

Predicting Sales in BigMart using Machine Learning Techniques

Aryan Gupta

Department Of Computer Science And
Engineering

Apex Institute Of Technology Chandigarh
University

Mohali -140413, Punjab

20BCS6656@cuchd.in

Ankith Raj

Department Of Computer Science And
Engineering

Apex Institute Of Technology Chandigarh
University

Mohali -140413, Punjab

20BCS6684@cuchd.in

Dr Alankrita Aggarwal

Professor, AIT-CSE

Chandigarh University

Mohali, Punjab

India

alankrita.e14496@cumail.in

Aryan Khushwaha

Department Of Computer Science And
Engineering

Apex Institute Of Technology Chandigarh
University

Mohali -140413, Punjab

20BCS6691@cuchd.in

Sheikh Shahnawaz Hussain

Department Of Computer Science And
Engineering

Apex Institute Of Technology Chandigarh
University

Mohali -140413, Punjab

20BCS6628@cuchd.in

Introduction- Sales forecasting involves analyzing a company's historical sales data to make predictions about its future sales performance. It is an essential aspect of financial planning that helps companies plan for short- and long-term growth. Like all forecasting processes, there is a level of risk and uncertainty involved, so it is important for sales forecasting teams to acknowledge this uncertainty in their forecasts. Sales forecasting is a widely practiced corporate strategy that involves setting objectives, creating action plans, and allocating budgets and resources to achieve those objectives. Several techniques are commonly used for sales forecasting, including Linear Regression, Decision Tree, Random Forest, and XG BOOST.

Abstract— The article contains a prediction of the value of precise sales forecasting and the drawbacks of using conventional techniques is done. and a new strategy is proposed by using machine learning methods to provide more precise sales estimates. The algorithm used is decision trees and random forests specifically to demonstrate that the suggested approach is more accurate than conventional approaches and more resistant to changes in market conditions. The Proposed method will be beneficial to the companies that depend on sales estimates to make wise decisions, the results of this study have significant ramifications. The researchers, practitioners, and students who are interested in sales forecasting will find this paper to be useful by studying the implementation of the paper.

Keywords— sales forecasting, Label encoder, Linear Regression, K-Neighbors Regressor, XGBoost Regressor, Random Forest Regressor Introduction

A. Problem Definition

Sales forecasting involves analyzing a company's historical sales data to make predictions about its future sales performance. It is an essential aspect of financial planning that helps companies plan for short- and long-term growth. Like all forecasting processes, there is a level of risk and uncertainty involved, so it is important for sales forecasting teams to acknowledge this uncertainty in their forecasts. Sales forecasting is a widely practiced corporate strategy that involves setting objectives, creating action plans, and allocating budgets and resources to achieve those objectives. Several techniques are commonly used for sales forecasting, including Linear Regression, Decision Tree, Random Forest, and XG BOOST. Using Sales Forecasting the management of the enterprise can take decision regarding operations planning, scheduling, production programming inventories of various types, physical distribution and operating profits on the basis of sales forecasts. It also contains some additional benefits like deciding investment proporsals like modernization, expansion of existing units etc. Sales forecasts are essential to make proper arrangement for training the man-power in its own unit or sending them to other industries in the country or abroad to meet the future needs of expertise. Targeting the market audience is the main business objective of industries. Therefore, it is very important for businesses to be able to achieve this using the

Forecasting System. The forecasting process involves the analysis of data from various sources, such as market trends, consumer behaviour and other factors. The analysis will also help 4,444 businesses to manage their financial resources effectively. The forecasting process can be used for several purposes, including: forecasting future demand for a product or service, forecasting the number of products that will be sold in a given time period. This is where machine learning can be leveraged in beautiful way. Machine learning is where machines gain the ability to outperform humans in specific tasks. They are used to perform certain specialized tasks in a logical manner, and perform better for the advancement of today's society. The foundation of machine learning is the mathematical art, and with the help of, various paradigms can be formulated to approximate the optimal output of. In the case of Sales Forecast, machine learning also turned out to be a godsend. This work proposes a machine learning algorithm for data collected from past sales in a grocery store

B. DATA VISUALIZATION

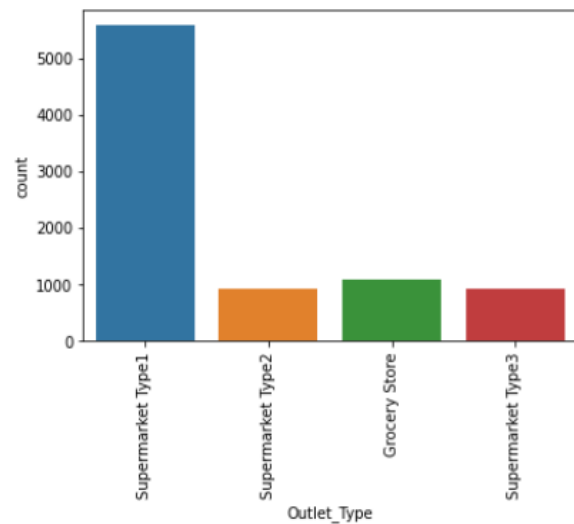
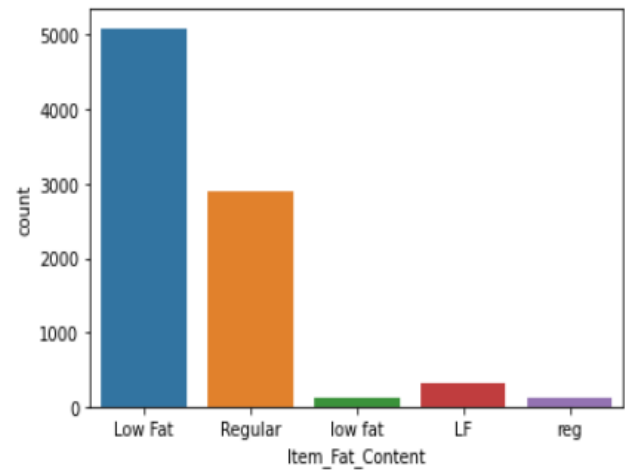
Heat map for determining correlation between the dataset attributes:

Over the years, a large amount of data has been generated A heat map, part of a data visualization library called Seaborn, is the color-coded matrix used here to depict the correlation between the target variable and the remaining attributes. The higher the color intensity of a feature relative to the target variable, the less the target variable depends on the corresponding feature. Target variable, Item_Outlet_Sales is least dependent on Item_Visibility and most dependent on Item_MRP.

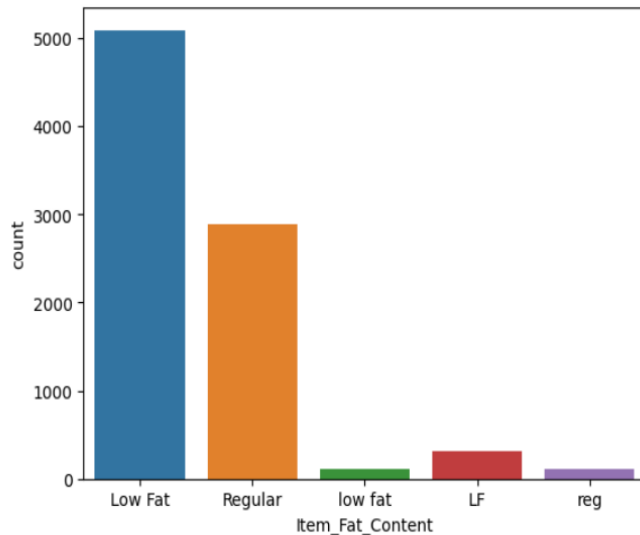


Count Plots

The distribution of various Outlet Types i.e., Supermarket Type1, Supermarket Type2, Grocery Store, Supermarket Type3 is plotted. It is observed that maximum outlets are of Supermarket Type1.



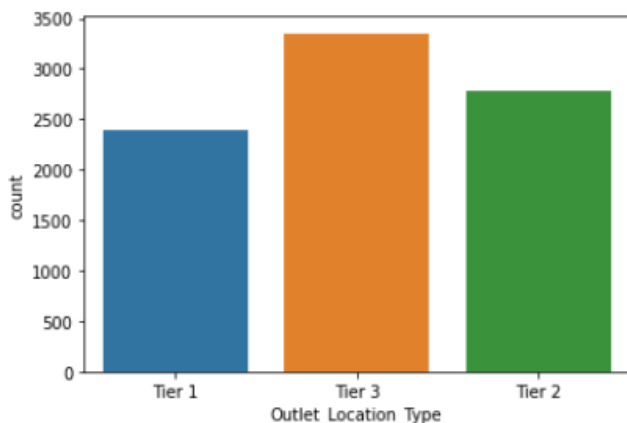
The count plot for Item Fat Content is plotted which consists of two categories Low Fat and Regular Fat written in different manners. It is observed that most of the items have low fat content.



The distribution of each item type is depicted by the following plot. Most of the items are Fruits and Vegetables followed by Snack Foods. In contrast to this, Seafood is least in count.:



As per the dataset used, there are three categories of Outlet Location Type i.e., Tier1, Tier2 and Tier3. Maximum number of outlets belongs to Tier3 type location.



C. METHODOLOGY USED



Following Steps have been used in Sales Prediction:

1) Data Pre-Processing: In machine learning algorithm, data can't be used in its normal form as it is the as the way it is obtained, so the data needs to be devised before employing it in machine learning models. This technique is used to solve problems that are not yet known by the knowledge extractor. For example, if you have a problem with your car and want to know how much oil it takes to drive, you can ask the car's driver to tell you the exact amount of oil needed. This is called preprocessing work. The goal of preprocessing is to find out what kind of information the car needs before making any decisions about whether to use it or not. Proper formatted and cleaned data is essential for preprocessing

2) Data Analysis: The process of cleansing, converting, and modelling data in order to find relevant information for business decision-making is known as data analysis. Extracting usable information from data and making decisions based on that analysis are the goals of data analysis.

Every time a decision is made in daily life, as a simple example of data analysis, what happened previously or what would happen if that particular choice is made is taken into consideration. This is nothing but examining our past or future and making judgements based on it. For that, memories of past or dreams of future is needed. Thus, all that is is data analysis. Today, data analysis refers to the same process an analyst uses for commercial goals.

3) Algorithms Used:

Linear Regression: The most popular and widely used algorithm for machine learning is linear regression. It is employed to create a linear relationship between the response or independent variables and the target or dependent variable. This algorithm's primary goal is to

identify the line that fits the target variable and the data's independent variables the best.

It is accomplished by determining all of the values that are most ideal.

The predicted value should be the closest it can be to the real numbers and have the least amount of error.

The error is the separation between the data points and the regression line that was fitted.

K-Neighbors Regressor

A supervised learning strategy is used in the KNN method for regression. Based on the resemblance with other available cases, it forecasts the target. The similarity is calculated using the distance measure, with Euclidian distance being the most common way.

Finding the K examples from the total dataset that are most similar to the testing point, or its neighbours, is how predictions are made. KNN method uses the Euclidean distance formula to calculate the distance between these mathematical values.

$$\sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

The value of K that should be chosen shouldn't be too low because doing so could cause the data to overfit and introduce noise.

Reserving a portion of the data to evaluate the model's accuracy is the general solution. Choose K=1 after which you may calculate the prediction accuracy using all the samples in the test set by using the training portion of the model. Continue this procedure while increasing K, then select K so that it is ideal for the model.

XG Boost Regressor: Extreme Gradient Boosting, or XGBoost, has been used to create a model with great computational speed and efficacy. The ensemble technique, which models the anticipated errors of some decision trees to optimise previous forecasts, is used by the formula to create predictions. The development of this model also reports the importance of each feature's contributions to the prediction of the final construction performance score. This feature value shows the impact each attribute has on predicting school achievement in absolute terms. By building decision trees concurrently, XGBoost aids in parallelization. This algorithm can evaluate any huge and complex model, which makes distributed computing another important characteristic it possesses.

Random Forest Regressor: Random Forest is described as a group of decision trees that uses a bagging technique to help get accurate results. Two of the most popular ensemble strategies that aim to address higher variability and higher prejudice are boosting and bagging. There are numerous base learners, or base models, which take different random samples of records from the training dataset during bagging. The basic learners in Random Forest Regressor decision trees are trained using the data that they have gathered. Decision trees themselves are not accurate learners because,

when used to their greatest extent, there is a substantial likelihood of overfitting, which results in high training accuracy but low real accuracy. As a result, each model has been trained using all of these data files, and anytime test data is given to any of the trained models already in existence, their predictions are integrated in such a way that the output is the mean of all the outcomes produced. Here, the individual findings are aggregated, a process called as aggregation. The number of decision trees that must be taken into account in order to construct a random forest is the hyperparameter that needed to control in this algorithm.

Evaluating Results:

Results of the model can be evaluated by the use of following commonly used Distance Metrics:

Root Mean Squared Error:

Root Mean Square Error (RMSE) is the residuals' standard deviation (prediction errors). The distance between the data points and the regression line is measured by residuals, and the spread of these residuals is measured by RMSE. In other words, it provides information on how tightly the data is clustered around the line of best fit. In climatology, forecasting, and regression analysis, root mean square error is frequently used to validate experimental results.

R2 SCORE:

For assessing the effectiveness of a machine learning model based on regression, the R2 score is a crucial indicator. It is also referred to as the coefficient of determination and is pronounced as R squared. It operates by calculating the variation in the predictions that the dataset can explain.

II. LITERATURE REVIEW

A. EXISTING SYSTEM SUMMARY

Typically, sales forecasting involves gathering sales data from a store over a specific time period and making predictions using a variety of prediction approaches. Many factors, including as direct and indirect competition, state and local holidays, demographic fluctuations, sales incentives, etc., have an impact on sales forecasting. The aforementioned elements lead to a significant variance in sales forecasting in the current method, which does not deliver accurate results as anticipated. For some algorithms, the confidence level has not been implemented. Holiday considerations, which are crucial for predicting sales, are not taken into account. As a result, adopting various machine learning algorithms affects sales differently.

“Intelligent Sales Prediction Using Machine Learning Techniques”:

The fashion store dataset for sales data over three consecutive years is utilised. The sales forecast is over the following three years, 2015–2017. The data mining paradigm goes through several stages of exploratory analysis, including data interpretation, preparation, modelling, evaluation, and deployment. The forecast is made up of a smoothed average that has been linear trend-adjusted. The forecast is then seasonality-adjusted as well. Future sales are predicted using machine learning methods

like the Generalized Linear Model (GLM), Decision Tree (DT), and Gradient Boost Tree (GBT). Gradient Boost Algorithm offers 98% overall accuracy, Decision Tree Algorithms come in second with roughly 71% overall accuracy, and Generalized Linear Model comes in third with 64% accuracy, according to performance. Eventually, if the three algorithms are empirically evaluated, Gradient Boosted Tree, which offers the highest prediction accuracy of all the algorithms, is shown to be the best fit for the model.

“Machine Learning Models for Sales Forecasting”

To forecast future sales, the "Rosemann Store Sales" dataset is used. The primary Python packages used for the calculations were pandas, sklearn, numpy, keras, matplotlib, and seaborn. The regression algorithm detects patterns across the entire collection of products or businesses. The research takes into account factors including the previous data's mean sales value, state and school holiday flags, the distance between the store and the competitor's store, and the selection style of the retailer when making predictions. The data is analysed using a variety of machine learning models, including Random Forest, Neural Network, Lasso Regularization, Arima model, and ExtraTree model. Results from the first level's models (ExtraTree, Lasso, and Neural Network) have non-zero coefficients. Three level models are the foundation of the solution. Several models at the first level were built using the XGBoost machine learning method. Models from the Python scikit-learn package, Extra tree model, linear model, and neural network model are included in the second level. On the third level, the results from the second level were added with weights. When compared to time series methods, the application of regression models for sales forecasting can produce better results. In comparison to previous methods, the ExtraTree method offers greater stacking weights for regressors.

"Walmart's Sales Data Analysis- A Big Data Analytics Perspective"

Each of the 45 Walmart stores includes 99 departments and is located in a variety of different geographic regions. The dataset includes sales data for each store location over a three-year period as well as information on variables that may effect sales, such as temperature, fuel prices, unemployment rates, and holidays. This work makes use of Apache data science platforms, libraries, and tools. The data is analysed and visualised using tools like Hadoop Distributed File Systems (HDFS), Hadoop MapReduce framework, Apache Spark, and high-level programming languages like Scala, Java, and Python. To forecast future sales, a machine learning library and a straightforward regression model are used. The results are forecasted by data analysis, and the retailers must plan and assess their strategies in light of the market-driving elements, including but not limited to, weather, gasoline prices, holidays, human resources, geographic location, and a host of other variables. In particular, arranging sales at several locations requires effective and efficient supply chain, inventory, and human resource management to maintain a competitive advantage in the market.

“Forecast of Sales of Walmart Store Using Big Data Applications”

Walmart sales information is used in the forecasting process. The various store kinds, including convenience stores, department stores, luxury stores, super markets, shopping malls, etc., aid in choosing business models and operational strategies. The procedure includes steps like figuring out the dependent and independent variables, Create forecasting procedures, choose a forecast analysis method, collect and analyse data, provide data-related hypotheses, create and finalise forecasts, and assess outcomes. The plan calls for gathering vast amounts of sales-related data, which is then sent to HDFS (Hadoop's distributed file system) for map reduction. To forecast sales, the Holt Winters algorithm is employed. The algorithm exhibits seasonality, trend, and randomness. The final results show that the numerical depiction of the forecasted sales and the accuracy of sales predicted are measured by 80% low confidence sales, 80% high confidence sales, and 95% low confidence sales and 95% high confidence sales. Error factor can be found between the predicted sales and the observed sales data, i.e. to find the error factor of month June in both the predicted sales and the observed sales data, then the difference between predicted and observed sales should be taken into account.

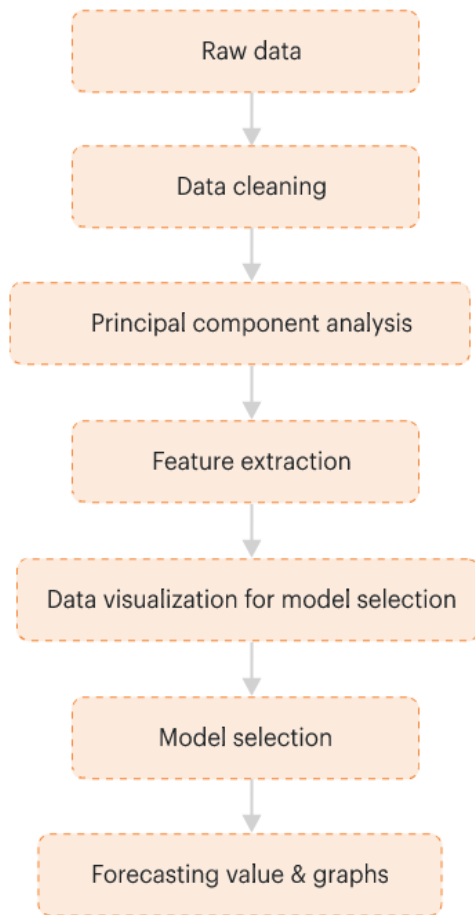
D. PROPOSED WORK




The proposed study is to conduct research that will lead to the creation of a method for predicting sales. The Sales Forecasting Application, which is the proposed project, will be accomplished by separating it into the following goals:

Obtain a Big Mart Dataset from Kaggle and refine it with real-world samples. Develop machine learning models to achieve optimum accuracy. We'll add another function to the algorithm that will present the analysis as a bar chart, line chart, or pie chart, giving the user a better understanding of future sales analysis. Flask will be used to create the GUI. Installation and hands-on experience on existing approaches of Sales Forecasting will be done. Relative pros and cons will be identified. Various parameters will be identified to evaluate the proposed system.


Comparison of newly implemented approach with existing approaches will be done.

FLOW CHART



	and sharing documents that contain live codes, visualisations, text, and equations. There are over 100 kernels other than IPython available for use.	
Atom Text Editor	Atom is a text and source code editor which works across all operating systems. It speeds up find-and-replace operations by an order of magnitude and improves performance of files	
Visual Studio Code	Visual studio code is an open-source code editor built for Windows, Mac OS, Linux which can be used for various programming languages like Java, JavaScript, Python, C, C++, Node.js.	
Flask	Flask is a micro web framework written in Python. It is classified as microframework because it does not require particular tools or libraries. It has no database abstraction layer, formvalidation, or any other components where pre-existing third-party libraries provide common functions.	

The software tools that will be utilised in the development of this project are as follows:

Software Tool Used	Description	Logo
Jupyter Notebook	Jupyter Notebook is a web-based open-source application that is used for editing, creating, running,	

E. RESULTS:

The sales of several Big Mart outlets have been predicted using machine learning methods like Linear Regression, KNearest Neighbors, XGBoost, and Random Forest. For each of the four methods, many metrics that affect the accuracy of results are tabulated, including Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracy, and more.

Big Mart Sales Prediction

Created By © Aryan Gupta

Enter Item Weight

Enter Item Visibility

Enter Item MRP

Outlet Establishment Year (YYYY)

← → ↻ ⓘ http://127.0.0.1:5000/predict

```
{"Prediction":4972.7197082824705}
```


F. CONCLUSION:

The usage of machine learning approaches proves to be a crucial feature for designing business plans while taking into consideration consumer buying habits, as traditional methods are not very helpful to commercial organisations in revenue growth. Businesses can adopt effective tactics for expanding sales and stepping unafraid into the competitive world by using sales predictions based on a variety of factors, including past sales.

F. FUTURE SCOPE:

- 1) **Real-time prediction:** The ability to predict sales in real-time can be a game-changer for businesses, allowing them to quickly adjust their strategies based on market trends and customer behavior. User can Upload there CSV Files of any organization and get the future sales
- 2) Expand this research to a larger scale in the future so that different embedding models can be considered on a wider range of datasets
- 3) To create an android application based on this machine learning model.
- 4) To integrate this model to an offline based system so that it runs completely without an internet connection

REFERENCES

- [1] Singh Manpreet, Bhawick Ghutla, Reuben Lilo Jnr, Aesaan FS Mohammed, and Mahmood A. Rashid. "Walmart's Sales Data Analysis-A Big Data Analytics Perspective." In 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 114-119. IEEE, 2017..
- [2] Sekban, Judi. "Applying machine learning algorithms in sales prediction." (2019)
- [3] Panjwani, Mansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar, and Rupali Hande. Sales Prediction System Using Machine Learning. No. 3243. EasyChair, 2020..
- [4] Cheriyan, Sunitha, Shaniba Ibrahim, Saju Mohanan, and Susan Treesa. "Intelligent Sales Prediction Using Machine Learning Techniques." In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 53-58. IEEE, 2018.
- [5] Giering, Michael. "Retail sales prediction and item recommendations using customer demographics at store level." ACM SIGKDD Explorations Newsletter 10, no. 2 (2008): 84-89. Clark, B. (2018).
- [6] Baba, Norio, and Hidetsugu Suto. "Utilization of artificial neural networks and GAs for constructing an intelligent sales prediction system." In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, vol. 6, pp. 565-570. IEEE, 2000.

- [7] Ragg, Thomas, Wolfram Menzel, Walter Baum, and Michael Wigbers. "Bayesian learning for sales rate prediction for thousands of retailers." *Neurocomputing* 43, no. 1-4 (2002): 127-144.
- [8] Fawcett, Tom, and Foster J. Provost. "Combining Data Mining and Machine Learning for Effective User Profiling." In *KDD*, pp. 8-13. 1996
- [9] Brownlee, J. (2018). *How to Develop a Deep Learning Model to Achieve World-Class Forecasting Accuracy. Machine Learning Mastery*
- [10] Cheng, Y., Xu, L., & Wang, J. (2018). Sales forecasting in retail industry using deep learning. In *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)* (pp. 3594-3597). IEEE.
- [11] Gudmundsson, K., & Helgason, H. (2020). Machine learning in sales forecasting. Available at: <https://www.arnia.com/blog/machine-learning-in-sales-forecasting>
- [12] Hu, Y., Yang, H., & Chao, C. (2018). Sales Forecasting with a Hybrid Method of ARIMA and XGBoost. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2813-2818). IEEE.
- [13] Sarkar, S., Rana, S., & Kumar, S. (2021). Sales forecasting using machine learning algorithms: A comprehensive review. *Journal of Retailing and Consumer Services*, 59, 102376.
- [14] Wang, H., Lu, H., & Gao, Y. (2018). Sales forecasting with neural networks and feature engineering. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 1158-1163). IEEE.
- [15] Zeng, L., & Zhao, Y. (2017). Sales forecasting based on machine learning in e-commerce. In *2017 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)* (pp. 58-63). IEEE.
- [16] Jiang, Y., Hu, J., Zhang, Q., & Zhang, Y. (2019). Sales forecasting of online retailing using LSTM neural network. *Journal of Physics: Conference Series*, 1159(3), 032012.
- [17] Zhang, Q., Ren, Y., Xu, Z., & Zhang, J. (2021). Sales forecasting of online shopping.