**Walmart Sales Forecasting**

**Introduction**

***Background***

Walmart is one of the biggest retail companies worldwide, with many stores in the U.S. and around the globe. The company sees changes in sales for several reasons, including seasonal trends, the state of the economy, special promotions, and external factors like fuel prices and unemployment rates. Getting sales forecasts right is essential for better managing inventory, streamlining the supply chain, and making strategic choices. By examining past sales data, Walmart can understand sales trends, recognize what affects them, and create plans to handle the risks that come with unexpected changes in demand.

***Motivation***

Sales forecasting is an important aspect of retail management. By predicting future sales accurately, businesses can reduce overstocking and understocking issues, allocate resources more efficiently, and improve overall profitability. For Walmart, effective sales forecasting helps in optimizing store operations, managing workforce requirements, and enhancing customer satisfaction. A robust predictive model will not only assist in decision-making but also allow the company to respond proactively to changing market conditions.

With the rise of data-driven decision-making, employing machine learning models to predict sales trends has become an industry standard. Companies like Walmart invest in analytics to enhance their competitive advantage, improve customer experience, and increase profitability. Predictive analytics in sales forecasting allow businesses to make informed decisions regarding product assortment, pricing strategies, and marketing campaigns. This study aims to leverage historical sales data to develop a predictive model that can help Walmart anticipate weekly sales trends and make informed business decisions.

***Research Question***

The research question for this study is: "Can historical sales data and external economic factors be used to accurately predict Walmart’s weekly sales?" This question will be explored by analyzing various independent variables such as holiday periods, temperature fluctuations, fuel prices, CPI, and unemployment rates to determine their impact on weekly sales. The objective is to build a predictive model that can forecast sales with high accuracy and provide insights into the key factors influencing sales trends. Furthermore, this study seeks to understand which variables have the most significant influence on sales performance and how Walmart can use this information to optimize inventory and pricing strategies.

**Data Summary**

***Data Source***

The data set used in this study is obtained from [Kaggle’s Walmart Sales Forecasting dataset](https://www.kaggle.com/datasets/yasserh/walmart-dataset/data). This dataset comprises historical sales data from multiple Walmart stores across different locations in the United States. It includes weekly sales records along with various external economic and environmental factors that could impact sales performance. The dataset has been widely used in time-series forecasting and machine learning applications to understand sales trends and predict future revenue streams.

***Data Variable Definition***

The dataset includes multiple variables that can be classified as follows:

**Dependent Variable (Target Variable):**

* **Weekly\_Sales:** Represents the weekly revenue generated by Walmart stores. This is the variable we aim to predict.

**Independent Variables (Predictors):**

1. **Date:** Represents the time component, essential for trend and seasonal analysis.
2. **Holiday\_Flag:** A binary variable indicating whether the week includes a holiday (1 = Holiday, 0 = No Holiday). Special holidays often influence consumer spending habits.
3. **Temperature:** Measures the weather conditions, which may affect consumer shopping behavior.
4. **Fuel\_Price:** Represents the price of gasoline, which could influence consumer travel and purchasing power.
5. **CPI (Consumer Price Index):** A measure of inflation that reflects changes in purchasing power and cost of living.
6. **Unemployment:** Represents the unemployment rate, which could impact consumer spending patterns.

***Statistical Summary***

Table 1 presents the descriptive statistics for key variables in the dataset, including Weekly Sales, Temperature, Fuel Price, Consumer Price Index (CPI), and Unemployment Rate. The average Weekly Sale across all stores is $1,046,964.88, with a standard deviation of $564,366.62, indicating significant variability in sales performance. The minimum recorded weekly sales are $209,986.25, while the maximum reaches $3,818,686.45, suggesting a wide difference in sales across different stores and time periods. The median weekly sales value of $960,746.04 suggests that half of the stores report sales below this amount, while the other half exceeds it.

Regarding environmental and economic factors, the average temperature is 60.66°F, with a standard deviation of 18.44°F, reflecting different climatic conditions across store locations. The minimum recorded temperature is -2.06°F, while the maximum is 100.14°F, indicating the broad range of weather conditions affecting store operations. The Fuel Price has a mean of $3.36 per gallon, with a standard deviation of $0.46, suggesting moderate price fluctuations. The minimum recorded fuel price is $2.47, while the maximum reaches $4.47 per gallon.

The CPI (Consumer Price Index), which represents inflation trends, has a mean value of 171.58, with recorded values ranging from 126.06 to 227.23. This range reflects the changes in consumer purchasing power over time. Similarly, the Unemployment Rate averages 7.99%, with a standard deviation of 1.88%, indicating some variation in labor market conditions. The unemployment rate fluctuates between 3.88% and 14.31%, highlighting the economic cycles that may influence consumer spending patterns.

**Table 1: Descriptive Statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Weekly Sales** | **Temperature** | **Fuel Price** | **CPI** | **Unemployment** |
| **count** | 6435 | 6435 | 6435 | 6435 | 6435 |
| **mean** | 1046964.878 | 60.66378244 | 3.358606838 | 171.5783938 | 7.999151049 |
| **std** | 564366.6221 | 18.44493288 | 0.459019707 | 39.3567123 | 1.875884782 |
| **min** | 209986.25 | -2.06 | 2.472 | 126.064 | 3.879 |
| **25%** | 553350.105 | 47.46 | 2.933 | 131.735 | 6.891 |
| **50%** | 960746.04 | 62.67 | 3.445 | 182.6165205 | 7.874 |
| **75%** | 1420158.66 | 74.94 | 3.735 | 212.7432935 | 8.622 |
| **max** | 3818686.45 | 100.14 | 4.468 | 227.2328068 | 14.313 |

**Correlation Matrix**

Table 2 presents the Pearson correlation coefficients between Weekly Sales and the independent variables. The correlation between Weekly Sales and Unemployment Rate is negative and weak (r = -0.11), suggesting that as unemployment increases, consumer spending declines slightly, leading to lower sales. Similarly, Temperature has a weak negative correlation with Weekly Sales (r = -0.06), indicating that colder weather may slightly reduce sales, potentially due to decreased shopping activity. Fuel Price shows a near-zero positive correlation with Weekly Sales (r = 0.01), suggesting that fluctuations in fuel prices have little to no direct impact on Walmart’s revenue. In contrast, CPI has a weak negative correlation with Weekly Sales (r = -0.07), implying that as the cost of goods and services increases, weekly sales tend to decline slightly. Additionally, the correlation between CPI and Unemployment Rate is moderately negative (r = -0.30), indicating that higher unemployment rates are generally associated with lower inflation levels.

**Table 2: Correlation Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Weekly Sales** | **Temperature** | **Fuel Price** | **CPI** | **Unemployment** |
| **Weekly Sales** | 1 |  |  |  |  |
| **Temperature** | -0.0638 | 1 |  |  |  |
| **Fuel Price** | 0.0095 | 0.1450 | 1 |  |  |
| **CPI** | -0.0726 | 0.1769 | -0.1706 | 1 |  |
| **Unemployment** | -0.1062 | 0.1012 | -0.0347 | -0.3020 | 1 |

**Data Cleaning Process**

Before proceeding with analysis, data cleaning was necessary to ensure accuracy and consistency. The following steps were undertaken:

1. **Removing Outliers:** Outliers were identified using the **Z-score method**, and extreme values (greater than 3 standard deviations from the mean) were removed to prevent distortions in the predictive model.
2. **Encoding Categorical Variables:** The ***Holiday Flag***variable, which indicates whether a given day is a holiday (1) or not (0), was converted into an integer format to ensure proper inclusion in the regression model.

**Methodology and Results**

**Model Selection**

The research question focuses on predicting Walmart’s sales, making regression models an appropriate choice. A Linear Regression model was selected for Walmart sales forecasting due to its simplicity, interpretability, and efficiency in handling structured numerical data. This model helps establish a direct relationship between the target variable, weekly sales, and multiple influencing factors, such as holiday periods, temperature fluctuations, fuel prices, the Consumer Price Index (CPI), and unemployment rates. The choice of linear regression is justified by its ability to provide clear insights into how each predictor impacts sales. The dependent variable (Weekly Sales) was regressed against independent variables such as Temperature, Fuel Price, CPI, and Unemployment.

The linear regression model follows the equation:

**Model Results**

Table 3: Regression Result

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dependent variable: | | |
|  | Weekly Sales |  | |  |
| CPI | 952.469\*\*\* | |  | |
|  | (175.660) |  | |  |
| Fuel Price | 252449.710\*\*\* | |  | |
|  | (11954.658) | |  | |
| Holiday Flag | 117551.176\*\*\* | |  | |
|  | (32323.210) | |  | |
| Temperature | -501.224 |  | |  |
|  | (463.707) |  | |  |
| Unemployment | 5666.446 |  | |  |
|  | (3758.690) | |  | |
| Observations | 5148 |  | |  |
| R2 | 0.764 |  | |  |
| Adjusted R2 | 0.763 |  | |  |
| F Statistic | 3323.619\*\*\* | |  | |
| Note: | \*p<0.05; \*\*p<0.01; \*\*\*p<0.001 | | |

**Interpretation of Results**

Table 3 presents the results of the linear regression analysis examining the relationship between Weekly Sales and the predictor variables CPI, Fuel Price, Holiday Flag, Temperature, and Unemployment Rate. The model demonstrates a strong explanatory power, with an R² of 0.764, indicating that approximately 76.4% of the variance in Weekly Sales is explained by the independent variables. The adjusted R² is 0.763, confirming the model’s robustness. The overall model is statistically significant (F = 3323.619, p < .001), suggesting that the predictors collectively influence weekly sales.

The CPI (B = 952.469, p < .001) has a statistically significant positive impact on Weekly Sales, indicating that for every one-unit increase in the CPI, weekly sales increase by approximately $952.47. Similarly, Fuel Price (B = 252,449.710, p < .001) has a strong positive relationship with Weekly Sales, suggesting that higher fuel prices are associated with increased sales.

The Holiday Flag (B = 117,551.176, p < .001) is also significantly positive, implying that weekly sales experience a substantial increase of approximately $117,551.18 during holiday periods compared to non-holiday weeks.

Temperature (B = -501.224, p = 0.281) does not have a statistically significant effect on Weekly Sales, suggesting that fluctuations in temperature do not strongly impact sales in this dataset. Similarly, Unemployment (B = 5,666.446, p = 0.129) is not statistically significant, indicating that changes in unemployment rates do not have a strong direct effect on weekly sales.

**Model Evaluation**

To assess model performance, we calculated the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):

* **MAE** = 469043.07
* **RMSE** = 569992.55

The MAE is 469,043.07, indicating that, on average, the model's predictions deviate from actual sales figures by approximately $469,043. The RMSE is 569,992.55, suggesting that the model's prediction error is around $569,993.

**Conclusion and Future Works**

**Conclusion**

This study analyzed the factors influencing Walmart's weekly sales using a linear regression model. The findings indicate that CPI, Fuel Price, and Holiday periods significantly impact weekly sales, while Temperature and Unemployment Rate do not have a statistically significant effect. The model demonstrates strong explanatory power, with an R² of 0.764, meaning that approximately 76.4% of the variation in sales is explained by the independent variables. Model evaluation metrics, such as MAE (469,043.07) and RMSE (569,992.55), indicate that while the model is reasonably accurate, prediction errors remain substantial. Conclusively, the results suggest that economic conditions and holiday periods are key drivers of sales performance at Walmart.

**Limitations**

Despite the model’s strong predictive performance, several limitations should be acknowledged:

1. The model does not incorporate factors such as store-specific promotions, competitor pricing, or consumer sentiment, which could further enhance prediction accuracy.
2. The study assumes a linear relationship between the independent variables and weekly sales, potentially overlooking complex interactions and non-linear effects.
3. The dataset may not capture all external economic shocks, such as sudden policy changes or supply chain disruptions, which could influence sales patterns.

**Future Works**

To improve the robustness and applicability of sales forecasting models, future research could explore the following:

1. Techniques such as Random Forest, Gradient Boosting, or Neural Networks could be used to capture non-linear relationships and improve predictive accuracy.
2. Future studies could include external influences such as marketing expenditures, competitor strategies, and social media trends to enhance prediction models.