

Capstone Project - PROJECT WALMART DATASET

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons for Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same 10.

Future Possibilities of the Project

11. Conclusion
12. References.

Problem Statement

A retail company with multiple outlet stores is having poor revenue returns from the stores with most of them facing bankruptcy. This project undertakes to review the sales records from the stores with a view to provide useful insights to the company and also to forecast sales outlook for the next 12-weeks.

Project Objective

The retail company with multiple outlets across the country are facing issues with inventory management. The task is to come up with useful insights using the provided data and make prediction models to forecast the sales the next twelve weeks.

Data Description

The available dataset contains 6,435 records (rows) and eight features (columns) as shown in the Table below:

| 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 |

Data description, various insights from the data.

From the given dataset of the company, it is observed that the data consists of six thousand four hundred and thirty-five (6,435) records with seven features (captured weekly) as follows:

1. 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 Pre-processing Steps and Inspiration

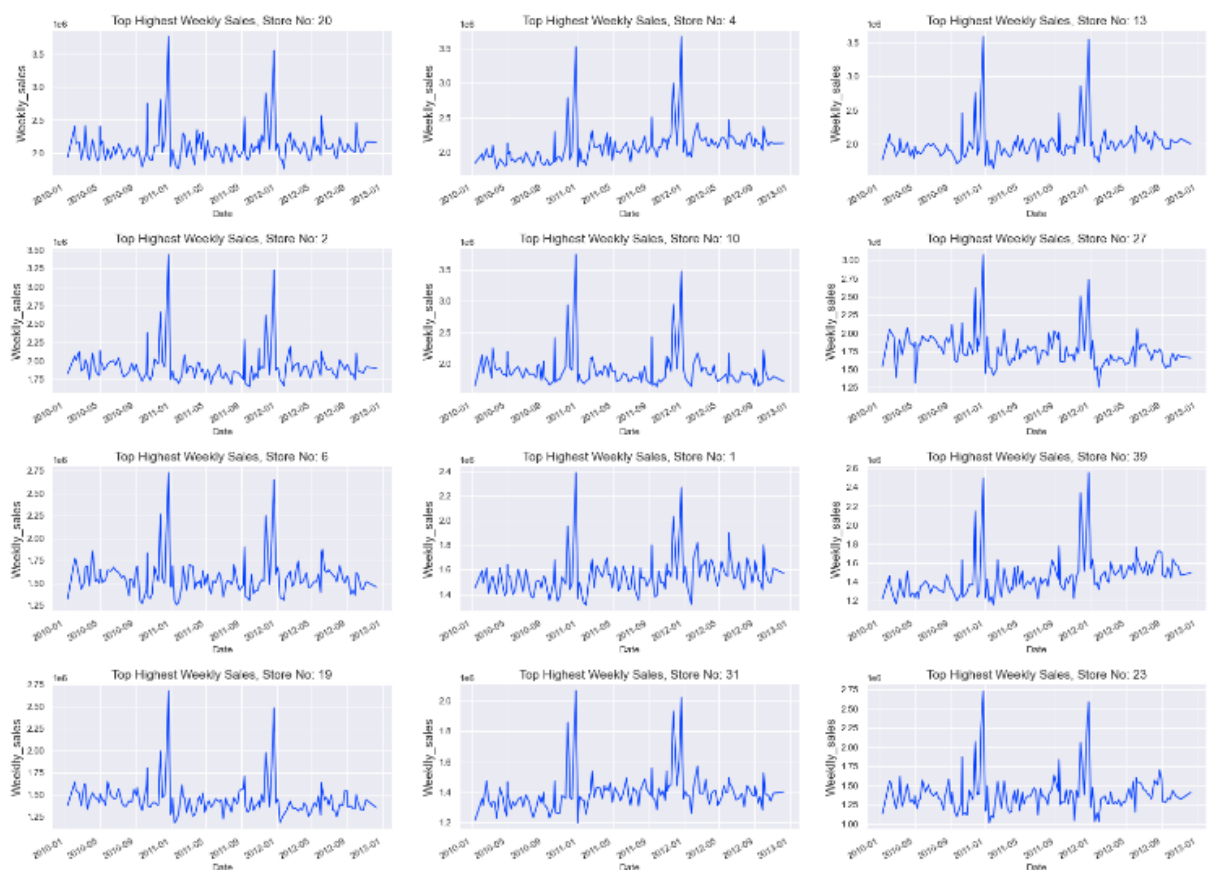
The preprocessing of the data included the following steps:

1. Step 1: Load data
2. 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. Check for outliers that are known to distort predictions and forecasts

3. Step 3: Model Predictions, two approaches:
 - a. Time Series Model (ARIMA)
 - b. Linear Regression Model(s)
4. Step 4: Forecast
5. Step 5: Compare Results from the different models.

Model Evaluation and Technique

Top Weekly Sales against Time



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 (ARIMA, SERIMA, SERIMAX) 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

The ARIMA model:

Autoregressive Integrated Moving Average (ARIMA) is defined as a statistical analysis model that uses [time series data](#) to either better understand the data set or to predict future trends ([Autoregressive Integrated Moving Average \(ARIMA\) Definition\(investopedia.com\)](#)).

A statistical model is autoregressive if it predicts future values based on past values.

ARIMA model is based on a number of assumptions including:

1. Data does not contain anomalies,
2. Model parameters and error term is constant,
3. Historic timepoints dictate behaviour of present timepoints,
4. Time series is stationary

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/adaboost-algorithm-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 is 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-linear-regression-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:
- i. **Linear Relationship:** The relationship between the independent and dependent variables should be linear.
 - ii. **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.
 - iii. **No Multicollinearity:** There is little or no multicollinearity in the data which means that the independent variables should have minimal correlation with each other.

- iv. **No Autocorrelation:** There is little or no autocorrelation in the data where the data values of the same column are related to each other.
 - v. **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>).

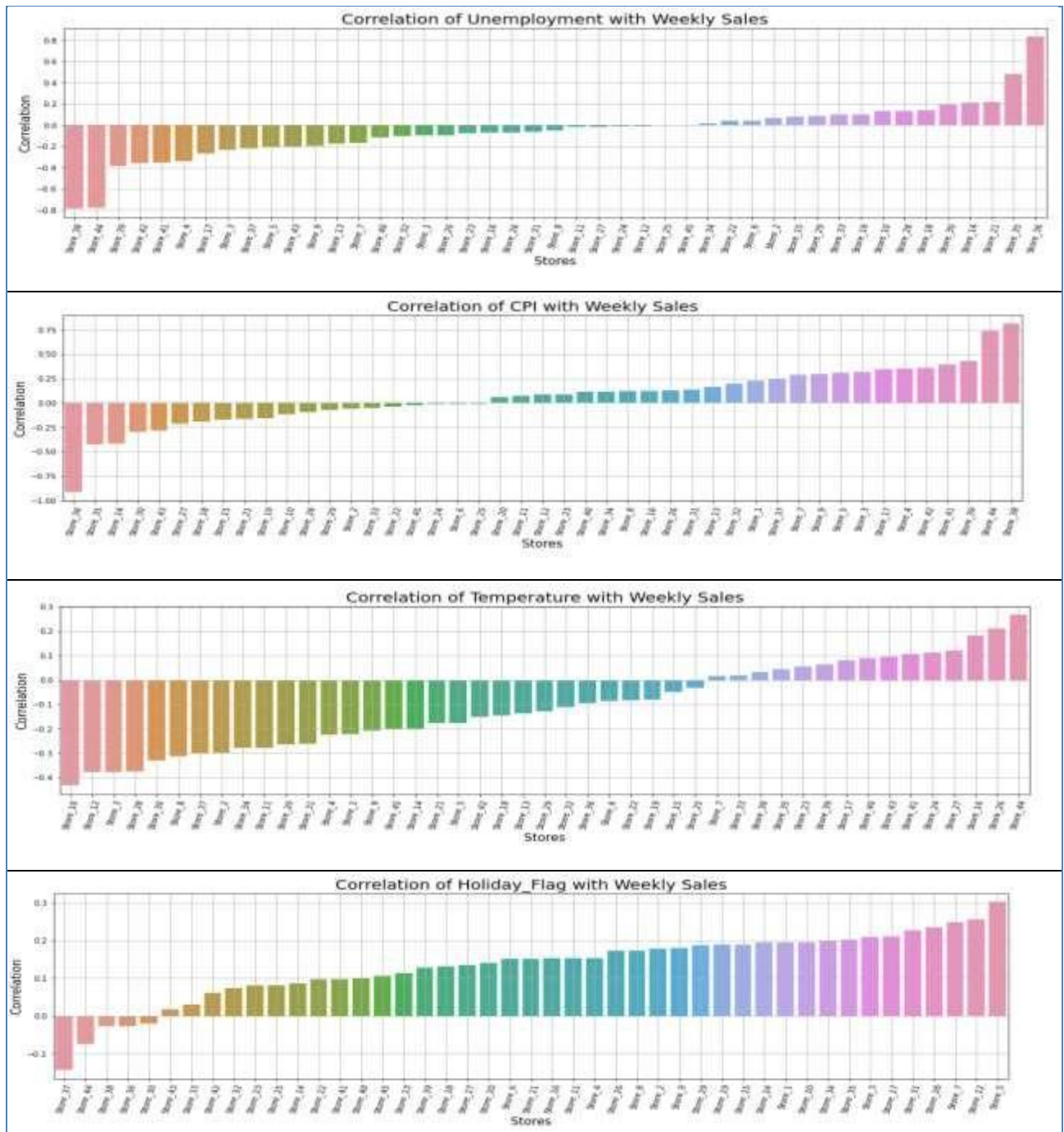
Model Evaluation:

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. For example, as

shown above, the effects of unemployment vary by the stores whereas it appears to have positive effects on some and negative effects on others. The same is also true for Temperature, CPI and to some extent, the Holiday_Flag.

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, ARIMA.

- The first step for this model is to check the stationarity of the dataset (p-value less than 0.05).
- Next is to find the best ARIMA order for the model
- Using the best ARIMA order, make predictions for the selected stores.
- Forecast using SERIMAX
 - Detrend the dataset, if necessary,
 - Using SERIMAX estimate a 12-weeks weekly sales forecast

2. Regression Models:

- Regression model, Gradient_Boosting, Linear_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

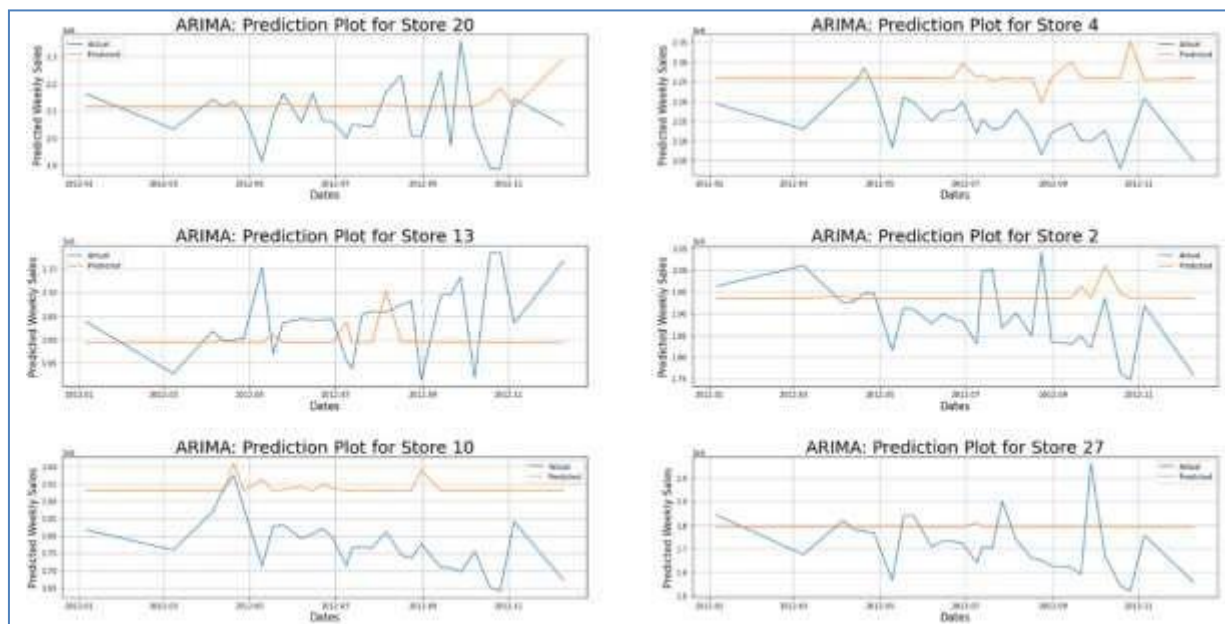
Model Results:

1. ARIMA model:

a. Predictions:

Predictions were performed for six stores (stores: 20, 4, 13, 2, 10 and 27 in order of decreasing Weekly_Sales sales revenue). The predictions results are summarized in the Table and graphs below:

| | Store_20 | Store_4 | Store_13 | Store_2 | Store_10 | Store_27 |
|-------------------------|----------|---------|----------|---------|----------|----------|
| Median Error (%) | 3.42 | 5.83 | 2.90 | 3.28 | 9.35 | 5.14 |
| Mean Error (%) | 4.80 | 5.37 | 3.57 | 3.94 | 9.18 | 6.94 |



b. Forecast:

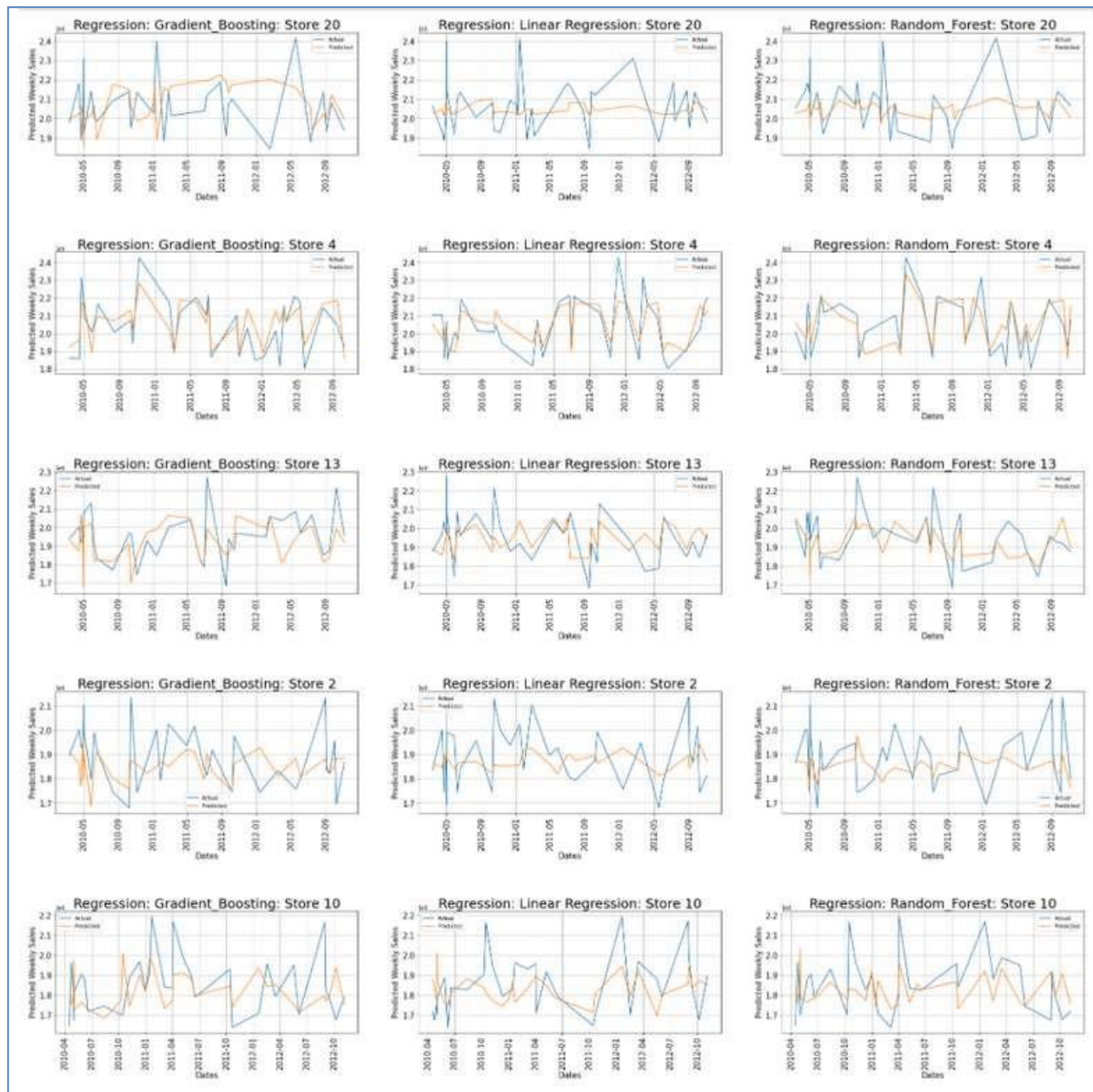
The initial results of the forecast (see plots below) are not very good showing evidence of noise which maybe as a result of the trends, and

the observed outliers in the dataset which are distorting the forecasts. As a result, the dataset was detrended and the forecast repeated.

The forecast after detrending shows the anticipated variabilities as observed in the Weekly_Sales history. However, the overall projected sales outlook for the next 12-weeks is down for all of the stores studied!

2. Regression Models:

The prediction results from the three chosen regression models: Gradient_Boosting, LinearRegression, and Random_Forest are summarized in the chart below:



Model Evaluation:

1. ARIMA/SERIMAX Models

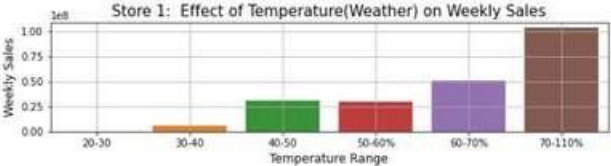
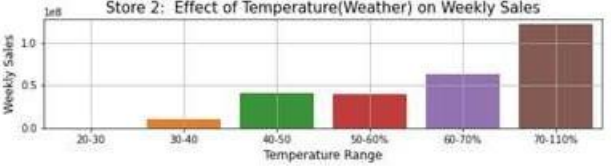
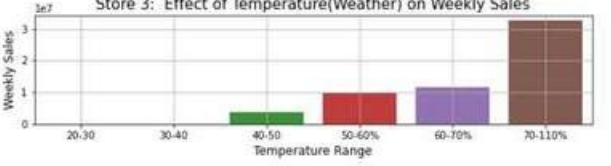
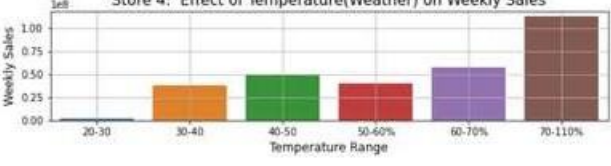
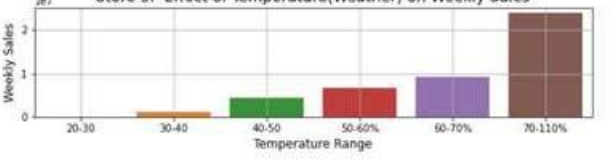
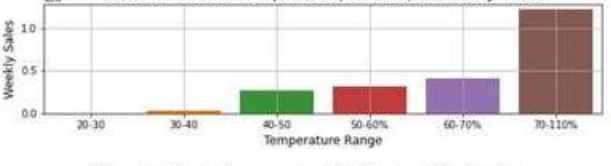
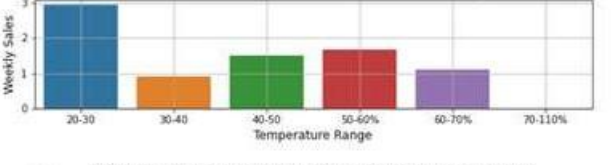
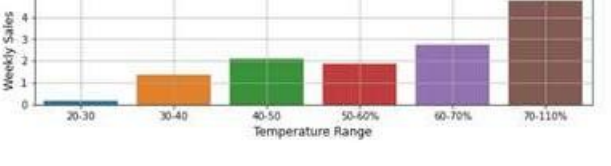
The model predictions for the selected stores were okay. The forecast after detrending was also okay showing variabilities of the weekly sales inline with the sales history.

2. The Regression Models: The regression model results are summarized below (Table & chart):

| | regressor_name | rmse |
|----|--------------------------------|--------------|
| 0 | Random Forest Regression | 1.473277e+05 |
| 1 | Boosted Tree Regression | 1.865208e+05 |
| 2 | Decision Tree Regression | 1.913309e+05 |
| 3 | K-Nearest Neighbour Regression | 3.750646e+05 |
| 4 | Spline Regression | 4.704331e+05 |
| 5 | Polynomial Regression | 4.841134e+05 |
| 6 | Ridge Regression | 5.231625e+05 |
| 7 | Lasso Regression | 5.231627e+05 |
| 8 | Linear Regression | 5.231628e+05 |
| 9 | Elastic Net Regression | 5.280729e+05 |
| 10 | Support Vector Regression | 5.687756e+05 |
| 11 | Neural Network Regression | 1.186783e+06 |

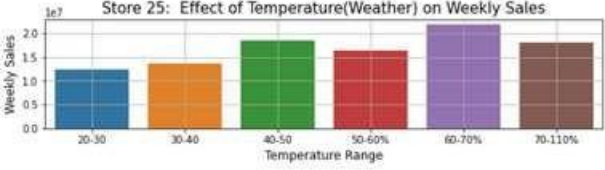
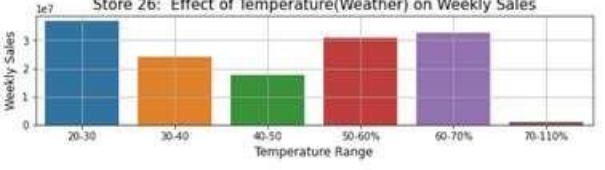
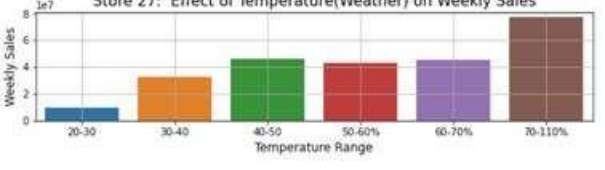
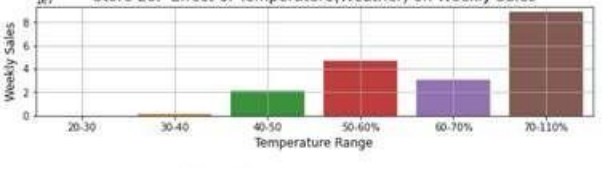
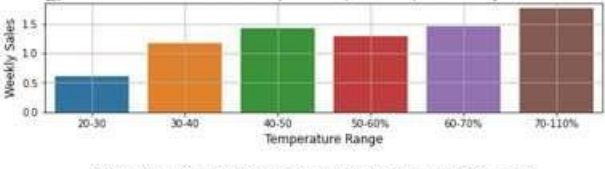
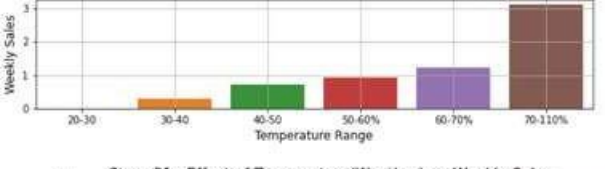
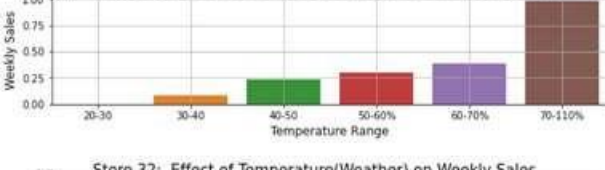
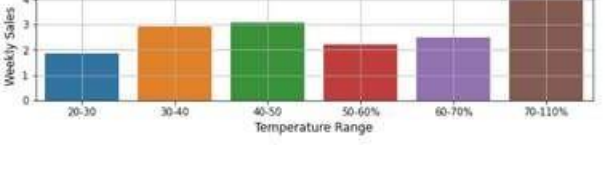
The above table shows that Random Forest Regressor outperformed all the regressors with RMSE of 1.17e+05. This provides a good estimate for future sales as it has about 12% average error compared to the typical median sale.

Temperature Effects on Weekly Sales: evaluation of how changes in temperature effects the weekly sales revenue is presented below:

| Store | Outlook – Recommendation(s) | |
|-------|--|--|
| 1 |  <p>Store 1: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 2 |  <p>Store 2: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 3 |  <p>Store 3: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 4 |  <p>Store 4: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 5 |  <p>Store 5: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 6 |  <p>Store 6: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 7 |  <p>Store 7: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in winter cold weather months – must shore-up inventory for winter</i> |
| 8 |  <p>Store 8: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |

| Store | Outlook – Recommendation(s) | |
|-------|---|--|
| 9 | <p>Store 9: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 10 | <p>Store 10: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 11 | <p>Store 11: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 12 | <p>Store 12: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 13 | <p>Store 13: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 14 | <p>Store 14: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 15 | <p>Store 15: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 16 | <p>Store 16: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in winter cold weather months – must shore-up inventory for winter</i> |

| Store | Outlook – Recommendation(s) | |
|-------|---|--|
| 17 | <p>Store 17: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in winter cold weather months – must shore-up inventory for winter</i> |
| 18 | <p>Store 18: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months (Fall & Summer) – must shore-up inventory for warm weather</i> |
| 19 | <p>Store 19: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months (Fall) – must shore-up inventory for warm weather</i> |
| 20 | <p>Store 20: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months (Summer) – must shore-up inventory for summer</i> |
| 21 | <p>Store 21: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months (Summer) – must shore-up inventory for summer</i> |
| 22 | <p>Store 22: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for summer</i> |
| 23 | <p>Store 23: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months – must shore-up inventory for warm weather</i> |
| 24 | <p>Store 24: Effect of Temperature(Weather) on Weekly Sales</p> | <i>Most sales in summer warm weather months (Fall & Summer) – must shore-up inventory for warm weather</i> |

| Store | Outlook – Recommendation(s) | |
|-------|---|--|
| 25 |  <p>Store 25: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Mild sales in very cold weather Significant sales in cold and warm weather Most sales in warm weather <p>→ Must shore-up inventory for cold and warm weather</p> |
| 26 |  <p>Store 26: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> No sales activity in hot weather Significant sales in warm weather Most sales in winter cold weather <p>→ Must shore-up inventory for cold and warm weather</p> |
| 27 |  <p>Store 27: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Mild sales in very cold weather, Moderate sales in cold and warm weather, Most sales in hot weather months (summer) <p>→ Must shore-up inventory for hot weather</p> |
| 28 |  <p>Store 28: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> No sales activity in very cold weather Moderate sales in warm weather Most sales in summer hot weather months (Summer) <p>→ Must shore-up inventory for hot weather</p> |
| 29 |  <p>Store 29: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Mild sales in very cold weather Significant sales in cold and warm weather Most sales in hot weather <p>→ Must shore-up inventory for cold and warm weather</p> |
| 30 |  <p>Store 30: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> No sales activity in very cold weather (winter) months Mild sales in early winter (cold weather) months Most sales in hot weather months (summer) <p>→ Must shore-up inventory for hot weather</p> |
| 31 |  <p>Store 31: Effect of Temperature(Weather) on Weekly Sales</p> | <p>→ Must shore-up inventory for hot weather</p> |
| 32 |  <p>Store 32: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Mild to moderate sales in cold and warm weather Most sales in summer hot weather months (Summer) <p>→ Must shore-up inventory for hot weather</p> |

| Store | Outlook – Recommendation(s) | |
|-------|---|---|
| 33 | <p>Store 33: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> No sales activity in cold weather (winter) months Mild sales activity in warm weather Most sales in summer hot weather months (summer) <p>→ Must shore-up inventory for hot weather</p> |
| 34 | <p>Store 34: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Mild to moderate sales activity in cold weather (winter) months Moderate sales activity in Fall Most sales in summer hot weather months (Fall) <p>→ Must shore-up inventory for hot weather</p> |
| 35 | <p>Store 35: Effect of Temperature(Weather) on Weekly Sales</p> | |
| 36 | <p>Store 36: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> No sales activity in very cold weather (winter) months Moderate sales activity in Fall Most sales in summer hot weather months (Fall) <p>→ Must shore-up inventory for hot weather</p> |
| 37 | <p>Store 37: Effect of Temperature(Weather) on Weekly Sales</p> | |
| 38 | <p>Store 38: Effect of Temperature(Weather) on Weekly Sales</p> | |
| 39 | <p>Store 39: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Significant sales activity in very cold weather (winter) months Moderate sales in early winter & late Fall Most sales activity in warm weather (early summer) months <p>→ Must shore-up inventory for cold and warm weather</p> |
| 40 | <p>Store 40: Effect of Temperature(Weather) on Weekly Sales</p> | |

| Store | Outlook – Recommendation(s) | |
|-------|---|---|
| 41 | <p>Store 41: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Moderate sales in hot & warm weather months Most sales in cold weather months <p>→ Must shore-up inventory for cold and warm weather months</p> |
| 42 | <p>Store 42: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> No sales activity in cold weather (winter) months Moderate sales activity in warm weather (Fall) Most sales in summer hot weather months (summer) <p>→ Must shore-up inventory for hot (summer) weather</p> |
| 43 | <p>Store 43: Effect of Temperature(Weather) on Weekly Sales</p> | <p>→ Must shore-up inventory for hot (summer) weather</p> |
| 44 | <p>Store 44: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Mild sales activity in very cold weather (winter) months Moderate sales activity in warm weather (Fall) and early winter months |
| 45 | <p>Store 45: Effect of Temperature(Weather) on Weekly Sales</p> | <ul style="list-style-type: none"> Most sales in summer warm weather months <p>→ Must shore-up inventory for hot, warm and cold weather months</p> |

Future Possibilities

One area that future studies could explore is the relationship between festive sales and profit margins. By augmenting the dataset with expenses data, it would be possible to investigate whether larger festive sales always translate to larger profit margins. This can inform decisions about marketing and pricing strategies.

- The analysis showed a 500% difference in sales between the top performing and lowest performing stores, which is a significant difference. This suggests that there may be underlying factors contributing to the performance of these stores. To better understand the reasons behind the performance of these stores, it is necessary to gather additional data and parameters that may be influencing the sales of the top selling products.

- Hyperparameter tuning involves adjusting the parameters of a model in order to improve its performance on a given dataset. By iteratively adjusting the parameters of the best model, it is possible to achieve an even better model.

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 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 through 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, tradeshow, 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. [Kaggle](https://www.kaggle.com/datasets/yasserh/walmart-dataset)
2. - [B2B International (2018). Sales Forecasting: The Importance and Benefits.] (https://www.b2binternational.com/resources/blog/sales-forecasting-importance-benefits/)
3. - [Business News Daily (2021). Sales Forecasting: The Importance of Accurate Sales Forecasts.] (https://www.businessnewsdaily.com/5659-sales-forecasting.html)
4. - [Small Business Administration (2021). The Importance of Sales Forecasting.] (<https://www.sba.gov/blogs/importance-sales-forecasting>).

