# Exploring Models to Predict Walmart Weekly Sales

Manvi Goyal, Khyati Jariwala, Benjamin Kanarick, Rochan Nehete, Teja Sirigina

Intro to ML, Summer 2022

# Overview

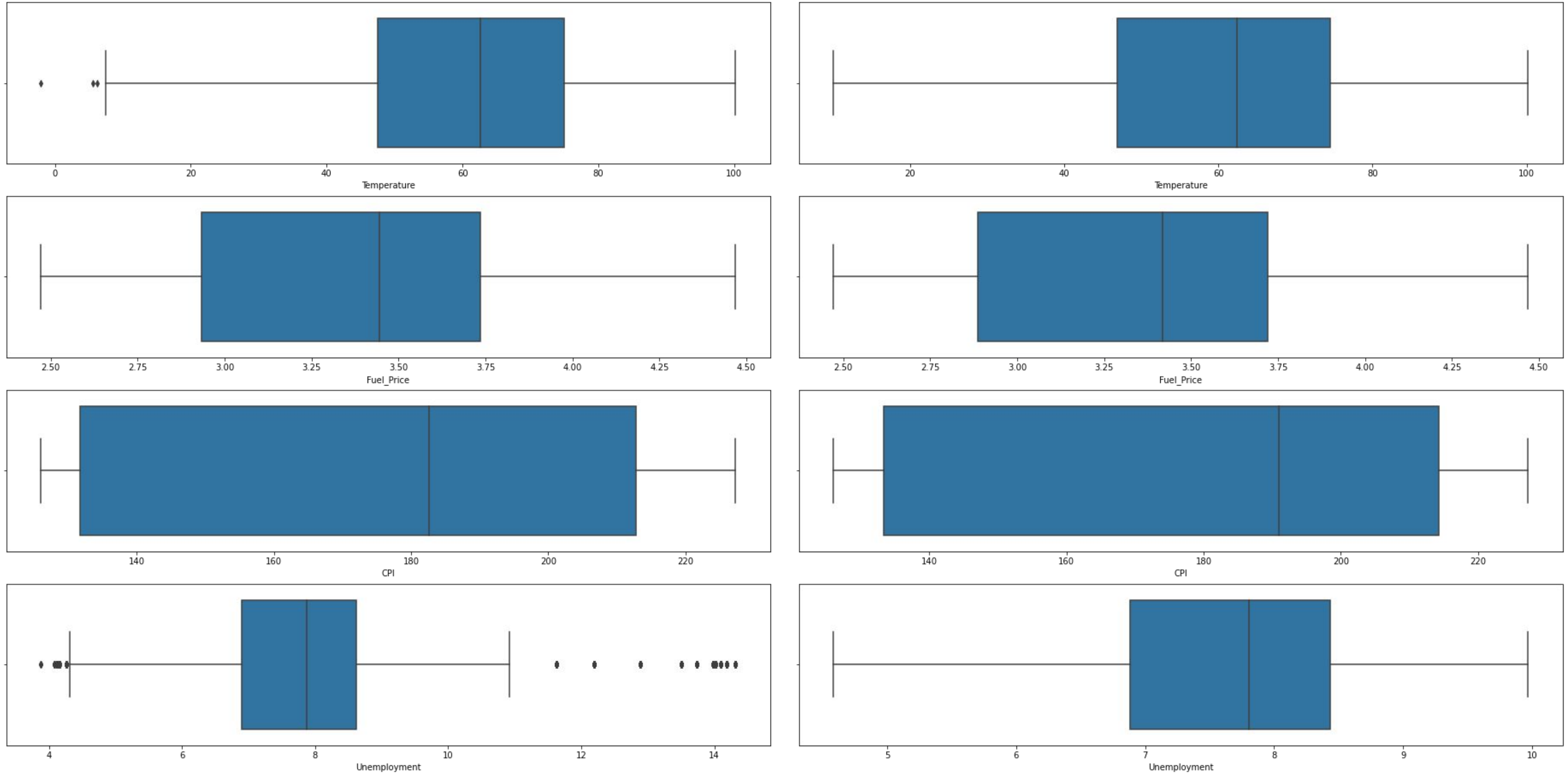
# Problem Statement: Develop the best possible model for predicting Walmart weekly sales.

- Split data into a training set and a test set
- Use the training set to fit the data to the selected model
  - Multiple Linear Regression(parametric)
  - Boosting(non-parametric)
  - Random Forest Regression(non-parametric)
- Use the test set to measure effectiveness of selected model
  - Accuracy
  - RMSE
  - Top predictors
- Compare selected models

# Overview of the Data

- Walmart’s historical retail data
    - Weekly sales from February 5th, 2010 to November 1st, 2012
- Predictors
    - Store
    - Date (broken into Day, Month, and Year)
    - Holiday_Flag
    - Temperature
    - Fuel_Price
    - CPI (Consumer Price Index)
    - Unemployment

# Outliers



0 20 40 60 80 100
Temperature
20 40 60 80 100
Temperature
2.50 2.75 3.00 3.25 3.50 3.75 4.00 4.25 4.50
Fuel_Price
2.50 2.75 3.00 3.25 3.50 3.75 4.00 4.25 4.50
Fuel_Price
140 160 180 200 220
CPI
140 160 180 200 220
CPI
4 6 8 10 12 14
Unemployment
5 6 7 8 9 10
Unemployment

# Model 1 - Multiple Linear Regression

# Multiple Linear Regression: What's The Purpose?

## BIG PICTURE:

- Allows for estimation of how a dependent variable changes as independent variable(s) change

- Using independent variables whose values are known to predict the value of a single dependent value

## IN CONTEXT:

- Useful to see the individual significance for each predictor when estimating weekly sales for Walmart

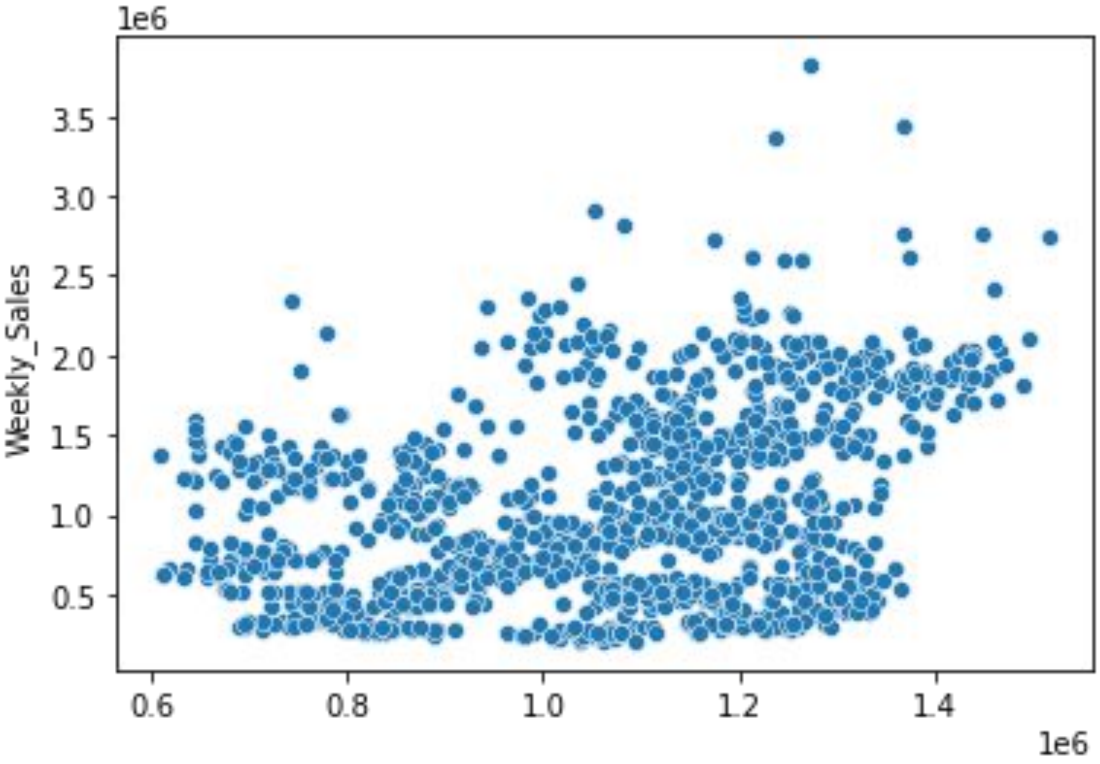
- Least complex model, allows for interpretability and understanding the drivers and relevance of all predictors relative to one another

# Interpreting Our Results

- Multiple R-squared 0.104
  - Not much variability explained by the model
- Certain T-values near 0, not much relevance for some predictors
- Uses the most predictors, subsequent models will have less / perform better

```
> summary(lm.fit)

Call:
lm(formula = Weekly_Sales ~ ., data = Walmart.train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1013297  -418481   -78899   409130  2683419 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  84139931.0 37409881.3   2.249 0.024552 *  
Store          -17792.8      940.2 -18.924  < 2e-16 ***
Holiday_Flag    44805.9    35346.7   1.268 0.205001    
Temperature      -520.2      483.7  -1.075 0.282219    
Fuel_Price      77435.0    31944.8   2.424 0.015389 *  
CPI             -3310.7      251.4 -13.168  < 2e-16 ***
Unemployment    34523.2     9300.1   3.712 0.000208 ***
Day              -914.1      954.3  -0.958 0.338164    
Month           11181.2     2724.0   4.105 4.12e-05 ***
Year           -41106.5    18640.2  -2.205 0.027486 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 558400 on 4516 degrees of freedom
Multiple R-squared:  0.104,	Adjusted R-squared:  0.1022 
F-statistic: 58.26 on 9 and 4516 DF,  p-value: < 2.2e-16
```

# Plotting The Results




**Takeaways:**

Predicted vs. Actual

Accuracy: 13.186%

Less accuracy for more interpretability

Not all predictors are relevant

Parametric → Non-Parametric models

# Model 2 - Boosting

# Boosting:
# What's The Purpose?

## BIG PICTURE:

- Improves predictions from trees by combining trees to produce an overall fit

- To learn slowly from previous trees that have already been grown

## IN CONTEXT:

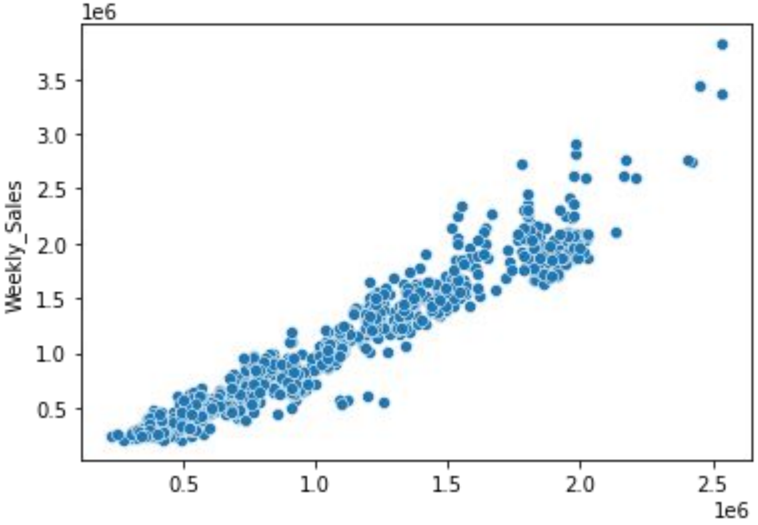
- Visually displays relative importance of each predictor in estimating Walmart weekly sales

- Most complex model, not very interpretable but tells us which predictors are "important"

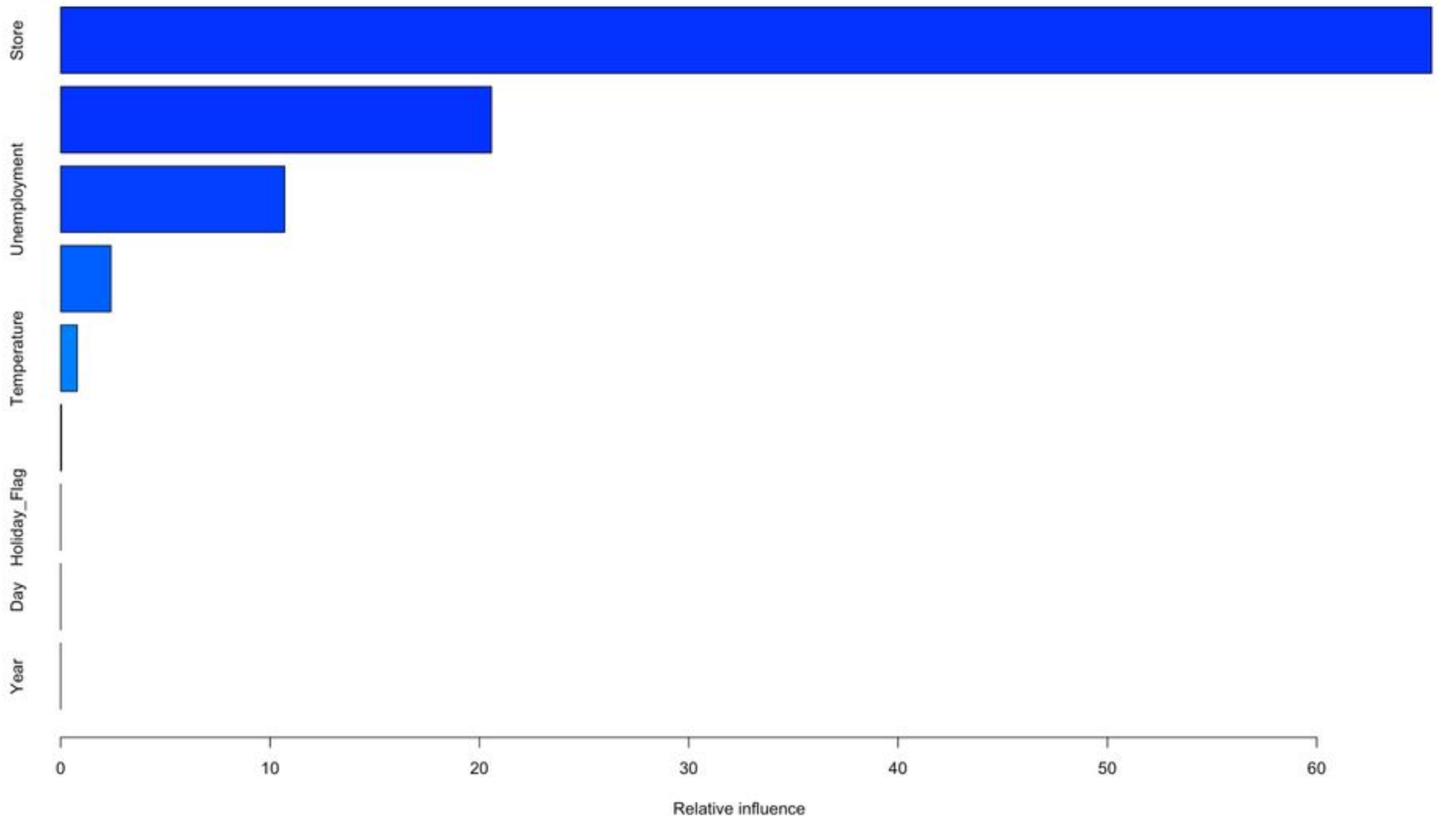
# Boosting Model

- Accuracy: 91.155%
  - Higher than MLR
  - Sacrifice interpretability for accuracy

- RMSE: $175,029.67



# Boosting Predictors



## Boosting Predictors

| | var | rel.inf |
|---|---|---|
| Store | Store | 65.49191141 |
| CPI | CPI | 20.57712006 |
| Unemployment | Unemployment | 10.69869842 |
| Month | Month | 2.40290305 |
| Temperature | Temperature | 0.79112991 |
| Fuel_Price | Fuel_Price | 0.03823715 |
| Holiday_Flag | Holiday_Flag | 0.00000000 |
| Day | Day | 0.00000000 |
| Year | Year | 0.00000000 |

# Model 3 - Random Forests

# Random Forest Regression: What's The Purpose?

## BIG PICTURE:

- A non-parametric model which works well with non-linear datasets, more variability

- Compare variable values using trees, average the outputs from all trees to obtain prediction

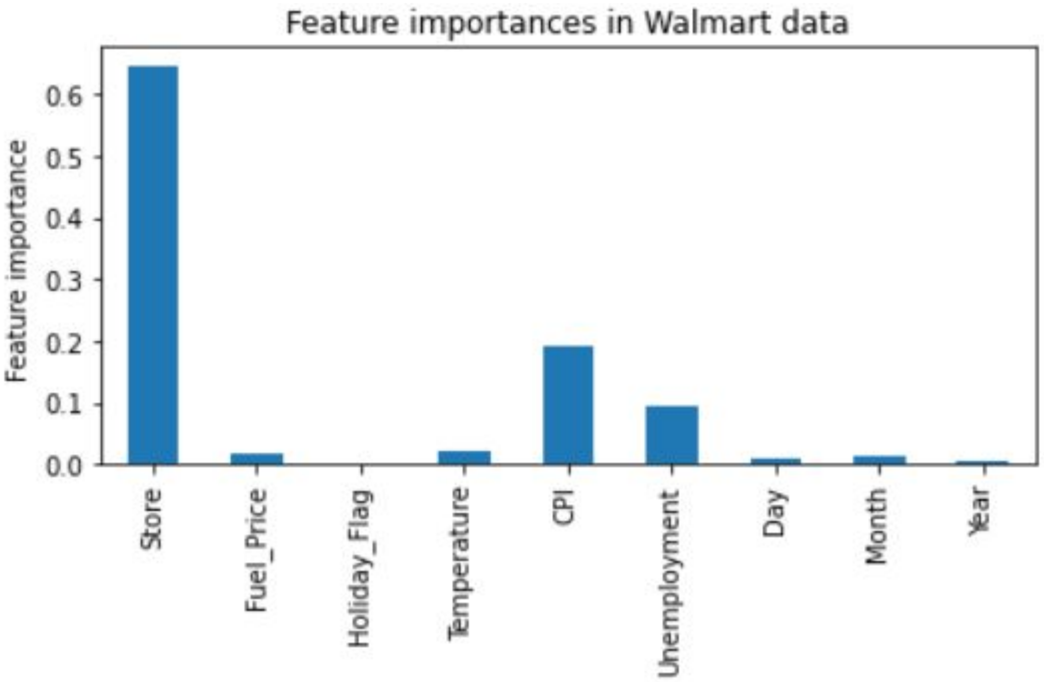
## IN CONTEXT:

- Performance can be tuned by using parameters such as n_estimators, max_depth

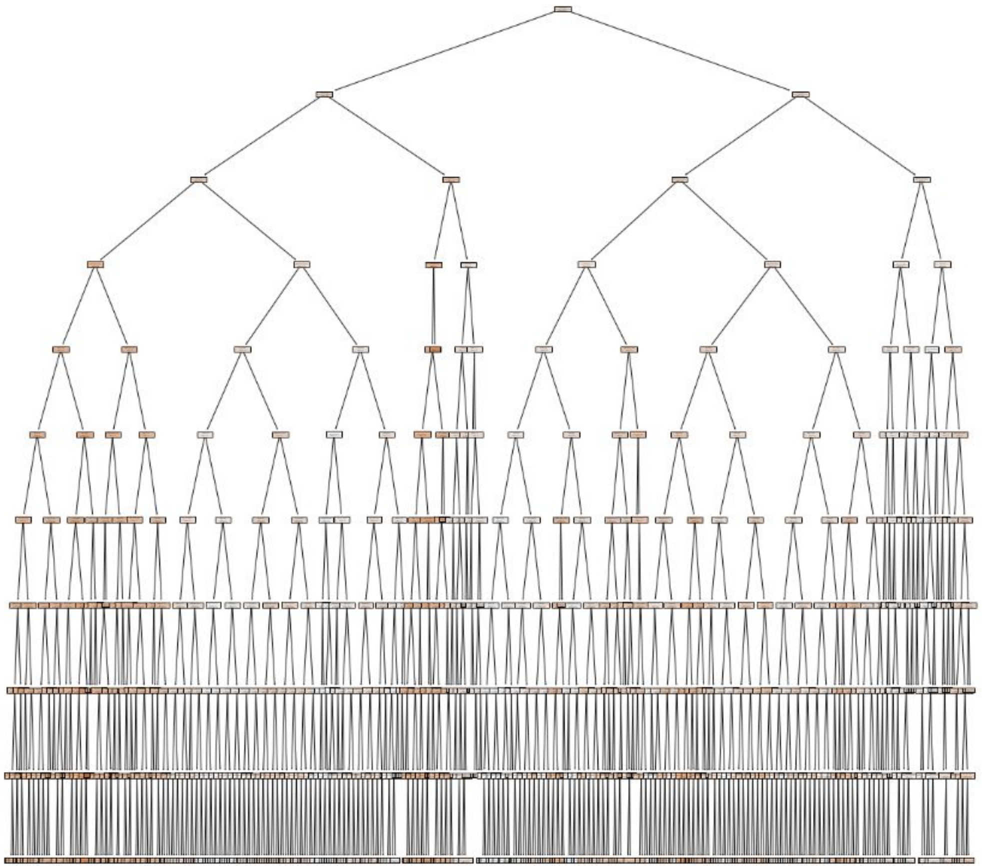
- Trees compute importance of each variable thereby evaluating the most significant features

# Random Forests Predictors

```
Parameters currently in use:

{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': 10,
 'max_features': 5,
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
{'bootstrap': [True, False],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}
```

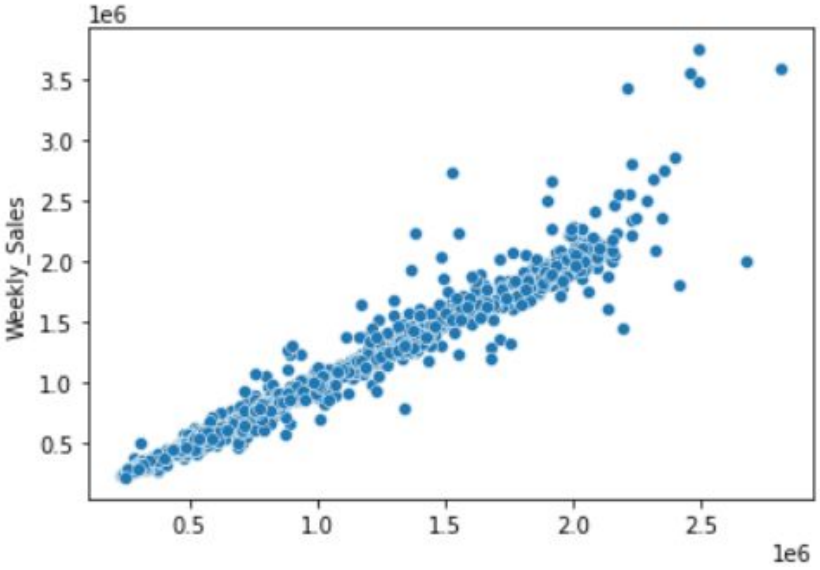


```python
plt.figure(figsize=(20,20))
_ = sklearn.tree.plot_tree(rfr.estimators_[0], feature_names=X_train.columns, filled=True)
```



# Random Forests Model

- Accuracy: 92.36%
- Using Sales, RMSE turns out to be $162,320 on a data ranging from ($209k,$3.8M)
- R squared value turns out to be 0.9236



# Conclusion

# Predictive Model Comparisons

## 13%
### Multiple Linear Regression

Not accurate

Not considered

## 91%
### Boosting

Very Accurate

RMSE: $175,029

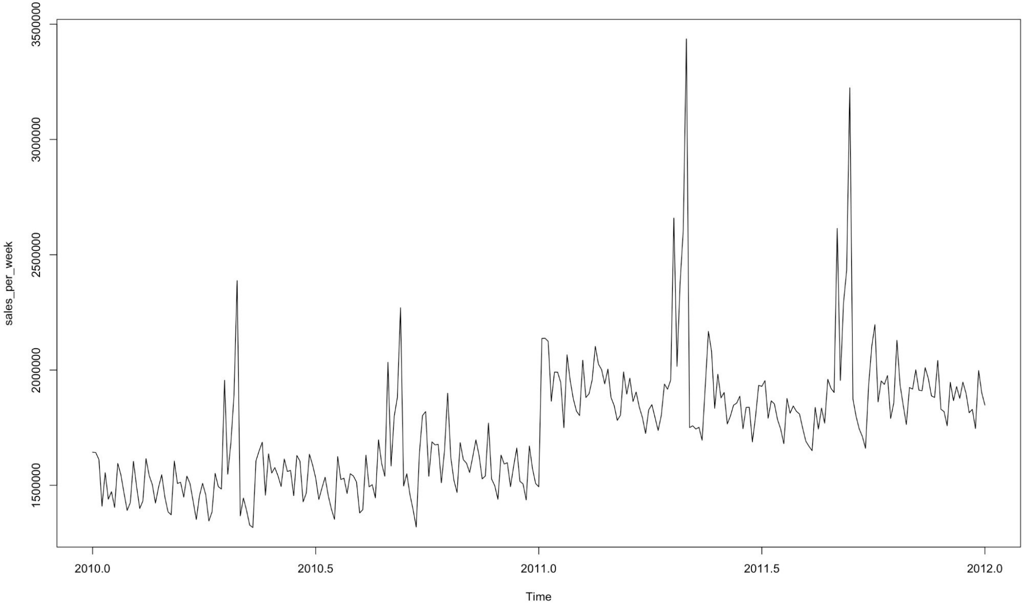
## 92%
### Random Forests

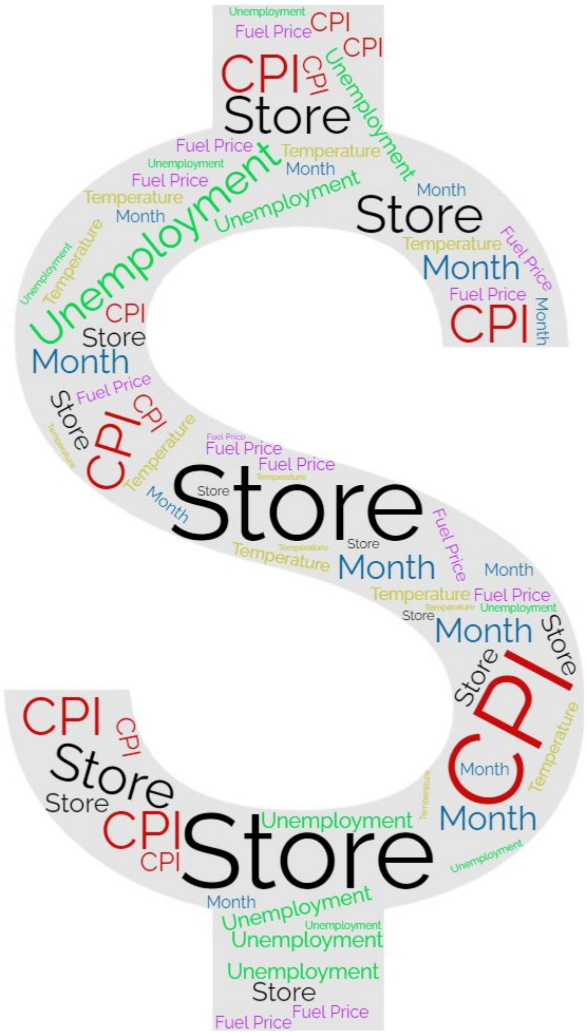
Precise and Very Accurate

RMSE: $162,030

# of Variables (mtry) was 5

# Significant Variables and Further Experimentation



# Summary and Q&A

- We were able to predict Walmart’s Weekly Sales data at 45 stores across 143 weeks.
- We explored and demonstrated the effectiveness of multiple different models (LR, Boosting, and RF) and selected the most accurate one (RF) to be the best choice to predict Weekly Sales.