PROJECT REPORT

GROUP MEMBERS

Neha Afaque Khan	(22F-Bsai-56)
Sabeeh ur Rehman Khan	(22F-Bsai-64)
Taha Saeed	(22F-Bsai-83)
Saad bin Haseeb	(22F-Bsai-87)

—

COURSE TITLE

Machine Learning

—

TEACHER NAME

Sir Hamza Farooqui

Topic: Sales Forecasting and Anomaly Detection

Dataset

- Walmart sales dataset
- Dataset size: **6,000 records**
- Target Feature: **Weekly_Sales**
- No missing values

Methodology

Overall Workflow:

The project uses two parallel pipelines with shared preprocessing, EDA, and distinct models.

i) Sales Forecasting Methodology

- Date features were transformed, and new features such as lag and rolling features were engineered to help the model learn temporal patterns.
- Rows with missing values from lag and rolling features were removed.
- Data was split into **80%** training and **20%** testing sets.
- Feature scaling was applied for the **linear model**.
- **Linear Regression** and **Random Forest** models were used for forecasting.

ii) Anomaly Detection Methodology

- Feature engineering was similar to sales forecasting, with additional features for anomaly detection: Z-score, sales deviation, and week-over-week (WoW) change.
- **K-Means** and **Isolation Forest** models were used for anomaly detection.
- Feature scaling was applied for K-Means.
- **Hierarchical clustering** technique was applied for exploratory analysis to understand data structure.
- The **optimal number of clusters for K-Means** was determined using the **elbow method**.

Introduction

This project focuses on building a Sales Forecasting and Anomaly Detection system using Machine Learning techniques. The system predicts future sales based on historical data and identifies unusual or abnormal sales patterns. A user-friendly interface is developed to perform live anomaly detection, where users can enter records and instantly view results.

Metrics

- RMSE
- MAE
- MAPE
- R^2

Results

- Linear Regression: **97% training, 96% testing** accuracy.
- Random Forest: **99% training, 97% testing** accuracy.
- Random Forest **slightly outperformed** Linear Regression.
- Elbow method determined **4** as the optimal K-Means clusters.
- Both K-Means and Isolation Forest detected **318 anomalies (5% of data).**

Conclusion

The project successfully detects anomalies in sales data and predicts future sales based on past data. Visualizations were created to better interpret the results, and a user interface (UI) was developed to enable real-time anomaly detection as new records are entered, making monitoring and analysis more interactive and effective.