RETAIL SALES FORECASTING ANALYSIS

LEVERAGING MACHINE LEARNING TO PREDICT WEEKLY SALES

By : Abhishek Kumar & Team

Date=05-05-2025

PROJECT OVERVIEW

Objective:

Forecast weekly sales and identify key drivers (e.g., promotions, holidays, economic factors).

Key Questions:

- What factors drive sales?
- How do markdowns impact sales?
- Can historical data predict future sales accurately?

Success Metrics:

- R^2 score
- Actionable business insights

DATA SOURCES

Datasets Used:

1. **Features.csv**
   - Environmental/economic data (Temperature, Fuel Price, CPI, Unemployment, MarkDowns, IsHoliday).
   - Size: 8,190 rows × 12 columns.
2. **Stores.csv**
   - Store characteristics (Type, Size).
   - Size: 45 rows × 3 columns.
3. **Train.csv**
   - Historical sales data (Store, Dept, Weekly_Sales, Date).
   - Size: 421,570 rows × 5 columns.

Merged Dataset:

- Final size: 421,570 rows × 16 columns.

DATA CLEANING & PREPROCESSING

Steps Taken:

1. Handled Missing Values:
    - Filled missing MarkDown values with 0 (no discount applied).
2. Data Transformation:
    - Log-transformed Weekly_Sales and MarkDown columns to reduce skewness.
3. Outlier Treatment:
    - Clipped extreme values in Unemployment using IQR.
4. Feature Engineering:
    - Extracted date components (Year, Month, Week, Day).

Before vs After:

- Skewness reduced from 3.26 to near-normal for Weekly_Sales.

EXPLORATORY DATA ANALYSIS (EDA)

Key Insights:

1. Sales Distribution:
   - Right-skewed; most sales are low, with few high outliers.
2. Store Types:
   - Type A stores dominate and have higher sales (likely due to larger size).
1. Holiday Impact:
   - No significant sales boost during holidays.
2. Correlations:
   - Positive: Size, MarkDowns.
   - Negative: Unemployment, CPI.

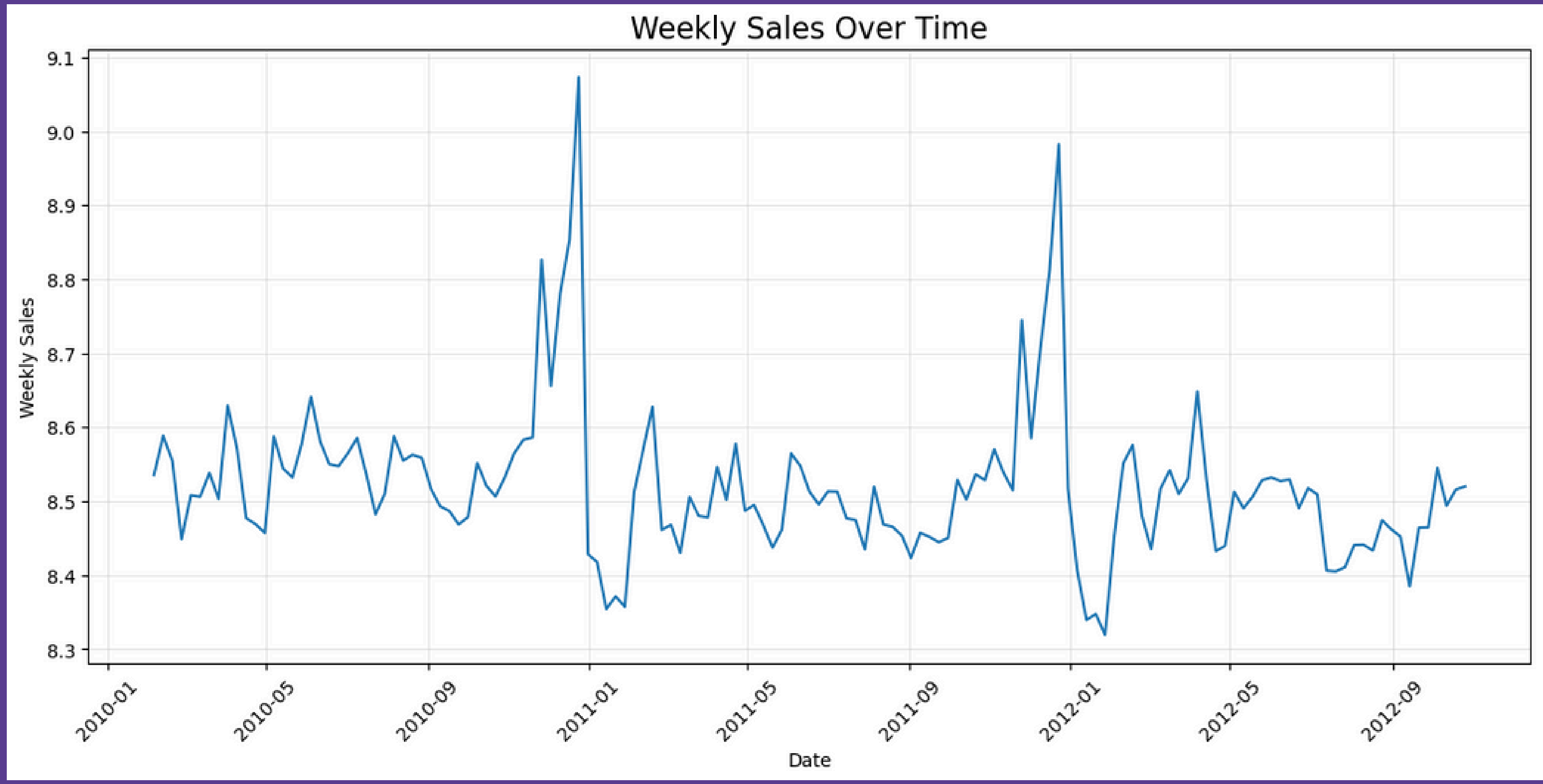
FEATURE SELECTION & VIF

Multicollinearity Check:

- Dropped high-VIF features (Month, Week, Day, MarkDown1, MarkDown4).
- Final VIF < 5 for all retained features.

Retained Features:

- Store, Dept, IsHoliday, Temperature, Fuel_Price, MarkDown5, CPI, Unemployment, Type, Size, Year.



MODEL BUILDING

Algorithm: Random Forest Regressor

Hyperparameters:

- n_estimators=40, max_depth=10.

Performance Metrics:

- Train R^2: 0.89
- Test R^2: 0.82
- Interpretation: Model generalizes well with minimal overfitting.

FORECASTING EXAMPLE

Sample Prediction:
Predicted Sales for custom data: [6543.21]
Input Features:

- Store=1, Dept=1, IsHoliday=0, Temperature=50°F, Fuel_Price=3$, MarkDown5=0`.

CONCLUSION & NEXT STEPS

Summary:

- Achieved 82% accuracy in forecasting sales using Random Forest.
- Identified key drivers: store size, markdowns, and economic indicators.

Next Steps:

1. Incorporate external data (e.g., local events, competitor pricing).
2. Test advanced models
3. Deploy as a dashboard for real-time decision-making.

Q&A

" **Thank you** for your time and attention! "