Analyzing Walmart Sales

Data Science Programming

How do outside factors like unemployment and weather impact Walmart's sales?

Our Team

Sarah Stephens

Brooks Li

Alex Imhoff

John Izzo

Sidharth Saha

Agenda

What we'll discuss today

01 Background & EDA

02 Feature Engineering & Data Splitting

03 Our Models

04 Analysis/Takeaways

05 Conclusion

Background

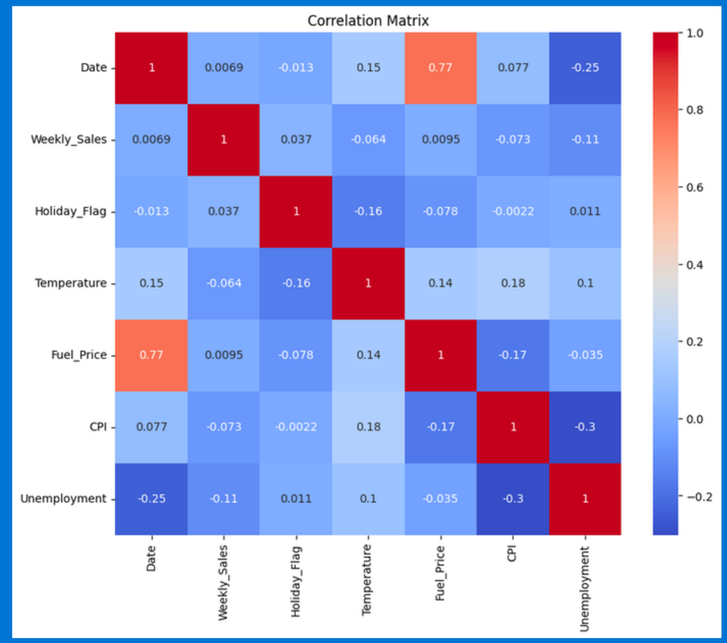
- **Dataset Information:** "Walmart Dataset (Retail)" by Rutu Patel from Kaggle
- **Variables Include:** Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment
- **Research Objective:** Identify patterns in Walmart’s weekly sales influenced by factors like weather and national economic conditions.
- **Store-Specific Analysis:** Explore patterns specific to different store locations based on the stratified data.

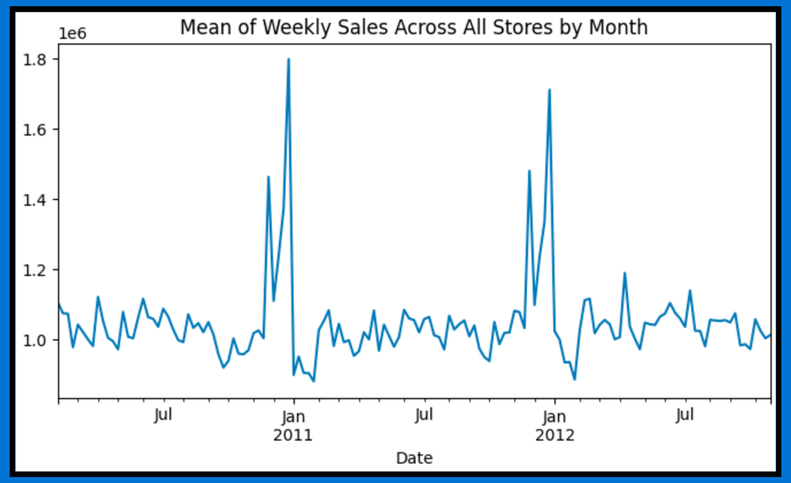
https://www.kaggle.com/datasets/rutuspatel/walmart-dataset-retail

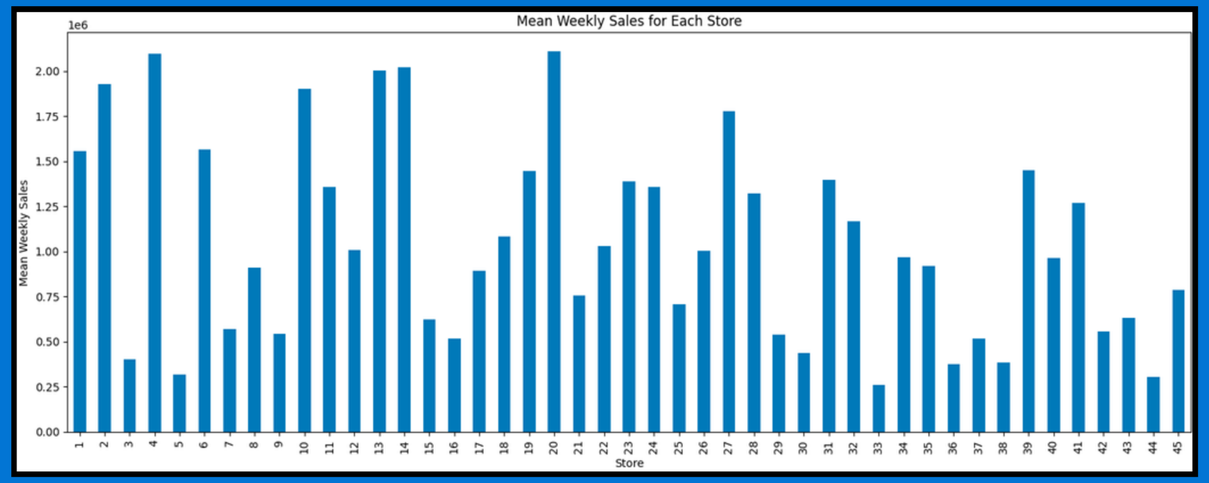
	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
unique	143	NaN	NaN	NaN	NaN	NaN	NaN
top	5/2/2010	NaN	NaN	NaN	NaN	NaN	NaN
freq	45	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	NaN	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	NaN	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	NaN	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	NaN	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	NaN	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	NaN	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

Exploratory Data Analysis

- Seasonality (holiday spikes)
- Significant variation in weekly sales depending on store number
- Correlations
  - Fuel price & date
  - Unemployment & date
  - CPI & unemployement







Feature Engineering & Data Splitting

1

Distance variables

Difference between individual store observations & store averages

Difference between individual store observations & national averages

2

Log Terms

Transformed weekly sales response variables into log terms to reduce skewness given wide range

3

One-Hot Encoding

Converted each unique store from a categorical variable into a dummy variable

4

Test-Train Split

Applied 80-20 split

Resulted in MSE in billions...

Reevaluate!

5

Revise Train/Test Split

Recoded train/set split to ensure inclusion of all store numbers in test set that were represented in train set

Multiple Regression

OLS Regression Results

Dep. Variable:	Weekly_Sales	R-squared:	0.958
Model:	OLS	Adj. R-squared:	0.958
Method:	Least Squares	F-statistic:	2367.
Date:	Sat, 10 Aug 2024	Prob (F-statistic):	0.00
Time:	16:21:03	Log-Likelihood:	3598.1
No. Observations:	5112	AIC:	-7096.
Df Residuals:	5062	BIC:	-6769.
Df Model:	49		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5503	0.002	709.394	0.000	1.546	1.555
Holiday_Flag	0.0573	0.007	8.540	0.000	0.044	0.070
Temperature	0.0027	0.000	13.530	0.000	0.002	0.003
Fuel_Price	2.9887	0.003	995.883	0.000	2.983	2.995
CPI	0.0107	4.54e-05	235.082	0.000	0.011	0.011
Unemployment	0.0105	0.001	9.804	0.000	0.008	0.013
temp_distance	-0.0029	0.000	-13.075	0.000	-0.003	-0.003
fuel_price_distance	-1.5088	0.003	-475.468	0.000	-1.515	-1.503
cpi_distance	-0.0041	0.000	-9.506	0.000	-0.005	-0.003
unemployment_distance	-0.0139	0.002	-7.312	0.000	-0.018	-0.010
fuel_price_distance_national_avg	-1.5088	0.003	-475.468	0.000	-1.515	-1.503
cpi_distance_national_avg	-0.0041	0.000	-9.506	0.000	-0.005	-0.003
unemployment_distance_national_avg	-0.0139	0.002	-7.312	0.000	-0.018	-0.010
Store_1	0.5067	0.011	44.710	0.000	0.484	0.529
Store_2	0.7257	0.011	64.583	0.000	0.704	0.748
Store_3	-0.8815	0.011	-79.179	0.000	-0.903	-0.860

Root Mean Squared Error: 184,783
R-squared: 0.90

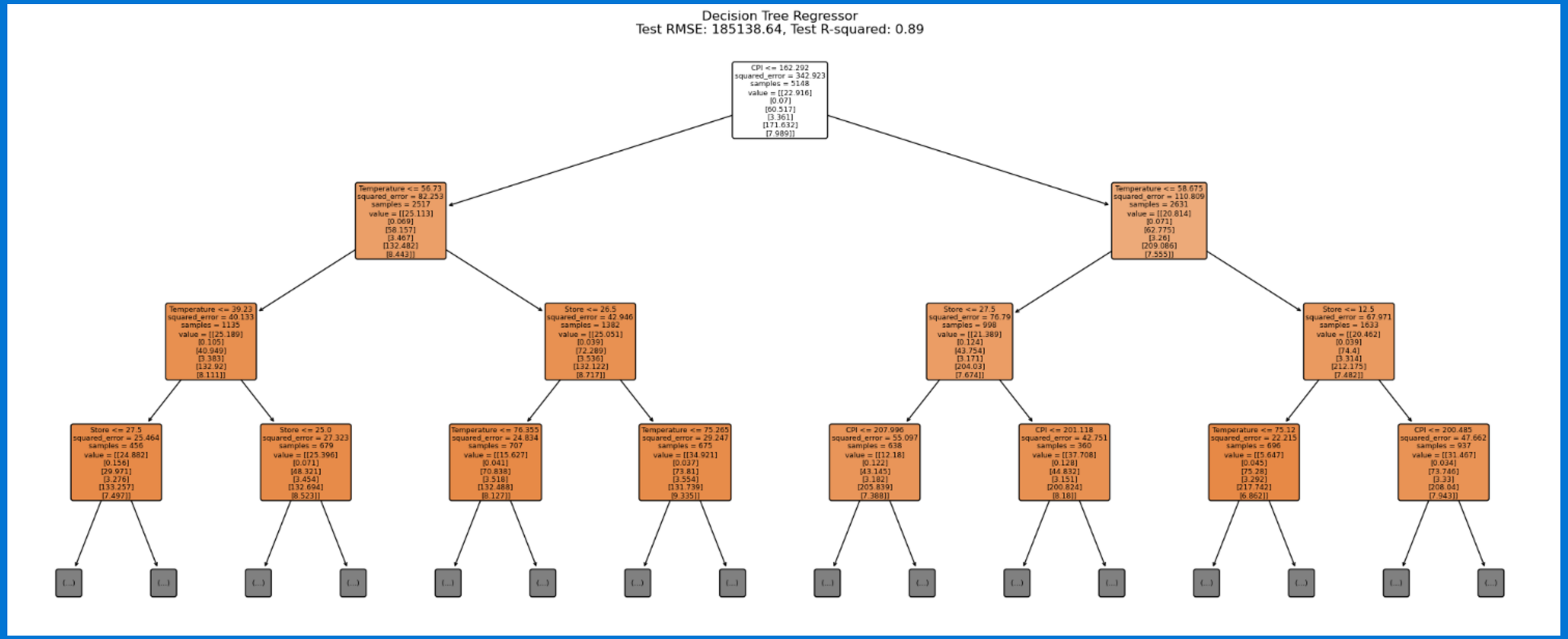
90% of variability in weekly sales is explained by the predictors

- We also tried LASSO, but gave worse result. So, we decided to leave that model out.

Regression Tree

Root Mean Squared Error: 185,139
R-squared: 0.89

89% of variability in weekly sales is explained by the predictors


Decision Tree Regressor
Test RMSE: 185138.64, Test R-squared: 0.89
CPI <= 162.292
squared_error = 342.923
samples = 5148
value = [[22.916]
[0.07]
[60.517]
[3.361]
[171.632]
[7.989]]
Temperature <= 56.73
squared_error = 82.253
samples = 2517
value = [[25.113]
[0.069]
[58.157]
[3.467]
[132.482]
[8.443]]
Temperature <= 58.675
squared_error = 110.809
samples = 2631
value = [[20.814]
[0.071]
[62.775]
[3.26]
[209.086]
[7.555]]
Temperature <= 39.23
squared_error = 40.133
samples = 1135
value = [[25.189]
[0.105]
[40.949]
[3.383]
[132.92]
[8.111]]
Store <= 26.5
squared_error = 42.946
samples = 1382
value = [[25.051]
[0.039]
[72.289]
[3.536]
[132.122]
[8.717]]
Store <= 27.5
squared_error = 76.79
samples = 998
value = [[21.389]
[0.124]
[43.754]
[3.171]
[204.03]
[7.674]]
Store <= 12.5
squared_error = 67.971
samples = 1633
value = [[20.462]
[0.039]
[74.4]
[3.314]
[212.175]
[7.482]]
Store <= 27.5
squared_error = 25.464
samples = 456
value = [[24.882]
[0.156]
[29.971]
[3.276]
[133.257]
[7.497]]
Store <= 25.0
squared_error = 27.323
samples = 679
value = [[25.396]
[0.071]
[48.321]
[3.454]
[132.694]
[8.523]]
Temperature <= 76.355
squared_error = 24.834
samples = 707
value = [[15.627]
[0.041]
[70.838]
[3.518]
[132.488]
[8.127]]
Temperature <= 75.265
squared_error = 29.247
samples = 675
value = [[34.921]
[0.037]
[73.81]
[3.554]
[131.739]
[9.335]]
CPI <= 207.996
squared_error = 55.097
samples = 638
value = [[12.18]
[0.122]
[43.145]
[3.182]
[205.839]
[7.388]]
CPI <= 201.118
squared_error = 42.751
samples = 360
value = [[37.708]
[0.128]
[44.832]
[3.151]
[200.824]
[8.18]]
Temperature <= 75.12
squared_error = 22.215
samples = 696
value = [[5.647]
[0.045]
[75.28]
[3.292]
[217.742]
[6.862]]
CPI <= 200.485
squared_error = 47.662
samples = 937
value = [[31.467]
[0.034]
[73.746]
[3.33]
[208.04]
[7.943]]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Random Forest

MAE: 113,643

RMSE: 196,575

R-squared: 0.89

89% of variability in weekly sales is explained by the predictors

Gradient Boosting

```
param_grid = {
    'learning_rate': [0.01, 0.05, 0.1],
    'n_estimators': [100, 300, 500],
    'subsample': [0.8, 1.0],
    'max_depth': [3, 5, 7],
    'alpha': [0.1, 0.2, 0.3]
}
```

Root Mean Squared Error: 130,876
R-squared: 0.94

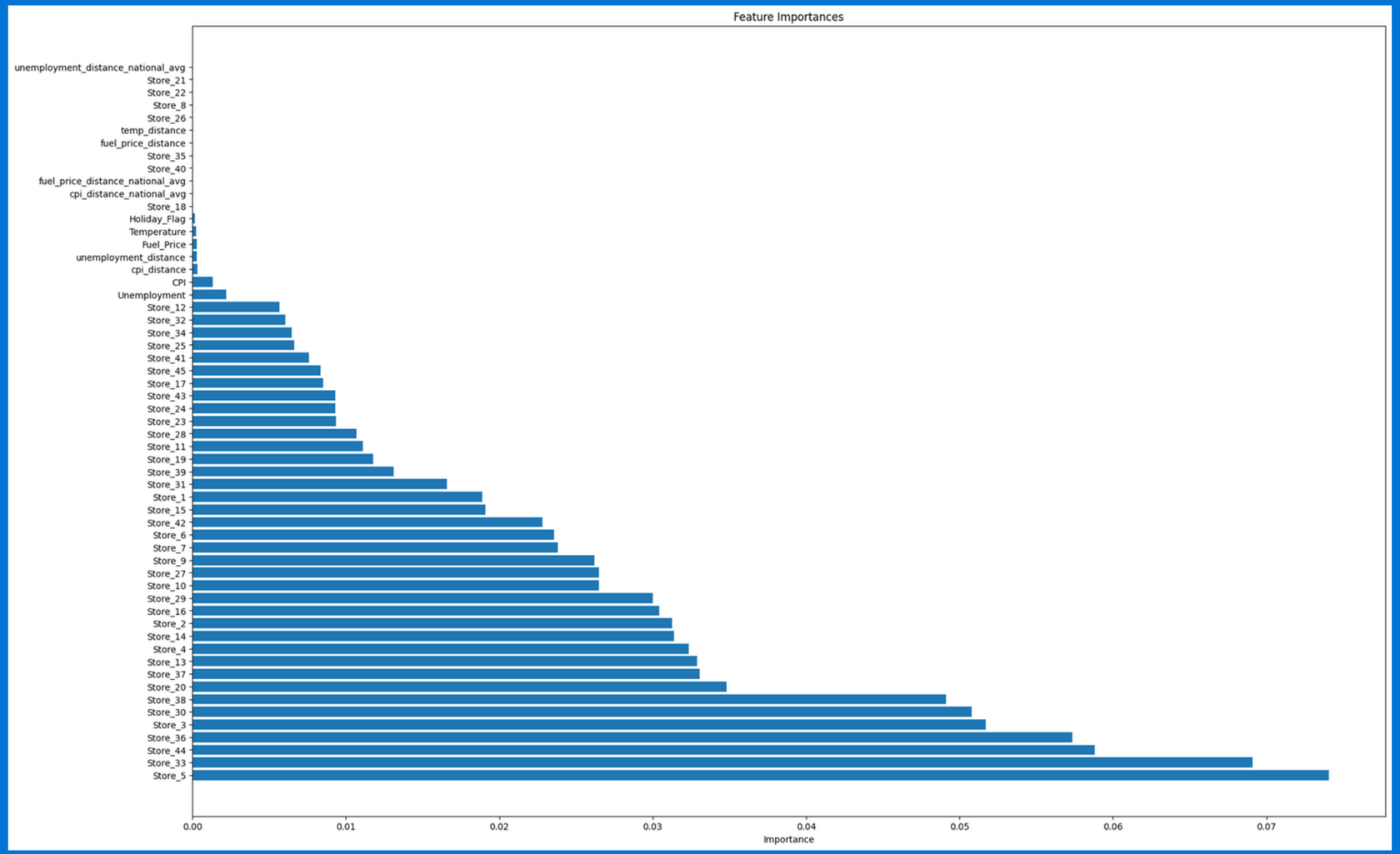
94% of variability in weekly sales is explained by the predictors

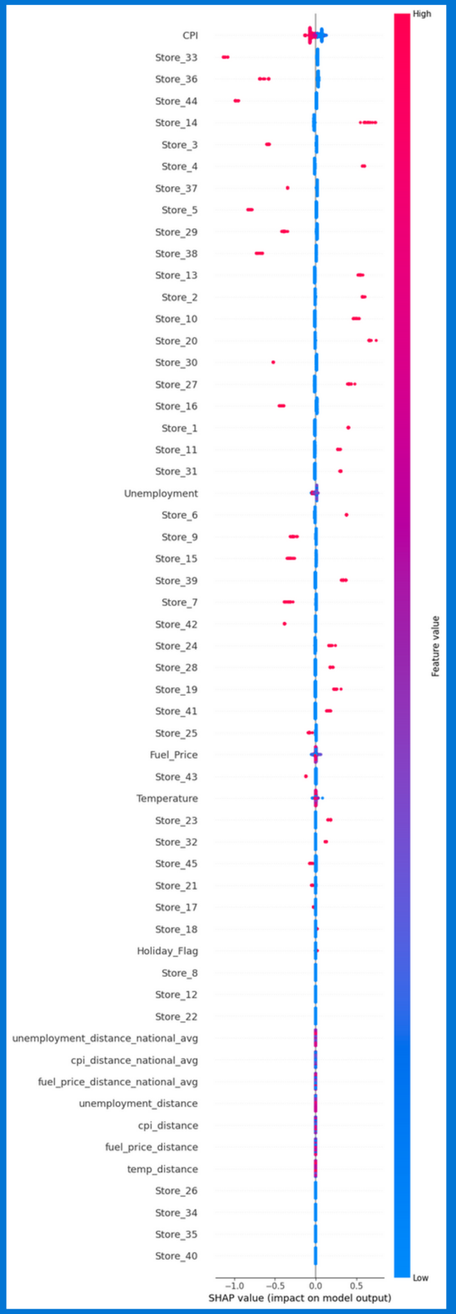
Xgboost

Root Mean Squared Error: 222,087
R-squared: 0.86

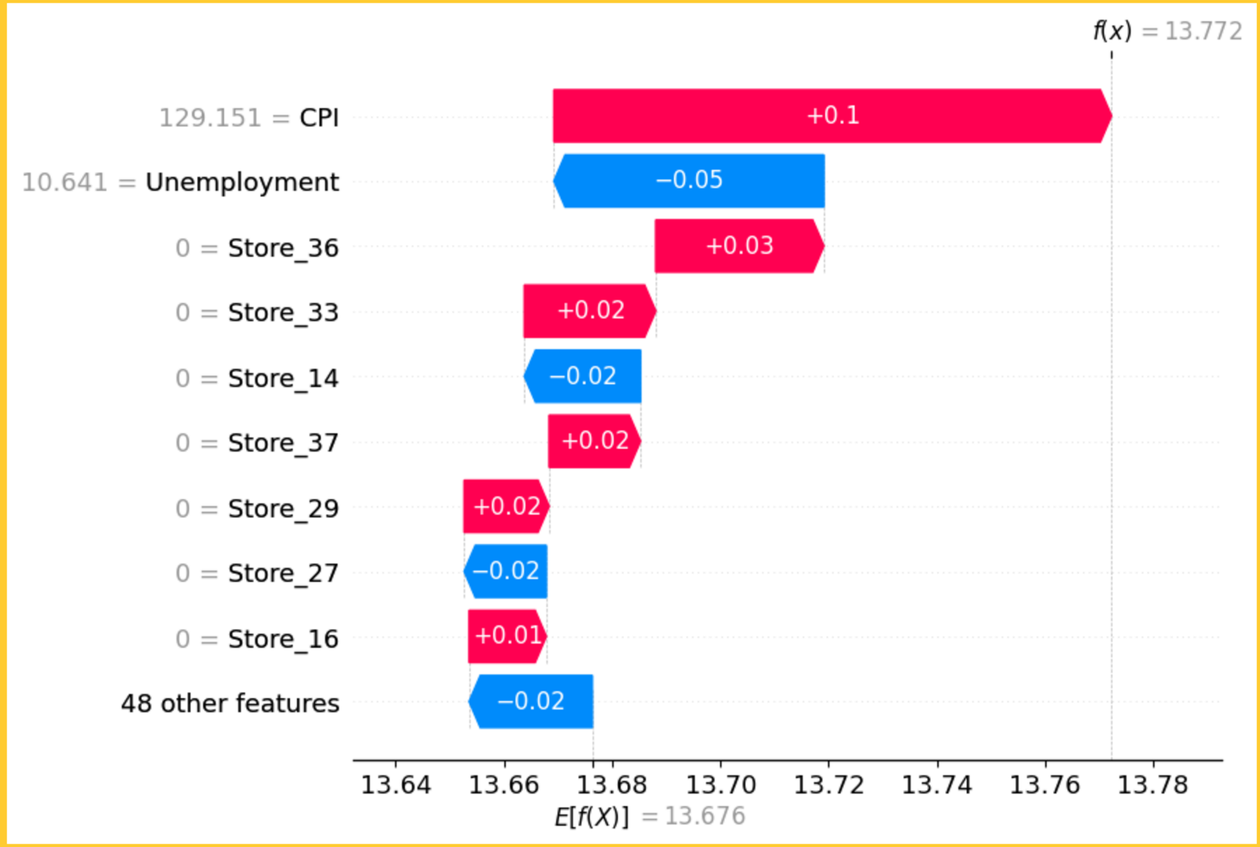
86% of variability in weekly sales is explained by the predictors

Feature Importance

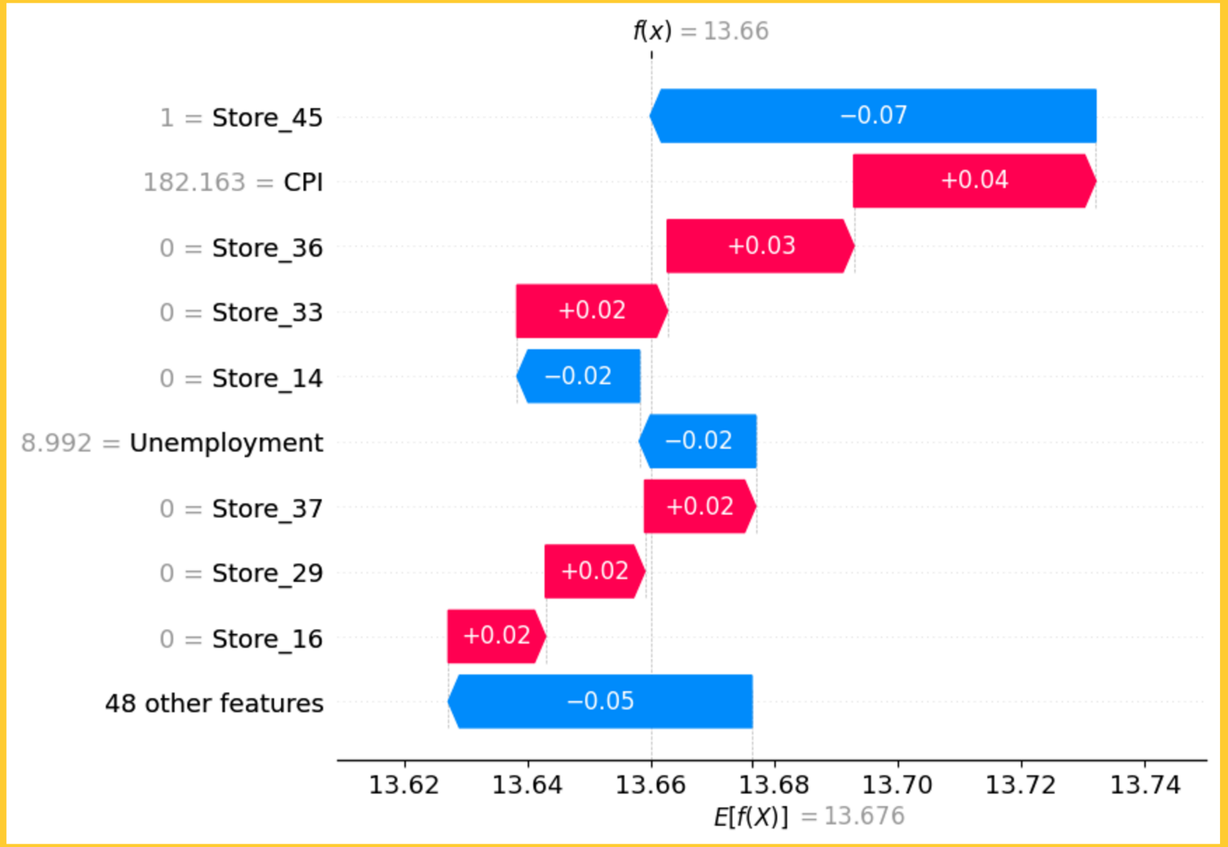




Shapley Waterfall Plots



Store 34, 7/22/11



Store 45, 3/12/10

Model Type	RMSE ($)	R-squared
Multiple Regression	184,783	0.90
Regression Tree	185,139	0.89
Random Forest	196,575	0.89
Gradient Boosting	**130,876**	**0.94**
Xgboost	222,087	0.86

Takeaways

- **Best Model:** Gradient boosting
- **Key Finding:** All models heavily relied on the store number
- **Implication:** External factors related to unique store locations, such as population density and median income, might better capture the underlying drivers of Walmart sales.
- **Important Metrics:** CPI and unemployment were identified as the most significant factors beyond store numbers, directly influencing consumer behavior and spending power.
- **Recommendation:** Walmart should prioritize hyperlocal trends over national macroeconomic measures to improve accurately capture the factors driving sales, considering the significant variation across different stores

THANK YOU