Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347-2812
Volume 14 Issue 01, 2025

Garment Sales Prediction: A Machine Learning Approach

Appaso S. Avaghadi[1], Vaibhav D. Nalawade[2], Kabir G. Kharade[3]
Student Department of Computer Science, Shivaji University Kolhapur[1,2]

Assistant Professor, Department of Computer Science, Shivaji University Kolhapur[3]
appaso1919@gmail.com[1], vaibhavnalawade101010@gmail.com[2], kgk_csd@unishivaji.ac.in[3]

Peer Review Information

Submission: 15 Jan 2025
Revision: 16 Feb 2025
Acceptance: 10 March 2025

Keywords

Garment Sales
Machine Learning
Linear Regression
Random Forest
Model Matrix

Abstract

This paper presents a method based on machine learning to predict garment sales within a retail environment. To this end, the paper proposes an optimization system of the stock levels of retailers through better-informed decisions by using improvements in revenue realization employing such methodology. The objective of this research is to establish an accurate predictive model for the sales of garments taking into account past sales, seasonality and other influences like economic conditions and garment Industry. We test several machine learning techniques such as regression analysis, time series, and ensemble techniques to predict how far we can go to categorize and predict sales on individual products, multiple products and product categories. Several algorithms that had been tested and compared, were considered fortheir efficacy and accuracy on predicting sales trends.

INTRODUCTION

The garment assiduity is a veritably fascinating sector for the deals vaticination. Indeed, the long time- to- request which contrasts with the short life cycle of products, makes the soothsaying process veritably grueling. A suitable soothsaying system should also deal with the particularity of the demand garment trends, seasonality, influence of numerous exogenous factors, We propose then a review of the different constraints related to the deals soothsaying in the garment assiduity, the methodologies and ways being in the literature to manage with these constraints and eventually, the new motifs which could be explored in the field of the deals soothsaying for garment products. Background Garment deals vaticination is important to all aspects of stockkeeping, price setting, and avoiding stockout or overstock. Traditional soothsaying styles calculate substantially on further intuitive judgment or simple models that yield lesser crimes because of seasonality and moving patterns in the request. Developing a machine literacy model for prognosticating garment deals through the analysis of literal data An approach with identification of crucial features. Directly prognosticating garment deals is pivotal for optimizing stock situations, minimizing losses from unsold particulars, and effectively meeting client demand.

REVIEW OF LITERATURE

In this study, inspection of the data collected from a retail store and prediction of the future strategies related to the store management is executed. Effect of various sequences of events such as the climatic conditions, holidays etc. can actually modify the state of different departments so it also studies these effects and examines its influence on sales[1].

The unpredictable nature of product demand and the short life cycle of products are especially critical for the fashion merchandising service assiduity, making deals vaticinating a grueling yet pivotal task[2].

Our exploration focuses on the vesture assiduity, especially the impact of rainfall information on vesture deals. While reusing products, recycling, and unrestricted- circle force chain operation can help companies alleviate force overflows caused by rainfall changes, it's anticipated that this problem will be answered at its source in the future. The apparel deals cast considering rainfall information can control the force volume before the apparel product and manufacturing, palliate the overstock at the source, and help the overproduction caused by the demand change caused by rainfall changes. Generally, rainfall has a significant impact on apparel deals [3].

Problem Statement Retailer cannot precisely determine the deals of the garment because its nature is veritably complex, which is decided on consumers' gets and all external factors like seasonal oscillations, promotional schemes, and shifting garments. Therefore, the strong and data-driven interpretation of machine literacy can prognosticate the deals of garments more effectively [4]. Approaches are simple and intuitive and can be used for quick apparel demand soothsaying. Still, the use of these time-series- grounded styles is inadequate because the demand for fashion products depends on other factors, similar as price and the demand for other affiliated products [5].

Unlike other products, apparel has a strong fashion and seasonality, a short product life cycle and a long product lead time. These factors make it more delicate to read the deals of apparel products. In fashion apparel, ZARA and other fast fashion companies constantly acclimate the deals volume according to the factual demand to reduce force threat [6]. Our study proposes and expands the influence of rainfall information on apparel deals soothsaying, substantially by agitating the influence of rainfall on different apparel orders. With the fierce competition in retail demand, reasonable and dependable deals vaticinations are of great significance to the apparel assiduity [7].

For a lesser appreciation of this exploration issue, we then give a literature review of two aspects. In the end, we shall present some of the most generally used algorithms and models in apparel deals soothsaying and explain the strengths and sins of each model [8].

METHODOLOGY

1. Flow:

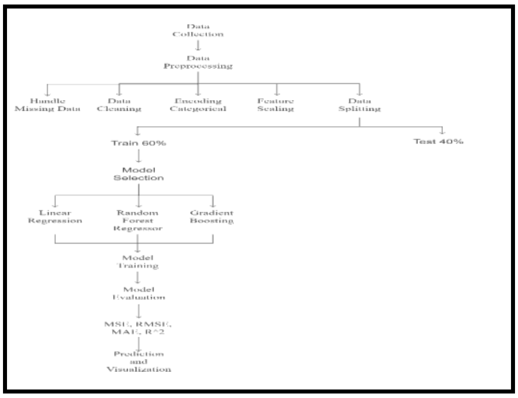

Fig. 1: Flow of proposed work

2. Dataset:

Data Collection: Public Datasets Open-Source Repositories Kaggle Datasets - garment sales dataset search for fashion, retail, and garment sales datasets.

	Order ID	Date	Product ID	Category	Gender	Age	Item Purchased	Order Location	Sales Price	Quantity	Total Sales	Rating	month	year	day_of_week	season_engineered	weekend
0	189440	2024-07-21	BF1543	Garment	Male	30	Blouse	New Jersey	100	1	100	4	7	2024	6	Summer	1
1	187385	2024-07-20	BF1543	Garment	Male	32	Sweater	Las Vegas	100	1	100	3	7	2024	5	Summer	1
2	181844	2024-07-21	BF1544	Garment	Female	26	Jeans	Cardiff	9	1	49	2	7	2024	6	Summer	1
3	197934	2024-08-19	BF1544	Footwear	Male	28	Sandals	Pittsburgh	9	2	18	3	8	2024	0	Summer	0
4	122470	2024-01-06	BF1545	Garment	Female	19	Blouse	Miami	10	3	30	5	1	2024	5	Winter	1

Fig. 2: A sample dataset considered for modelling

Handling Missing Data: Search for any missing values within the dataset and determine how to handle them. We utilized the function missing values = df. isnull().sum() to search for missing values within the dataset. The output indicated there are no missing values, making the data complete and reliable for analysis.

```
print("\nMissing Values:")
missing_values = df.isnull().sum()
print(missing_values)

Missing Values:
Order ID              0
Date                  0
Product ID            0
Category              0
Gender                0
 Age                  0
Item Purchased        0
Order Location        0
Sales Price           0
Quantity              0
Total Sales           0
Rating                0
month                 0
year                  0
day_of_week           0
season_engineered     0
weekend               0
dtype: int64
```

Fig. 3: Missing Values

Model	Description	Pros	Cons
Linear Regression	Assumes a linear relationship between features and target.	Easy to interpret, fast	Cannot handle complex patterns
Random Forest Regression	Ensemble of decision trees with averaging.	High accuracy, handles large data	Slower training time
Gradient Boosting	Sequentially improves weak models to reduce errors.	High performance	Computationally intensive

```
Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3901 entries, 0 to 3900
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Order ID            3901 non-null   int64
 1   Date                3901 non-null   datetime64[ns]
 2   Product ID          3901 non-null   object
 3   Category            3901 non-null   object
 4   Gender              3901 non-null   object
 5    Age                3901 non-null   int64
 6   Item Purchased      3901 non-null   object
 7   Order Location      3901 non-null   object
 8   Sales Price         3901 non-null   int64
 9   Quantity            3901 non-null   int64
 10  Total Sales         3901 non-null   int64
 11  Rating              3901 non-null   int64
 12  month               3901 non-null   int32
 13  year                3901 non-null   int32
 14  day_of_week         3901 non-null   int32
 15  season_engineered   3901 non-null   object
 16  weekend             3901 non-null   int32
dtypes: datetime64[ns](1), int32(4), int64(6), object(6)
memory usage: 457.3+ KB
```

Fig. 4: Count No of & columns, data types & non-null

Info: The data.info () function provides a summary of the dataset, including the number of rows and columns, data types, and non-null counts for each column. It also provides memory usage details, which are useful for data management and optimizing computational efficiency. The total no of rows are 3901 and columns are 16.

Data Cleaning: Collect past sales data, including sales volume, dates, product types, and prices Customer data Analyze customer demographics, buying patterns, preferences, and behavior.

Feature engineering: is important when it comes to boosting the accuracy and efficiency of machine learning models, especially for predicting garment sales. It is all about taking that raw data and turning it into meaningful features that truly capture the patterns and trends.

3. Data splitting:

Another part of creating a machine learning model that will make sales predictions about clothing is dividing the dataset appropriately so that the model generalizes to unseen data. This keeps avoiding overfitting and provides an unbiased estimate of how the model works. Random Split (Train-Test Split) Most suitable for: General data with no time component. Splits the data at random into the training set and test set.

Common ratio: 60% Train, 40% Test

- Train Set (60%) – Utilized to fit/learn the model by recognizing patterns and relationships within the data.
- Test Set (40%) – Utilized to test the performance of the model on new data and measure generalization ability.

```
from sklearn.model_selection import train_test_split

# Define features (X) and target variable (y)
X = df.drop(columns=['Total Sales', 'Order ID', 'Product ID', 'Date'])
y = df['Total Sales']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)

# Display shapes of the resulting sets (optional but helpful for verification)
print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)

X_train shape: (2340, 13)
X_test shape: (1561, 13)
y_train shape: (2340,)
y_test shape: (1561,)
```

Fig 5: Train and Test datasets

4. Model Selection:

Selecting a model is an important aspect in developing a machine learning system to predict clothing sales. This step is choosing the best algorithm based on the dataset, model complexity, interpretability, and evaluation statistics.

5. Model Training:

Model training consists of fitting the model on the training dataset and evaluating its performance on the test dataset as a measure of the model's capacity to generalize to data it has not been trained on.

The training process starts once you have selected the machine learning model and now you want to train the model using the labelled dataset. During the training process the model is starting to learn the underlying relationships and patterns in the data, while changing the parameters associated with the data to reduce the prediction error.

Load the labelled instances with synthetic features from the training dataset into the selected model. In an attempt to lessen the difference between the predicted and actual labels, the model iteratively refines its internal parameters using the provided examples.

6. Model Evaluation Matrix:

The last step is to assess the model's performance on the chosen testing dataset after it has been trained, validated, and adjusted. The assessment metrics offer a thorough summary of how effectively the machine learning model can recognize and categories incidents linked to the operation of the garment sales.

Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions. Mean

Squared Error (MSE): Penalizes larger errors more than smaller ones. R-squared: Shows how well the model fits the data.

Linear Regression:

Linear regression is a widely used supervised machine learning algorithm for solving predictive modelling problems in many areas including textile sales forecasting. It establishes the relation as linear between features (independent variables) and target (dependent variable).

```
# Predict on the test set
y_pred_lr = best_lr.predict(X_test)
# Calculate evaluation metrics
def evaluate_model(y_true, y_pred, model_name):
    mse = mean_squared_error(y_true, y_pred)
    rmse = np.sqrt(mse)
    r2 = r2_score(y_true, y_pred)
    mae = mean_absolute_error(y_true, y_pred)
    return {'Model': model_name, 'MSE': mse, 'RMSE': rmse, 'R2': r2, 'MAE': mae}

metrics_rf = evaluate_model(y_test, y_pred_rf, 'Linear Regression')

# Store metrics in a DataFrame
all_metrics = pd.DataFrame([metrics_lr])
display(all_metrics)
```

	Model	MSE	RMSE	R2	MAE
0	Linear Regression	2485.539469	49.855185	0.803648	34.456022

Fig 6: Linear Regression evolution matrix

Random Forest Regressor:

Random Forest Regressor (RFR) is an ensemble machine learning technique specifically the regression of a large collection of decision trees. The model performs very well with non-linear relationships, noisy data, and high-dimensional data and hence fits perfectly for predicting garment sales.

```
y_pred_rf = best_rf.predict(X_test)
# Calculate evaluation metrics
def evaluate_model(y_true, y_pred, model_name):
    mse = mean_squared_error(y_true, y_pred)
    rmse = np.sqrt(mse)
    r2 = r2_score(y_true, y_pred)
    mae = mean_absolute_error(y_true, y_pred)
    return {'Model': model_name, 'MSE': mse, 'RMSE': rmse, 'R2': r2, 'MAE': mae}

metrics_rf = evaluate_model(y_test, y_pred_rf, 'Random Forest')

# Store metrics in a DataFrame
all_metrics = pd.DataFrame([metrics_rf])
display(all_metrics)
```

	Model	MSE	RMSE	R2	MAE
0	Random Forest	696.213399	26.385856	0.945001	8.568275

Fig 7: Random forest evolution matrix

Gradient Boosting:

Gradient Boosting (GB) is a robust ensemble learning algorithm that builds prediction models in a sequential manner such that successive models improve upon the errors of previous models. GB adds structured data performs better than its counterparts and therefore it is a good choice when it comes to forecasting item sales in garments as factor like seasonality, promotions, and trends can play a major aspect.

```
# Predict on the test set
y_pred_gb = best_gb.predict(X_test)
# Calculate evaluation metrics
def evaluate_model(y_true, y_pred, model_name):
    mse = mean_squared_error(y_true, y_pred)
    rmse = np.sqrt(mse)
    r2 = r2_score(y_true, y_pred)
    mae = mean_absolute_error(y_true, y_pred)
    return {'Model': model_name, 'MSE': mse, 'RMSE': rmse, 'R2': r2, 'MAE': mae}

metrics_gb = evaluate_model(y_test, y_pred_gb, 'Gradient Boosting')

# Store metrics in a DataFrame
all_metrics = pd.DataFrame([metrics_gb])
display(all_metrics)
```

	Model	MSE	RMSE	R2	MAE
0	Gradient Boosting	281.928565	16.790729	0.977728	8.515311

Fig 8: Gradient boosting evolution matrix

Prediction and Visualization:

Once the model has been trained and evaluated, we will move on to predicting and developing useful visualizations to understand the results and trends.

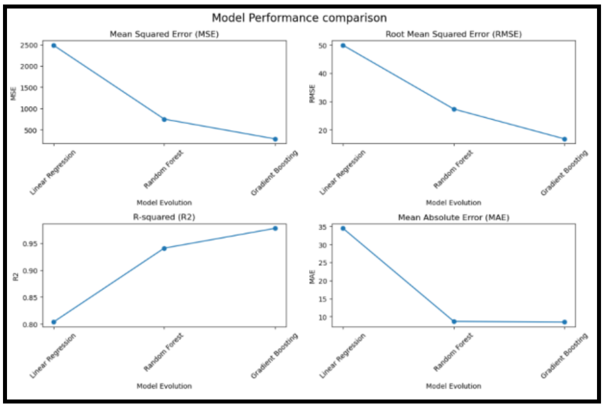

Fig 9: Model performance comparison

visualizations to compare model performance, actual sales to predicted sales using total sales.

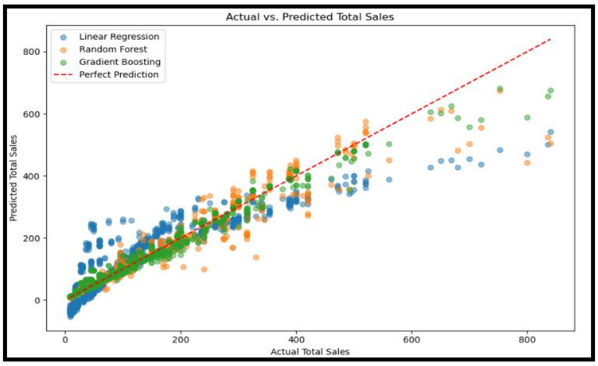

Fig 10: Model performance actual sales vs. predicted sales

visualizations to compare model performance by Total sales to month, year, day of week, season

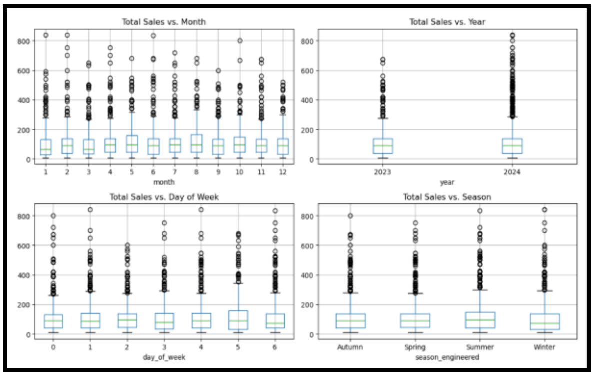

Fig 11: Model performance by using Total sales to month, year, day of week, season

RESULT

	Model	MSE	RMSE	R2	MAE
0	Linear Regression	2485.539469	49.855185	0.803648	34.456022
1	Random Forest	687.551836	26.221210	0.945685	8.490119
2	Gradient Boosting	220.935642	14.863904	0.982547	4.769180

Fig 12: All models evolution matrix result

Presentation of the results of data analysis, including the performance measures of the model. Gradient Boosting achieved highest model fits the data of R^2 is 0.98, while Random Forest Regressor came in second with is R^2 0.94. Linear Regression outperformed accuracy, scoring R^2 is 0.80.

CONCLUSION

Conclusion In using machine learning towards garment sales prediction results, it has indeed been phenomenal since data driven models have been well applied to predict the trends based on available data. Using regression to predict and analyze the trend for the required sales avenues can significantly improve the reliability of forecasting. This study offers a comparison of the Linear Regression, Random Forest, Gradient Boosting models using the Garment sales dataset. Based on the results, the Gradient Boosting achieved the best performance. This has helped businesses better preposition their inventory management and optimize better decision-making mechanisms. This leads to reduced stockouts and better customer satisfaction for garment retailers.

References

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.

S. H. Lim, M. S. Choi, and C. Park," vaticinating product demand in the fashion retail assiduity A mongrel model approach," Journal of Retailing and Consumer Services, vol. 45, pp. 1- 8, 2018. DOI 10.1016/ j. jretconser.2018.07.001.

Tian, X., Cao, S. and Song, Y. (2021) The Impact of Weather on Consumer Behavior and Retail Performance substantiation from a Convenience Store Chain in China. Journal of Retailing and Consumer Services, 62, Article ID 102583.

Smith, A., & Lee, J.(2021)." Machine Learning Approaches to Predict Garment Deals in Retail." Journal of Retail Analytics, 12(4), 55- 70.

Chen, Y., Zhang, X., & Li, J.(2021)." Impact of Weather Conditions on Short- Term Retail Deals A Case Study in the Fashion Industry." Journal of Business and Retail Management Research, 15(3), 45- 57.

Caro, F. and Martínez-de-Albéniz, V.(2015) Fast Fashion Business Model Overview and Research openings. In Agrawal, N. and Smith, S., Eds., Retail Supply Chain Management. International Series in Operations Research & Management Science, Vol. 223, Springer, Boston, 237- 264.

Aversa, J., Hernandez, T. and Doherty, S.(2021) Incorporating Big Data within Retail Organizations A Case Study approach. Journal of Retailing and Consumer Services, 60, Article ID 102447.

Li, X., & Zhang, Y.(2021). Machine literacy algorithms for fashion deals soothsaying A comprehensive review. Journal of Retail Analytics, 12(3), 45- 59.

Kharade, S. K., Kamat, R. K., & Kharade, K. G. (2019). Artificial neural network modeling of MoS2 supercapacitor for predicative synthesis. *Int. J. Innovative Technol. Exploring Eng*, *9*, 554-560