Artificial Intelligence Bootcamp: Project 2 - Group 3

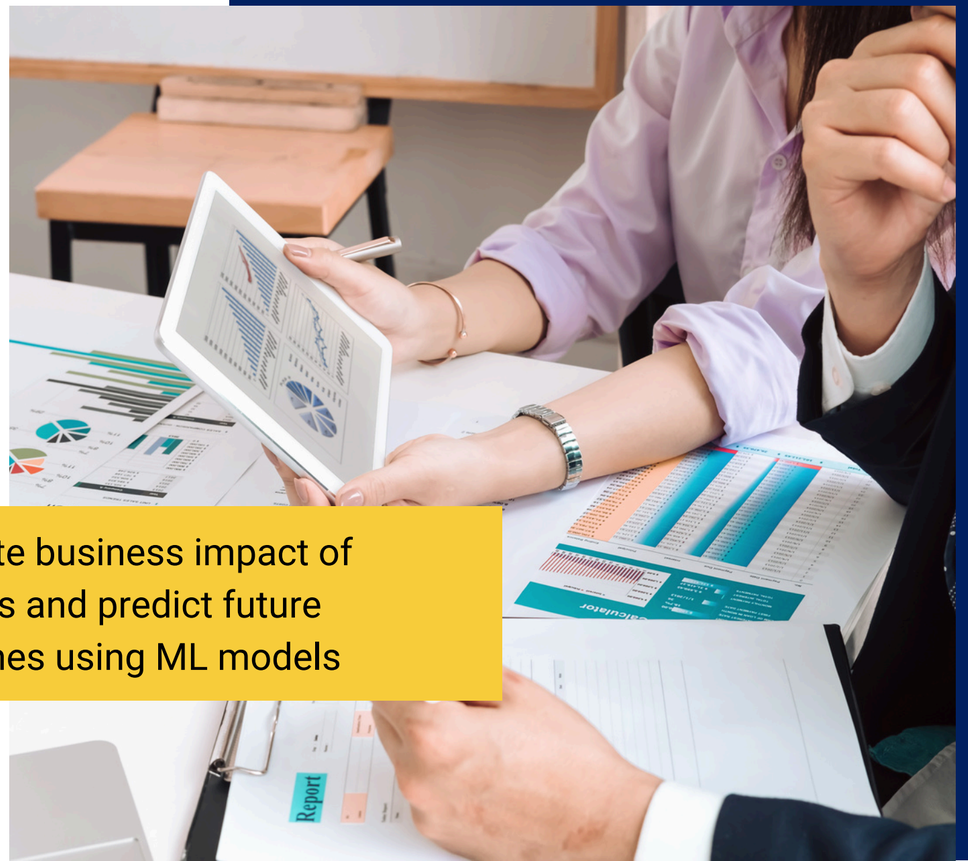
WALMART ANALYSIS

William Aromando, Dan Becker, Joanne Donohue, Vivin Rajagopalan, and Aaron Swan

UNDERSTANDING OUR BUSINESS VIA STORE SALES DATA & LOCAL ECONOMIC INDICATORS

Source: https://www.kaggle.com/datasets/yasserh/walmart-dataset/data

- 45 Stores, weekly sales
- Feb 5, 2010 to Oct 25, 2012
- Holiday Flag (Super Bowl, Labor Day,
- Thx Giving and Xmas only)
- Local CPI, Fuel Price, Temp, Unemploy %



Explore and Analyze data set w/ Pandas

Estimate business impact of features and predict future outcomes using ML models

Four diverse ways to analyze and action this data set, varying accuracy rates

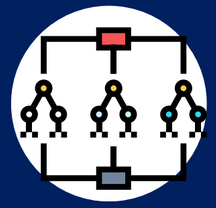
Through various machine learning techniques and visualization platforms, we were able to action the Walmart store level sales data to understand our business and make predictions to drive sales

PANDAS,SKLEARN

Analysis

KMEANS CLUSTERING

Unsupervised



RANDOM FOREST

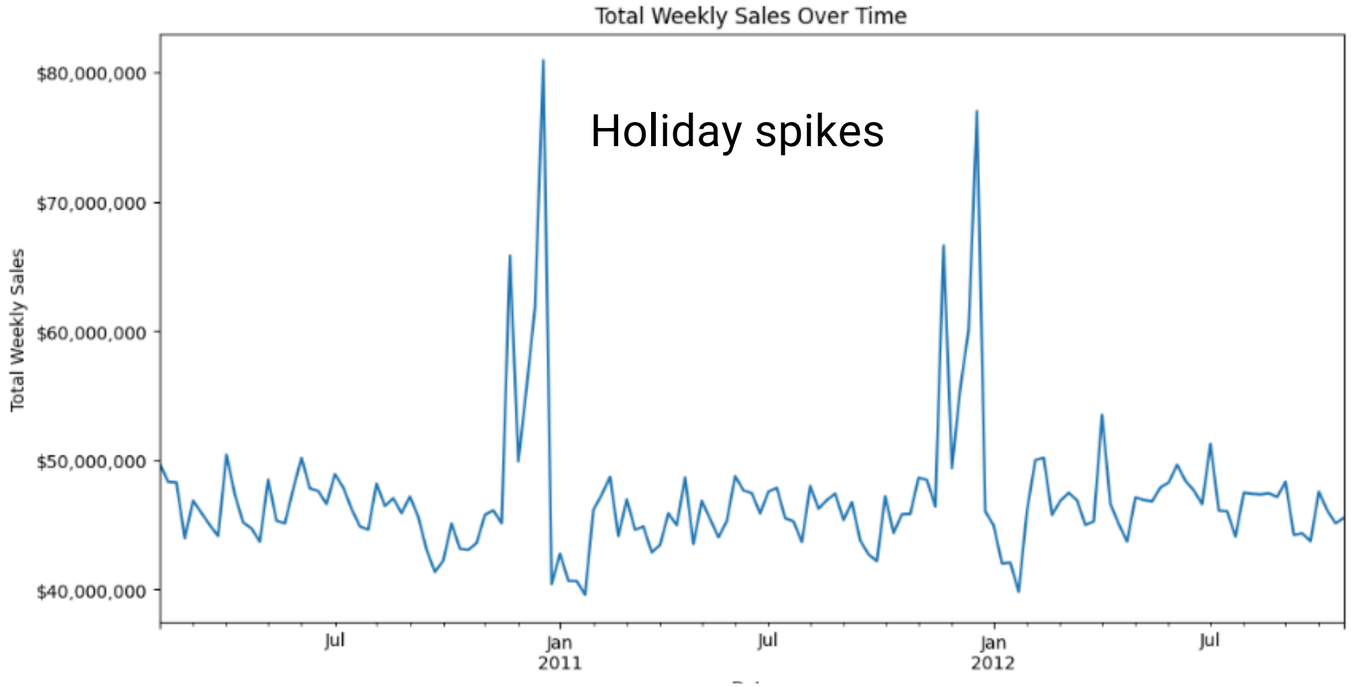
Supervised

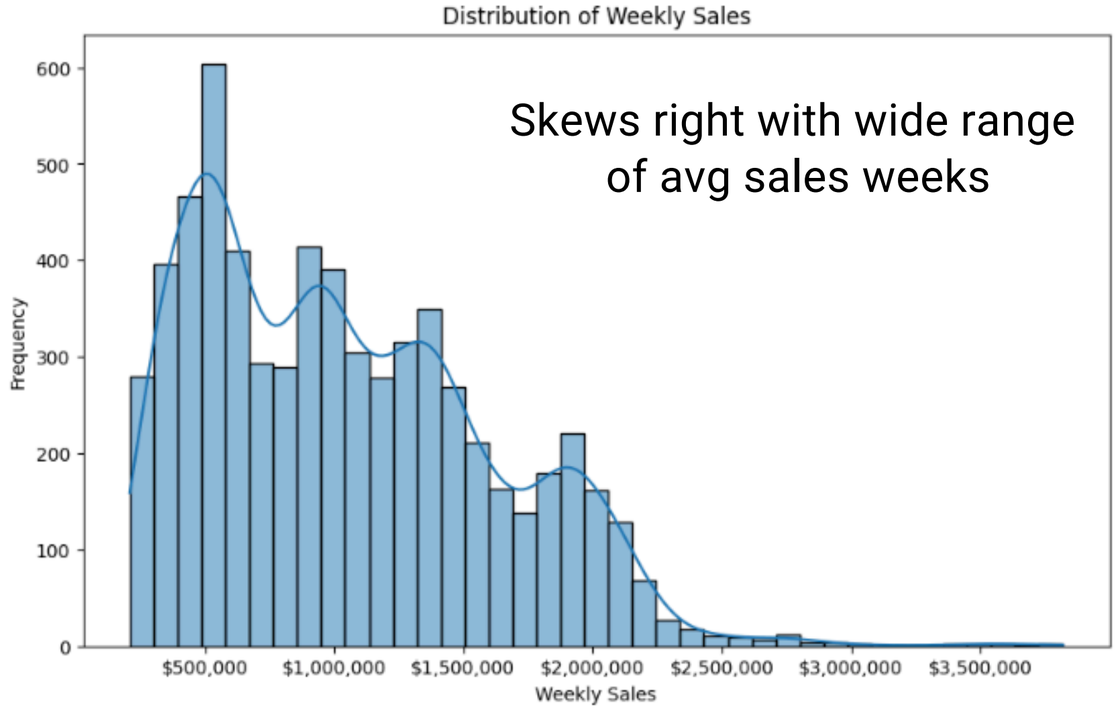
STREAMLIT APP

Application

Sales vary by location

Our 45 stores range in performance, we can leverage economic and weather indicators to better understand sales patterns and manage our business better.

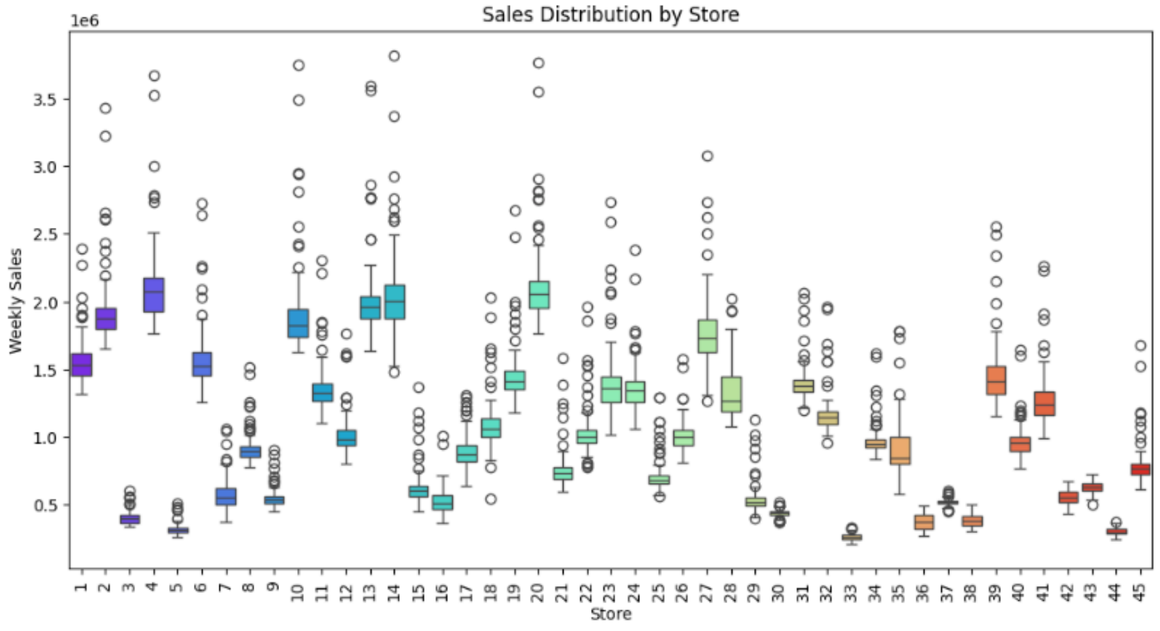


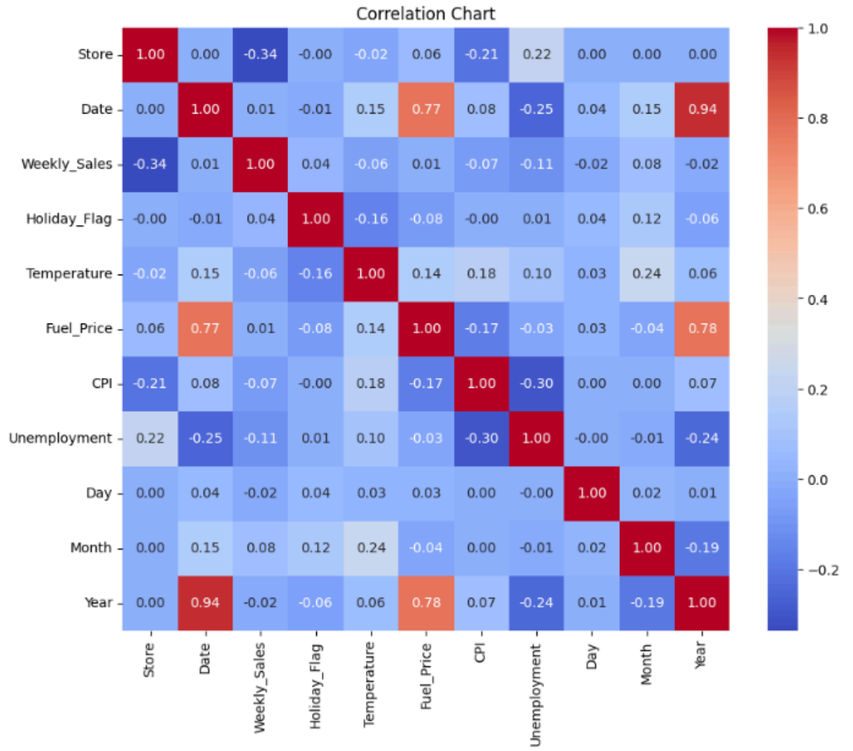


STORE LEVEL STATS DON'T TELL US MUCH, FEATURES TELL A STORY

PANDAS,SKLEARN

Analysis

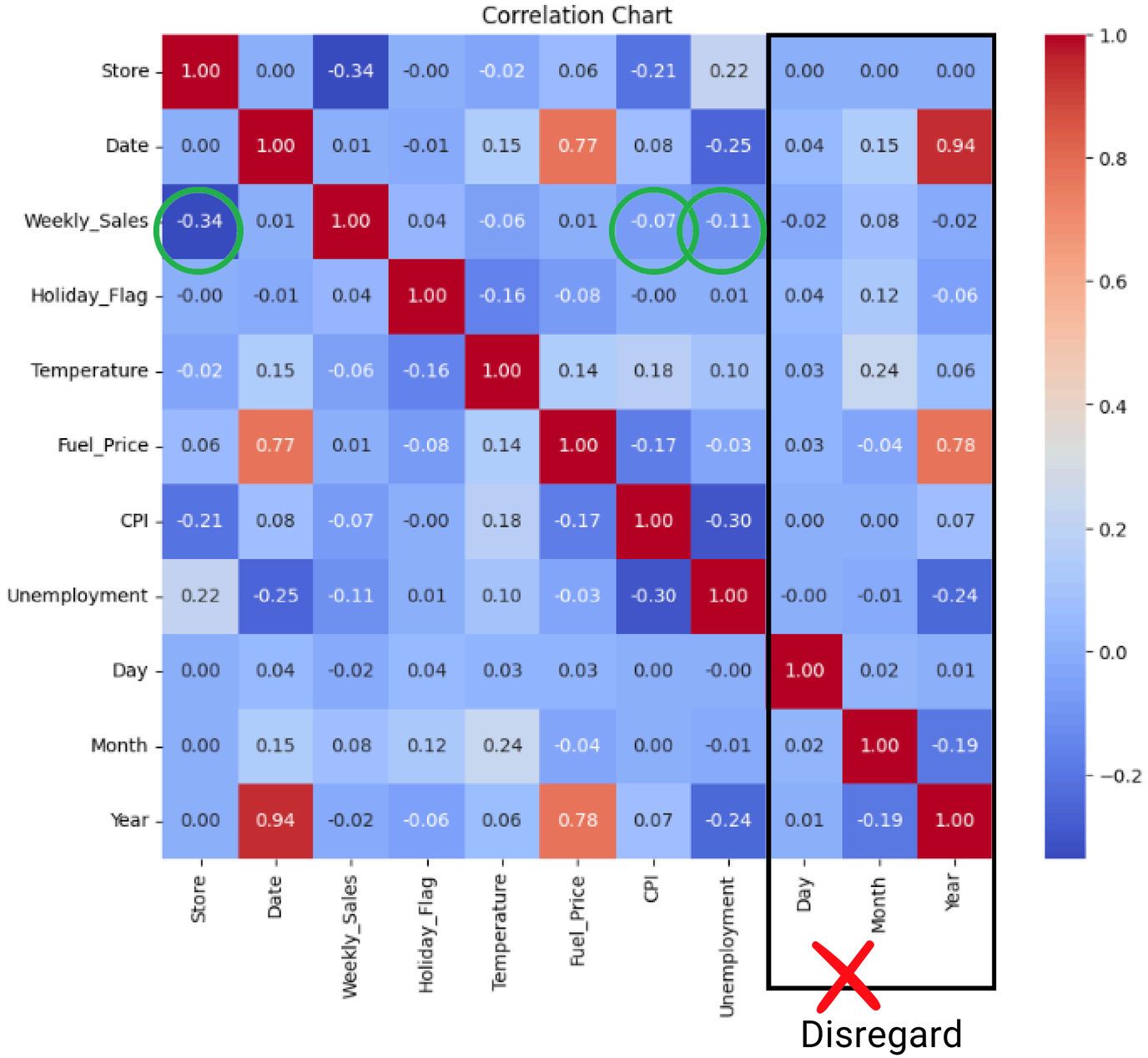




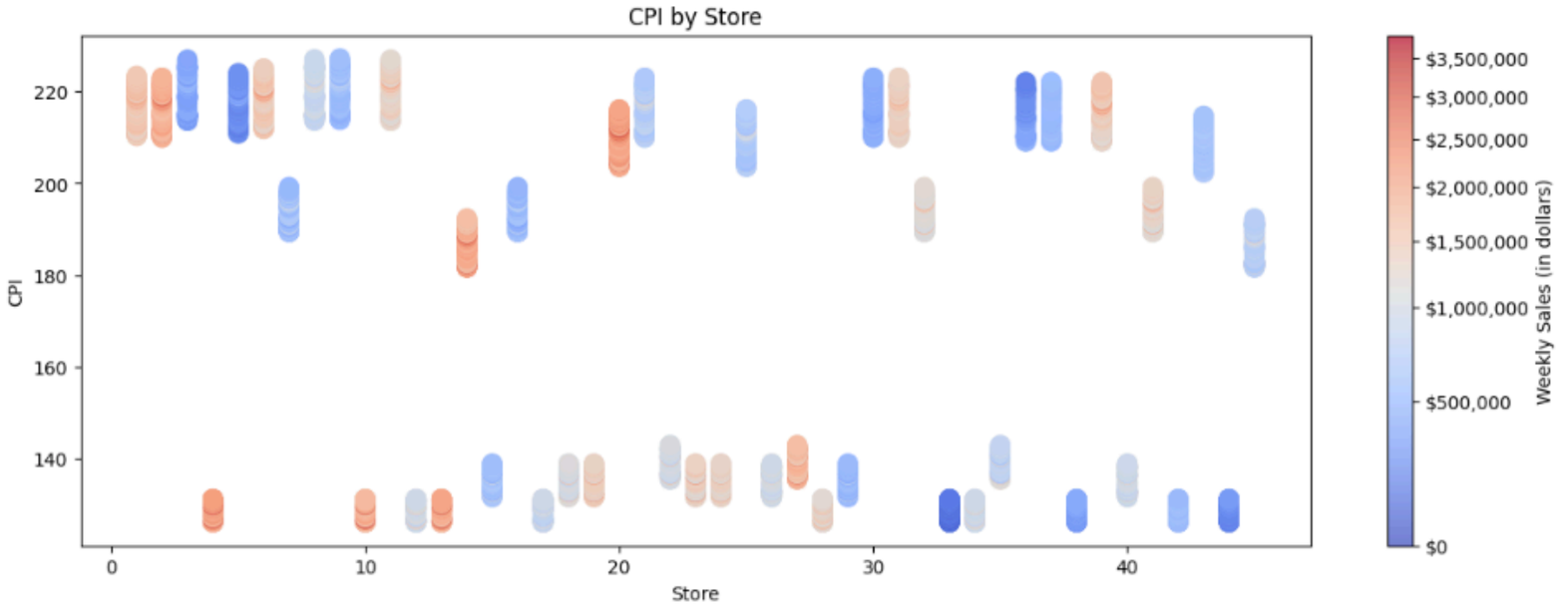
Local CPI and Unemployment Rate are indicative of weak negative relationship with weekly sales

Observations:

- Strongest useable correlation is between **Store # and Weekly Sales at 34%**, which indicates that stores are consistent in their sales performance
- Next best correlation is between Sales and **Unemployment at -11%**, followed by the **CPI at -7%**
- Somewhat surprisingly, **fuel price and holiday had a neutral correlation** with weekly sales.
  - This could be a function of few holiday weeks in the data and store proximity to home, work or public transportation.



Deep dive into important features w/ color emphasis confirms correlations



K MEANS CLUSTERING TO STRATEGICALLY DRIVE SALES INCREASES WHERE MARKET ALLOWS

Store	Weekly_Sales	Cluster
1	$1,555,264.40	2
2	$1,925,751.34	2
3	$402,704.44	2
4	$2,094,712.96	0
5	$318,011.81	2

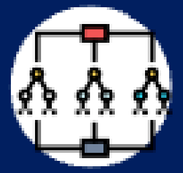
KMEANS CLUSTERING

Unsupervised

Implement Tiered Pricing Strategy

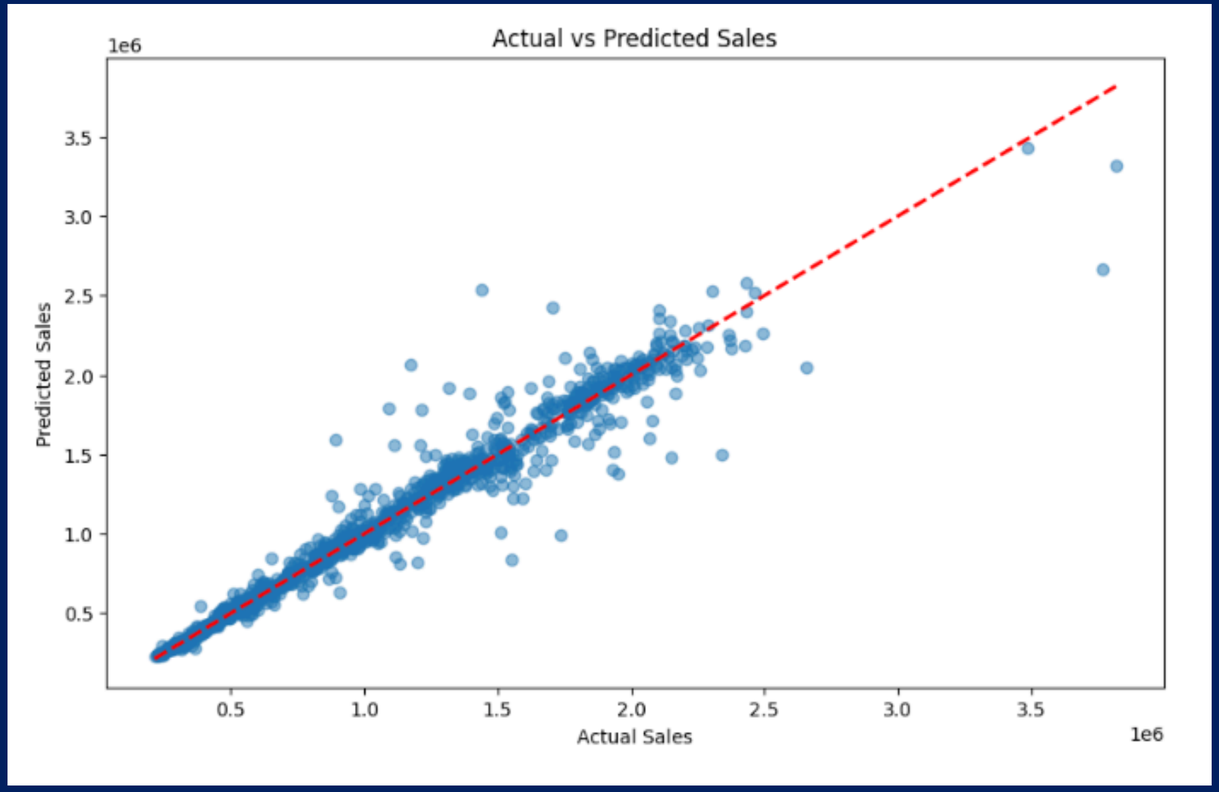
Use localized economic features to cluster stores into pricing tiers to increase price on items with low price elasticity - example, Every Day Low price on bottom tier, EDLP + 5% on next tier and so on to drive more slaes

```
Mean Squared Error: 13504777637.656
R-squared Score: 0.958079819067563
```

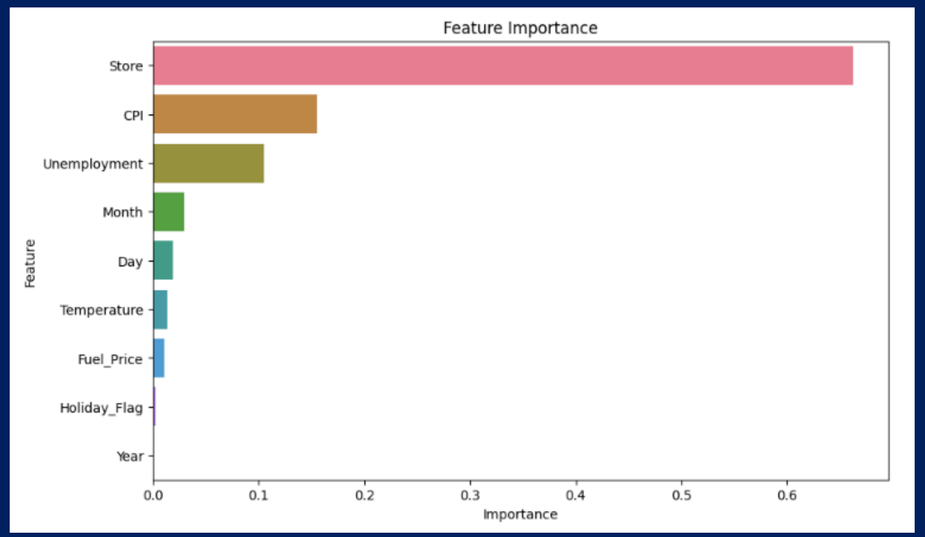


RANDOM FOREST

Supervised

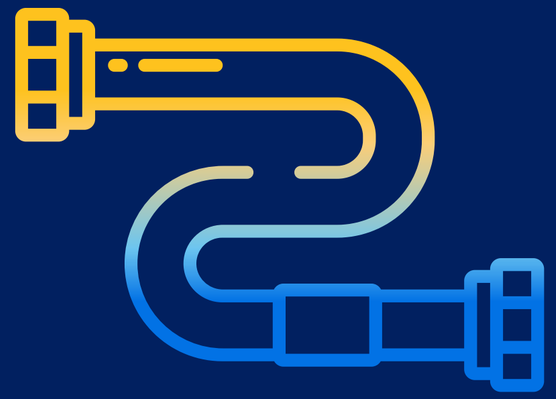


HIGH ABILITY TO PREDICT SALES USING RANDOM FOREST MODEL



Aspect	Predicting Holiday_Flag (Classification)	Predicting Weekly_Sales (Regression)
Y Value	Binary (0 or 1)	Continuous (Sales Value)
Nature of Problem	Classification (Binary)	Regression (Continuous)
Model Types	Logistic Regression, Random Forest Classifier, SVC, etc.	Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, etc.
Goal	To predict whether a week is a holiday or not	To predict the weekly sales for a given week
Metrics	Accuracy, Precision, Recall, F1-Score	MAE, MSE, R^2
Target Variable	Discrete (0 or 1)	Continuous (Sales figures)

MODELING PIPELINE



V1 Model Pipeline:

Feature Engineering: Lag features and rolling mean of weekly sales.

Models: Random Forest, Gradient Boosting, Voting Regressor.

V2 Model Pipeline:

Additional Features: Added holiday season flag and temperature × fuel price interaction feature.

Tuning: Hyperparameter tuning (XGBoost), introduction of Stacking Regressor.

Models: Random Forest, Gradient Boosting, XGBoost (with tuning), Stacking Regressor, Voting Regressor.

Model	MAE	MSE	R^2	CV R^2
Random Forest	55,709.95	9,049,825,670.59	0.97	N/A
Gradient Boosting	44,733.78	6,439,709,463.17	0.98	0.98
XGBoost	44,309.90	6,241,384,017.98	0.98	0.98
Voting Regressor	46,255.29	7,036,292,725.81	0.98	N/A

Key Insight:

Best Model: XGBoost with a high R^2 of 0.98 and low MAE.
Good Generalization: Gradient Boosting shows consistent cross-validation performance.

V1 Model Performance

COMPARING MODELS

Model	MAE	MSE	R^2	CV R^2
Random Forest	55,717.01	8,979,783,790.22	0.97	0.96
Gradient Boosting	45,579.62	6,646,940,248.57	0.98	N/A
XGBoost	42,841.63	5,807,657,025.75	0.98	N/A
Stacking Regressor	42,331.64	5,872,638,767.56	0.98	N/A
Voting Regressor	45,993.31	6,726,442,312.06	0.98	0.97

- Best Model: Stacking Regressor with lowest MAE of 42,331.64.
- Improvement: XGBoost and Stacking Regressor outperformed V1 versions with better accuracy after tuning.

V1 Model Performance

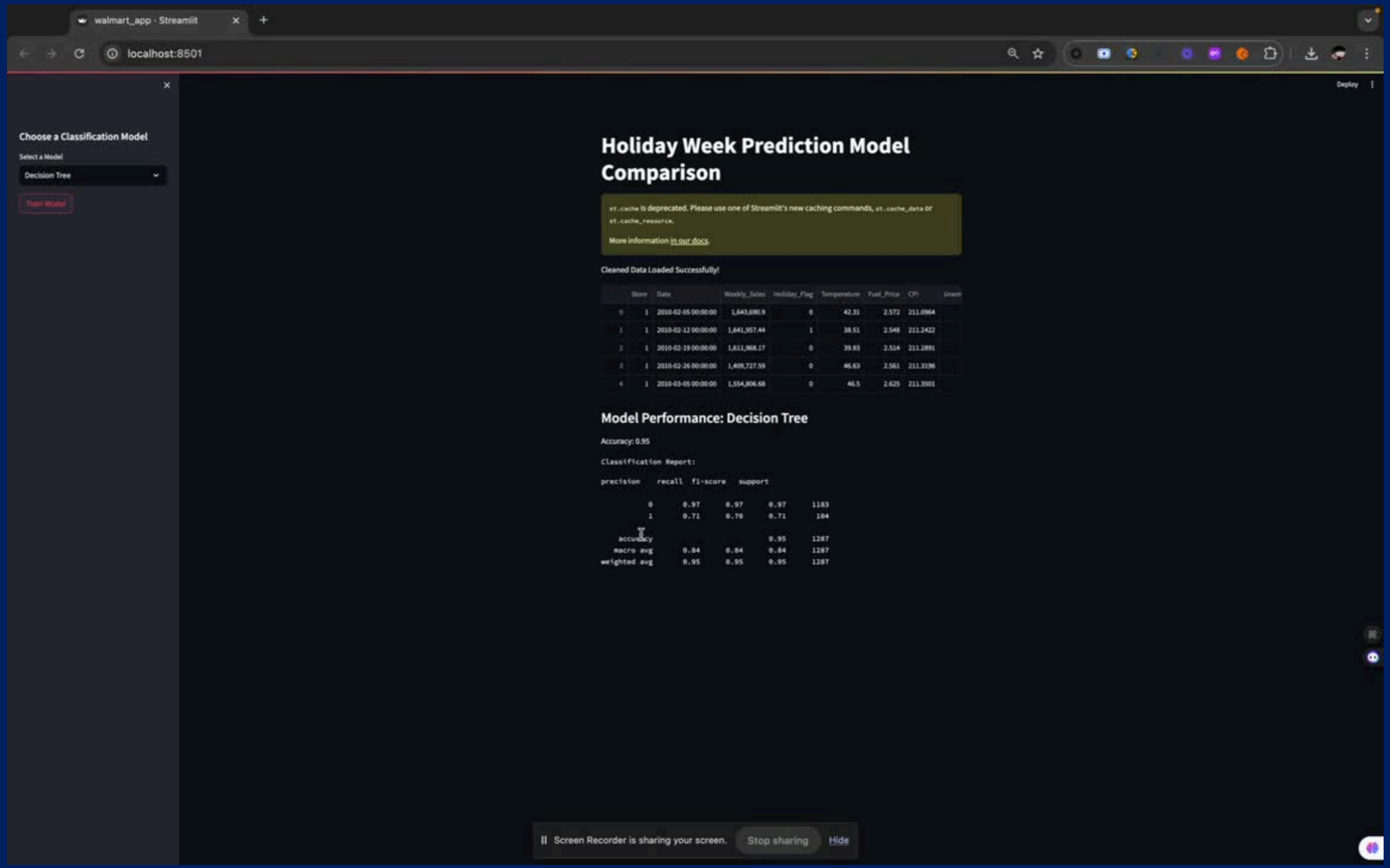
Key Insights & Takeaways

- Hyperparameter Tuning Works: XGBoost in V2 benefited significantly from hyperparameter tuning, leading to improved performance.
- Feature Engineering Matters: Adding the holiday season flag and interaction feature in V2 helped improve model accuracy.
- Best Performing Model: The Stacking Regressor in V2 outperformed all other models with the lowest MAE and high R^2.
- Model Generalization: Consistent cross-validation scores in V2 indicate that the models generalize well to unseen data.

Conclusions & Recos

- Conclusion: V2 models (especially Stacking Regressor and XGBoost) provide more accurate and consistent results due to improved feature engineering and hyperparameter tuning.

- Recommendation:
  - Use Stacking Regressor or XGBoost for future sales forecasting tasks, as they demonstrated the best performance.
  - Continue to explore additional features (e.g., external economic data) and further fine-tune hyperparameters for potential incremental improvements.



STREAMLIT APP

Application

Objectives

1. Predicting given all the data if its weekend
2. Predicting weekly_sales given all the data

FUTURE WORK

Prophet Model

Implement time series analysis techniques for more accurate forecasting

Promotional Modeling

Explore the impact of promotional events on sales and price elasticity modeling

More granular data

Source department-level or item-level sales information by store for indepth analysis

External data sources

Consider incorporating external data sources (e.g., proximity to competitors) to improve predictions


THANK YOU