Prophet Model With an Attention Mechanism," IEEE Trans. Ind. Informatics, vol. 3203, no. c, pp. 1–9, 2021, doi: 10.1109/TII.2021.3057922.
6. B. Sri, S. Ramya, and K. Vedavathi, "An Advanced Sales Forecasting Using Machine Learning Algorithm," Int. J. Innov. Sci. Res. Technol., vol. 5, no. 5, pp. 342–345, 2020, [Online]. Available: https://www.ijisrt.com/assets/upload/files/IJISRT20MAY134.pdf
7. A. Lasek, N. Cercone, and J. Saunders, "Sales and customer demand forecasting: Literature survey and categorization of methods," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 166, pp. 479–491, 2016, doi: 10.1007/978-3-319-33681-7_40.
8. J. T. Mentzer and M. A. Moon, "Sales forecasting management : a demand management approach," p. 347, 2005.
9. A. Tealab, "Time series forecasting using artificial neural networks methodologies: A systematic review," Futur. Comput. Informatics J., vol. 3, no. 2, pp. 334–340, Dec. 2018, doi: 10.1016/J.FCIJ.2018.10.003.
10. S. Kohli, G. T. Godwin, and S. Urolagin, "Sales Prediction Using Linear and KNN Regression," pp. 321–329, 2021, doi: 10.1007/978-981-15-5243-4_29.
11. J. Feizabadi, "Machine learning demand forecasting and supply chain performance," Int. J. Logist. Res. Appl., vol. 25, no. 2, pp. 119–142, 2022, doi: 10.1080/13675567.2020.1803246.
12. S. Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, and Y. Kazancoglu, "The derived demand for advertising expenses and implications on sustainability: a comparative study using deep learning and traditional machine learning methods," Ann. Oper. Res., 2022, doi: 10.1007/s10479-021-04429-x.
13. J. Chen, W. Koju, S. Xu, and Z. Liu, "Sales Forecasting Using Deep Neural Network and SHAP techniques," 2021 IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. ICBAIE 2021, no. Icbaie, pp. 135–138, 2021, doi: 10.1109/ICBAIE52039.2021.9389930.
14. A. Massaro, A. Panarese, D. Giannone, and A. Galiano, "Augmented data and Xgboost improvement for sales forecasting in the large-scale retail sector," Appl. Sci., vol. 11, no. 17, 2021, doi: 10.3390/app11177793.
15. Y. Kaneko and K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales," IEEE Int. Conf. Data Min. Work. ICDMW, vol. 0, pp. 531–537, 2016, doi: 10.1109/ICDMW.2016.0082.
16. Z. Xia, S. Xue, L. Wu, J. Sun, Y. Chen, and R. Zhang, "ForeXGBoost: passenger car sales prediction based on XGBoost," Distrib. Parallel Databases, vol. 38, no. 3, pp. 713–738, 2020, doi: 10.1007/s10619-020-07294-y.
17. Y. Ensafi, S. H. Amin, G. Zhang, and B. Shah, "Time-series forecasting of seasonal items sales using machine learning – A comparative analysis," Int. J. Inf. Manag. Data Insights, vol. 2, no. 1, p. 100058, Apr. 2022, doi: 10.1016/J.JJIMEI.2022.100058.
18. M. Seyedan, F. Mafakheri, and C. Wang, "Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach," Decis. Anal. J., vol. 3, p. 100033, Jun. 2022, doi: 10.1016/J.DAJOUR.2022.100033.
19. S. S. J. Nithin, T. Rajasekar, S. Jayanthy, K. Karthik, and R. R. Rithick, "Retail Demand Forecasting using CNN-LSTM Model," Proc. Int. Conf. Electron. Renew. Syst. ICEARS 2022, pp. 1751–1756, 2022, doi: 10.1109/ICEARS53579.2022.9752283.
20. H. Wei and Q. Zeng, "Research on sales Forecast based on XGBoost-LSTM algorithm Model," J. Phys. Conf. Ser., vol. 1754, no. 1, 2021, doi: 10.1088/1742-6596/1754/1/012191.
21. B. Kumar Jha and S. Pande, "Time Series Forecasting Model for Supermarket Sales using FB-Prophet," Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021, no. Iccmc, pp. 547–554, 2021, doi: 10.1109/ICCMC51019.2021.9418033.
22. C. CATAL, K. ECE, B. Arslan, and A. Akbulut, "Benchmarking of Regression Algorithms and Time Series Analysis Techniques for Sales Forecasting," Balk. J. Electr. Comput. Eng., vol. 7, no. 1, pp. 20–26, 2019, doi: 10.17694/bajece.494920.
23. P. F. Pai and C. H. Liu, "Predicting vehicle sales by sentiment analysis of twitter data and stock market values," IEEE Access, vol. 6, no. c, pp. 57655–57662, 2018, doi: 10.1109/ACCESS.2018.2873730.
24. J. Wang, G. Q. Liu, and L. Liu, "A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry," 2019 4th IEEE Int. Conf. Big Data Anal. ICBDA 2019, pp. 317–320, 2019, doi: 10.1109/ICBDA.2019.8713196.
25. P. Ranjitha and M. Spandana, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms," Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021, no. Iciccs, pp. 1416–1421, 2021, doi: 10.1109/ICICCS51141.2021.9432109.
26. Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," Proc. - 2020 Int. Conf. Big Data Artif. Intell. Softw. Eng. ICBASE 2020, pp. 458–461, 2020, doi: 10.1109/ICBASE51474.2020.00103.

27. Z. Qiao, "Walmart Sale Forecasting Model Based on LightGBM," Proc. - 2020 2nd Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2020, pp. 76–79, 2020, doi: 10.1109/MLBDBI51377.2020.00020.
28. R. Gustriansyah, "Integration of Decision-Making Method and Data-Mining Method as A Preliminary Study of Novel Sales Forecasting Method," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 5730–5735, 2020, doi: 10.30534/ijatcse/2020/227942020.
29. Y. Weng, X. Wang, J. Hua, H. Wang, M. Kang, and F. Y. Wang, "Forecasting Horticultural Products Price Using ARIMA Model and Neural Network Based on a Large-Scale Data Set Collected by Web Crawler," IEEE Trans. Comput. Soc. Syst., vol. 6, no. 3, pp. 547–553, 2019, doi: 10.1109/TCSS.2019.2914499.
30. B. Ratner, Statistical and machine-learning data mining. Accessed: Oct. 07, 2022. [Online]. Available: https://www.routledge.com/Statistical-and-Machine-Learning-Data-Mining-Techniques-for-Better-Predictive/Ratner/p/book/9780367573607