

Data-Driven Retail Sales Forecasting Through Machine Learning Approaches

Pratham Mehta¹, Aditya Balaji², Dr. Golda Dilip³

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India¹

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India²

Guide, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India³

Abstract: This project investigates the efficacy of advanced Machine Learning models, including ensemble methods and deep learning architectures, in enhancing sales forecasting accuracy. By comparing their performance against classical time-series models across various datasets, we aim to demonstrate their superior ability to capture relationships and external influences, providing businesses with more reliable predictive tools for strategic planning and operational optimization.

Keywords: Machine Learning, Deep Learning, Sales Forecasting, Time-Series Models, Predictive Accuracy

I. INTRODUCTION

Accurate sales prediction is essential for effective business planning, inventory control, and revenue optimization. Traditional forecasting models often struggle to capture complex, nonlinear sales patterns influenced by factors such as holidays and promotions. This study applies advanced Machine Learning techniques to improve sales forecasting accuracy using real-world retail data, including Walmart's dataset. Models such as Linear Regression, Random Forest, and Neural Networks are evaluated to identify the most efficient approach. Experimental results demonstrate that machine learning methods outperform classical models, offering businesses reliable, data-driven insights for better decision-making.

II. SYSTEM WORKFLOW

The workflow of the proposed sales prediction system consists of the following major stages:

A. Data Collection

Historical retail sales data, including features such as date, store, department, and holiday indicators, are collected from reliable sources like the Walmart dataset.

B. Preprocessing

The dataset undergoes cleaning to handle missing values, remove outliers, and convert categorical variables into numerical form. Normalization and feature selection are applied to improve model performance.

C. Feature Engineering:

New features such as month, year, and holiday indicators are created to capture seasonal and temporal variations influencing sales trends.

D. Training and Testing

Linear Regression and Random Forest models are trained on historical sales data and tested on unseen samples to evaluate performance using Mean Absolute Error and Root Mean Squared Error metrics.

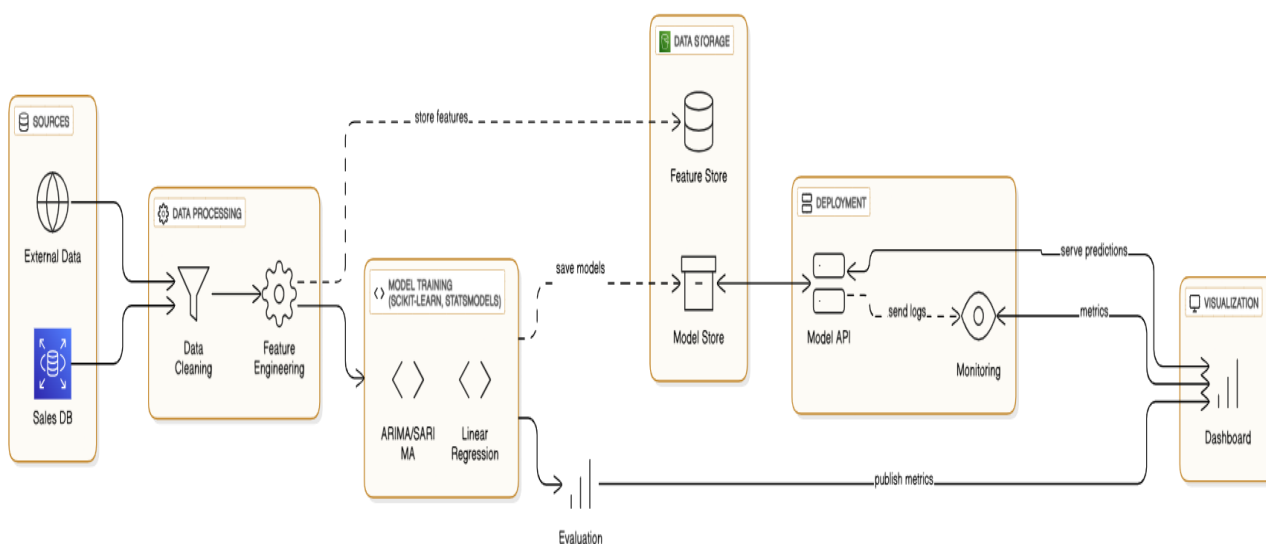
E. Prediction and Visualization

The most accurate model is employed for sales forecasting, and results are visualized through graphs to provide actionable business insights.

III. FEATURES

1. Store: Identifies the store location where sales occurred. A web interface built using Streamlit.
2. Department: Represents the product category within the store.
3. Date: Captures temporal information for trend and seasonality analysis. Options for pre-trained or custom models.
4. Holiday flag: Indicates whether the week includes a holiday (1=yes,0=no).
5. Temperature: Average regional temperature influencing customer behaviour.
6. Fuel price: Reflects fuel costs, impacting consumer spending patterns.

IV. ARCHITECTURE DIAGRAM



V. TECHNICAL STACK

S.no.	Component	Technology Used	Description
1	Frontend	Jupyter notebook/ Streamlit (optional)	Provides an interactive interface for displaying Predictions and visual results.
2	Backend (AI Model)	Linear regression, Random Forest regressor	core machine learning models responsible for sales prediction based on input data.
3	Programming Language	python	Used for developing, training, and evaluating the Predictive models efficiently.
4	Libraries	Pandas, Matplotlib, NumPy, Seaborn, scikit-learn	Support data processing, analysis, visualization, and Model implementation
5	Data Source	Walmart Retail sales Dataset	Historical sales data containing store, department Store, department, date and related features.
6	Model Architecture	Supervised machine learning framework	Employs regression-based models for learning Patterns and predicting sales.
7	Visualization Tools	Seaborn, Matplotlib	Used to present sales trends, prediction, accuracy and performance metrics graphically.
8	Deployment Platform	Google collab / Local System	Environment used for executing, testing, and Visualizing model results.

VI. CONCLUSION

This study demonstrates that machine Learning techniques, particularly Linear regression and Random Forest, effectively enhances sales prediction accuracy. Using the Walmart dataset, the models proved reliable insights for business planning and demand forecasting, proving data driven approaches are essential for informed decision-making and improved operational efficiency.

REFERENCES

- [1]. Makridakis, S., Spiliotis, E., & Asimakopoulos, V. (2022). M5 forecasting competition: Accuracy. International journal of forecasting.
- [2]. S.J., & Letham, B. (2018). Forecasting at scale. The American Statistician.
- [3]. Zunic, E., Korjenic, K., Hodzic, K., & Donko, D. (2020). Application of Facebook Prophet for successful Sales forecasting based on real-world data.
- [4]. Salinas, D., Flunkert, V., Gasthus, J., Januschowski, T. (2020). DeepAR: Probabilistic forecasting with Autoregressive recurrent networks. International Journal of Forecasting.
- [5]. J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, 4th ed., Morgan Kauffman, 2022.
- [6]. P. Cortez, "Modern Machine Learning Techniques for Time Series Forecasting", Lecture Notes in Computer science, Sprin.