Walmart Sales Prediction

Agreda Nicolas, Shainidze Ravil, Satarov Mansurbek

North American University (NAU)

COMP 4353 1 - Data Mining

Prof. Sabina Adhikari

4/21/2024

# Author Note

First paragraph: Introduction

Second paragraph: Background

Third paragraph: Findings

Fourth paragraph: Results

Last paragraph: Conclusion

# Introduction

This report encapsulates the findings from a detailed analysis of the Walmart dataset aimed at understanding the factors influencing weekly sales. Through a combination of data cleaning, exploratory analysis, and various machine learning models, we investigated how certain variables like temperature, fuel prices, unemployment rates, and Consumer Price Index (CPI) interact with sales dynamics.

**Background**

The dataset includes both categorical and numerical data, capturing a wide array of variables from store identifiers and dates to sales figures and economic indicators. Methods of data preprocessing like handling missing values, outlier detection, data transformation, and feature engineering were extensively used to prepare the dataset for analysis.

# Findings:

* **Exploratory Data Analysis**: Initial analysis showed that weekly sales data is right-skewed, suggesting variability in sales figures. The analysis of temperature and unemployment revealed normal distributions, while CPI and fuel prices showed bimodal distributions. Sales patterns vary significantly across holidays, seasons, and different years with notable peaks during certain months like April and December.
* **Bi-Variate Analysis**: The analysis indicated that while sales are generally higher on holidays, the total holiday sales are less due to the fewer number of holiday days. Significant variance was observed in weekly sales among different stores.

## Results:

**Data Correlation and Regression Analysis:**

* **Fuel Price vs. Weekly Sales**: No significant correlation was found, indicating that fuel prices are not reliable predictors of sales.
* **Unemployment vs. Weekly Sales**: A strong negative correlation suggests that higher unemployment rates lead to lower sales, making it a significant predictor.
* **CPI and Temperature vs. Weekly Sales**: Both showed some level of significant correlation with sales.

**Model Performance**

* **Linear Regression:** The model showed limited success with an R-Square score of 16.11%, suggesting the need for a more complex model.
* **Polynomial Regression**: Substantial improvements were observed after tuning with polynomial features, achieving an R-Square score of 98.61% and demonstrating high accuracy.
* **K-Nearest Neighbors (KNN):** After tuning, the model did not perform well, with a best R-Square score of around 51.87%.

# Conclusion

The comprehensive analysis of the Walmart dataset highlighted the complex interplay between various economic indicators and sales performance. While some variables like unemployment proved to be strong predictors of sales, others like fuel prices did not exhibit significant influence. The use of polynomial regression after the initial linear models provided a better fit and higher prediction accuracy, showcasing the importance of selecting appropriate modeling techniques based on the data characteristics. Future analyses could explore additional machine learning models and feature engineering techniques to further enhance predictive accuracy.