An Application of Machine Learning in Retail for Demand Forecasting

Muhammad Umer Farooq[1], Mustafa Latif[2]*, Waseem[1], Mirza Adnan Baig[3] , Muhammad Ali Akhtar[4] and Nuzhat Sana[1]

[1]Department of Computer Science and Information Technology, NED University of Engineering and Technology, Karachi, 75270, Pakistan.
[2]Department of Software Engineering, NED University of Engineering and Technology, Karachi, 75270, Pakistan.
[3]Department of Computer Science, IQRA University Karachi Pakistan.
[4]Department Of Computer And Information Systems Engineering, NED University of Engineering and Technology, Karachi, 75270, Pakistan.
*Corresponding Author: Email: mustafalatif@neduet.edu.pk

Abstract
Demand prediction is an essential component of any business or supply chain. Large retailers need to keep track of tens of millions of items flows each day to ensure smooth operations and strong margins. The demand prediction is in the epicenter of this planning tornado. For business processes in retail companies that deal with a variety of products with short shelf life and foodstuffs, forecast accuracy is of the utmost importance due to the shifting demand pattern, which is impacted by an environment of dynamic and fast response. All sectors strive to produce the ideal quantity of goods at the ideal time, but for retailers, this issue is especially crucial as they also need to effectively manage perishable inventories. In light of this, this research aims to show how Machine Learning approaches can help with demand forecasting in retail and future sales predictions. This will be done in two steps. One by using historic data and another by using open data of weather conditions, fuel, Consumer Price Index (CPI), holidays, any specific events in that area etc. Several machine learning algorithms were applied and compared using the r-squared and mean absolute percentage error (MAPE) assessment metrics. The suggested method improves the effectiveness and quality of feature selection while using a small number of well-chosen features to increase demand prediction accuracy. The model is tested with a one-year weekly dataset after being trained with a two-year weekly dataset. The results show that the suggested expanded feature selection approach provides a very good MAPE range, a very respectable and encouraging value for anticipating retail demand in retail systems.

Keywords:
Demand forecasting, time series forecasting, retail Industry, machine learning.

1. Introduction

A lot of things depend on accurate forecasts, which are considered the backbone of a successful retail organization. Retail demand forecasting is used to comprehend the purchasing habits of retail customers. Similar to predictive analysis, it aids in improved supply chain management, inventory control, and budgeting. The path to better decision-making is through demand forecasting.

The complexity of customer expectations and the rapid pace of change in the world makes it hard for retailers to predict demand using traditional methods. Machine learning is now playing a key role in demand prediction, assisting retailers in getting ready for future customer demand and market conditions. In the retail Industry, accurate demand forecasts enable the effective arrangement of resources, including personnel, raw materials, stock, human labor, etc.

Basically, forecasting is the process of analyzing historical or current data to make predictions about the future. In the majority of firms, the role of forecast accuracy is crucial since it gives a clear image of the plan, which helps to increase revenue.

The issue with demand prediction systems is that the majority of retail establishments still rely on traditional techniques to forecast demand, which causes significant financial losses because today's demand is reliant on a variety of factors, including seasons, events, inflation, weather, and more. Therefore, effective methods are needed for retail organizations to forecast demand in order to grow revenue.

Machine learning was used to address this problem and estimate demand more accurately or with fewer mistakes. Given how well its algorithm mimics, demand patterns can be thoroughly evaluated utilizing a variety of characteristics. The learning algorithm will automatically learn if a lot of data is provided, and it will automatically learn patterns from the data in less time with more accuracy.

The objective of this research is to evaluate different models that will enable the improvement of demand forecasting through the use of machine learning skills. This

Manuscript received August 5, 2023
Manuscript revised August 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.8.25

research mainly focuses on applying machine learning algorithms to estimate retail demand effectively and reliably. Due to this, considerable information was obtained from the provided data set and by utilizing this information, machine learning algorithms accurately estimated weekly sales.

The above objectives raise the following goals:

1. To conduct a thorough assessment of the literature that includes studies from the past ten years on the use of intelligent systems for demand forecasting.
2. To recognize fresh trends in demand forecasting. That includes two steps, the first one is to use open data and the other is to use historical data. Information about local events, holidays and weather conditions will also be included.
3. To employ a variety of techniques to forecast demand. Create, construct, and test a prototype intelligent system with the goal of enhancing retail demand forecasting.

This paper is comprised of the following sections. Section 2 lists the earlier pertinent works. The dataset is described in Section 3, along with data analysis and visualization. The evaluation of different models is presented in Section 4. The experimental setup necessary for the proposed framework is presented in Section 5, and the results are mentioned in Section 6. In Section 7, a conclusion is eventually reached.

2. LITERATURE REVIEW

Traditional forecasting methods have been used to anticipate demand since 2010; however, several machine learning algorithms are currently being utilized to forecast demand in various industries. Support vector regression (SVR), a method for time series forecasting that is frequently used with genetic algorithms (GA), was first introduced by M. Sarhani and A. El Afia in 2014. To optimize SVR, they used Particle Swarm Optimization (PSO), and they demonstrated that SVR-PSO performs better than SVR-GA [1].

It was suggested that linear regression could be combined with underlying models. It was found that, compared to linear models, machine learning models typically produce better out-of-sample fits without sacrificing the in-sample quality of fit. Any model in combination with LASSO, Forward stage-wise, Random forest, Stepwise, SVM, Linear, Logit, Bagging, and Combined—produces a better out-of-sample fit when all the models are combined linearly with nonnegative weights [2].

The ARIMA model was created by Fattah et al., they showed that future demand can be predicted with the help of historical demand data. They modeled demand forecasting for the final product in food production using the Box-Jenkins time series technique. They chose the ARIMA model, which minimizes the four preceding requirements (1,0,1). The outcomes demonstrate that this model can be applied to projecting and estimating future demand in the food production industry [3].

In 2019, various models, including LSTM, ARIMA, RNN, and SVM, performed in the five areas of predictive performance, generalization capability, runtime, expense, and accessibility. The research of Wang et al. demonstrates that SVM and LSTM generally do quite well in prediction tasks. With regard to handling perishable goods, LSTM, SVM, and RNN have high accuracy. However, ARIMA excels in terms of runtime and ease of use. Due to its best accuracy, SVM is more suited for perishable goods. Due to its reduced cost and superior prediction ability, LSTM is the most favored method for non-perishable goods [4].

Xue et al. created the demand forecasting method for retail supplier emergency management. It is built on the SVM algorithm and NRS algorithm, with genetic algorithms used to optimize the parameters (GA). The more accurate dynamic forecasting and increased operating rate of this model increase the competitiveness of the retail supply chain [5].

It was claimed that combining forecasts greatly increases accuracy by developing exponential smoothing models with selection criteria like Mean Absolute Deviation (MAD). It was recommended that the simple average of the forecasts produced by the three different models (4-week moving average, exponential smoothing, and ARIMA) might provide a workable solution to this forecasting challenge. To identify the optimum model, demand forecasting will see a lot of study in 2020 [6].

To make forecasts and achieve maximum accuracy, a number of algorithms, including MLR, SVM, and LSTM, were utilized. LSTM gives the best results. For predicting time series data, LSTM models are advised since they can monitor and take into consideration temporal gaps for a sequential series. This quality gives LSTM an advantage over competing strategies. It has been found that LSTM is quicker than SVM and regression methods [7].

Warnakulasooriya et al. compared LSTIM to SARIMA and found that SARIMA has the drawback of only being able to derive linear correlations from time series data. Because of their high nonlinear mapping capabilities and tolerance for complex forecasting data, artificial neural networks are potentially beneficial to time series forecasting techniques. Due to its lower MSE (mean squared error) value, the LSTM model was utilized to forecast demand and price [8]. In the same year, K. Chen conducted additional research on LSTIM. He proposed using a genetic algorithm to enhance the network's time step, hidden layer count, and training durations in order to increase the model's predictive accuracy. Additionally, his findings demonstrated that the AGA-LSTM model's prediction accuracy was significantly higher than that of the conventional LSTM model. [9].

Lakshmanan et al. also suggested long short-term memory (LSTM) networks, which use prior sales data of the products as input and estimate demand for each product for the upcoming three time periods. The suggested model outperformed all other machine learning models in terms of accuracy, scoring 96.77%. (like MLP regressor, linear regression, and multilayer backpropagation neural network) [10].

Sushil et al. published their research in which LSTM is combined with random forest (RF). Their suggested method outperforms existing forecasting techniques in terms of accuracy. The suggested approach, however, outperforms even a random forest in terms of performance. The hybrid technique combined the advantages of both random forest and LSTM, allowing it to take advantage of their complementing capabilities [11].

Huber et al. also utilized LSTM to forecast sales for several types of special days, comparing it to gradient-boosted regression trees, feed-forward ANNs (FNNs), and RNNs (GBRTs). The model parameters were optimized using a gradient-based methodology in all of these methods. Recurrent ANNs (LSTM) performed better than their feed-forward (MLP) counterparts. Therefore, it would seem advantageous to process the time series data sequentially [12].

In another study, Javad Feizabadi created hybrid demand forecasting techniques based on machine learning, such as ARIMAX and Neural Networks. The devised technique takes time series and explanatory factors into account. One financial and two non-financial performance criteria were examined by him. These research findings indicate the potential for improving operational and financial indicators by employing the ML-based forecasting approach (ARIMAX and NN) [13].

Using a supervised learning approach for one-dimensional (scalar) time series and recurrent neural networks, M. A. Khan examined the efficacy of time series and rule-based forecasting (RNN). For this time series prediction, the Deep AR algorithm from Amazon Sage Maker was employed. Forecasted calculations revealed that the DeepAR models' performance was accurate and comparable. Because the percentage error numbers were relatively less, DeepAR models offered a high percentage of forecasting accuracy. With a larger data set, the model's output got more accurate. [14].

In comparison to well-known statistical techniques like Croston's method and its modifications, machine learning (ML) techniques can generate forecasts that are substantially less biassed and more accurate. The research compared the forecasting abilities of various ML techniques taught both cross- and series-wise. Additionally, it is demonstrated that cross-learning can enhance NN predicting ability when taught series-by-series, beating conventional statistical standards. This is true when training the networks using time series characteristics in addition to historical data [15].

3. DATASET AND VISUALIZATION

The most crucial component of any application or model is its data. The model gathered information from different data sources, including stores, departments, inventories, weather, economical factors, and the calendar. These were the main sources of data for forecasting. A well-known retail store dataset was used for this research. On the dataset, weekly demand was forecast, the dataset contained three years of weekly data from 2010 to 2012.

The dataset is composed of different sources, as shown in Figure 1.

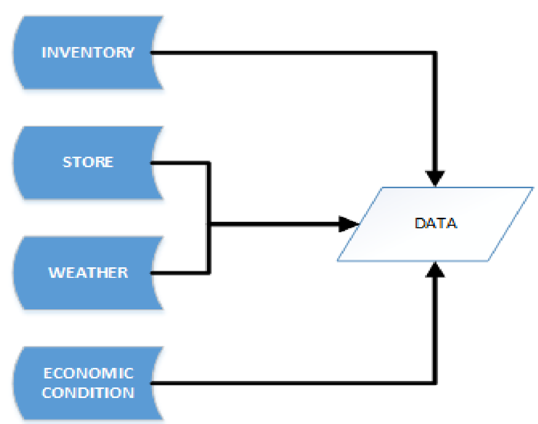

Fig. 1: Detailed overview of data source.

To see how features are related with one another, data visualization was done on the resulting dataset. Data was analyzed by using different plots against features, as shown in Figure 2. It was concluded from these plots that seasonality in the data was present, as evidenced by the clear peaks near the end of the year due to Christmas and

New Year's activities, which can be seen when comparing Date against the column Weekly Sales.

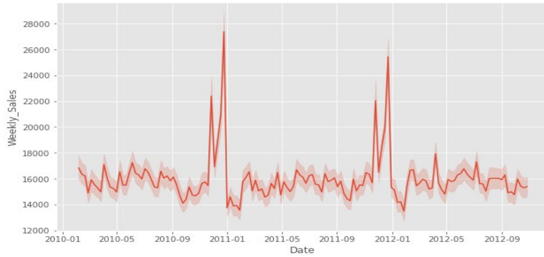

Fig. 2: Showing Seasonality in Dataset

The weekly sales column in the plot (Figure 3) also demonstrated a linear link between store sizes and weekly sales each week which showed that store size is directly proportional to weekly sales.

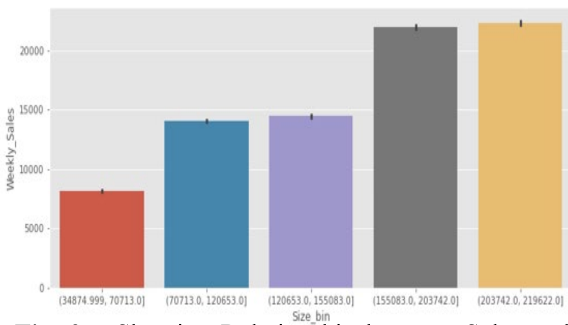

Fig. 3: Showing Relationship between Sales and Stores Size

4. METHODOLOGY

This section represents how demand prediction was carried out in this research. The main purpose of the proposed framework was to predict demand using different feature sets and different models. Figure 4 describes the proposed framework and the specific interactions between each component, and how each component was implemented.

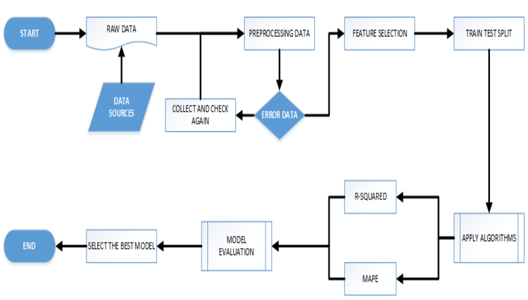

Fig. 4: Detailed Overview of Proposed Framework

Components of the proposed framework

The proposed framework has the following components.

- **Data Preprocessing**

In order to make the dataset acceptable for machine learning models, which are data sensitive, data preprocessing was conducted on the dataset. During the data analysis phase, data discrepancies were discovered in the dataset, such as negative sales and missing values. To avoid these discrepancies, data cleaning and transformation activities were carried out. The datasets for stores, inventory, weather, and economic factors were later combined through data integration, leading to the creation of a merged feature space. Since the resulting dataset contained values that fall within different ranges, data transformation and standardization were carried out on it. With the help of the standard scaler, continuous data's big values were transformed into regularly distributed values as part of the data pretreatment process.

The standard score of a sample x was calculated using Eq. 1

- $z = (x - u) / s$ (1)
- where
- s is the sample standard deviation
- u is the sample mean

- **Feature Selection**

Two different feature spaces were created, one with basic data sources and the other with external data sources. This was done to analyze which feature contributed more to demand prediction and how external data sources impacted demand prediction in retail.

The extended feature set was created through a fundamental analysis of the attributes in which demand varies in retail systems, such as stores, their sizes, their relationship to demand and some external factors too. The weather, fuel costs, the CPI, unemployment, and holidays are examples of external sources. As per the literature review we have not found any dataset which can cover these external attributes for demand prediction. They all used internal attributes to predict demand.

Weather parameters, seasonal (or calendar), and economic variables (fuel price, unemployment, CPI) make up the entire extended feature set for the retail demand forecasting model in this feature space (FS), and the basic feature space consists of the attributes of inventory and stores. The basic and extended feature space extracted from the provided dataset for the FS work in this research is shown in Tables 1 and 2.

Table. 1: Basic Feature Space of the Problem

Index	Feature	Data Category
1	Store	Numeric
2	Department	Numeric
3	Date	Date
4	Weekly Sales	Numeric
5	Holiday	Boolean
6	Store Size	Numeric

Table. 2: Extended Feature Space of the Problem

Index	Feature	Data Category
1	Store	Numeric
2	Department	Numeric
3	Date	Date
4	Weekly Sales	Numeric
5	Holiday	Boolean
6	Store Size	Numeric
7	Temperature	Numeric
8	Fuel Price	Numeric
9	CPI	Numeric
10	Unemployment	Numeric
11	Day	Numeric
12	Month	Numeric
13	Year	Numeric

- **Train Test Split**

The train-test split was an approach for assessing how well a machine learning system performed. The dataset was split between two subsets, the training dataset, and the testing dataset. The Machine Learning Models were trained by training dataset, and their performance was evaluated by test dataset. Seventy-five percent of the data was used for training and twenty-five percent for testing.

- **Machine Learning**

After creating appropriate feature sets, modeling machine learning methods was the next step. To hunt for patterns, insights, and forecasts use of proper machine learning techniques or algorithms was required. Weekly demand, which is a continuous value, was predicted. Regression-based and time series Machine Learning algorithms were used for modeling. In order to investigate the effects of other data sources, such as store sizes, weather aspects, and economic factors on weekly demand, these attributes were incorporated in the extended feature set. To observe the behavior of internal and external attributes, a number of different models were tested out to evaluate their accuracy.

Time Series and Regression based, both Machine Learning Algorithms, i.e., Autoregressive Integrated Moving Average, or ARIMA, Multilinear Regression, Random Forest, Gradient Boosting, Decision tree regression, and Facebook Prophet were trained for weekly demand prediction using two feature sets in this research. After training, a detailed comparison was made of their performance.

- **Evaluation Metrics**

The final step in this research is the evaluation of a classifier, which enables to deliver reliable information about the classifiers. It provides information about how well the classifiers performed. The models in the envisioned system were evaluated for performance using the R-Squared and mean absolute percentage error (MAPE).

5. EXPERIMENTAL SETUP

For the suggested model, data visualization, data preprocessing, model training, and evaluation were performed using Jupyter Notebook using Anaconda Navigator (Anaconda3) and Python version 3. Additionally, because several models were trained and evaluated required GPU processing, Google Colab was used. Google Colab is a free tool of Google Research to run Python code for tasks involving machine learning and data analytics.

6. RESULTS

The models were trained on two feature sets, basic and extended. In the basic dataset, attributes were inventory and stores; in the extended data set, attributes inventory, stores, weather, and economic conditions were the attributes. Results are shown in Figures 5 and 6 and in Table 3.

Figure 5 illustrates the R-Squared and MAPE evaluation metrics for all Machine Learning Models employed in this study. Based on the findings presented in Figure 5, it can be inferred that the basic Feature Set yields the lowest Mean Absolute Percentage Error (MAPE) when utilizing the Facebook Prophet model.

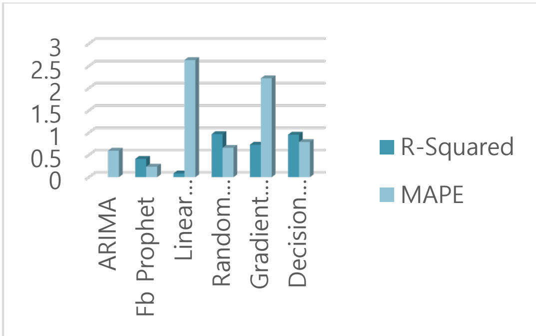

Fig. 5: Evaluation of Models on Basic Feature Set

Table. 3: Evaluation of Models

Models	R-Squared	MAPE
On Basic Feature		
ARIMA		0.594
Fb Prophet	0.411	0.236
Linear Regression	0.084	2.632
Random Forest	0.968	0.66
Gradient Boosting	0.729	2.221
Decision Tree	0.955	0.790
On Extended Feature		
Fb Prophet	0.941	0.245
Linear Regression	0.088	2.867
Random Forest	0.966	0.754
Gradient Boosting	0.734	2.071
Decision Tree	0.944	1.013

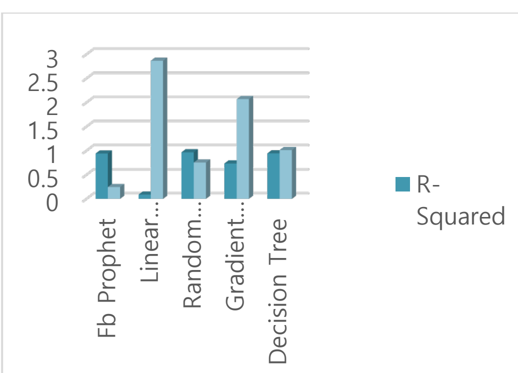

Fig. 6: Evaluation of Models on Extended Feature Set

Figure 6 illustrates the R-Squared and MAPE evaluation metrics for all Machine Learning Models employed in this study. The analysis presented in Figure 6 indicates that the extended Feature Set yields the lowest Mean Absolute Percentage Error (MAPE) value when utilizing the Facebook Prophet model.

• 7. CONCLUSION

In this research, two different feature sets were selected for model training to predict demand in the retail Industry. For this purpose, different time series and regression-based models were used. The Results concluded that Facebook Prophet is better for predicting demand as it has both seasonality and trend implemented under it, and our data have clear seasonality; that's why we get minimum MAPE on Facebook Prophet, the next algorithm which performs well in the dataset is random forest because it has second lowest MAPE value.

REFERENCES

[1] Sarhani, Malek, and Abdellatif El Afia. "Intelligent system based support vector regression for supply chain demand forecasting." 2014 second world conference on complex systems (WCCS). IEEE, 2014.

[2] Bajari, Patrick, et al. "Machine learning methods for demand estimation." American Economic Review 105.5 (2015): 481-85.

[3] Fattah, Jamal, et al. "Forecasting of demand using ARIMA model." International Journal of Engineering Business Management 10 (2018): 1847979018808673.

[4] Wang, Jiaxing, G. Q. Liu, and Lu Liu. "A selection of advanced technologies for demand forecasting in the retail industry." 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA). IEEE, 2019.

[5] Xue, Hong, et al. "Research on demand forecasting of retail supply chain emergency logistics based on NRS-GA-SVM." 2018 Chinese Control And Decision Conference (CCDC). IEEE, 2018.

[6] Silva, Juliana C., Manuel C. Figueiredo, and Ana C. Braga. "Demand forecasting: A case study in the food industry." International Conference on Computational Science and Its Applications. Springer, Cham, 2019.

[7] Palkar, Anish, et al. "Demand Forecasting in Retail Industry for Liquor Consumption using LSTM." 2020

International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.

[8] Warnakulaooriya, Hashini, et al. "Supermarket Retail–Based Demand and Price Prediction of Vegetables." 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2020.

[9] Chen, Keqiao. "An Online Retail Prediction Model Based on AGA-LSTM Neural Network." 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2020.

[10] Lakshmanan, Balakrishnan, Palaniappan Senthil Nayagam Vivek Raja, and Viswanathan Kalathiappan. "Sales Demand Forecasting Using LSTM Network." Artificial Intelligence and Evolutionary Computations in Engineering Systems. Springer, Singapore, 2020. 125-132.

[11] Punia, Sushil, et al. "Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail." International journal of production research 58.16 (2020): 4964-4979.

[12] Huber, Jakob, and Heiner Stuckenschmidt. "Daily retail demand forecasting using machine learning with emphasis on calendric special days." International Journal of Forecasting 36.4 (2020): 1420-1438.

[13] Feizabadi, Javad. "Machine learning demand forecasting and supply chain performance." International Journal of Logistics Research and Applications 25.2 (2022): 119-142.

[14] Khan, Muhammad Adnan, et al. "Effective demand forecasting model using business intelligence empowered with machine learning." IEEE Access 8 (2020): 116013-116023.

[15] Spiliotis, Evangelos, et al. "Comparison of statistical and machine learning methods for daily SKU demand forecasting." Operational Research (2020): 1-25.