Report On

Time Series Analysis on Sales using Random Forest

Submitted in partial fulfillment of the requirements of the Machine Learning Course project in
Semester VII of Fourth Year Computer Engineering

by
Dream Patel (6)
Hardik Nikam (5)
Jay Prajapati (10)

Mentor
Dr. Megha Trivedi

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(A.Y. 2023-24)

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

CERTIFICATE

This is to certify that the Course Project entitled "**Time Series Analysis on Sales using Random Forest"** is a bonafide work of **Dream Patel(6)**, **Hardik Nikam(5)**, **Jay Prajapati(10)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in Semester VII of Fourth Year **"Computer Engineering" .**

Dr. Megha Trivedi
Mentor

Dr Megha Trivedi
Head of Department

Dr. H.V. Vankudre
Principal

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Course Project Approval

This Course Project entitled "**Time Series Analysis on Sales using Random Forest"** by **Dream Patel (6), Hardik Nikam (5), Jay Prajapati (10)** is approved for the degree of **Bachelor of Engineering** in in Semester VII of Fourth Year **Computer Engineering .**

Examiners

1……………………………………..
(Internal Examiner Name & Sign)

2………………………………………
(External Examiner name & Sign)

Date:

Place:

Contents

Abstract

This project focuses on developing a Time Series Forecasting system tailored to Walmart, a major multinational retail corporation. Accurate forecasting is crucial for optimizing inventory, demand planning, and supply chain operations. We employ diverse machine learning models with libraries like pandas, and sklearn to predict sales data over time. Walmart's vast dataset, spanning multiple stores, product categories, and regions, presents data management challenges. Ensuring data quality is vital, as missing data and anomalies can affect accuracy. Evaluation metrics such as MAE, MSE, and RMSE gauge forecast accuracy, aiming to enhance supply chain efficiency and reduce costs.

ACKNOWLEDGEMENT

We would like to express our special thanks and gratitude to our Institute, **Vidyavardhini's College of Engineering and Technology**, our principal **Dr. H.V. Vankundre**, our Head of Department **Dr. Megha Trivedi** and our Project Guide **Dr. Megha Trivedi** who gave us this valuable opportunity to develop this course project on the topic: Time Series Analysis on Sales using Random Forest. This project has greatly helped us in expanding our core of knowledge in Machine Learning. It has provided us a precious opportunity to take a hands-on experience and showcase our skills. We are also thankful to each of us because everyone of us aided to complete this project in a limited frame of time.

1. Introduction

1.1 Introduction

- Walmart is an American multinational wholesale retail corporation. Time series forecasting is a critical task for large retail organizations like Walmart. Accurate forecasting helps in inventory management, demand planning, and optimizing supply chain operations. In this report, we will outline the steps and methodologies for time series forecasting for Walmart.

- Time series is a series of data points recorded over even intervals in time. For e.g. Sales records, CPI, Unemploy, Fuel Price and much more. Just seeing the examples, you can also get an understanding of the importance of analysing time series and forecasting (predict) the data.

- This project covers different machine learning models for the forecasting of Time Series Sales Data using different libraries like pandas, sklearn, etc.

1.2 Problem Statement & Objectives

❖ Problem Statement

- Walmart's dataset is extensive, comprising sales data from thousands of stores, various product categories, and diverse regions. Managing and processing this data efficiently is a challenge.

- Ensuring data quality and consistency is vital for accurate forecasting. Missing data, outliers, and data anomalies can significantly impact model performance.

- Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to measure the accuracy of sales forecasts.

- Increased supply chain efficiency and cost savings.

❖ Objectives

The main objective of this project is to develop Time Series Forecasting for Walmart using various Machine Learning models:

- Linear Regression Model
- Random Forest Regression Model
- K Neighbors Regression Model
- XGBoost Regression Model

The ultimate goal is to develop a reliable and scalable time series forecasting model that empowers Walmart to make data-driven decisions, optimize operations, and improve customer experiences while maintaining cost-efficiency.

1.3 Scope

The scope of the "Time Series Forecasting for Walmart" project can be adapted and expanded based on the specific goals, resources, and priorities of the organization:

- Forecast demand for specific products, enabling better procurement and distribution planning. This can lead to cost savings and reduced wastage.

2. Literature Survey

2.1 Survey of Existing System

The existing system for time series analysis on Walmart's sales typically relies on traditional statistical methods like exponential smoothing. While these methods are effective for capturing basic sales patterns, they often struggle with complex seasonality and external factors. Machine learning techniques, such as Random Forest, have gained popularity for their ability to capture more intricate relationships in the data. However, integrating Random Forest into the existing system may require overcoming challenges related to data preprocessing, model tuning, and ensuring that the system can effectively handle Walmart's extensive sales dataset.

2.2 Limitation Existing system

The existing system for time series analysis on sales at Walmart using Random Forest has certain limitations. Firstly, it may not effectively capture complex seasonality and external factors influencing sales. Random Forest, while powerful, might struggle with modeling intricate temporal patterns. Additionally, the model's performance may be impacted by changes in inventory, promotions, and market trends, which it may not adequately account for. The system's scalability and efficiency for processing large volumes of data could also be a concern, given Walmart's vast sales data. Addressing these limitations is crucial for a more accurate and robust sales forecasting system.

2.3 Mini Project Contribution

Time series data, encompassing a variety of domains such as weather, sales figures, economic indicators, and fuel prices, is systematically logged at regular time intervals. This project employs a diverse array of machine learning models, making use of libraries like pandas and scikit-learn (sklearn), to predict future sales data in time series. It underscores the pivotal role of analyzing and projecting these data points to support well-informed decision-making.

The fundamental objective of this project is to develop a robust Time Series Forecasting system for Walmart, leveraging a wide spectrum of Machine Learning models, which includes Linear Regression, Random Forest, K Neighbors, and XGBoost.

The ultimate aim is to create an adaptable and reliable forecasting tool that enhances

Walmart's data-driven decision-making capabilities, streamlines operations, and boosts customer satisfaction while maintaining cost-efficiency

3. Proposed System

3.1 Introduction

- Time series data, such as weather, sales, economic indicators, and fuel prices, is recorded at regular time intervals. This project employs diverse machine learning models, utilizing libraries like pandas, and sklearn, to forecast time series sales data. It underscores the critical role of analyzing and predicting these data points for informed decision-making. [2]

- The core aim of this project is to create a robust Time Series Forecasting system for Walmart, employing a range of Machine Learning models, including Linear Regression, Random Forest, K Neighbors, XGBoost. [1]

- The end goal is an adaptable and dependable forecasting tool that enhances Walmart's data-driven decision-making, streamlines operations, and enhances customer satisfaction without compromising cost-efficiency.

3.2 Block Diagram

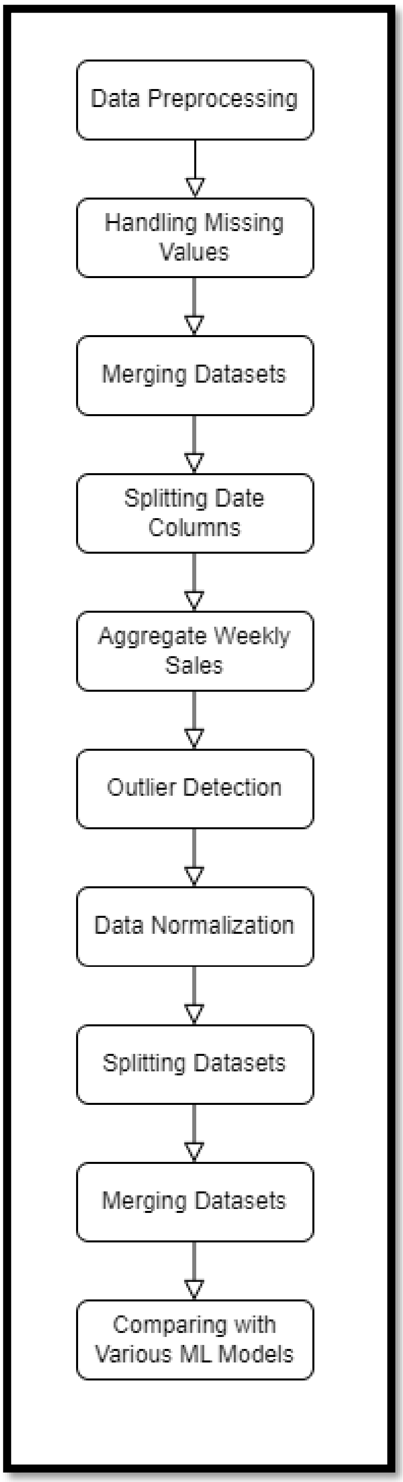

Fig 1: Block Diagram for the proposed model

Data Preprocessing: Prepare data by cleaning, transforming, and scaling it to ensure it's suitable for analysis and modeling.

Handling Missing Values: Deal with missing data through imputation or removal to ensure complete and accurate datasets.

Merging Datasets: Combine data from multiple sources using common identifiers for a comprehensive analysis.

Outlier Detection: Identify and manage unusual data points that can skew forecasts, ensuring better model accuracy.

Comparing Models: Evaluate different forecasting models to select the most accurate and suitable one for time series forecasting.

3.3 Algorithm & Process Design

Machine Learning Models

- Linear Regression Model
- Random Forest Regression Model
- K Neighbors Regression Model
- XGBoost Regression Model

Data Preprocessing

First of all, we have to handle the missing values from the dataset.

Handling Missing Values

CPI, Unemployment of features dataset had 585 null values.

- MarkDown1 had 4158 null values.
- MarkDown2 had 5269 null values.
- MarkDown3 had 4577 null values.
- MarkDown4 had 4726 null values.
- MarkDown5 had 4140 null values.

All missing values were filled using fillna() with the median of respective columns.

Merging Datasets

- Main Dataset merged with stores dataset.
- Resulting Dataset merged with features dataset.
- Total 421570 data rows and 15 attributes.
- Date column converted into the DateTime data type.
- Set Date attribute as the index of the combined dataset.

Splitting Date Column

Using the Date column, three more columns are created Year, Month, Week.

Aggregate Weekly Sales

The median, mean, max, min, std of weekly_sales are calculated and created as different columns.

Outlier Detection and Other abnormalities

- Markdowns were summed into Total_MarkDown.
- Outliers were removed using z-score.
- After outliers removal, 375438 Data rows, and 20 columns.
- Negative weekly sales were removed.
- After removal, 374247 Data rows and 20 columns.

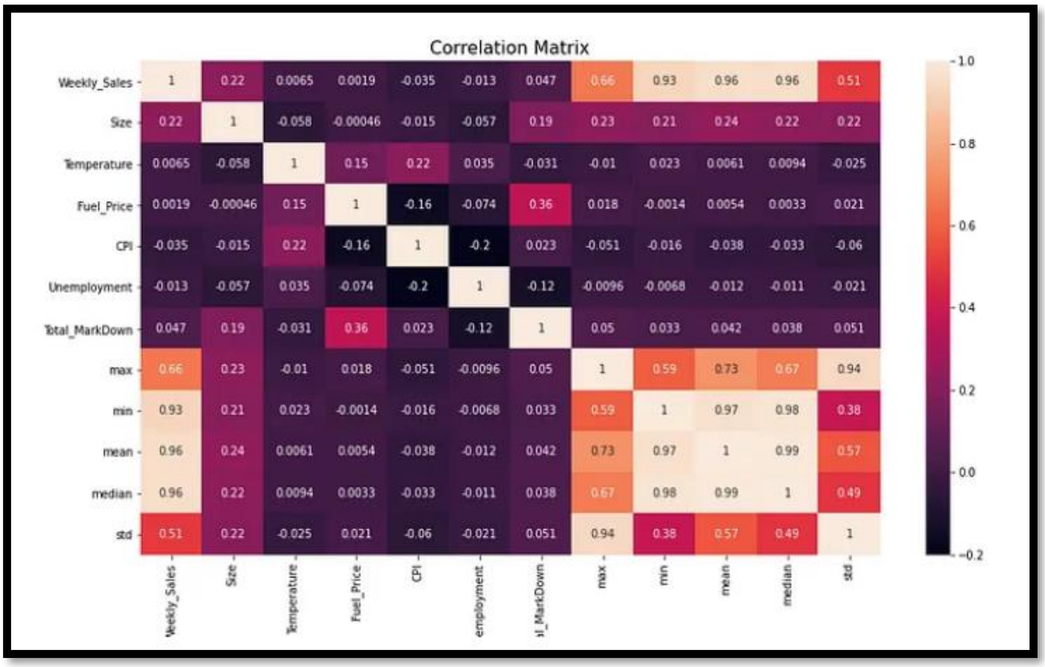

Fig 1. Correlation Matrix represented as Heatmap

We used heatmap to check if there was any collinearity in the dataset according to this dataset we concluded that there was not collinearity so we didn’t drop any rows.

3.4 Details of Hardware & Software

❖ Hardware:

- Ram should be 4gb or more
- Processor should be Intel i3 or above.
- SSD is recommended for faster data access.
- Stable and fast internet connection.

❖ Software:

- System should have following software's and libraries installed:
- Python 3.6 or up
- TensorFlow
- Pandas
- Tkinter
- Sklearn

3.5 Experiment & Results for validation and Verification

Linear Regression

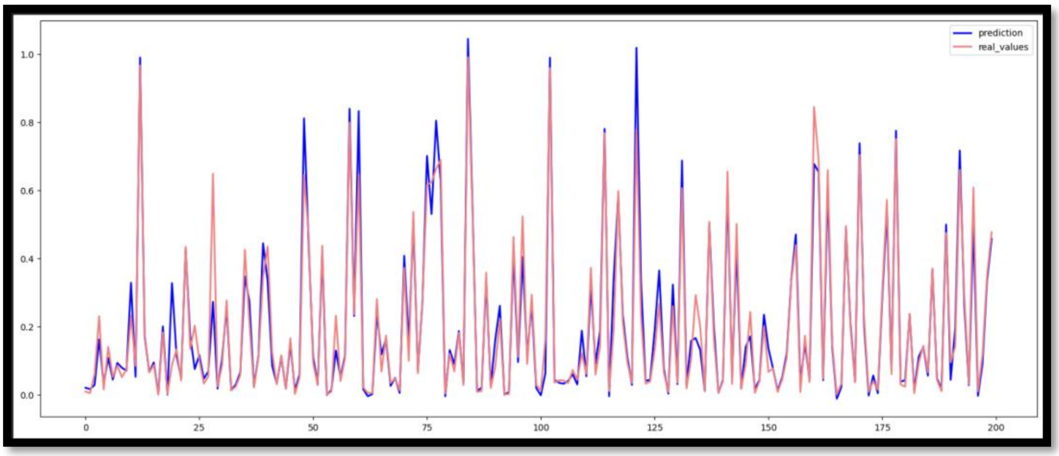

Fig 2. Linear Regression models Prediction vs actual values

Linear Regressor Accuracy - 91.5229
MAE 0.03218626970399393
MSE 0.003815897383902121
RMSE 0.06177295026062881
R2 0.9152297406179852

Random Forest

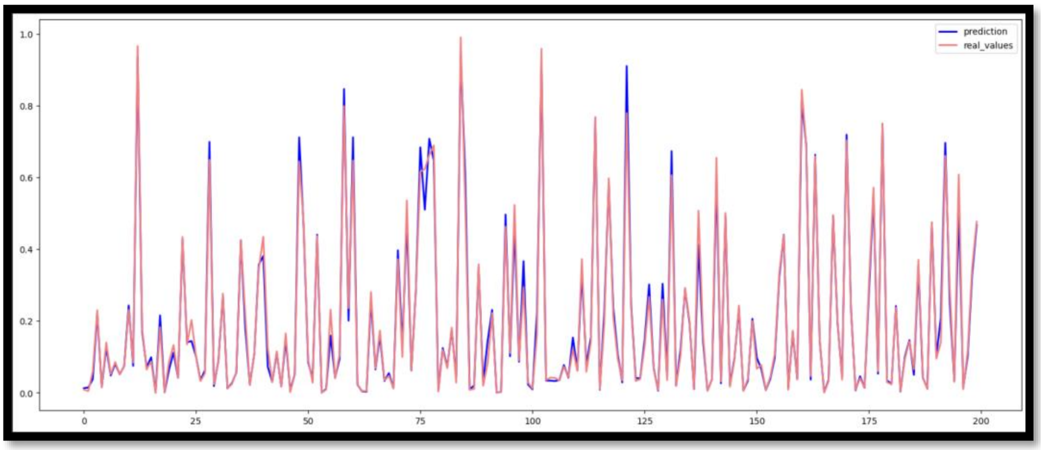

Fig 3. Random Forest models Prediction vs actual values

Random Forest Regressor Accuracy - 97.5484
MAE 0.016598546872087633
MSE 0.0011035547893315492
RMSE 0.03321979514282936
R2 0.9754871330315386

K Neighbors Regressor Model

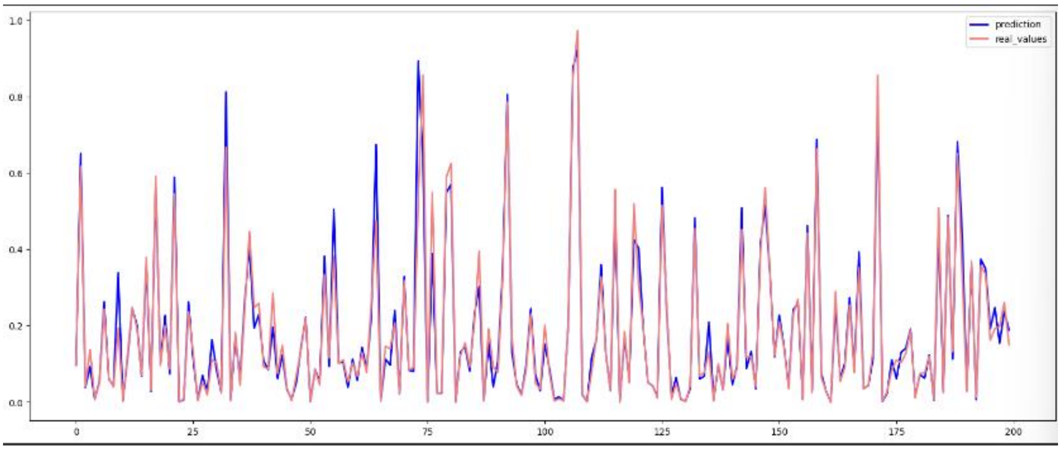

Fig 4. KNN models Prediction vs actual values

KNeigbhbors Regressor Accuracy - 95.2193
MAE 0.022166538871300 51
MSE 0.0021519916995465754
RMSE 0.0463895645543971
R2 0.9522094416852463

XGboost Model

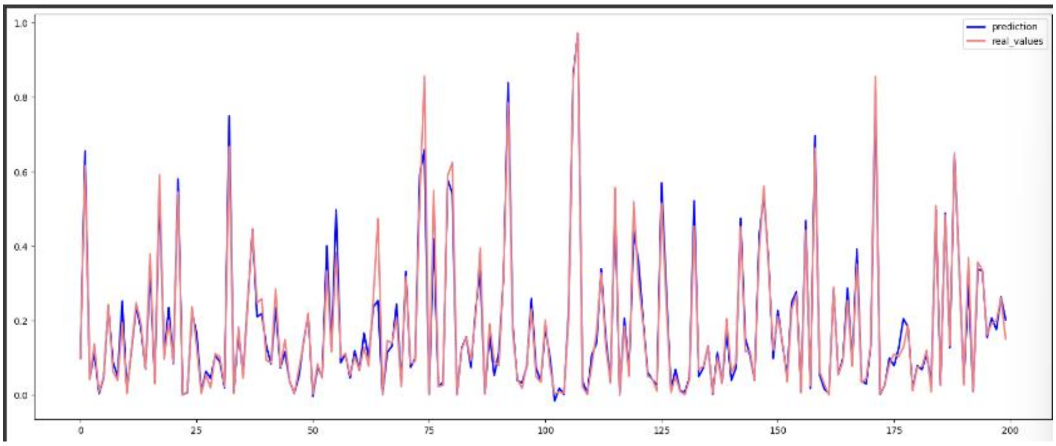

Fig 5. XGboost models Prediction vs actual values

XGBoost Regressor Accuracy - 97.4072
MAE 0.019325806281225714
MSE 0.0011671123631343267
RMSE 0.034163026258432184
R2 0.9740731945232541

Comparing Models

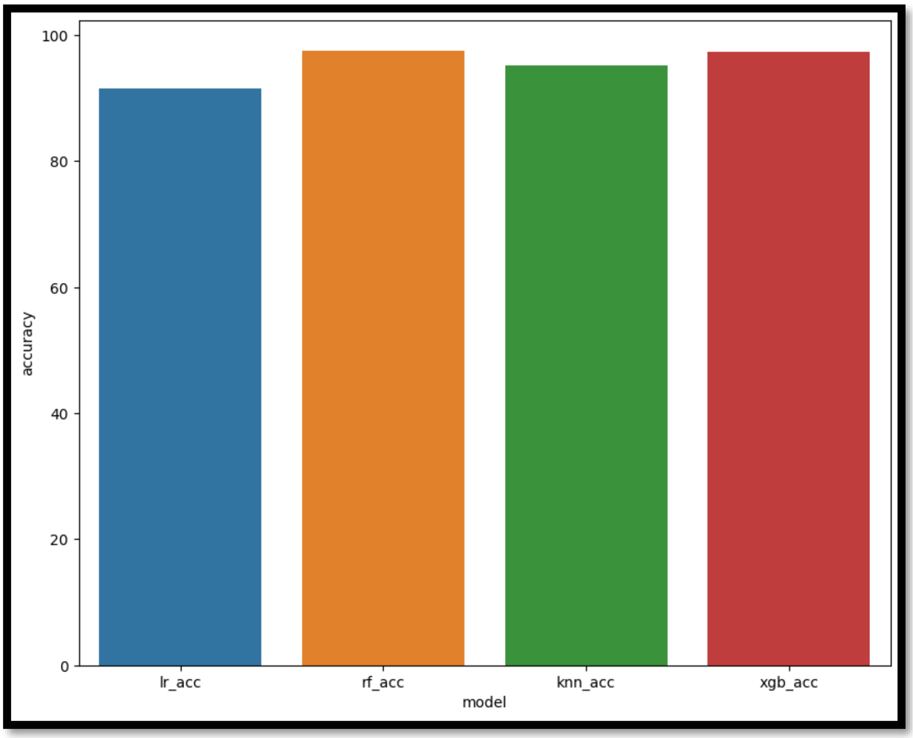

Fig 6. Comparison between models - Random Forest, Linear regression, Knn, XgBoost

Random Forest provides better result than other algorithms.

3.6 Conclusion & Future Work

❖ Conclusion

This machine learning project employing Random Forest for time series analysis on Walmart's sales represents a significant step toward enhancing sales forecasting accuracy. It addresses the limitations of traditional methods and can capture more complex sales patterns. However, the project's success hinges on meticulous data preprocessing, feature selection, and model optimization to maximize Random Forest's potential.

❖ Future Work

For future work, expanding the scope to consider external variables, like economic indicators or local events, could improve forecasting accuracy. Additionally, leveraging deep learning models or hybrid approaches may provide even more robust predictions. The project's scalability to handle Walmart's extensive data and real-time updates is another avenue for further exploration, ensuring that the system remains efficient and reliable in a dynamic retail environment.

❖ References

[1]. A comprehensive survey on sales forecasting models using machine learning *algorithms*. (2022, December 26). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10060168

[2].Time Series Analysis sales of sowing crops based on machine learning methods. (2018, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8633610