Bellevue University

# Predicting Future Walmart, Inc. Sales

Rachel Young & Maddie Bauer
DSC 630 Predictive Analytics

February 28, 2021

## Executive Summary

The purpose of this report is to discuss and identify machine learning models that will best predict future weekly store sales for 45 different Walmart stores that appear in the dataset. Walmart is a popular international company currently with 11,443 stores worldwide. Being able to predict weekly sales is a critical piece for retail companies as it can help determine how much inventory to purchase, how many people to have on staff, and improve overall decision making.

The Walmart dataset has data from February 5, 2010 through November 1, 2012. Retail sales fluctuate between weeks, months and years due to different economic factors, such as Consumer Price Index, unemployment rates, average fuel prices, as well as average temperatures, and whether or not there is a holiday within the given week. Through the use of predictive analytics and machine learning, patterns and relationships among the features in the dataset were analyzed. Models such as linear regression, K Nearest Neighbor, Random Forest and Gradient Boosting were all developed as well as a time series approach with an ARIMA model. The linear regression, K-Nearest Neighbor and Random Forest all had above a 90% accuracy rate. However, in the end the ARIMA model performed the best at predicting future sales.

**Abstract**

In order for a company to make smarter decisions about their sales, they need to explore the potential factors that can cause their sales to vary. This research papers seeks to understand how factors such as holidays, economic factors, temperature, fuel prices, Consumer Price Index (CPI) and unemployment may or may not impact the trends in sales. A retail business will be able to create a blueprint of future plans and predict future sales when they have all the known variables and past data. Various models will be reviewed and tested so that the best method for predicting future Walmart sales will be selected.

**Introduction/Problem Statement**

Walmart, Inc. is part of the retail and wholesale business and is based in Bentonville, Arkansas. The President, Chief Executive Officer, and Director is C. Douglas McMillon. They specialize in merchandise of consumer products and services. Walmart operates Walmart, Walmart Neighborhood Market, Wal-Mart, Walmart.com, and Sam's Club. The market cap is $421.3 billion and as of 2019, there are approximately 2.20 million fulltime, part time, and temporary employees (CNN Business, 2020).

Retail companies commonly have issues with predicting sales accurately throughout the days, months, and years ahead. There are many varying factors that can cause issues with predicting sales such as holidays, economic factors, temperature, fuel prices, Consumer Price Index (CPI), and unemployment. If the varying factors are not predicted correctly, then there could be staffing issues at stores,

financial implications, and the business could become obsolete if customer satisfaction goes down. The goal of this analysis is to predict future sales for the Walmart stores based on the varying factors mentioned above.

**Methods**

The Walmart dataset was retrieved from Kaggle. The data ranges from February 5, 2010 through November 1, 2012. Forty-five stores are represented in the data and the target variable for the analysis is the weekly sales variable.

The overall plan with this data is to analyze the trends of weekly sales during this time period. This will help with understanding how the sales will increase or decrease in the future. It will be necessary to analyze and determine which features are the strongest predictors for weekly sales, why they are important, and their relationship with one another.

Below is the process that will be used to create, train, test, and evaluate machine learning models.

- Data exploration – Dataset is analyzed, and questions are identified
- Data preparation – Dataset is put into a usable format
- Research and create a machine learning model and set the baseline model
- Train and test the machine learning model
- Compare the test data and the predicted data from the model
- Update and rerun the machine learning model as needed
- Report and interpret the data to identify if the model is the best fit

There are five methods that will be used when identifying the best methods to predict future Walmart sales. The methods are Auto Regressive Integrated Moving Average (ARIMA) time series, multiple linear regression, random forest regression, K-Nearest Neighbor (KNN) regression and gradient boosting.

The ARIMA time series model is used for forecasting data into the future. The article "How to Create an ARIMA Model for Time Series Forecasting in Python" shows what the acronyms mean and the parameters of the model (Brownlee, 2017).

- **AR**: *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations.
- **I**: *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
- **MA**: *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

- **p**: The number of lag observations included in the model, also called the lag order.
- **d**: The number of times that the raw observations are differenced, also called the degree of differencing.
- **q**: The size of the moving average window, also called the order of moving average.

ARIMA is the underlying process that generates the observations for time series.

"Multiple Linear Regression is a statistical technique that uses a number of explanatory variables to predict the outcome of the response variable" (Kenton, 2020). A multiple regression model is based on a few assumptions. This includes there is a linear relationship between the independent and dependent variables, the independent variables are not too highly correlated with one another, observations are selected randomly and independently, and the residuals should be normally distributed (Kenton,

2020). The statistical metric used to measure how much variation is explained by the independent variables is called $R^2$. $R^2$ ranges from 0, where the outcome cannot be predicted by any of the independent variables, to 1, where the outcome can be predicted without error by the independent variables (Kenton, 2020).

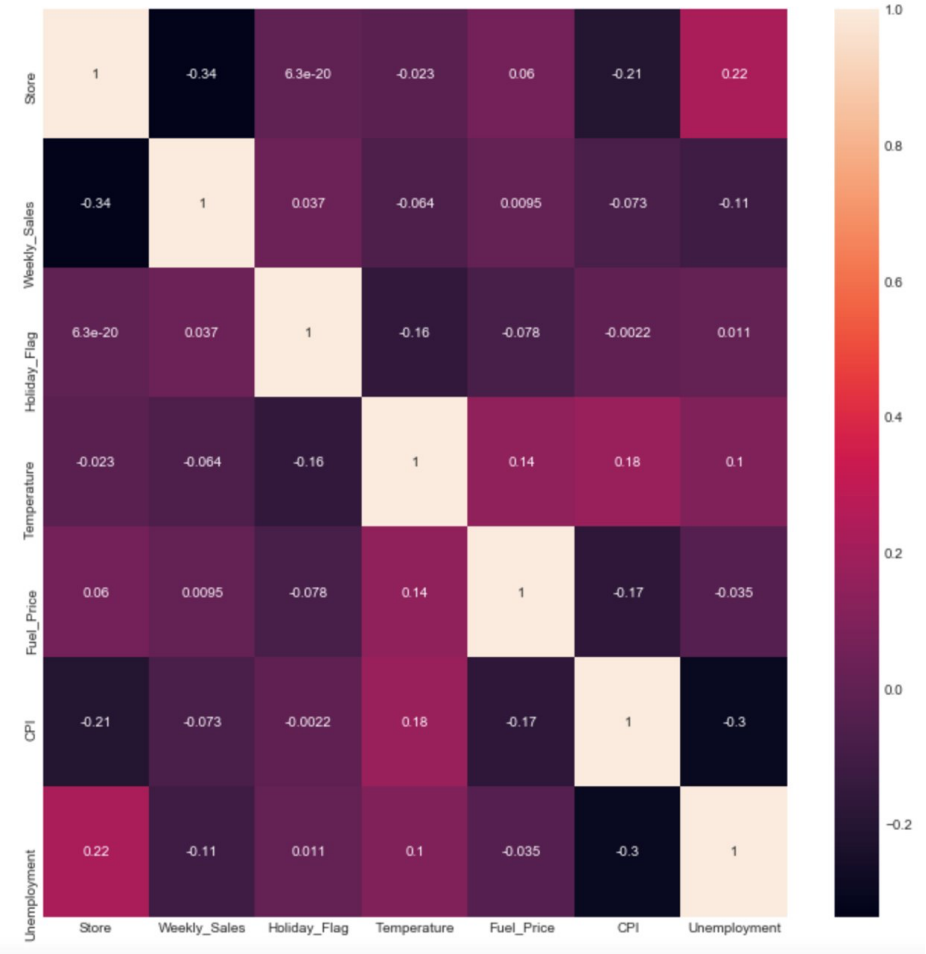
Random forest regression is a technique that can perform classification and regression tasks through the use of multiple decision trees. This is the base learning model for decision trees and performs random feature and row sampling (GeeksforGeeks, 2020).

"KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood" (BIO Statistics Collaboration of Australia, 2020). The person that is doing the analysis would need to set the size of the neighborhood or cross-validation can be used to decide on the size.

The goal of gradient boosting is to turn weak learners into strong learners. This technique produces a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. Gradient boosting trains several models in a gradual, additive and sequential manner by using the loss function (Singh, 2018). As the models continue to be combined, the prediction error decreases and a new and better model is created.

## Results

*Correlation*: By looking at the correlation matrix below, we can see that none of the features are highly correlated with one another. The matrix also illustrates that the correlations are pretty evenly split between positive and negative values. We chose to keep all features in the models used below.



*ARIMA Model for Time Series:*

Before applying the ARIMA model for time series, the data was split up by store number. Below is a comparison of two different stores and the applied model. In figure 1, we can see that the ARIMA model predicted the future sales relatively well for store number 2. The predicted weekly sales follow a similar pattern as the actual data. In figure number 2, we can see that the ARIMA model clearly predicted lower sales than the previous trends for store number 14.

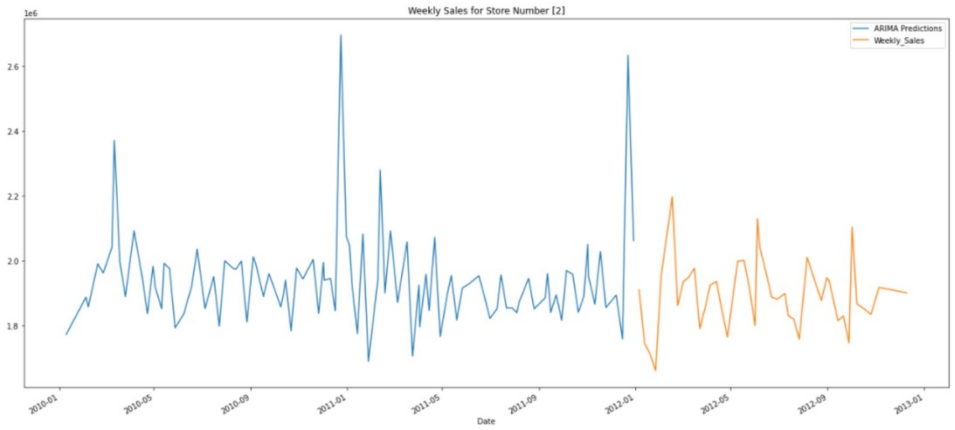



Figure 1

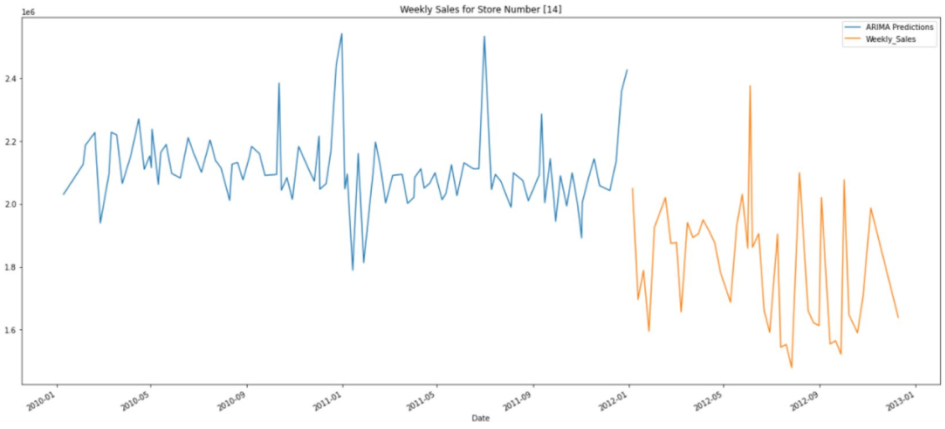



Figure 2

*Linear Regression/K-Nearest Neighbors/Random Forest/Gradient Boosting:*

Below is a table with the results from the machine learning algorithms used.

| **Model** | **Accuracy ($R^2$)** |
|---|---|
| Linear Regression | 0.92172 |
| K-Nearest Neighbors | 0.93460 |
| Random Forest | 0.94517 |
| Gradient Boosting | 0.82879 |

The random forest model performed the best with roughly 94.5% of the variance in weekly sales being accounted for by the independent variables while the gradient boosting had the worst performance (~82.9%). These models took into account all the features in the dataset, which included store number, if it was a holiday or not, temperature, fuel price, CPI, and unemployment rates.

**Conclusion**

Overall, we have gained a lot of knowledge while evaluating and identifying various models that could potentially be used to predict the weekly sales for Walmart, Inc. Through our analysis, the random forest model was identified as the best regression model. However, the regression models are not the first choice when it comes to forecasting weekly sales data. After reviewing all of the different models, the best model for this data would be the ARIMA time series model. The evaluation tool for regression models is a percentage of accuracy whereas the ARIMA model aims to

predict values. In the end, we were able to successfully predict weekly sales data for Walmart, Inc.

**Acknowledgments**

We would like to thank Walmart, Inc. for providing their data. Also, we would like to thank our Professor, Brett Werner, from Bellevue University as well as the students of DSC630 for their support throughout the course. Finally, we would like to express our gratitude for Kaggle.com, Machinelearningmastery.com, Medium.com, Geeksforgeeks.com and their countless authors for their research and posts on the subject matter. This report would not have been possible without their contributions and sharing of information.

**References**

BIO Statistics Collaboration of Australia. 2020. Machine Learning for Biostatistics. Retrieved on January 9, 2021 from https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html

Brownlee, Jason. April 27, 2018. How to Calculate Correlation Between Variables in Python. Retrieved on January 9, 2021 from https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/

Brownlee, Jason. January 9, 2017. How to Create an ARIMA Model for Time Series Forecasting in Python. Retrieved on January 9, 2021 from https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

CNN Business. 2020. Walmart Inc. Retrieved on December 7, 2020 from https://money.cnn.com/quote/profile/profile.html?symb=WMT

GeeksforGeeks. May 28, 2020. Random Forest Regression in Python. Retrieved on January 9, 2021 from https://www.geeksforgeeks.org/random-forest-regression-in-python/

Kaggle. 2020. Retail Analysis with Walmart Data. Retrieved on December 7, 2020 from https://www.kaggle.com/aditya6196/retail-analysis-with-walmart-data

Kenton, W. September 21, 2020. Multiple Linear Regression (MLR). Retrieved on January 11, 2021 from https://www.investopedia.com/terms/m/mlr.asp

Singh, H. November 3, 2018. Understanding Gradient Boosting Machines. Retrieved February 24, 2021 from https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab