Predicting Future Walmart, Inc. Sales

Rachel Young & Maddie Bauer

DSC630 Predictive Analytics

Introduction

- Walmart, Inc is part of the retail and wholesale business
- Specializes in merchandise of consumer products and services
- 2019 – has a market cap of $421.3 billion
- Approximately 2.2 million employees
- One of the largest retail and wholesale businesses

https://money.cnn.com/quote/profile/profile.html?symb=WMT
https://www.bestdesigns.co/best-design/walmart

Problem Statement

- Retail companies have issues with predicting sales throughout the coming days, months, and years
- Varying factors that can cause potential issues with predicting sales:
  - Holidays
  - Economic factors
  - Temperature
  - Consumer Price Index (CPI)
- Potential factors if sales are predicted correctly:
  - Store staffing issues
  - Financial implications
  - Business becomes obsolete
  - Customer satisfaction
- Outcome: Predict future sales for the Walmart stores in this dataset

Walmart, Inc Data

- Walmart dataset was retrieved from Kaggle
- Data ranges from February 5, 2010 through November 1, 2012
- 45 stores represented in the dataset
- Features of the dataset:
  - Store – Store number
  - Date – The week of sales (Saturday to Friday)
  - Weekly_Sales – A store's sales for the given week in USD
  - Holiday_Flag – Boolean value and shows whether a week has a holiday or not
  - Temperature – Average weekly temperature in Fahrenheit
  - Fuel_Price – Average weekly cost of fuel
  - CPI (Consumer Price Index) – Measure of inflation or deflation
  - Unemployment – Unemployment rate for a given week

https://www.kaggle.com/aditya6196/retail-analysis-with-walmart-data

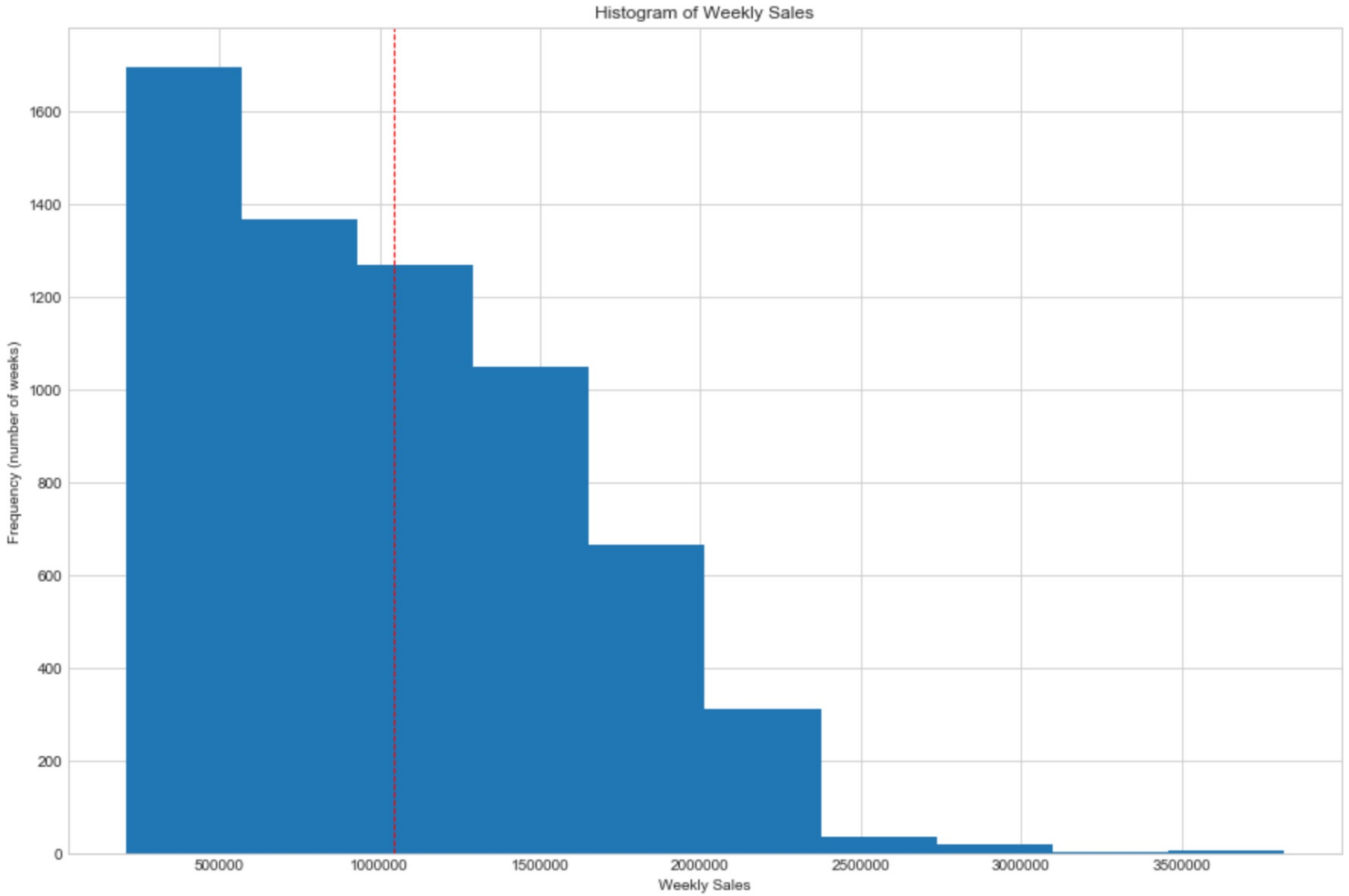
Walmart, Inc Data

- Steps for creating machine learning models
  - Data exploration – Analyze and identify questions about the dataset
  - Data preparation – Putting the data into a usable format
  - Research and start creating possible machine learning models
  - Train and test the machine learning models
  - The test data is compared with the predicted data from the model
  - Models would need to be updated and rerun if the models did not perform as well as a person hoped
  - A final report is created, and a model is identified that would best predict the future data

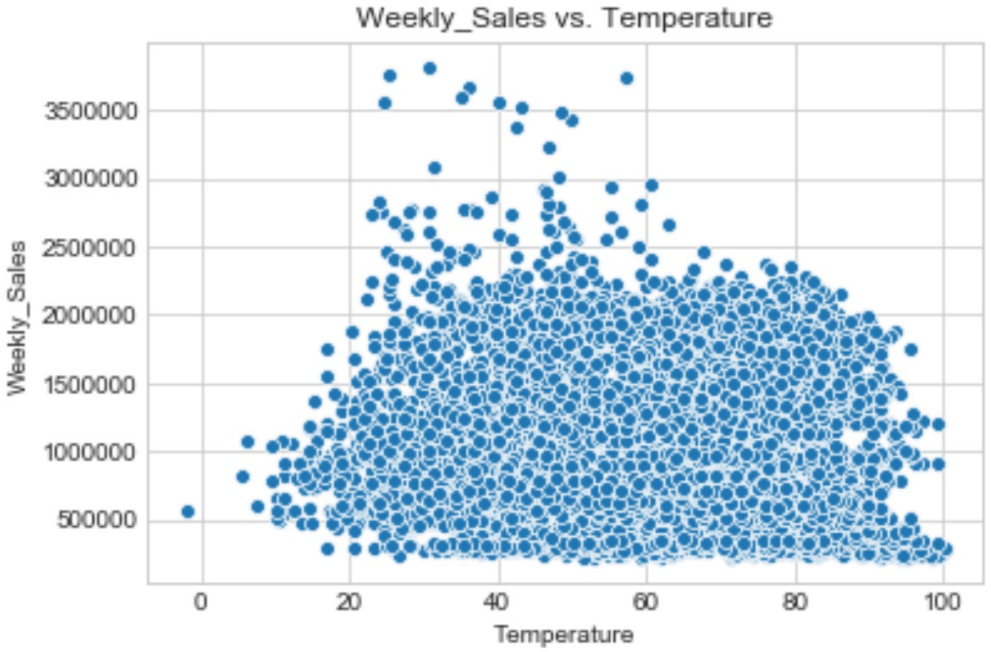
https://www.kaggle.com/aditya6196/retail-analysis-with-walmart-data

Walmart, Inc Data

- The mean Weekly Sales for all 45 stores in the dataset is $1,046,964.88.

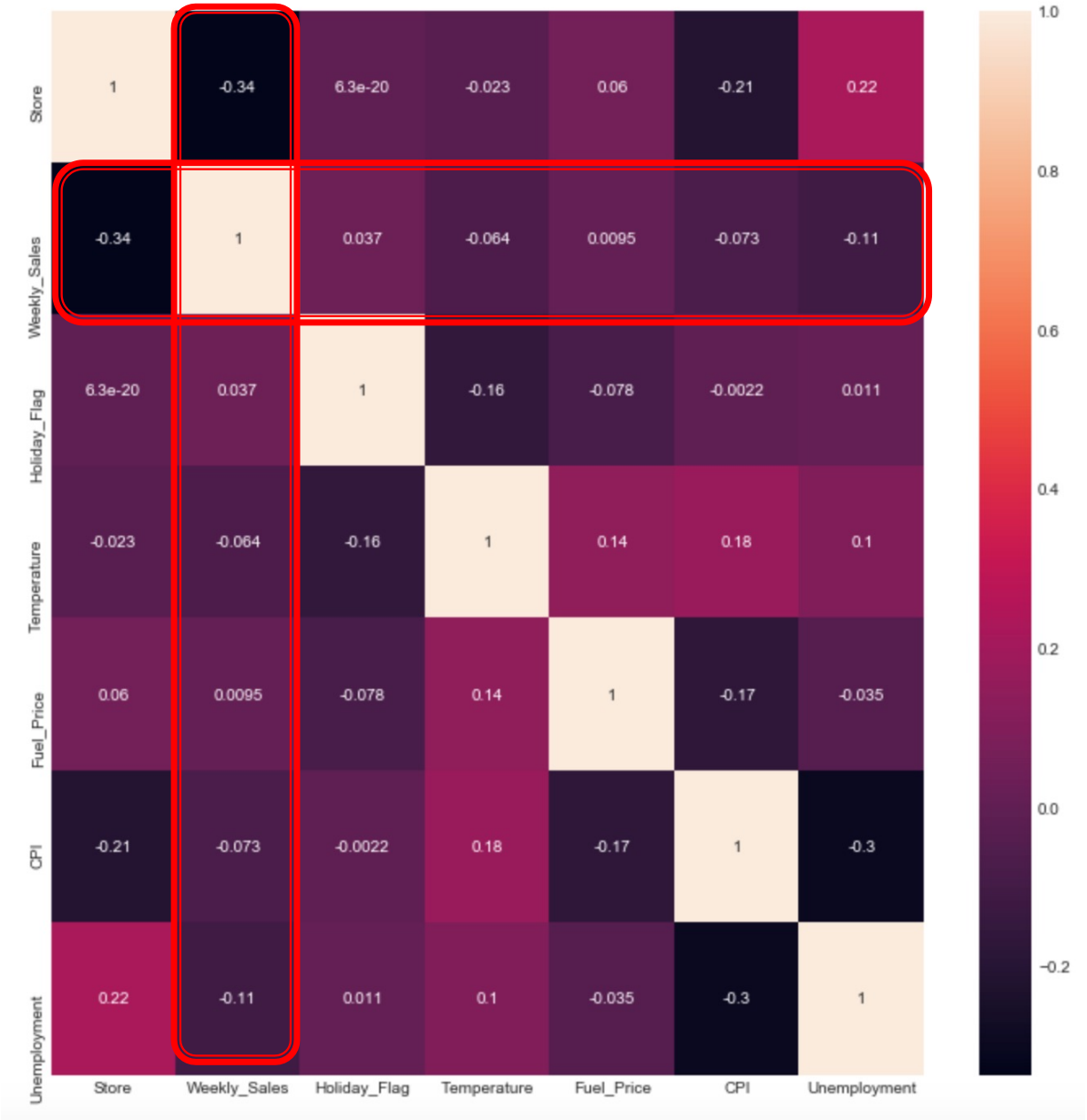


Walmart, Inc Data



Methods: Correlation

- Notable Correlations With Target Variable:
  - Unemployment: -0.11
  - CPI: -0.073
  - Fuel Price: 0.0095
  - Temperature: -0.064
  - Holiday Flag: 0.037
  - Store: -0.34



Models Used

- Linear Regression
- K-Nearest Neighbors
- Random Forest
- Gradient Boosting
- Auto Regressive Integrated Moving Average (ARIMA) Time Series

https://www.kaggle.com/aditya6196/retail-analysis-with-walmart-data

Features & Target Variable for Regression Models

- *Independent Variables*
  - Store Number
  - Holiday
  - Temperature
  - Fuel Price
  - CPI
  - Unemployment
  - Month
  - Year

- *Dependent Variable*
  - Weekly Sales

Results:

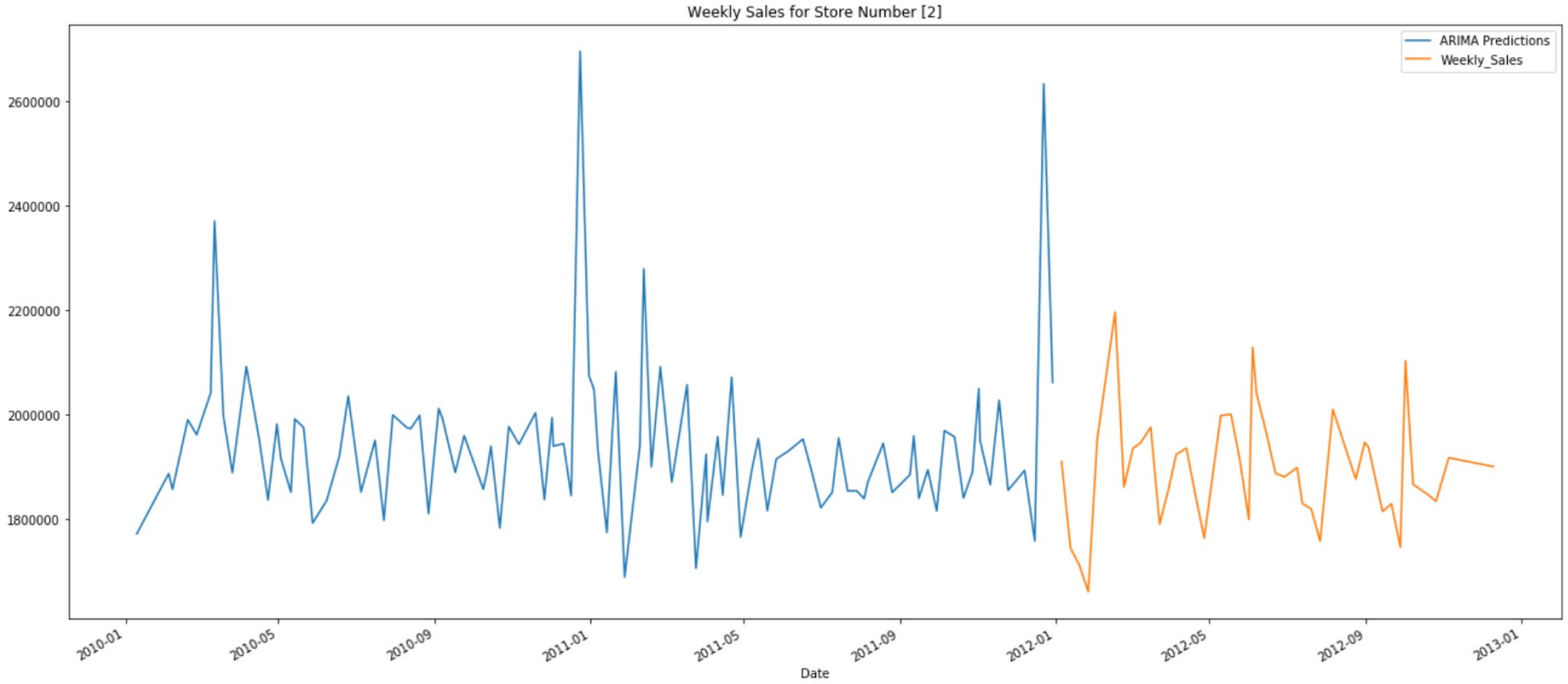
Linear Regression R^2 Score: 0.92172

K-Nearest Neighbors R^2 Score: 0.93460

Random Forest R^2 Score: 0.94517

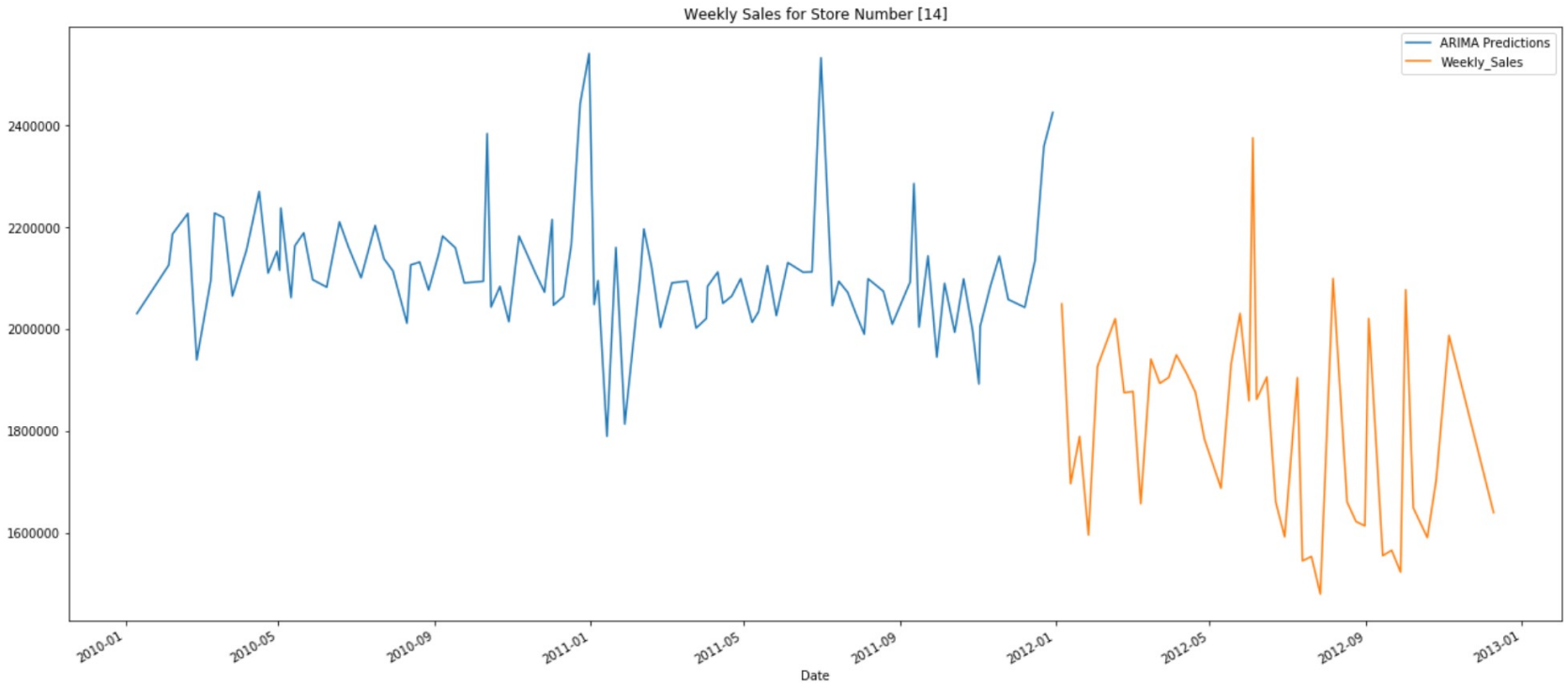
Gradient Boosting R^2 Score: 0.82879

Results: ARIMA Forecasting Model



- Store 2
- Stores Sales in the Millions
- Sales are predictable

Results: ARIMA Forecasting Model



- Store 14
- Stores Sales in the Millions
- Weekly Sales are predicted lower than in the past, which seems something happened when ARIMA forecasted

Issues/Challenges

- We are assuming that any trends in Weekly Sales during this specific time frame are stable and will continue over time
- We are also limited to only 2 years, 8 months, and 28 days in our data set
- R^2 can be an issue if the number is too high. There are variables that are included that may not correlate well to the Weekly_Sales variable
- ARIMA Model is hard to compare with the regression models as their results aim towards different evaluation tools (accuracy vs. predicted values)

Conclusion

- Random forest was the best model to use out of all the regression models
- Regression models are not the first choice when it comes to predicting weekly sales data
- The ARIMA model predicts weekly sales well when the correct independent variables are used within the model
- After reviewing all of the models, the best model for this data would be the ARIMA model because it is geared towards time series