International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Sales Forecasting in Retail using Machine Learning: A Case Study on Big Mart Dataset

Rishul, Antriksh Baluni, Jatin Rajput

Department of Computer Science and Engineering (Artificial Intelligence)

IIMT College of Engineering, Greater Noida, India

Emails: rishulbosatta@gmail.com , antriksh.baluni@gmail.com, rajputjatin@gmail.com

Abstract

Stores really need to guess accurately how much they'll sell to handle what they have in stock, how they price things, and their special offers. This paper looks at using smart computer programs (machine learning) to guess these sales using old sales info from Big Mart stores. The study used different guessing methods like basic ones (Linear Regression, Ridge Regression) and a better one (Random Forest) to predict how much of each item would sell in each store. The results showed that the smarter method, Random Forest, was better at guessing and working with new data. The paper also talks about getting the data ready, making new useful information from it, and checking how good the guesses were, to give a full picture of how to predict sales in stores.

Keywords—Machine Learning, Sales Forecasting, Retail Analytics, Big Mart Dataset, Regression Models, Data Preprocessing

1. Introduction

Knowing how much stuff will sell in stores is super important for making money, keeping customers happy, and using resources wisely. Older ways of guessing sales often struggle with complicated real-world data. Because we now use data a lot to make decisions, machine learning programs have become popular because they can learn from old data and make good guesses.

In this study, we looked at using machine learning to guess how much of each item would sell in Big Mart stores. The goal was to see which of the different guessing methods worked best for store sales. This project shows how even simple but effective computer learning methods can help get useful information from store data that doesn't just look at trends over time.

2. Related Work

Many studies before this one have tried to guess sales using both math and machine learning.

Traditional time-series models such as ARIMA often underperform in retail sales forecasting due to their reliance on stationary data assumptions, a limitation highlighted by Aggarwal & Sharma (2021) in dynamic retail environments.

On the other hand, machine learning guessing methods like Linear Regression, Ridge, and Random Forest have become popular. For example, a study by Aggarwal et al. (2021) found that combining different guessing methods usually works better in stores. Other studies show that making the data better and changing categories into numbers helps the guesses.

This paper builds on these ideas by using data that anyone can get and seeing how well these methods work.

3. Dataset Description

The study analyzes the Big Mart Sales dataset, a publicly available retail dataset comprising 8,523 product entries from multiple store outlets. Each record includes 12 attributes that capture product-specific and store-specific features, categorized as follows:

Independent Variable

1. **Product Identification & Attributes**

- **Item_Identifier:** Alphanumeric code distinguishing individual products.
- **Item_Weight:** The mass of the item (in grams), critical for logistics and pricing analysis.
- **Item_Fat_Content**: Binary classification (Low Fat/Regular) for dietary categorization.

2. **Store Attributes**

- **Outlet_Identifier**: Unique ID for each retail outlet.
- **Outlet_Establishment_Year**: Operational inception year (e.g., 1999, 2007).
- **Outlet_Size:** The size of the store in terms of ground area covered.

3. **Spatial features**

- **Item_Visibility:** Relative shelf visibility (0–1 scale), calculated as a percentage of total store display space.
- **Outlet_Location_Type:** The type of city in which the store is located

Data includes both numerical and categorical features, requiring preprocessing before feeding into ML models.

4. Methodology

A. Data Preprocessing

- **Handling Missing Data:** To address missing entries in the Item_Weight column, we replaced null values with the mean weight of products in the dataset, a common practice for numerical data imputation

- **Categorical Encoding Transformation:** Label Encoding and One-Hot Encoding were applied to convert categorical features to numerical form.

- **Feature Engineering:** New features such as Item_Category (derived from Item_Identifier) were created to improve model performance.

- **Scaling:** Standard Scaler was used for models sensitive to feature magnitudes like Ridge Regression.

B. Model Selection and Training

- **Linear Regression:** Simple baseline model assuming linear relationships.

- **Ridge Regression**: Regularized versions of linear regression to prevent overfitting.
- **Random Forest Regressor:** Utilized 100 decision trees (max_depth=10) to capture non-linear patterns and feature interactions..

C. Evaluation Protocol

- Performance measured via 5-fold cross-validation to mitigate overfitting.
- Cross-validation was also used to check for overfitting and ensure the robustness of the models.

5. Results and Discussion

Model Performance:

Model	Root Mean Squared Error	R^2 Score
Linear Regression	1172.34	0.51
Ridge Regression	1135.68	0.53
Random Forest	980.23	0.64

While Random Forest outperformed linear models, its higher computational cost may deter small retailers. Future work could explore lightweight alternatives like Gradient Boosting for resource-constrained scenarios.

Linear and Ridge models offered simplicity and faster training but were limited in accuracy. Ridge performed slightly better than plain Linear Regression due to regularization.

6. Conclusion

Our analysis demonstrates that machine learning models, particularly Random Forest, significantly enhance sales prediction accuracy in retail settings, as evidenced by the 64% R^2 score achieved on the Big Mart dataset. More advanced methods, especially Random Forest, gave the best guesses. Getting the data ready and creating new useful information were really important for getting good results. This study shows that even if you don't have information about sales over time, you can still make pretty good guesses using information about the products and the stores. Future work could look at even more advanced methods like deep learning, adding time-based information, or making the guesses easier to understand.

7. References

1. Aggarwal, R., & Sharma, M. (2021). Ensemble machine learning approaches for retail sales prediction: A comparative analysis. Journal of Data Science and Analytics, *15*(3), 112–130.

2. Brownlee, J. (2016). Practical machine learning for predictive analytics: A Python-based guide.