Walmart Store Sales Forecasting 2010 to 2012

Department store sales trends and analysis

Juyang Hu

OVERVIEW

This project is to uncover trends and patterns in the historical weekly sales data from 45 Walmart stores in Texas, Wisconsin and California from 2010 to 2012. I examined relationships between sales and Markdown events, holiday and temperature.

DATA DESCRIPTION

This compiled dataset was pulled from three other datasets linked by Date and Store ID, and was built to explore what other factors, other than the seasonal variations, contribute to store sales.

The datasets include the following information that we used to analyze factors related to store sales:

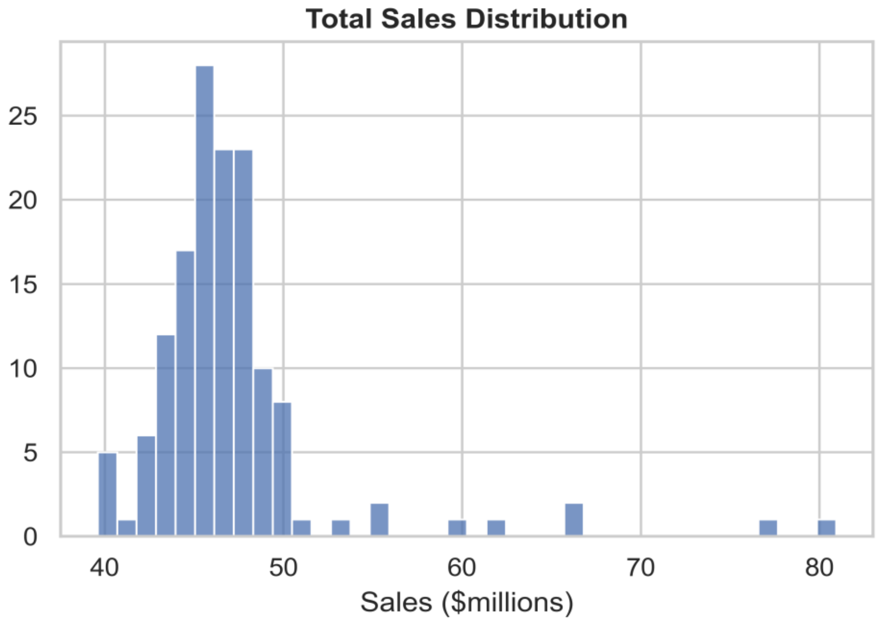
- Date
- Promotion Type
- Economic Data
- Store Type/Size
- Temperature

DATA WRANGLING

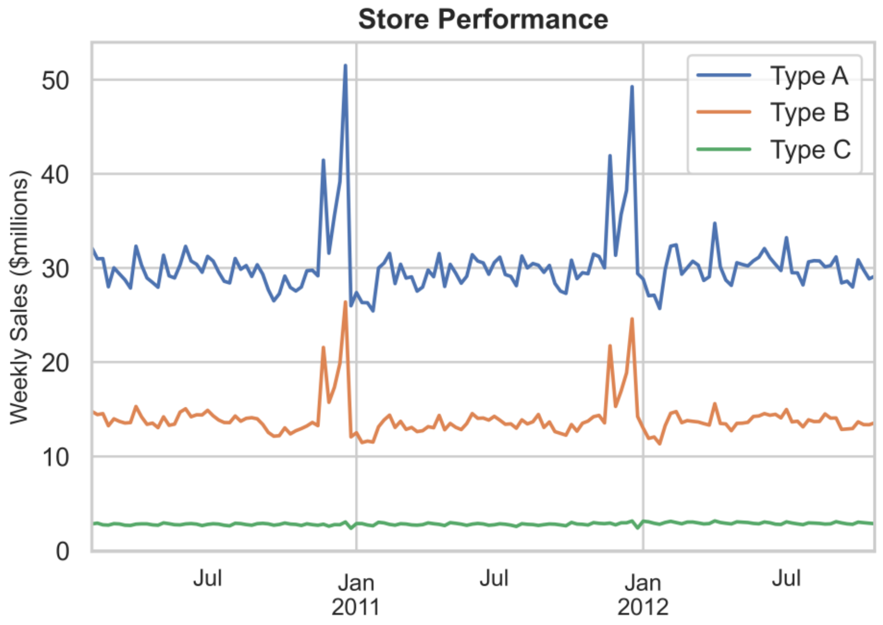
Temperature °F	Sales Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
39.93	✓					
46.63			✓			
54.58				✓		
36.39						
67.41	✓					✓
72.55					✓	
80.44		✓				
78.69					✓	
82.11		✓				
59.61				✓		
49.27			✓			

- Dropped unnecessary columns in the DataFrame
- Set up several bins (groups) for analysis
- Transformed categorical data to numeric data
- Decomposed date into greater granularity: Year, Month and Day for better holiday effect mapping

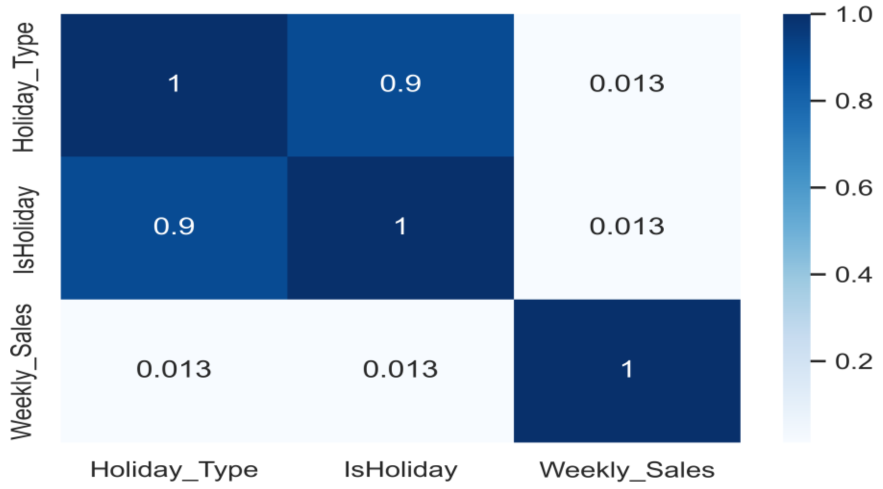
Exploratory Data Analysis (EDA)



The majority of weekly sales fall between 40 to 50 million



Type C stores shows no apparent seasonality or holiday effects



Breaking down holiday to specific events does not yield a higher correlation to sales than a binary holiday indicator

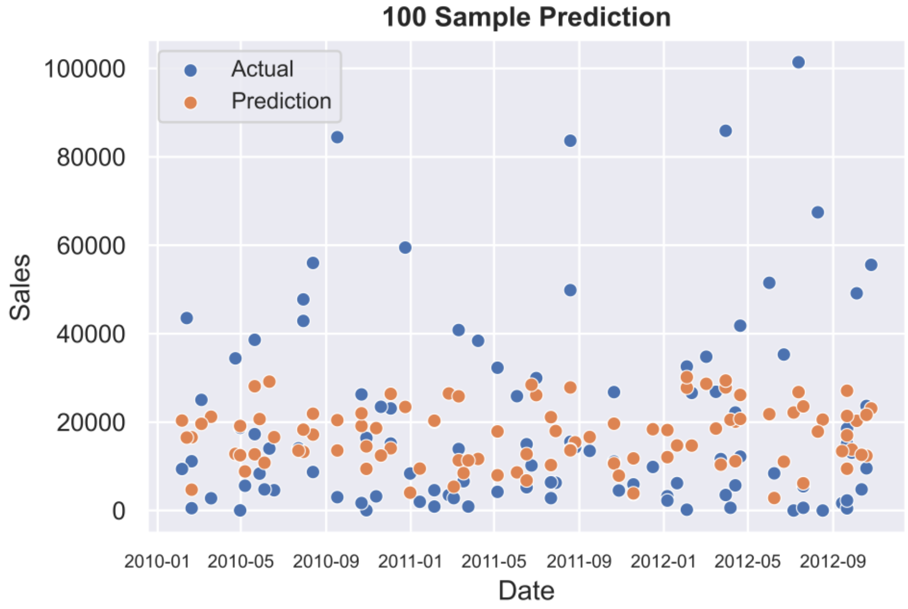
Benchmark Model Performance

The Linear Regression Model is used to serve as the benchmark model. A total of 15 variables are used to predict weekly sales. The variables are as follow:

['Store', 'Dept', 'Date', 'IsHoliday', 'Temperature', 'Fuel Price', 'MarkDown1', 'MarkDown2', 'MarkDown3', 'MarkDown4', 'MarkDown5', 'CPI', 'Unemployment', 'Type', 'Size']

- The model has a R-squared of 0.062
- The WMAE (Weighted Mean Absolute Error) is **14756**

This means the benchmark Linear Regression Model has a rather weak predictive power



- The Linear Regression Model made poor prediction especially during holidays when sales were at their peak level
- The variables do not seem to have a linear relationship with sales

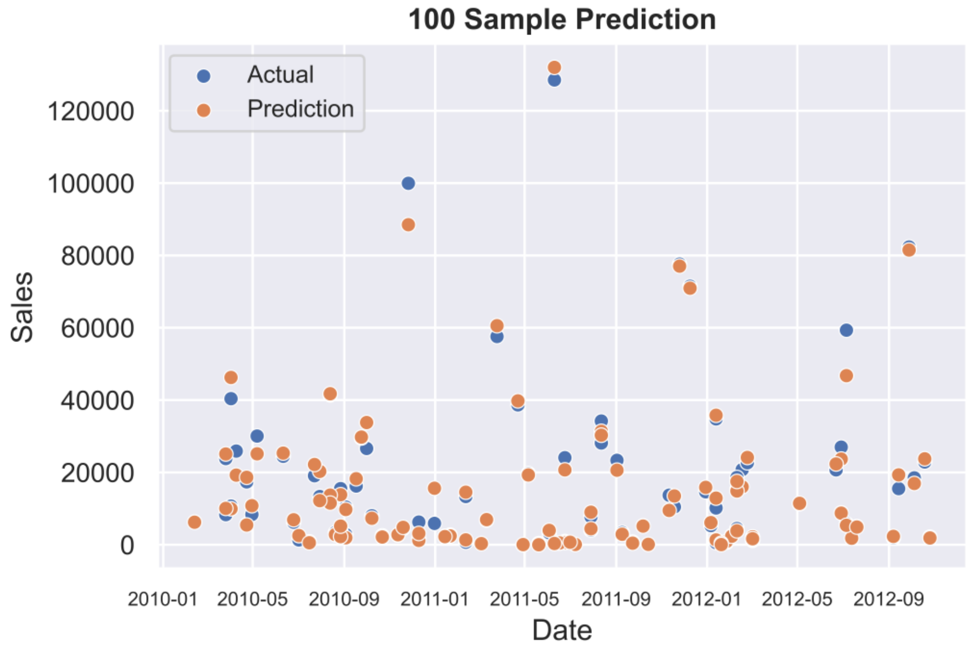
Random Forest Model

The Random Forest Model is used to compare with the benchmark model. Similarly, the same set of 15 variables was fed.

2/3 of the data were used to train the model, and 1/3 were used to test the model

- The WMAE (Weighted Mean Absolute Error) is **2031**

Prediction Result from random 100 samples



- The Random Forest Model shows a significant improvement in prediction accuracy
- The model handles holiday season peak sales prediction fairly well
- **Feature Importance Rank**

Department	0.62
Store Size	0.19
Store Type	0.05
CPI	0.01
Date	0.01