# Ablation Study on Decision Tree and ARIMA in Walmart Sales Prediction

**Wahub Ahmed, Joseph Kim, Matthew Heinen, Abel Kelbessa, Andrew Buck**
Department of Computer Science
Haverford College
Haverford, PA 19041
{wahmed, jkim5, akelbessa, mheinen, dbuck}@haverford.edu

## Abstract

Sales prediction is a valuable tool for businesses, as it allows them to make informed decisions about their operations in order to maximize profits and minimize costs. The use of machine-learning algorithms for generalization enables stakeholders and corporations to make predictions even with a limited amount of historical sales data. In this paper, we refined machine-learning models that are well-suited for time series datasets, such as Decision Tree Regression and ARIMA (Autoregressive Integrated Moving Average), to test the accuracy of sales forecasts from April to October 2012. The dataset for this analysis was obtained from Kaggle and covered a period from February 5, 2010 to October 26, 2012, containing weekly sales data for 45 Walmart stores across 99 different departments. The results showed that the Decision Tree model had a lower MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) compared to ARIMA, indicating a more accurate prediction of sales.

## 1 Introduction

Machine-learning generalization is particularly useful in capturing patterns in a data set as a whole. Accurate predictions made using machine learning can save companies time, money, and resources. In particular, the operations pipeline can use predictions of how social, economic, and natural factors impact sales to directly affect the supply chain, such as by ensuring that stores are stocked with the right items and choosing the most profitable items to sell. These trends can be analyzed using a variety of machine-learning models and regression techniques. Optimizing these models can also help to reduce environmental issues such as food waste and shortages. In order to make time series predictions based on past data, this paper analyzed two methods: ARMIA (Autoregressive Integrated Moving Average) and a Decision Tree Regression model. The Naïve approach of using the previous day's value was also analyzed for comparison.

### 1.1 Overview of ARIMA

ARIMA is a model that uses past data in order to predict future data. It takes into account the dependencies between observations and incorporates information about trends and seasonality, which can improve the accuracy of predictions for time series data. This method assumes that we are using data that is stationary so that the model is not predicting based on time passed but only on previous values. Thus in order for ARIMA to be successful there needs to be adequate training data over a period of time.

### 1.2 Overview of Decision Tree

A decision tree is a machine learning model that is used to make predictions based on how a previous set of questions have been answered. Decision trees begin at a root node and split into decision nodes with the model always choosing one of the nodes to choose from. Within this tree-like structure, each node represents a decision, and each branch represents an outcome to that decision. These splits are determined by the model based on the attributes of the data, and the resulting tree can be used to make predictions about new data points by following the decisions made at each split. Decision Tree models are often used for forecasting sales, because they can handle complex data with multiple variables and can provide insight into the importance of different factors in determining the outcome. Additionally, decision trees are easy to interpret and can be visualized, which makes them useful for explaining the predictions to stakeholders.

## 2 Related Works

The 2019 research paper by Catal et al. compared different regression models including linear regression, neural network regression, decision tree regression, and more, along with various time series methods. After comparing both the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) for each model, it was found that Boosted Decision Tree Regression algorithm was the best predictor for sales forecasting (Catal et al., 2019).

Another paper was published in 2019 that considered various machine learning models and concluded that sales forecasting is better treated as a regression problem than a time series problem. Similar to the first paper mentioned, it showed that using regression approaches for sales forecasting can often provide improved results compared to time series methods. Furthermore, the paper stated that one of the assumptions of regression models is that patterns in the historical data will be repeated in the future (B. M. Pavlyshenko, 2019).

## 3 Analysis of Dataset

The dataset contained weekly sales for 45 Walmart stores across 98 different departments. It was obtained from Kaggle[1] and ranged from February 5, 2010 to November 1, 2012, as shown in Table 1. In addition, it also contained binary information on whether the week was a special holiday week in the column `isHoliday`. Within the data processing pipeline, `Date` was sorted in order and set as the index of the dataset so as to analyze the time series data chronologically as the independent variable. The cutoff of the training and test data for the machine learning models was April 13th, 2012, respectively splitting the data in an approximate 75:25 ratio. The data processing was pipelined respectively into three groups: the Naïve method, Decision Tree, and ARIMA.

Table 1: Initial dataset retreived from Kaggle

| | Store | Dept | Date | Weekly_Sales | isHoliday |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 2010-02-05 | 24924.50 | False |
| 2 | 1 | 1 | 2010-02-12 | 46039.49 | True |
| 3 | 1 | 1 | 2010-02-19 | 41595.55 | False |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 421568 | 45 | 98 | 2012-10-19 | 760.01 | False |
| 421569 | 45 | 98 | 2012-10-26 | 1076.80 | False |

### 3.1 Decision Tree

The day and month were respectively extracted from the date as features for the weekly sales, as in the initial analysis, dates surrounding holiday seasons such as Christmas played a significant factor in increasing the sales. The stores and departments were also analyzed separately initially, to provide Decision Tree with additional information to determine the weekly sales more accurately. Because

[1] Link to dataset: https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast?select=train.csv

the department directly reflects the type of product, and the store reflects the geographical location of the Walmart store, weekly sales necessarily have a degree of dependency on these factors that must be taken into account. Therefore, marginalizing across the data and calculating the mean weekly sales across all stores and departments to provide one particular week to focus on sales would lose crucial information in the Decision Tree Regression. Holidays were similarly considered rather than marginalizing out, since the presence of holidays impact the sales similar in fashion to the day and month (Table 2). The features and weekly sales were added in their own respective dataframes, which were used to fit and predict the decision tree regression. After all the predictions were made, the results were summed across all the stores and departments to yield one cumulative weekly sale prediction per date.

Table 2: Cleaned dataset post-processing

| Date | Store | Dept | Weekly_Sales | isHoliday | Month | Day |
|---|---|---|---|---|---|---|
| 2010-02-05 | 1 | 1 | 24924.50 | False | 2 | 5 |
| 2010-02-05 | 29 | 5 | 15552.08 | False | 2 | 5 |
| 2010-02-05 | 29 | 6 | 3200.22 | False | 2 | 5 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2012-10-26 | 19 | 33 | 5740.14 | False | 10 | 26 |
| 2012-10-26 | 45 | 98 | 1076.80 | False | 10 | 26 |

### 3.2 ARIMA

The Weekly Sales were for each date across all departments and all stores were calculated for the ARIMA model, as shown in Table 3. This was because the ARIMA mainly uses the past values to accurately predict the future values for the model. In the future, more exogenous features could be added in the dataset to see if it will improve the model, but for now, only the weekly sales were used as a feature.

Table 3: Dataframe of cumulative weekly sales

| Date | Weekly_Sales |
|---|---|
| 2010-02-05 | 49750740.50 |
| 2010-02-12 | 48336677.63 |
| 2010-02-19 | 48276993.78 |
| ⋮ | ⋮ |
| 2012-10-19 | 45122410.57 |
| 2012-10-26 | 45544116.29 |

### 3.3 Naïve Approach

The same processed dataset for the ARIMA model was also used in the Naïve approach. For this model, the total weekly sales of the previous week was used to predict this week's sale, as it would make sense that the weekly sales would depend on the previous week's sale (Table 4). For example, if the total weekly sale for 2012-09-28 was 43734899.40, this value would be the predicted sales value for 2012-10-05. Because no data before February 5th, 2010 was available, the initial data entry was removed to manage the NaN in the Naïve approach.

## 4 Results

We evaluated the results of our 3 models with Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). For both metrics, Decision Tree had the best performance, with the ARIMA coming second. The baseline approach also performed the worst. (Table 5) This was the expected outcome, both due to previous related research and due to the fact that Decision Tree had much more features for training. Since the ARIMA was not far from the Decision Tree for both metrics, it would

Table 4: Dataframe for Naïve Approach analysis

| Date | Weekly_Sales | Baseline |
|---|---|---|
| 2010-02-12 | 48336677.63 | 49750740.50 |
| 2010-02-19 | 48276993.78 | 48336677.63 |
| 2010-02-26 | 43968571.13 | 48276993.78 |
| ⋮ | ⋮ | ⋮ |
| 2012-10-19 | 45122410.57 | 46128514.25 |
| 2012-10-26 | 45544116.29 | 45122410.57 |

Table 5: Dataframe for Naïve Approach analysis

| | Baseline | Decision Tree | ARIMA |
|---|---|---|---|
| RMSE | 2419672.68 | 1224013.04 | 1443200.43 |
| MAE | 1648526.29 | 984759.38 | 1034870.54 |

be interesting to see how much better it would perform when provided more features such as store, holiday, or department.

**Appendix A** contains the graphical representations of the predictions for each model. All the models predicted the fluctuating pattern of sales quite well. Unfortunately, our data set stopped at October of 2012, but it would be interesting to see how the three models predict the sales of the Christmas season, as there is a huge spike in sales during that time each year.

## 5 Limitations

One limitation of the decision tree model is that it can be prone to overfitting, which occurs when the model is overly complex and captures too much of the random noise in the data rather than the underlying relationships. This can lead to poor performance on new, unseen data, and can make the model less reliable for forecasting sales.

ARIMA models, on the other hand, have the limitation of requiring stationary data in order to produce accurate forecasts. If the data is not stationary, the model may not be able to adequately capture the underlying trends and patterns, leading to inaccurate predictions. Additionally, ARIMA models can be difficult to fit to data with multiple seasonal components or other complex patterns, which can make them less effective in some retail settings.

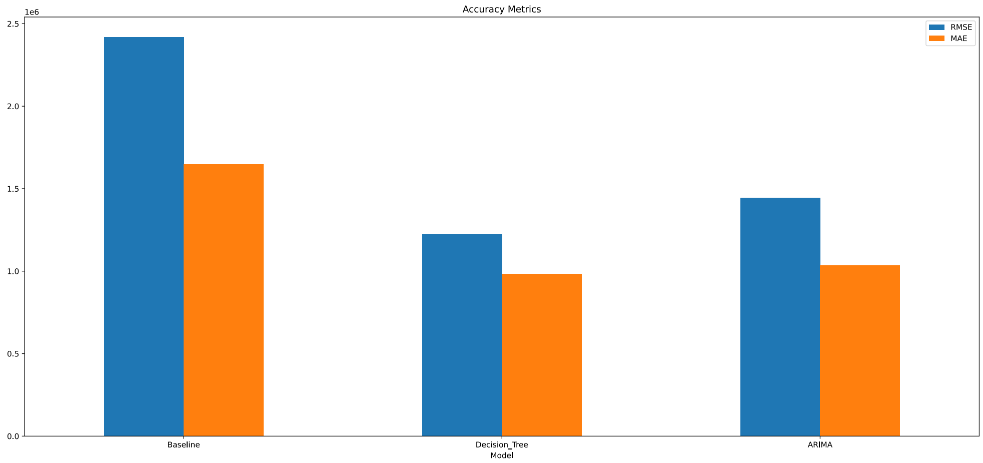


Figure 1: Accuracy Metrics of Models

These limitations in Decision Trees and ARIMA models are important to consider because they can affect the accuracy and reliability of the models for predicting unit sales in Walmart. If these limitations are not taken into account when reproducing the models on a larger scale, the models may produce significantly inaccurate forecasts, which could lead to poor decision-making and potentially costly consequences for the business.

## 6 Conclusion

Our study shows that Decision Trees outperform ARIMA for sales forecasting. Decision Trees are effective at handling complex data and provide insight into important factors. In comparison, ARIMA models use past data and seasonality to predict future values. Our findings suggest that Decision Trees should be considered as a valuable tool for sales forecasting, and have important implications for time series analysis and the development of future machine learning models. They suggest that Decision Trees may be a more effective tool for time series forecasting, and provide valuable insight into the design of machine learning models for this task. Further research can include examining the performances of Boosted Decision Trees and Random Forests with original models on this dataset.

## References

[1] C. Catal, K. Ece, B. Arslan, & A. Akbulut. (2019) Benchmarking of regression algorithms and time series analysis techniques for sales forecasting. *Balkan Journal of Electrical and Computer Engineering*, pp. 20 – 26.

[2] B. M. Pavlyshenko. (2019) Machine-learning models for sales time series forecasting. *Data* vol. 4, no. 1. [Online]. Available: https://www.mdpi.com/2306-5729/4/1/15

## A Appendix

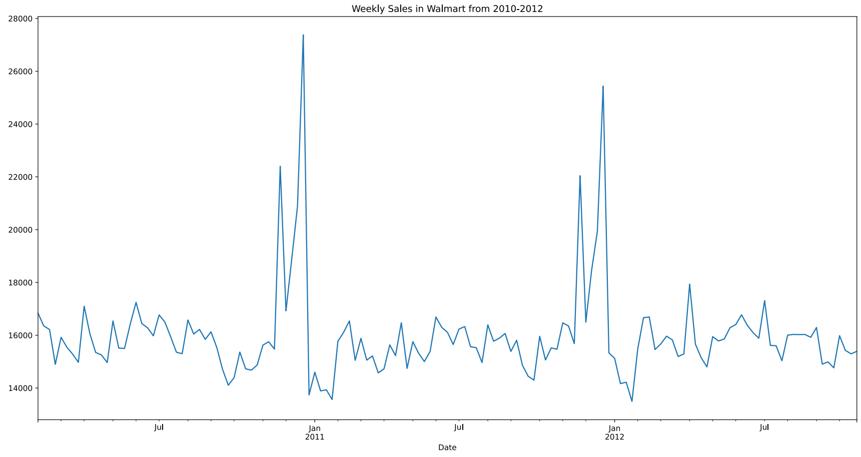
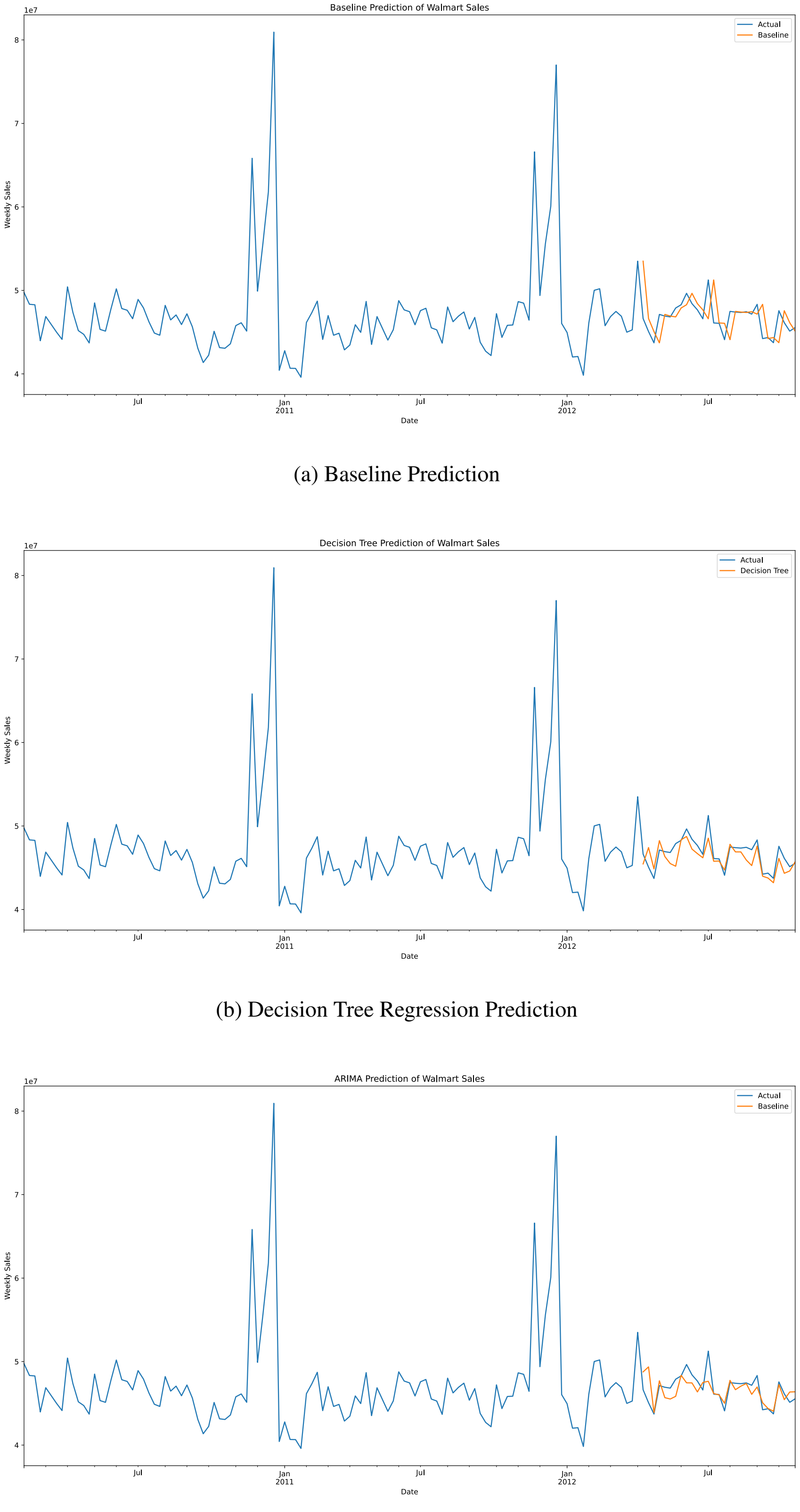


Figure 2: Actual Weekly Sales in Walmart



(a) Baseline Prediction

(b) Decision Tree Regression Prediction

(c) ARIMA Prediction

Figure 3: Comparisons of Model Predictions