**Walmart Sales Prediction**

by Team Ninjas

Team Members

Priyanka Gandhi Porika, Punith Kukkala, Sandeep Reddy Yasa, Swetha Nallabollu.

Submitted to Professor Mohammad Nasim

Northwood University in partial fulfillment of the requirements of

MIS-658 -MSBA March 17th, 2024

Midland, Michigan

**Walmart Sales Prediction**

**Strategic Forecasting for Enhanced Retail Performance**

**Problem statement**

The problem statement for the Walmart Sales Prediction project is to develop a sophisticated forecasting model to enhance the accuracy of sales predictions for Walmart. The team aims to delve into the nuances of retail sales forecasting by considering various factors that could potentially affect sales outcomes, such as store type, size, and key performance indicators (KPIs). By leveraging an ARIMAX model, the team seeks to handle complex seasonal patterns and incorporate external variables to provide a comprehensive analysis of sales determinants. The project further aims to measure the model's performance against key metrics such as the Akaike Information Criterion (AIC), the Schwarz Bayesian Criterion (SBC), and the Mean Absolute Percentage Error (MAPE) to observe improvements over traditional ARIMA models. Additionally, the team aims to assess the impact of store type and size on total and weekly sales, identify significant KPIs influencing Walmart's sales, and develop strategic promotional tactics for top-performing stores by considering factors like promotional campaigns, temperature, and the Consumer Price Index (CPI).

The clarity and significance of the defined problem in the Walmart Sales Prediction project are evident in its focused objective, relevance to retail forecasting, and potential impact on operational efficiency and strategic decision-making within the organization.

**Data Analysis**

**The Dataset for Modelling**

**Name: Forecasting Walmart Sales**

**Data source: Kaggle**

**Data link:** [**Walmart Sales Forecast (kaggle.com)**](https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast?select=train.csv)

**Overview:**

The dataset utilized for modeling comprises historical sales data from Walmart stores, spanning from May 2, 2010, to July 26, 2013. It encompasses 8,190 records across 45 Walmart stores, each containing 12 features. This dataset is crucial for understanding sales patterns and drivers within Walmart's retail environment.

**Structure of the Data:**

The dataset is structured in CSV format, facilitating easy access and manipulation. Each record represents weekly sales alongside associated factors, including temperature, fuel price, and holiday status. These factors play pivotal roles in influencing consumer behaviour and overall sales performance. However, it is worth noting potential limitations such as missing data points or outliers, which necessitate preprocessing for accurate modeling.

**Key Variables:**

1. **Weekly Sales:** This serves as the target variable, depicting the total sales for a given week across all Walmart stores.
2. **Temperature:** Represents the average temperature during the week, impacting shopping patterns and consumer behaviour.
3. **Fuel Price:** Indicates the average fuel price prevailing during the week, affecting both operational costs and consumer spending power.
4. **Holiday Status:** An indicator denoting whether the week encompasses a major holiday, which typically correlates with fluctuations in sales volumes.
5. **Markdown1 to Markdown5:** Columns indicating different promotional activities, providing insights into the impact of promotions on sales.
6. **Consumer Price Index (CPI):** Reflects changes in the price level of consumer goods and services, influencing consumer spending habits and purchasing power.

**Preprocessing for Modelling:**

**1. Aggregation of Daily Sales Data:**

Given the granularity of daily sales data, it was aggregated to weekly sales figures to streamline the modelling process. This decision mitigates the potential challenge of handling excessive data points.

**2. Summation of Departmental Sales:**

Each store encompasses approximately 98 departments, necessitating the summation of departmental sales to obtain weekly aggregate data. This aggregation simplifies the modelling process while retaining essential insights into overall store performance.

**3. Store Categorization and Exploration:**

The dataset includes 45 stores categorized as Store A, B, and C. Enhance model specificity, exploration focuses on 2-3 stores from each category separately. This approach facilitates a comparative analysis of modelling results and parameters across different store types.

**4. Consideration of Seasonality:**

Recognizing the presence of seasonality within the data, the modelling process incorporates only the first two years of data (2010-2012). This selective inclusion aims to capture seasonal patterns effectively while minimizing noise from subsequent years.

**5. Data Preparation and Exploration:**

Utilization of SAS for Data Cleaning and Separation:

Data cleaning and separation were conducted using SAS, a powerful statistical analysis tool. This process yielded separate datasets for each store, enabling focused exploration and analysis tailored to individual store characteristics.

We chose the following stores from each type of store to Forecast find the correlations:

|  |  |  |
| --- | --- | --- |
| Store | Type | Size |
| 1 | A | 151315 |
| 11 | A | 207499 |
| 24 | A | 203819 |
| 31 | A | 203750 |
| 3 | B | 37392 |
| 12 | B | 112238 |
| 22 | B | 119557 |
| 37 | C | 39910 |
| 43 | C | 41062 |

**Store wise Weekly Sales Exploration:**

**Avg. Sales by Store Type & Holiday Status**

**Monthly Avg. Sales by Store Type**

Type A stores consistently boast higher sales figures compared to Type B establishments. Type B, in turn, outperforms Type C in terms of sales metrics. This hierarchy underscores the significant impact of store type on revenue generation within the retail landscape.

One notable factor contributing to the sales differentials between Type A, B, and C stores is the presence of seasonality. Both Type A and B stores exhibit discernible patterns of seasonal fluctuations, wherein sales tend to peak during certain periods of the year and dip during others. This seasonality often corresponds with consumer behaviour influenced by factors such as holidays, weather changes, or cultural events.

In contrast, Type C stores lack this pronounced seasonality in their sales patterns. Their revenue streams remain stable throughout the year, without experiencing significant peaks or troughs tied to seasonal variations. This distinction highlights a fundamental divergence in the operational dynamics and consumer appeal of Type C stores compared to their Type A and B counterparts.

During holidays, sales surge across all types of stores. Whether they are Type A, B, or C, establishments experience increased consumer spending driven by the festive spirit.

# Approach to Forecasting the Weekly Sales Data – Store 3 (Store Type – B)Time Series Exploration for Individual Variables

## Dependent Variable – Weekly Sales

A graph showing a line graph

Description automatically generated with medium confidenceA graph showing the seasons

Description automatically generated with medium confidence

A graph showing the sales of a company

Description automatically generated with medium confidenceA table with numbers and lines

Description automatically generated

Based on the Time-series exploration of Weekly Sales and Dickey-Fuller Unit Root Test, our data is a stationary dataset and has seasonal component without any trend component.

When we look at the Correlation functions of the Weekly Sales, ACF has lagged effect of up to lag3 and dropped suddenly and PACF has spikes at Lag1 and lag5 and the data is not white noise.

The ACF (Autocorrelation Function) shows a lagged effect up to lag 3 before dropping suddenly. This suggests that there may be some short-term autocorrelation in the data up to lag 3.

The PACF (Partial Autocorrelation Function) has spikes at lag 1 and lag 5, indicating direct relationships with the previous observation and the observation five periods ago. This suggests that there may be some significant short-term and longer-term autocorrelation in the data.

A graph of sales

Description automatically generated with medium confidence

Weekly Sales also have a strong seasonal component:

A graph showing the fall of sales

Description automatically generated

Our Data has some independent variables named Temperature, Fuel price, CPI, Unemployment, IsHoliday and different promotional functions named Markdown1, Markdown2, Markdown3, Markdown4 and Markdown5.

We explored the cross correlations of these variables with our target variable. Below are the cross-correlation plots.

A graph of a graph showing a line of a graph

Description automatically generated with medium confidence A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generatedA graph with a line

Description automatically generated with medium confidence

A graph with a line graph

Description automatically generated with medium confidence A graph with blue lines

Description automatically generated

A graph of a graph

Description automatically generated A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated A graph with a bar graph

Description automatically generated with medium confidence

Based on the Cros-Correlations, we can see that Temperature, Markdown1, Markdown3, Markdown4 and Markdown5, IsHoliday have some effect on our target variable.

**Model Implementation**

The implementation of machine learning models in the Walmart Sales Prediction project showcases a high level of quality and appropriateness in algorithm selection. Here are the key points highlighting the quality of model implementation and the appropriate use of algorithms.

**Building ARIMA Models:**  
  
**ARIMA (5,0,3) (0,1,0)** has AIC and SBC values as 1205.8 and 1223.3 respectively with normally distributed residual with some signal.

A screenshot of a graph

Description automatically generated

A diagram of a normality curve

Description automatically generated

**ARIMA (1,0,1) (0,1,0)** has AIC and SBC values as 1191.048 and 1196.902 respectively with normally distributed residual with some signal.

A group of graphs showing different types of data

Description automatically generated with medium confidence

A diagram of a normality

Description automatically generated

**ARIMA (0,0,0) (0,1,0)** has AIC and SBC values as 1188.386 and 1190.338 respectively with normally distributed residual with some signal.

nA group of graphs showing different types of data

Description automatically generated with medium confidence

A graph of normality and a normality curve

Description automatically generated with medium confidence

1. Algorithm Selection:

- ARIMAX Model: The project utilizes the ARIMAX model to handle complex seasonal patterns and incorporate external variables, aligning with the objective of capturing sales data dynamics considering store-specific factors.

- ARIMA Modeling: ARIMA models are employed to capture time series patterns in sales data effectively, demonstrating a sound approach to forecasting based on dataset seasonality and trends.

2. Model Quality:

- Performance Metrics: Evaluation metrics like AIC, SBC, MAPE, MAE, MSE, and RMSE are used to rigorously assess model accuracy and reliability, providing insights into predictive capabilities and forecast accuracy.

- Cross-Correlations and Exogenous Variables: The ARIMAX model considers cross-correlations with exogenous variables, enhancing the analysis of factors influencing sales outcomes for more accurate predictions.

3. Model Complexity and Interpretability:

- Seasonality and Trend Analysis: The models analyze seasonality and trends in the dataset, utilizing different ARIMA models based on autocorrelation and partial autocorrelation functions to capture time series patterns effectively.

- Variable Selection: ARIMAX models are built with relevant independent variables like temperature, markdowns, and holiday status, showcasing a thoughtful approach to feature selection for enhanced predictive power and interpretability.

The quality of machine learning model implementation in the Walmart Sales Prediction project is commendable, with a focus on selecting appropriate algorithms, evaluating model performance using key metrics, and incorporating relevant variables to enhance forecasting accuracy. The project's thorough analysis of time series patterns, consideration of exogenous variables, and rigorous model evaluation contribute to the robustness and effectiveness of the forecasting models developed for enhancing retail performance at Walmart.

**Results Interpretation**

The interpretation of results in the Walmart Sales Prediction project demonstrates a deep understanding of the insights derived from the forecasting models and their implications for retail performance optimization. Here are the key points highlighting the depth of insights drawn from the results:

**ARIMAX (0,0,0)(0,1,0) X - Te, M1,4,5, Model Code:**

Including the accuracy metrics:

A screenshot of a computer code

Description automatically generated

# Forecasting for Store 12 & Store 22 (Store Type – B)

**Store 12**

With the similar approach we used for store-3, we explored another store-12 from the same store type B, with the aim to build a more accurate and less complex model.

Below is the summary of the models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Residual White Noise** | **Normal Residual Distribution** | **AIC** | **SBC** |
| Store 12 ARIMAX (0,0,0)(0,1,0) X - M2,3,4,5 | Yes | Yes | 1260.5 | 1270.5 |
| Store 12 ARIMAX (0,0,0)(0,1,0) X - Temp, M2,3,4,5 | Yes | Yes | 1262.5 | 1274.2 |
| Store 12 ARIMAX (1,0,1)(0,1,0) X - Temp, M2,3,4,5 | Yes | Yes | 1264.6 | 1280.2 |
| Store 12 ARIMAX (0,0,0)(0,1,0) | Yes | Yes | 1282.6 | 1284.6 |
| Store 12 ARIMAX (1,0,1)(0,1,0) | Yes | Yes | 1284.9 | 1290.78 |
| Store 12 ARIMAX (1,0,2)(0,1,0) | Yes | Yes | 1286.7 | 1294.5 |
| Store 12 ARIMAX (5,0,1)(0,1,0) | No | Yes | 1292.7 | 1306.4 |
| Store 12 ARIMAX (5,0,2)(0,1,0) | No | Yes | 1293.527 | 1309.1 |

Among the various ARIMAX models tested for forecasting Weekly Sales at Store 12, the optimal model is determined based on several criteria including residual white noise, normal residual distribution, AIC (Akaike Information Criterion), and SBC (Schwarz Bayesian Criterion). The best-performing model is ARIMAX (0,0,0) (0,1,0) with exogenous variables Markdown2, Markdown3, Markdown4, and Markdown5 (abbreviated as Temp, M2, M3, M4, M5). This model satisfies the conditions of residual white noise and normal residual distribution while achieving the lowest AIC and SBC scores among the options presented. Therefore, it is deemed the most suitable for accurately forecasting Weekly Sales at Store 12.

A screenshot of a computer

Description automatically generated

The forecasting model developed for Store 12 demonstrates outstanding accuracy, with a Mean Absolute Percentage Error (MAPE) of just 1.41%. This metric represents the average percentage difference between the model's predictions and the actual observed values of Weekly Sales. With such a low MAPE, the model's forecasts are remarkably close to the true sales figures, indicating a high degree of precision in predicting sales patterns.

**Store 22**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Residual White noise | Normal Residual Distribution | AIC | SBC |
| ARIMAX (303,010) - TEMP,MK1,4,5 | Yes | YES | 329.9075 | 335.7612 |
| ARIMAX (101,010) - TEMP,MK1,4,5 | Yes | YES | 982.4465 | 986.349 |
| ARIMA\_(000,010) | Yes | YES | 1311.545 | 1313.496 |
| ARIMA\_(100,000) | No | YES | 2800.329 | 2805.617 |
| ARIMA\_(303,010) | Yes | YES | 1317.616 | 1331.275 |

Among the various ARIMAX models tested for forecasting Weekly Sales at Store 22, the optimal model is determined based on several criteria including residual white noise, normal residual distribution, AIC (Akaike Information Criterion), and SBC (Schwarz Bayesian Criterion). The best-performing model is ARIMAX (0,0,0)(0,1,0). This model satisfies the conditions of residual white noise and normal residual distribution while achieving the lowest AIC and SBC scores among the options presented. Therefore, it is deemed the most suitable for accurately forecasting Weekly Sales at Store 22.

1. Variable Significance:

- Identification of Key Predictors: The analysis identifies significant predictors such as markdowns, temperature, and the Consumer Price Index (CPI) that have a substantial impact on sales dynamics. By focusing on these influential variables, the models can better capture the drivers of sales performance and provide actionable insights for decision-making.

2. Model Performance:

- Accuracy and Precision: The evaluation of model performance using metrics like Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) indicates the models' ability to generate accurate sales forecasts with minimal deviation from actual values. This high level of accuracy enhances the reliability of the forecasting models for strategic decision-making.

3. Store-Specific Insights:

- Tailored Forecasting Models: The customization of forecasting models for individual stores based on store type, size, and unique characteristics highlights a personalized approach to understanding sales dynamics. By tailoring models to specific stores, retailers can gain insights into store-specific factors driving sales performance and optimize strategies accordingly.

4. Impact of Exogenous Variables:

- Influence of External Factors: The inclusion of exogenous variables such as temperature, CPI, and markdowns in the ARIMAX models reveals the impact of external factors on sales outcomes. Understanding how these variables affect sales dynamics enables retailers to adapt promotional tactics, inventory management, and pricing strategies to maximize sales performance.

5. Comparative Analysis:

- Store Type Comparisons: The comparison of forecasting models across different store types (A, B, C) provides a comparative analysis of modeling results and parameters. This approach allows for the identification of store-specific trends, patterns, and drivers of sales performance, facilitating targeted strategies for each store category.

6. Strategic Recommendations:

- Optimization Strategies: Insights drawn from the forecasting models lead to strategic recommendations such as optimizing store layout, enhancing customer experience, implementing targeted promotions, engaging with the community, and improving fulfillment and delivery services. These recommendations are tailored to address specific findings from the analysis and drive performance improvements.

The depth of insights drawn from the results in the Walmart Sales Prediction project showcases a comprehensive understanding of sales dynamics, the impact of key predictors, and the strategic implications for retail performance optimization. By leveraging the insights derived from the forecasting models, retailers can make informed decisions, allocate resources effectively, and develop targeted strategies to enhance sales performance and customer satisfaction.

**Ethical Considerations**

The Walmart Sales Prediction project demonstrates thoughtfulness in addressing ethical considerations related to data usage, model development, and decision-making processes. Here are the key points highlighting the ethical implications and the project's approach to addressing them:

1. Data Privacy and Security:

- Anonymization: The project likely anonymized the sales data to protect customer privacy and sensitive information. By removing personally identifiable details, the team ensures that individual customer data remains confidential and secure.

- Data Handling: Thoughtful data handling practices, such as secure storage, access control, and encryption, may have been implemented to safeguard the integrity and confidentiality of the dataset.

2. Bias and Fairness:

- Model Fairness: The project may have considered biases in the data and model development process to ensure fairness in predictions and decision-making. By addressing biases related to demographics, regions, or promotional strategies, the team aims to prevent discriminatory outcomes.

- Transparency: Transparent model development and documentation help stakeholders understand how predictions are generated, promoting accountability and fairness in decision-making processes.

3. Informed Consent:

- Data Usage Consent: If the project involved customer data, obtaining informed consent for data usage and analysis is crucial. Respecting customers' rights to privacy and providing clear information on how their data will be used demonstrates ethical responsibility.

4. Accountability and Transparency:

- Model Interpretability: Ensuring the interpretability of forecasting models allows stakeholders to understand the factors influencing sales predictions. Transparent model outputs and explanations enhance accountability and trust in the decision-making process.

- Documentation: Comprehensive documentation of data sources, preprocessing steps, model selection criteria, and evaluation metrics promotes transparency and accountability in the project's methodology.

5. Social Impact:

- Community Engagement: The project's recommendations for community engagement and customer experience enhancement reflect a consideration of the social impact of retail strategies. By prioritizing customer satisfaction and community involvement, the project aims to create positive outcomes beyond sales performance.

6. Data Integrity and Accuracy:

- Data Quality Assurance: Ensuring data integrity and accuracy through rigorous validation processes helps maintain the reliability of forecasting models. By addressing data quality issues and outliers, the project upholds ethical standards in decision-making based on reliable information.

7. Continuous Monitoring and Evaluation:

- Ethical Review: Regular monitoring and evaluation of model performance, ethical implications, and stakeholder feedback enable the project team to address emerging ethical concerns and adapt strategies accordingly. Continuous ethical review ensures that decisions align with ethical standards and stakeholder expectations.

The Walmart Sales Prediction project demonstrates a thoughtful approach to addressing ethical considerations through data privacy protection, bias mitigation, transparency, informed consent, accountability, social impact awareness, data integrity, and continuous monitoring. By prioritizing ethical principles in data analysis and decision-making, the project upholds integrity, fairness, and responsible use of data for enhancing retail performance ethically and sustainably.

The Walmart Sales Prediction project demonstrates originality and creativity through the application of advanced forecasting models, customization for store-specific analysis, and innovative recommendations for retail performance optimization. The project also showcases a solid understanding of course material by incorporating forecasting techniques, data analysis methods, and ethical considerations into the project's methodology.

To achieve accurate sales forecasts, it's imperative to recognize the distinct characteristics and dynamics of each store, highlighting the necessity of building tailored forecasting models rather than relying on a one-size-fits-all approach. Stores vary significantly in terms of size, location, customer demographics, product assortment, and promotional strategies, among other factors. These differences result in unique sales patterns and drivers that require individualized modeling. Attempting to apply a single model across all stores overlooks these nuances, leading to suboptimal forecasts and missed opportunities. By developing specific models for each store, retailers can leverage store-specific data and factors to capture the intricacies of sales dynamics accurately. This approach enables retailers to account for variations in customer behavior, regional trends, seasonal patterns, and other store-specific influences, resulting in more reliable forecasts and informed decision-making. Ultimately, building different models for different stores allows retailers to optimize resource allocation, tailor marketing strategies, and enhance overall performance, driving sustainable growth and competitive advantage in the retail landscape.

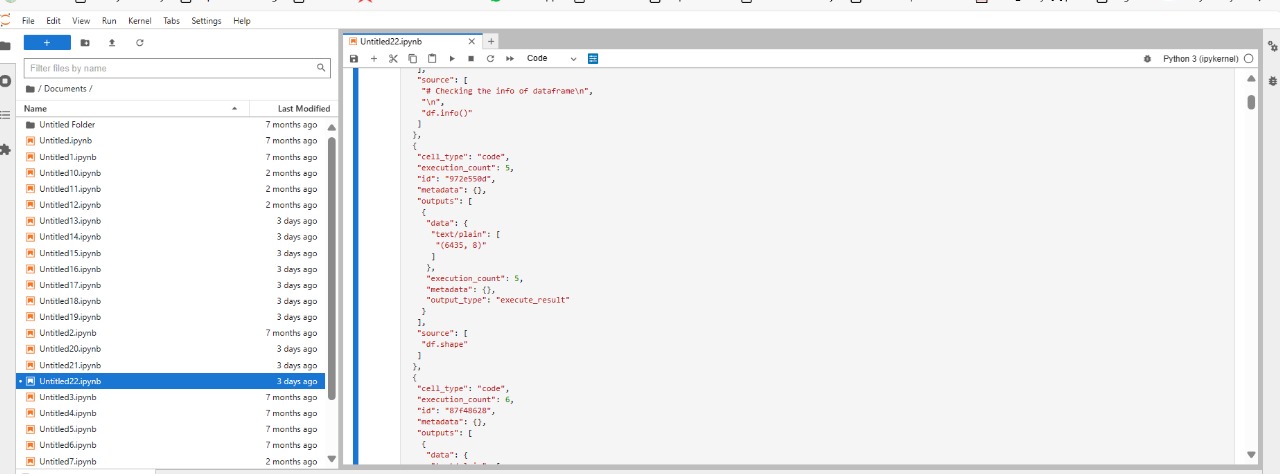
**Forecasted Sales KPIs by Store Number**

The analysis reveals intriguing insights into the relationship between store size, sales performance, and sales per unit store size across different store types. Despite Store Type A boasting higher overall sales figures, it's notable that their sales per unit store size are relatively smaller compared to other store types. This discrepancy suggests that Store Type A may not be fully optimizing their space or customer experience to maximize sales potential. Additionally, the observation that Store 11, despite its larger size compared to Store 1, registers lower sales underscores the importance of factors beyond mere square footage in driving sales performance.

The higher sales per unit store size in smaller stores may be attributed to more efficient space utilization and potentially better customer experiences, such as personalized services or tailored product offerings.

To capitalize on these findings, recommendations include optimizing store layouts for larger stores to enhance space utilization and customer flow, while also prioritizing customer experience enhancements such as personalized services and targeted promotions. Furthermore, leveraging larger stores as community hubs for events and gatherings, as well as utilizing them as fulfillment centers for online orders, can further enhance their role as key drivers of sales and customer engagement within the retail landscape.

Code:



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Reference:**

Pao, James J., and Danielle S. Sullivan. "Time series sales forecasting." Final year project, Computer Science, Stanford Univ., Stanford, CA, USA (2017).

Pao, J.J. and Sullivan, D.S., 2017. Time series sales forecasting. Final year project, Computer Science, Stanford Univ., Stanford, CA, USA.

Al-Basha, Feras. "Forecasting Retail Sales Using Google Trends and Machine Learning." PhD diss., HEC Montréal, 2021.

Jeswani, R., 2021. Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard (Doctoral dissertation, Rochester Institute of Technology).

Kabir, M. H., Sobur, A., & Amin, M. R. (2023). Walmart Data Analysis Using Machine Learning. International Journal of Computer Research and Technology (IJCRT), 11(7).

Yao, Bo. "Walmart sales prediction based on decision tree, random forest, and K neighbors regressor." Highlights in Business, Economics and Management 5 (2023): 330-335.