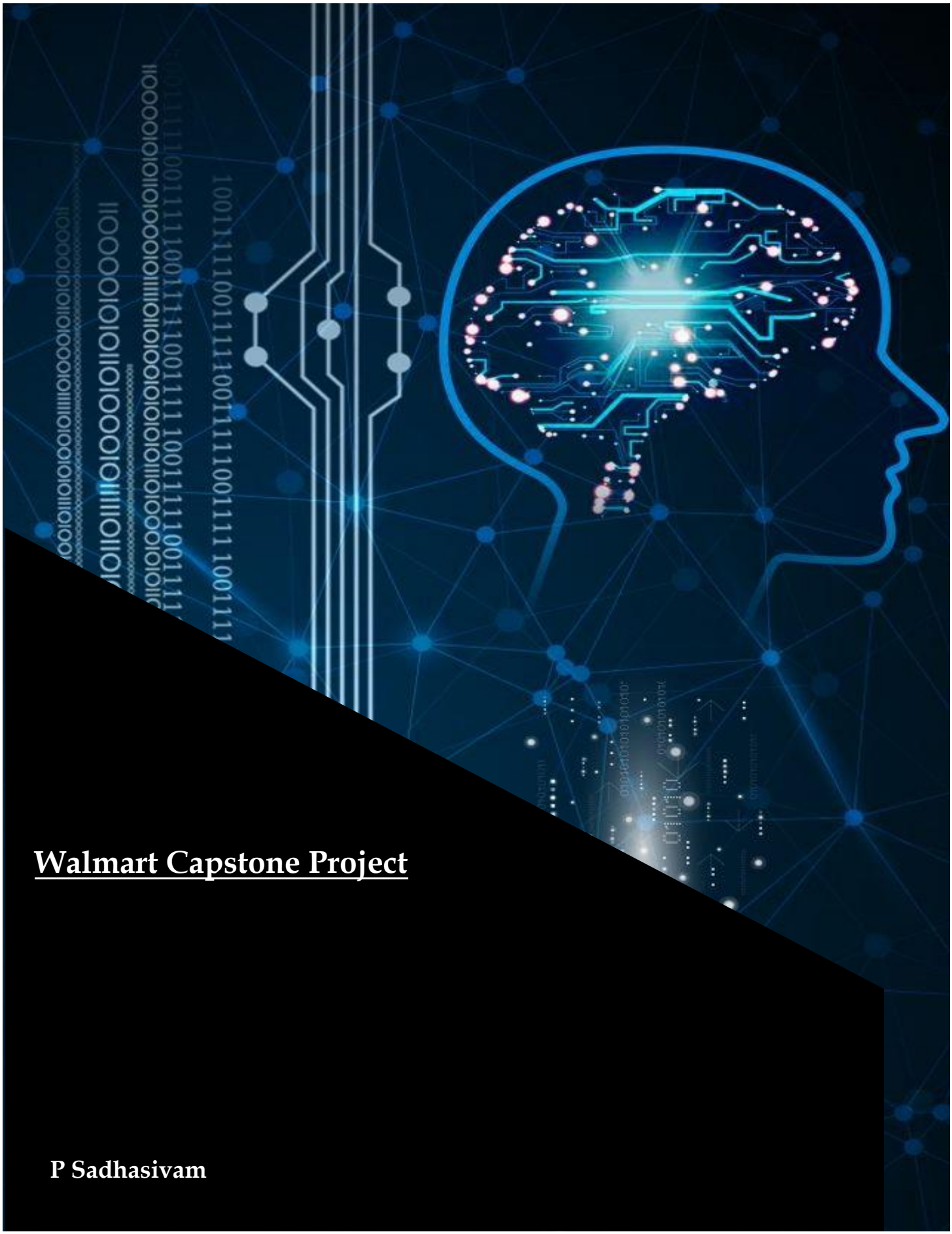


Walmart Capstone Project

P Sadhasivam

Contents

Problem Statement

A retail store chain with multiple outlets across the country is facing challenges in managing inventory to meet the demand while ensuring an optimal supply.
The project aims to provide valuable insights into the retail chain's operations and make predictions for future sales.

Project Objective

The primary objective of this project is to leverage data analysis and predictive modelling to address the inventory management challenges faced by a retail store chain with multiple outlets. Specifically, the project aims to achieve the following key objectives:

1. Data Analysis and Insights (EDA):
   - Conduct comprehensive exploratory data analysis (EDA) on the provided dataset to gain a deeper understanding of the retail store's sales dynamics.
   - Identify and analyse the impact of various factors, including unemployment rate, seasonality, temperature, and Consumer Price Index (CPI), on weekly sales.
   - Determine the top-performing and worst-performing stores based on historical sales data.
   - Provide actionable insights to assist in inventory management decisions and sales optimization.
2. Predictive Sales Forecasting:
   - Utilize predictive modelling techniques to forecast sales for each store for the next 12 weeks.
   - Develop accurate and reliable sales forecasts that can aid in inventory planning, resource allocation, and overall business strategy.
   - Evaluate the performance of the predictive models to ensure their effectiveness in providing accurate sales predictions.

By achieving these objectives, this project aims to empower the retail store chain with data-driven insights and forecasting capabilities, ultimately helping them optimize inventory management, enhance operational efficiency, and meet customer demand effectively.

Data Description

The dataset, named "walmart.csv," contains information related to the weekly sales and various attributes of a retail store chain with multiple outlets. The dataset comprises 6435 rows and 8 columns.

Feature Name	Description
Store	Store number
Date	Week of Sales
Weekly_Sales	Sales for the given store in that week
Holiday_Flag	If it is a holiday week
Temperature	Temperature on the day of the sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

Stores: there are 45 stores and each store has 143 recorded entries of:

a. Date of record (weekly),
b. Total sales record for the week,
c. Holiday flag for the week (1 or 0),
d. Temperature: average temperature recorded during the week,
e. Fuel Price: average fuel price for the week
f. CPI: average Consumer Price Index for the week
g. Unemployment: rate of unemployment for the week of record

Data Preprocessing Steps and Inspiration

The preprocessing of the data included the following steps:

Step 1: Load data

Step 2: Perform Exploratory Data Analysis

a. Confirm number of records in the data and how they are distributed
b. Check data types,
c. Check for missing data, invalid entries, duplicates
d. Examine the correlation of the independent features with the target (Weekly. Sales) variable.
e. Visualize the correlation, various plot of weekly sales with other features.
f. Check for outliers that are known to distort predictions and forecasts

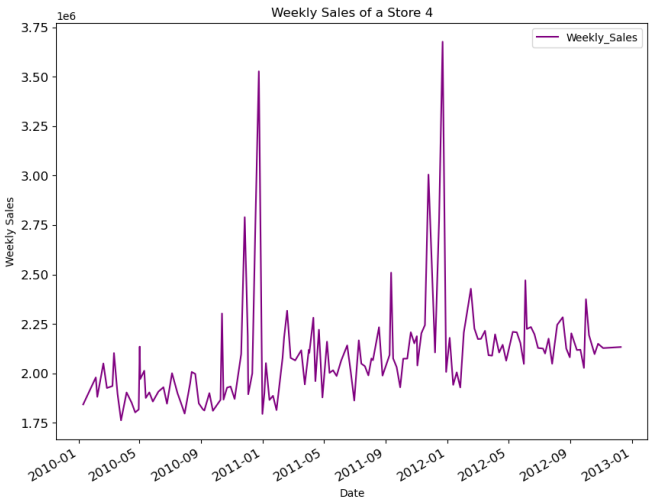
Step 3: Model Predictions, two approaches:

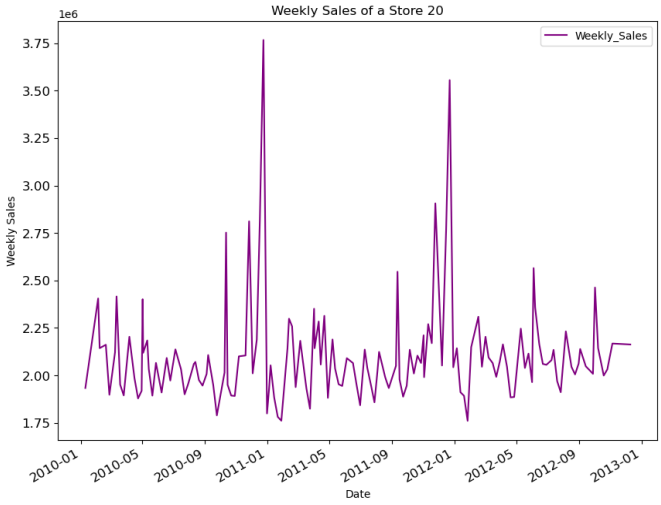
a. Time Series Model (SARIMA)
b. Linear Regression Model(s)

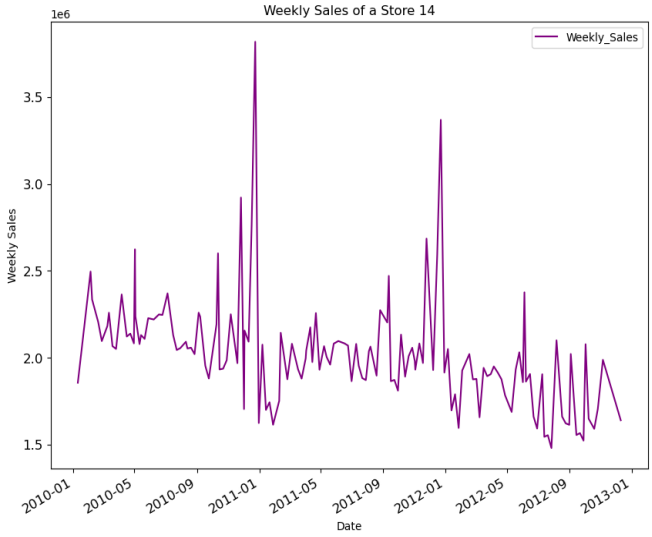
Step 4: Forecast

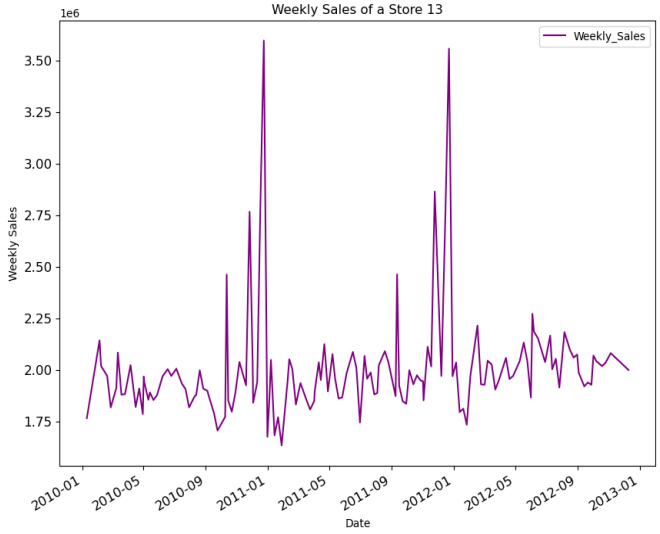
Step 5: Compare Results from the different models

Model Evaluation and Technique









Model Selection

Examination of the plot of the target feature, Weekly_Sales (as shown above) shows a continuously time varying data.
A Time Series (TS) model (SARIMA, SARIMAX) will be employed for the predictions and forecast. Attempt will also be made to use LinearRegression models (Gradient_Boosting, Linear Regression, Random_Forest) for prediction and compare the results with the TS predictions.

1) The SARIMA model:

Seasonal Autoregressive Integrated Moving Average (ARIMA) is defined as a statistical analysis model which is a widely used time series forecasting model that extends the capabilities of the ARIMA model by incorporating seasonality into the framework. SARIMA is particularly useful when dealing with time series data that exhibits a seasonal pattern, such as weekly, monthly, or yearly fluctuations.
A statistical model is autoregressive if it predicts future values based on past values.
SARIMA model is based on a number of assumptions including:

1. Stationarity: The time series data becomes stationary after differencing.
2. Independence: Observations in the time series are independent of each other.
3. Seasonality: The data exhibits a seasonal pattern captured by seasonal components.
4. Linearity: There's a linear relationship between current and past values.
5. Normality of Residuals: Residuals follow a normal distribution with a mean of zero.
6. No Multicollinearity: If exogenous variables are included, there is no high correlation among them.

2) Regression Models

a) Gradient_Boosting: Gradient boosting stands out for its prediction speed and accuracy, particularly with large and complex datasets (https://www.analyticsvidhya.com/blog/2021/09/adaboostalgorithm-a-complete-guide-for-beginners/). The algorithm has produced the best results from Kaggle competitions and machine learning solutions for business. In machine learning algorithm, two types of errors, otherwise called loss functions, are encountered, bias error and variance error. Gradient boosting algorithm i*s* based on minimizing the *bias error* or the loss function of the model. The gradient boosting algorithm is based on building models sequentially where the subsequent models try to reduce the errors of the previous model. The subsequent models are built on the errors or residuals of the previous model. The process is repeated until there is no more

significant change on the error.

b) Linear regression is a basic predictive analytics technique that uses historical data to predict an output variable (https://towardsdatascience.com/introduction-to-linearregression-in-python-c12a072bedf0). It is a popular algorithm employed to predict continuous (dependent) variables such as price, based on their correlation with other independent variables. It is based on the following assumptions:

- Linear Relationship: The relationship between the independent and dependent variables should be linear.
- Multivariate Normal: All the variables together should be multivariate normal, which means that each variable separately has to be univariate normal means, a bell-shaped curve.
- No Multicollinearity: There is little or no multicollinearity in the data which means that the independent variables should have minimal correlation with each other.
- No Autocorrelation: There is little or no autocorrelation in the data where the data values of the same column are related to each other.
- Homoscedasticity: There should be homoscedasticity or "Same variance" across regression lines. In other words, residuals are equal across regression line.

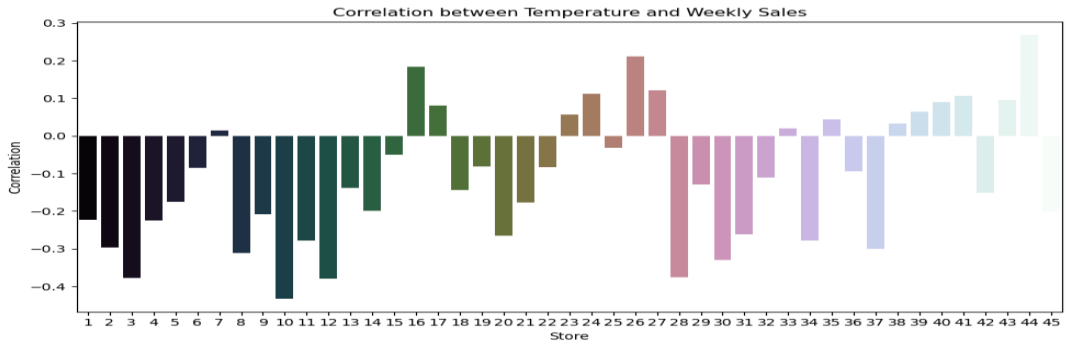
c) Random Forest: Random Forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems (https://www.ibm.com/cloud/learn/random-forest).

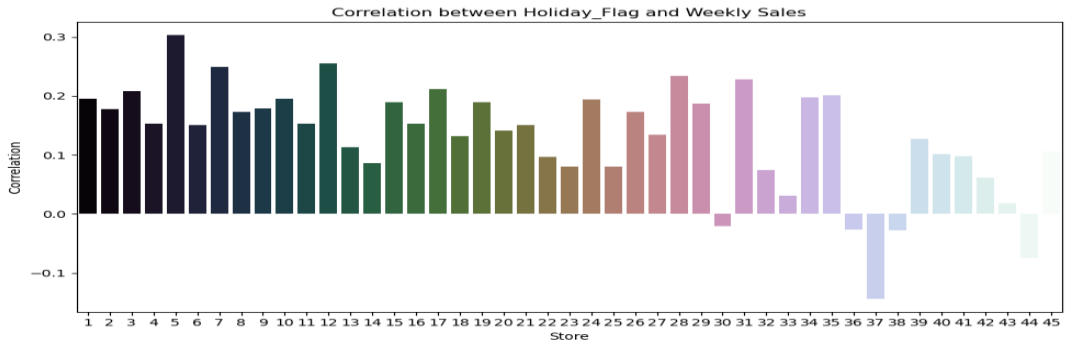
Model Evaluation

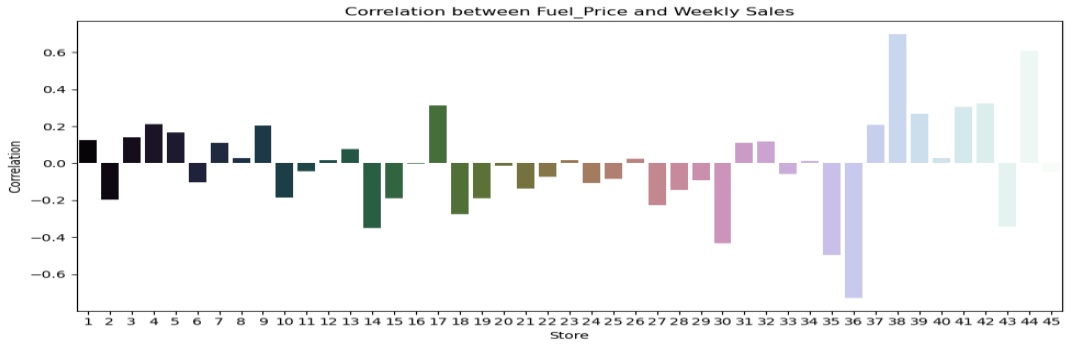
The following techniques and steps were involved in the evaluation of the model

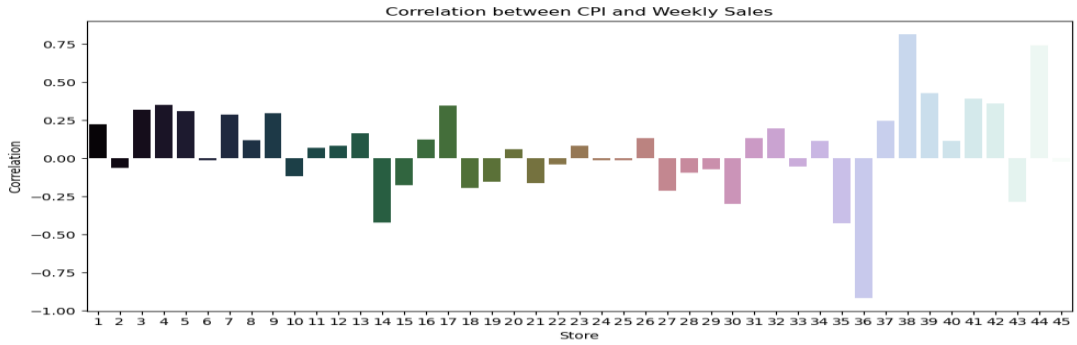
1. Load necessary Libraries
2. Load the dataset
3. Perform Exploratory Data Analysis (EDA) on the dataset
    a. Find the shape or size of the data
    b. Check for invalid and null entries
    c. Explore data description
    d. Examine the correlations of the independent variable to the target variable (Weekly_Sales)
    e. Line plot of the effects of the independent variables on the target variable
    f. Box plot of the features to identify outliers
4. Model Prediction
5. Forecast

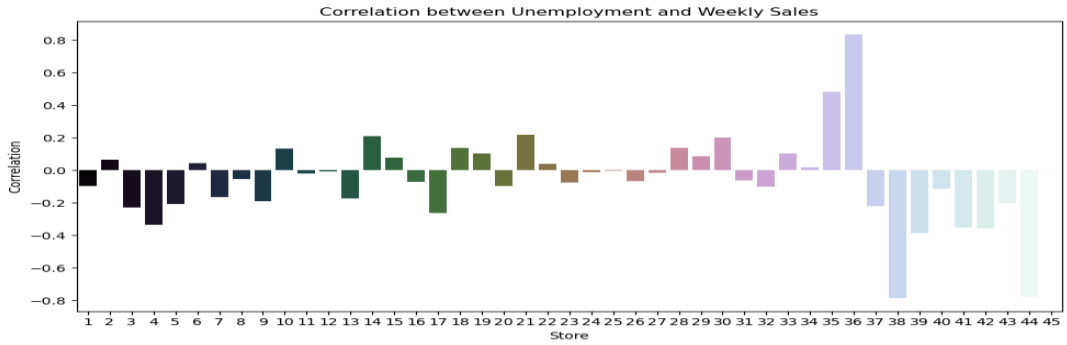
Model Design

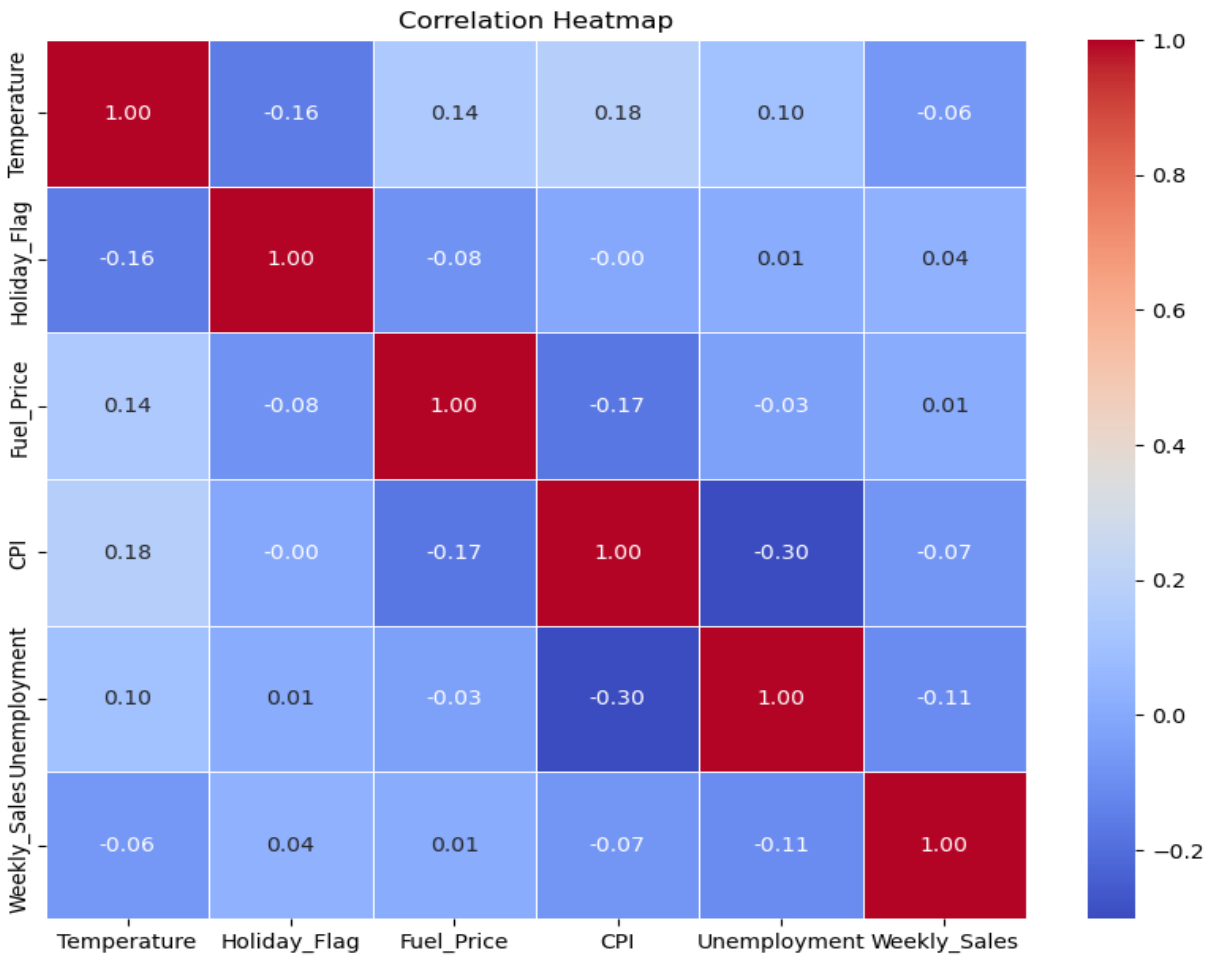












- It was observed from the EDA that the effects of the independent features (Unemployment, Temperature, Holiday_Flag, and CPI) on the target variable, Weekly_Sales differ greatly by the store.
- The correlation matrix gives a general relationship among features for all the stores, from correlation matrix it is observed that Temperature has a negative impact (inversely proportional), Holiday_flag has a positive impact, CPI has a negative impact and unemployment has negative impact on weekly sales in general for the 45 stores.
- A) From the bar plot we can see that Unemployment affects the stores 36(negative impact),38(positive impact), 44(positive impact) greatly.
  C) Temperature has a negative correlation on weekly sales.
  D) CPI has a negative correlation with weekly sales of stores in general

E) Top 10 Performing Stores:
20 ,4,14,13,2,10,27,6,1,39
F) Worst Performing Store:
Store 33: 37160221.96
Difference between Highest and Lowest Performing Stores:
264237570.49999997

- Premised on the findings, the decision was taken to handle the model predictions by the stores as a single prediction for all the stores may not be reasonable given the peculiar conditions prevalent in each region of the stores.
- For simplicity and ease of presentation, I have also decided to limit my predictions for the five stores with the highest Weekly_Sales revenue. That notwithstanding, the model could always be used to provide predictions for each of the store.

Model Approach:

1. TS Model, SARIMA.

   Stationarity Check:

   - The first step for this model is to check the stationarity of the dataset (p-value less than 0.05).
   - Decompose the time series to identify trends and seasonality.
   - Determine the seasonal period (s).

   SARIMA Model Selection:

   - Use auto_arima or manual grid search to identify the best SARIMA order (p, d, q, P, D, Q, s) for each store.

   Model Fitting:

   - Fit SARIMA models to each store's weekly sales data.

   Model Validation:

   - Validate model performance using historical data.
   - Calculate forecast errors (e.g., MAE, RMSE) for each store.

   12-Week Forecasting:

   - Forecast weekly sales for the next 12 weeks.

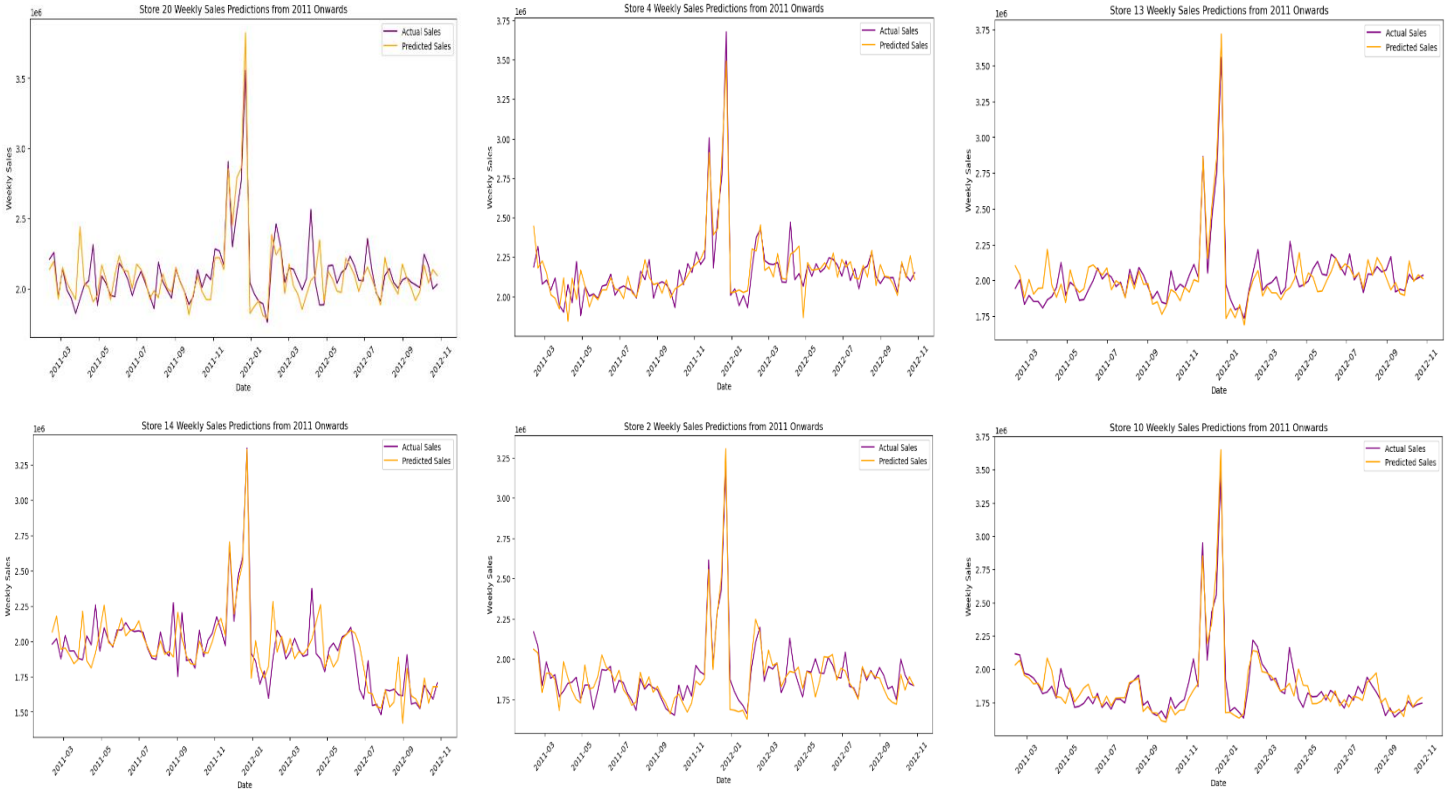
2. Regression Models:

   - Regression model, Gradient_Boosting, Linea Regression, and Random_Forest models were also used for the prediction. The best of the three predictions will then be compared to the predictions by ARIMA model predictions.

Inferences from the Project

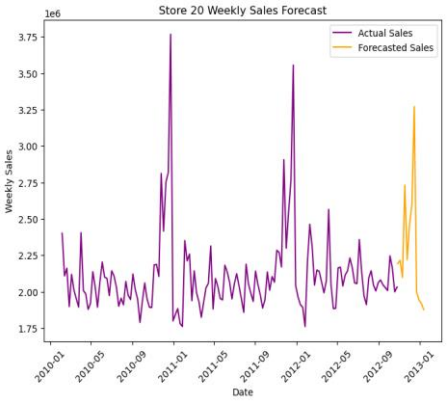
Model Results:

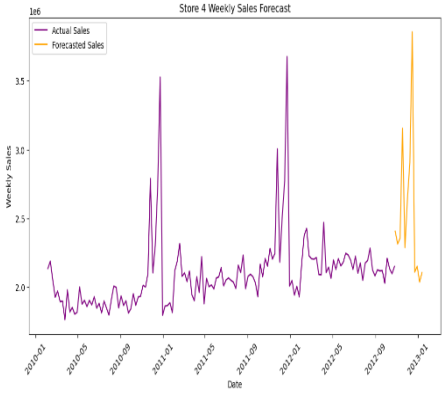
1. SARIMA model – 1) Predictions
   we present the evaluation of our time series forecasting model using SARIMA (Seasonal AutoRegressive Integrated Moving Average) for the 6 top-performing stores (20,4,13,14,2,10) in terms of weekly sales. Our objective was to assess the model's performance in predicting future sales trends without the need for extensive data preprocessing.

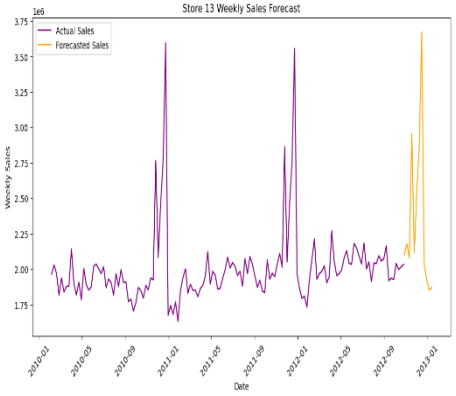


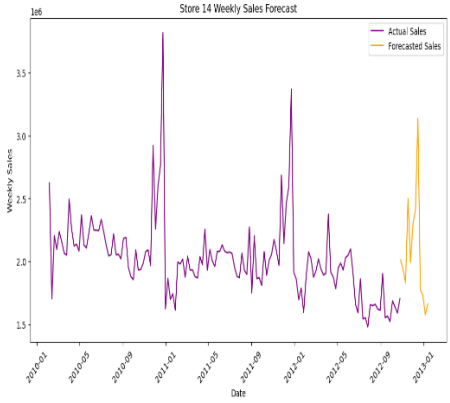
Error	Store 20	Store 4	Store 13	Store 14	Store 2	Store 10
Mean Error (%)	0.47	-0.28	0.37	-0.48	0.44	0.26
Mean Absolute Percentage Error (%)	4.32	3.51	3.59	5.40	3.09	3.53

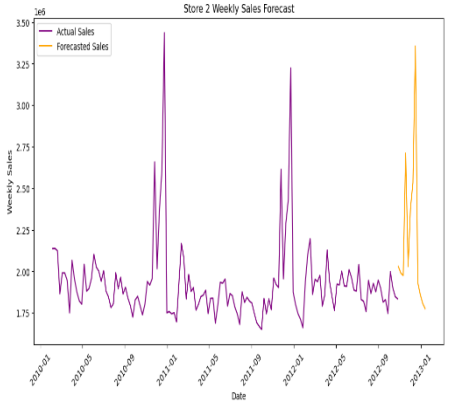
2) Forecast

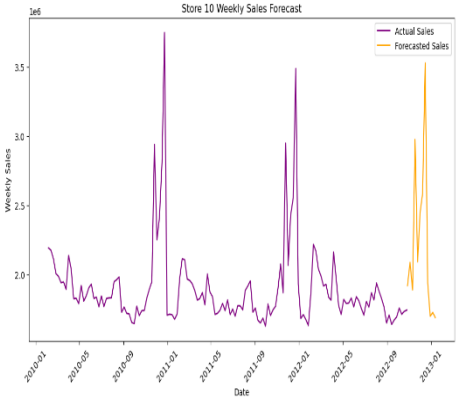








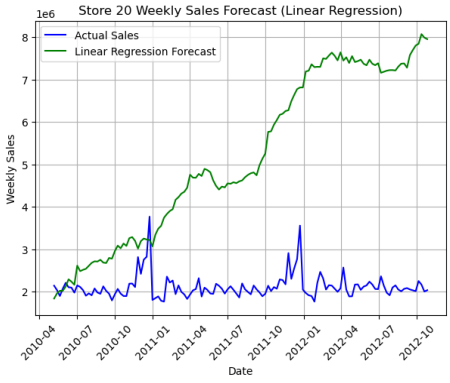


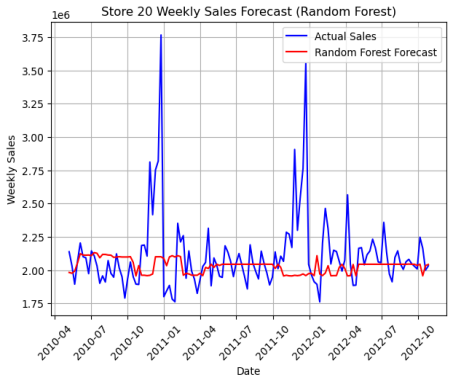


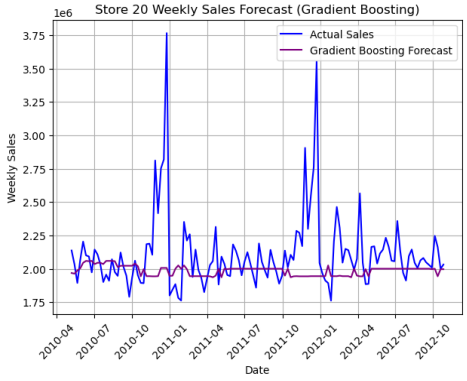
Error	Store 20	Store 4	Store 13	Store 14	Store 2	Store 10
Root Mean Square Error	613018.01	678459.82	653378.13	641573.61	581406.51	757364.38
Mean Absolute Error	432646.43	422054.78	399305.16	438238.58	358268.93	495956.13

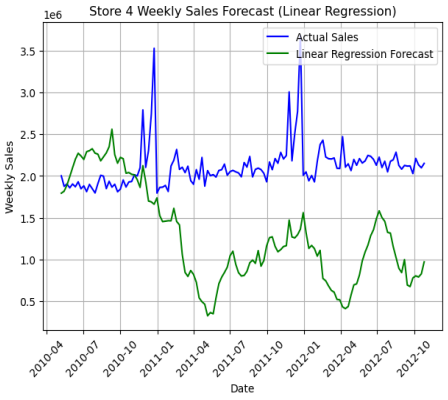
2) Regression Models

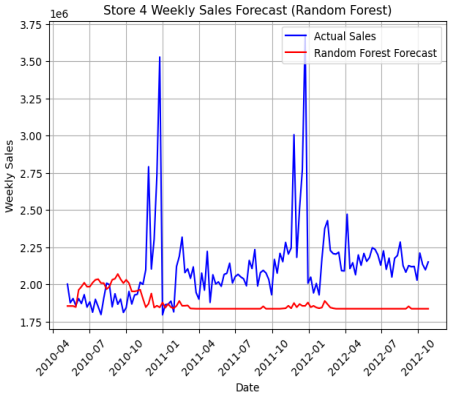
The prediction results from the three chosen regression models: Gradient Boosting, Linear Regression, and Random Forest are summarized below:

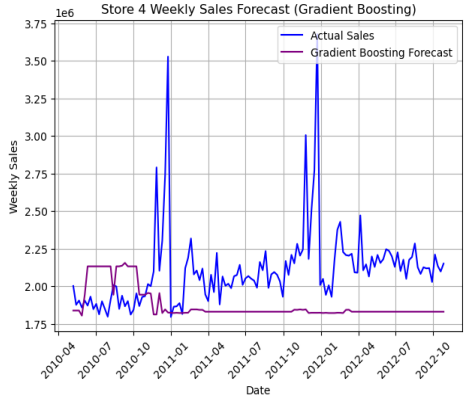


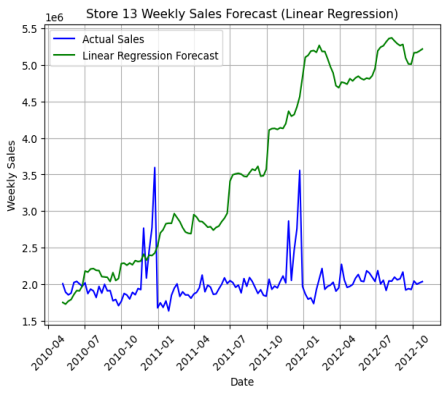


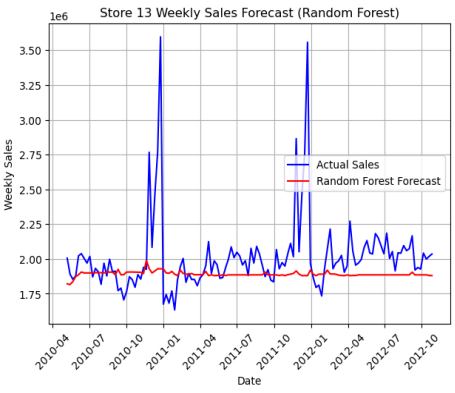


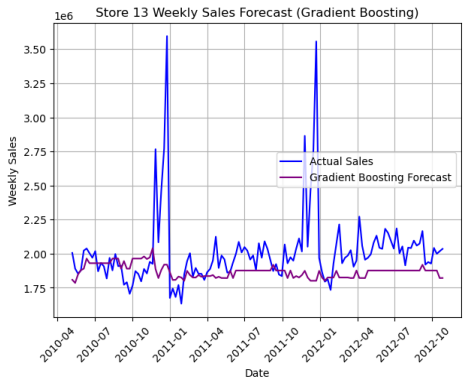


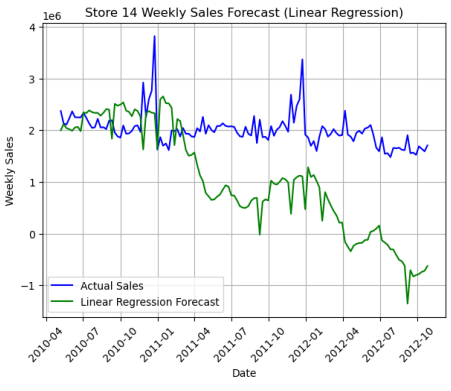


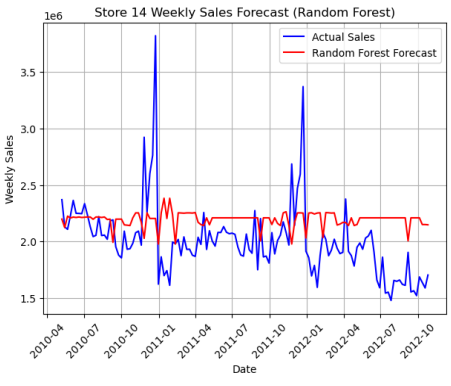


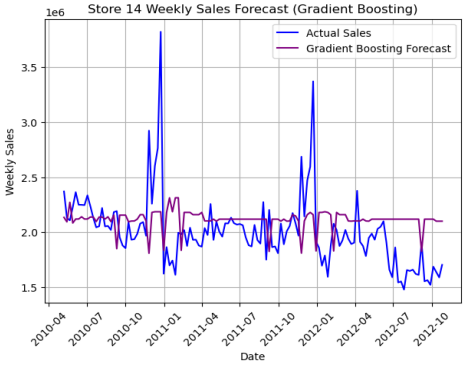


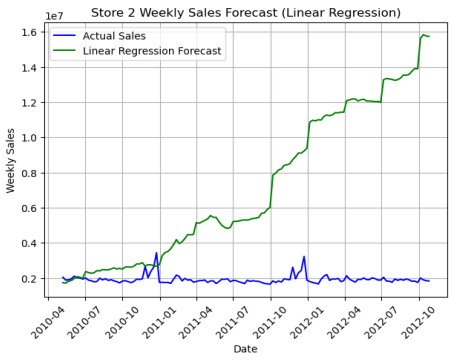


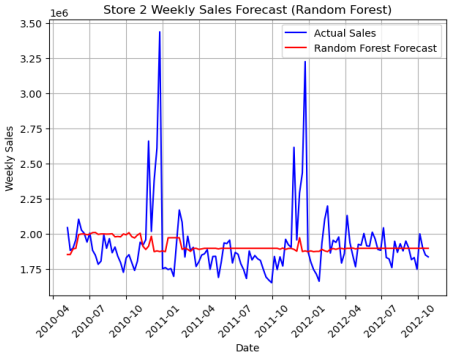


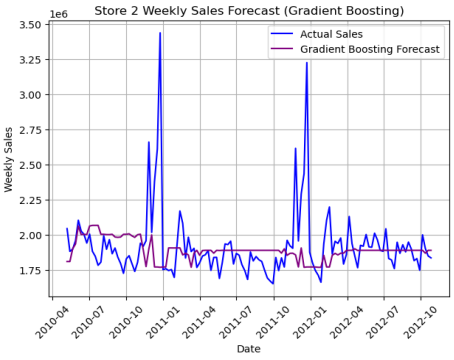


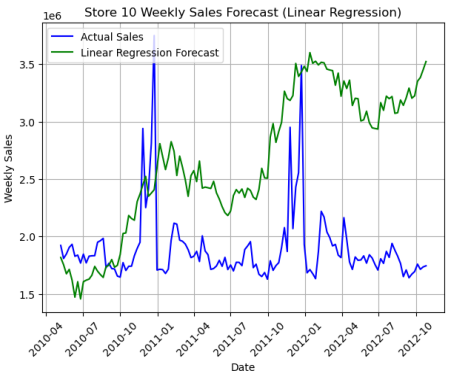


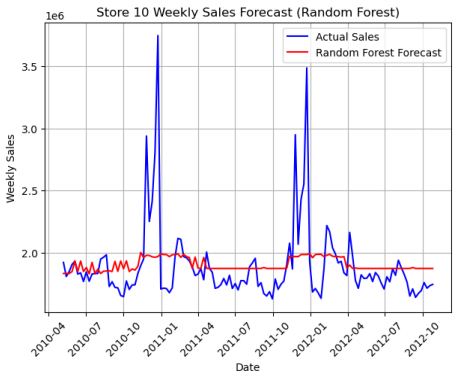


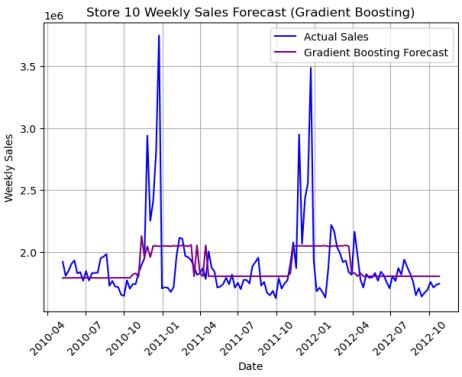












	Linear regression	Random Forest	Gradient boosting
Store - 20			
RMSE	3620538.45	308260.74	310165.89
MAE	3074490.35	180473.34	174761.87
Store - 4			
RMSE	1070416.24	381685.84	404322.69
MAE	943304.55	280648.78	309796.89
Store - 13			
RMSE	1987229.59	299219.67	313413.03
MAE	1632128.59	165979.55	184652.90
Store - 14			
RMSE	1368756.88	377470.57	354316.18
MAE	1144069.83	294350.50	258996.82
Store – 2			
RMSE	6763899.93	252746.73	146814.48
MAE	5247440.26	140962.03	272457.40
Store - 10			
RMSE	1000476.75	294009.54	281171.51
MAE	853477.83	159396.15	152546.28

Model Evaluation:

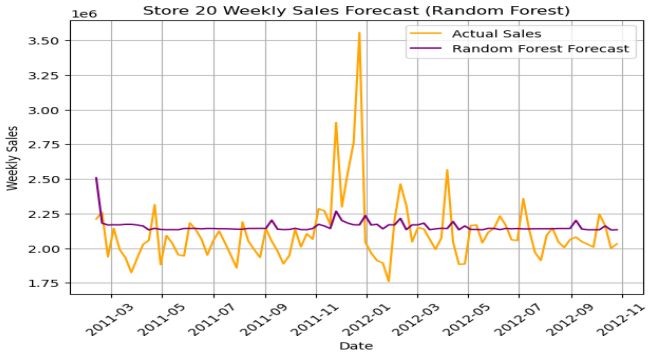
1. SARIMAX Model:

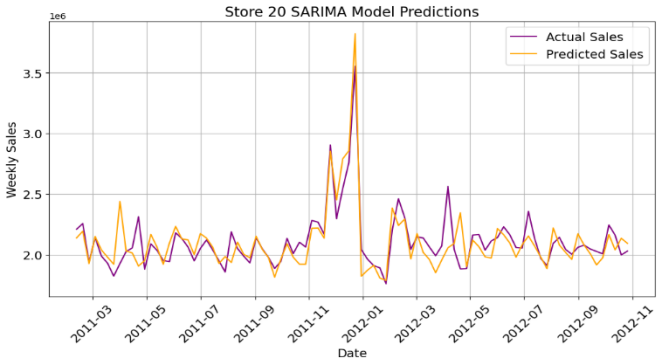
The model predictions for the selected stores were quite good. The forecast
were okay showing variabilities of the weekly sales in line with the sales history.

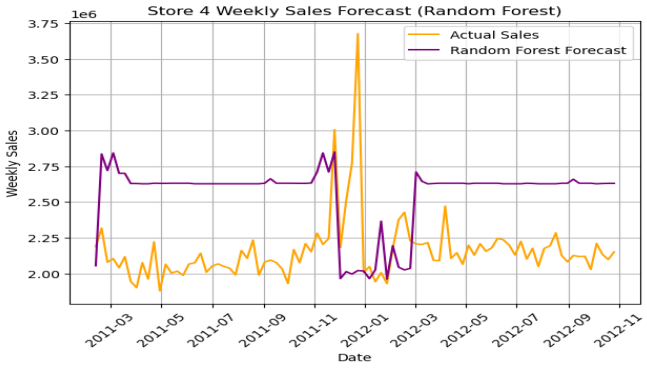
2. The Regression Model

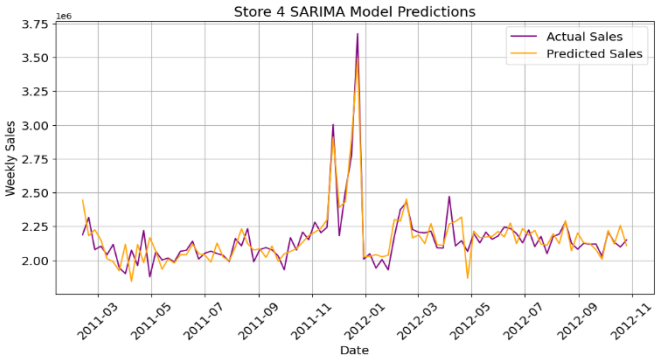
The summary of the evaluation report
root_mean_squared_errors and mean_absolute_percentage_errors,
is presented in the Table above.

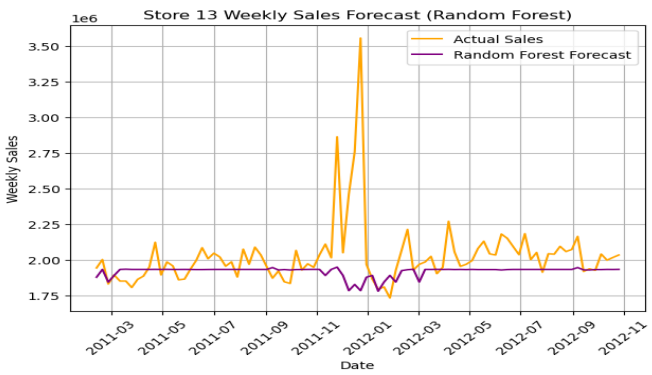
3. Comparing the Models:

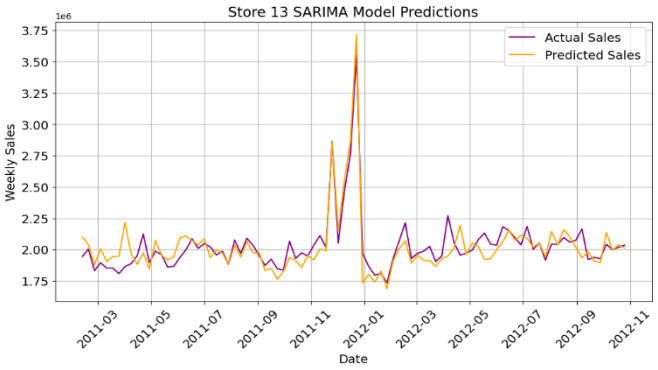


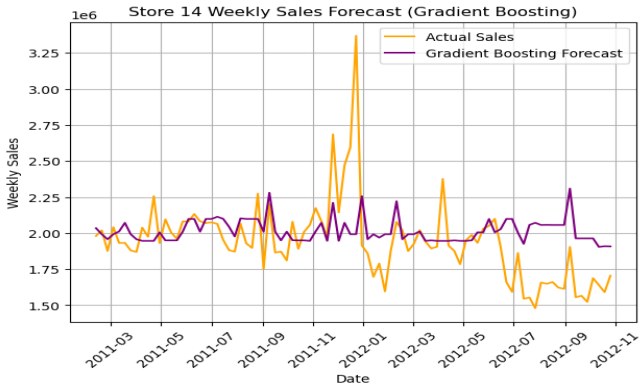


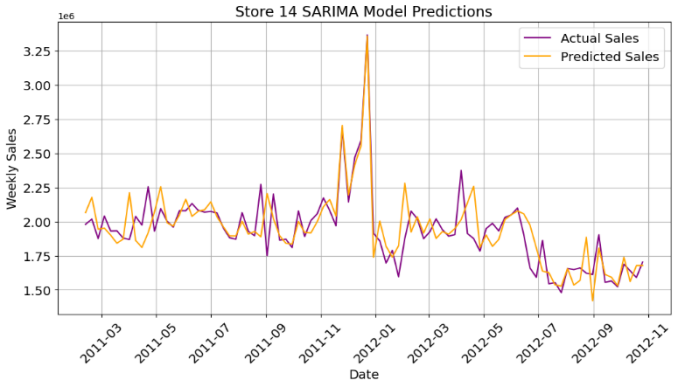


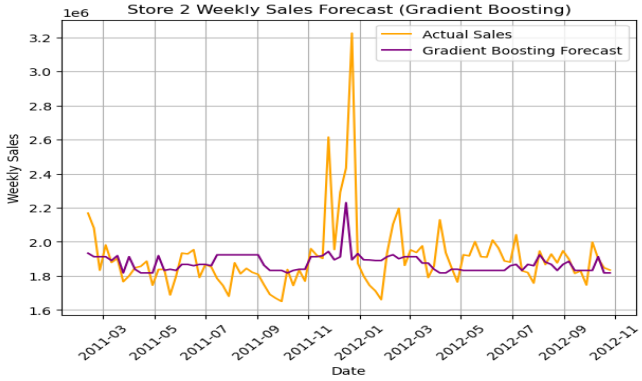


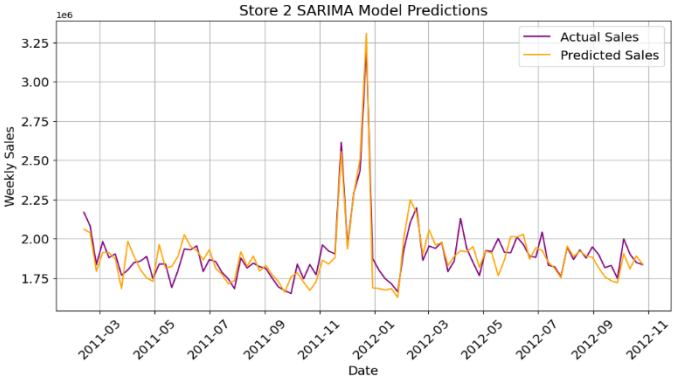


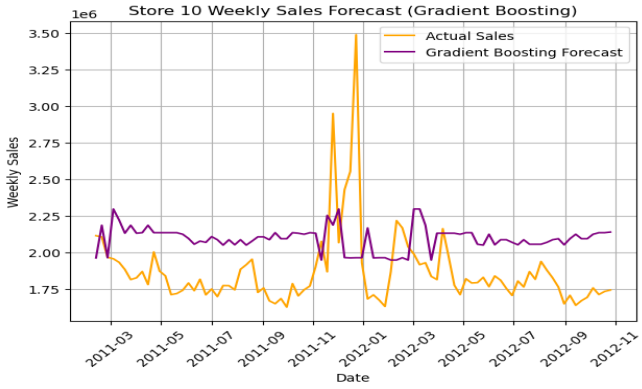


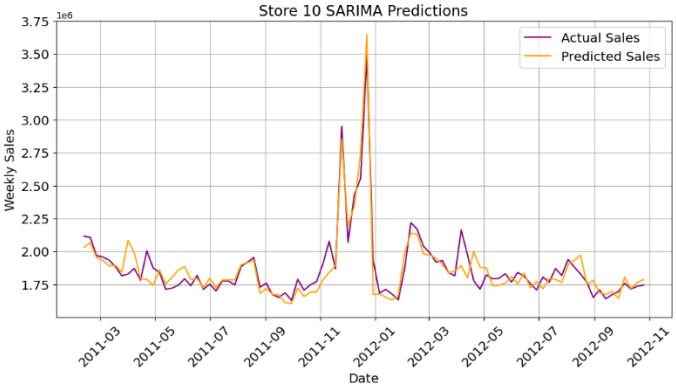












In our analysis, we conducted a thorough comparison of predictive models, including regression models and SARIMA (Seasonal Autoregressive Integrated Moving Average) models, to forecast weekly sales. Remarkably, the SARIMA model exhibited a remarkable level of accuracy, producing predictions that closely mirrored the actual weekly sales data. This exceptional predictive performance underscores the efficacy of SARIMA models in capturing the complex seasonal patterns and temporal dependencies present in our sales data, thereby providing valuable insights for future sales forecasting and strategic decision-making.

Future Possibilities

1. Model Enhancement and Advanced Techniques: In the future, the project has room for significant improvement through advanced model refinement. This includes fine-tuning model hyperparameters, exploring ensemble modelling techniques, and conducting sensitivity analyses to identify influential features. These efforts can enhance forecasting accuracy and reliability.

2. Incorporation of Additional Data Sources: To capture more nuanced sales patterns, consider integrating additional exogenous variables or external data sources. Variables such as macroeconomic indicators, competitor data, and weather information can provide valuable insights into the factors influencing sales. Expanding the scope of data can lead to more accurate predictions.

3. By enabling enterprises to better understand both the customers' and business functioning behaviour, Machine Learning has enabled companies to offer better/targeted customer service(s) leading to more loyal customers and ultimately improved sales revenue.

Conclusion

The project undertook a study of a retail company with 45 outlets stores. Some of the important findings from the report include the followings:

1. Sales revenue projections for the next 12-weeks shows similar seasonal trend and are down for most of the stores
2. Some of the stores have very weak or no sales activities during some period of the year.
3. To improve sales revenue, the following steps are recommended:
   a. Concerted efforts by the company to find out though local market surveys and past sales records what products are in high demand by the local population at any given period of the year and make efforts to replenish those stocks.
   b. Create increased local awareness of the products on offer at each store through commercial outreach: social media, television commercials, radio jingles, and print media, trade shows, to name a few, could help improve sales.
   c. Have detailed records of inventory of the items on offer at each store indicating amount and dates if sold as it is needed for effective inventory tracking.
   d. Explore other service options that have worked well for similar companies, such as same-day or next day home delivery.

It may just be that some stores may just have to be wound up if sales revenue does not improve.

References

1. Autoregressive Integrated Moving Average defined: *Autoregressive Integrated Moving Average (ARIMA) Definition (investopedia.com)*
2. Introduction to Linear Regression in Python: https://towardsdatascience.com/introduction-to-linear-regression-inpython-c12a072bedf0
3. Random Forest by IBM cloud Education: https://www.ibm.com/cloud/learn/random-forest
4. Future of Machine Learning From 2022 and Beyond: https://www.day1tech.com/what-does-the-future-of-machinelearning-look-like-2022-andbeyond/#:~:text=The%20future%20of%20ML%20clearly,instead%20of%20open%2Dsource%20platforms.