Walmart Sales Forecasting – Report

The Problem

With 56% of U.S. grocery market-share, 4,743 U.S. stores, 240M customers worldwide, and $559B in worldwide revenue it's easy to say that Walmart is huge (Statista 2021, April 27), and with Walmart's huge size comes huge amounts of data.

Thus, in 2014 they created a public competition for data-scientists to predict store sales using a selection of their anonymized datasets. The prizes were potential positions at Walmart.

Here, I accept the challenge.

The Data

Stored across multiple .csv files, once merged and cleaned the main features include:

- Date (2010-2013)
- Store (45 unique, numeric)
- Department (many unique per store)
- Weekly Sales (target feature, $)
- Holiday (T/F, this is of particular importance)
- Store Size (numeric)
- Etc.
    - Temperature, Fuel Price, Markdowns…

The Goal

Predict sales per store, per department, per year, per week.

The Metric of Success

Mean Absolute Error (hereafter MAE)

The Approach

This is a regression problem, thus experiment with many untuned regression machine-learning algorithms and see which perform best based on MAE.

Then, refine those best models with randomized-search cross-validation (hereafter RSCV). Randomized-search as opposed to grid-search so to reduce time and resources spent while also maximizing hyperparameter points explored. Performance is measured as the mean MAE over the cross-validation folds.

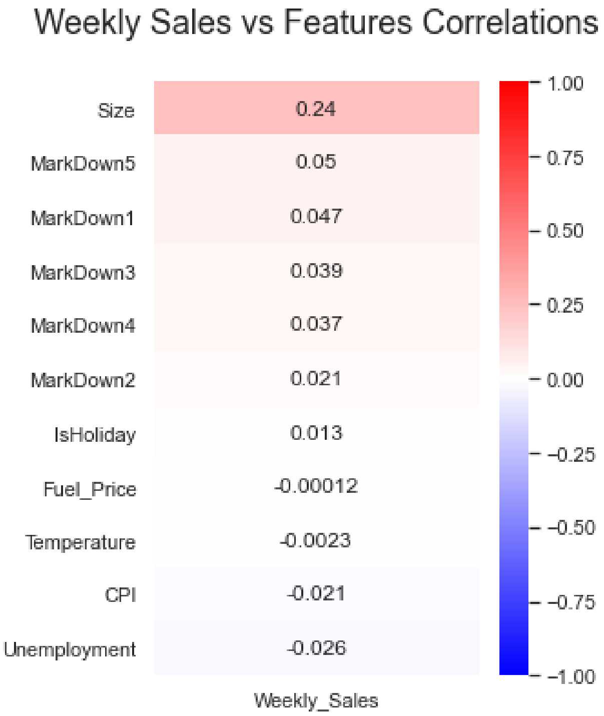
Of the refined models, choose the best performer on a basis of balance between cost and performance.

Finally, and again, refine the final chosen model's hyperparameters via RSCV and review performance.

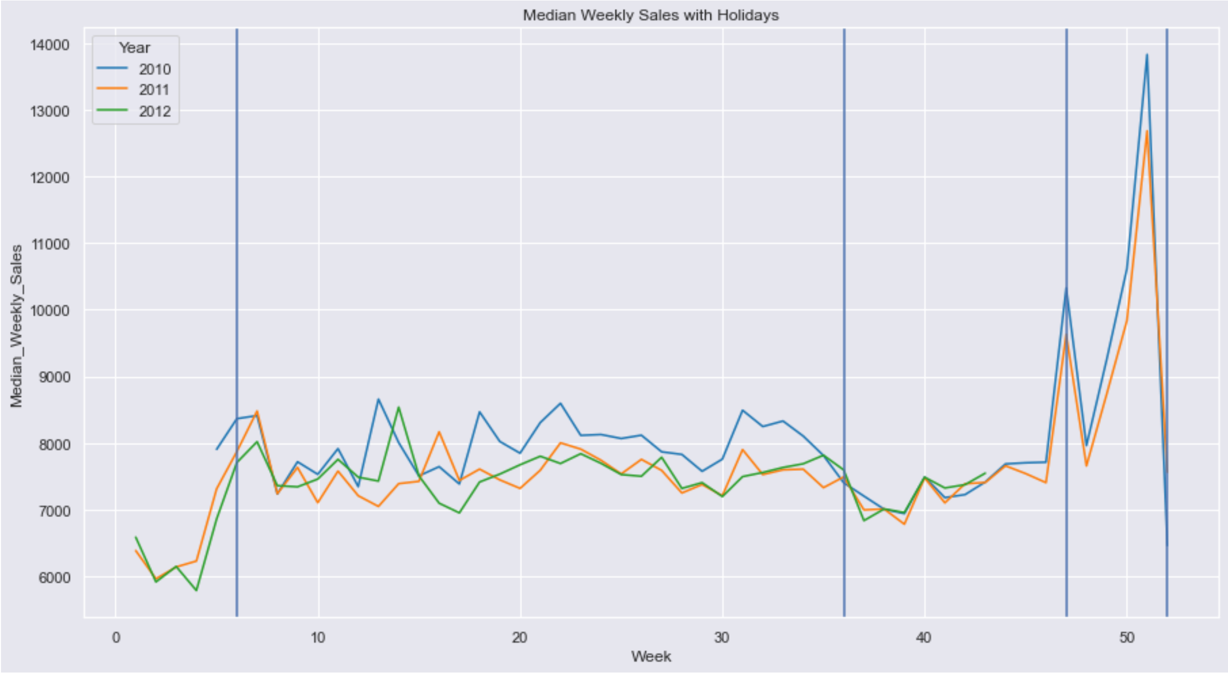
The Findings

No features highly correlated with weekly sales (the target feature). Though, store size correlated higher than the other features by a long shot with a value of 0.

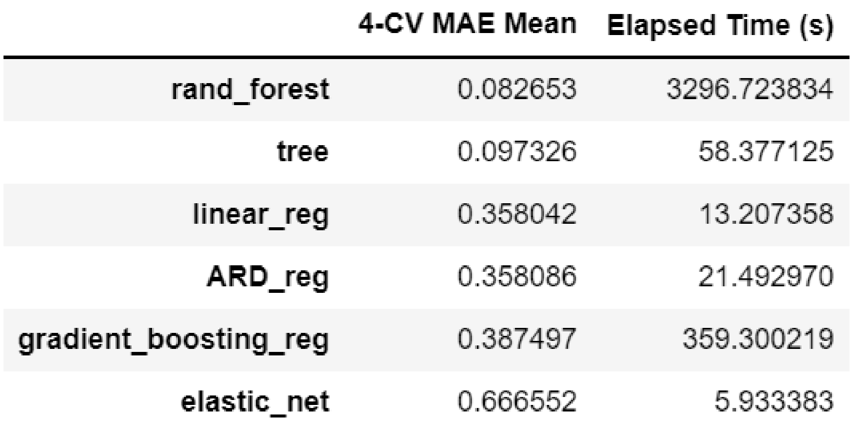
To clarify, a value of 1.00 is perfect positive correlation, -1.00 perfect negative correlation.



Trends in sales were visibly (and predictably) annual, with spikes in sales occurring around holidays. The vertical blue lines denote the Super Bowl, Labor Day, Thanksgiving (Black Friday), and Christmas from left to right. These trends may be obvious, but nonetheless interesting to visualize and model with.



Now, modelling. Tree based models significantly outperformed their counterparts, as seen in this table of model vs. model performance:



	4-CV MAE Mean	Elapsed Time (s)
rand_forest	0.082653	3296.723834
tree	0.097326	58.377125
linear_reg	0.358042	13.207358
ARD_reg	0.358086	21.492970
gradient_boosting_reg	0.387497	359.300219
elastic_net	0.666552	5.933383

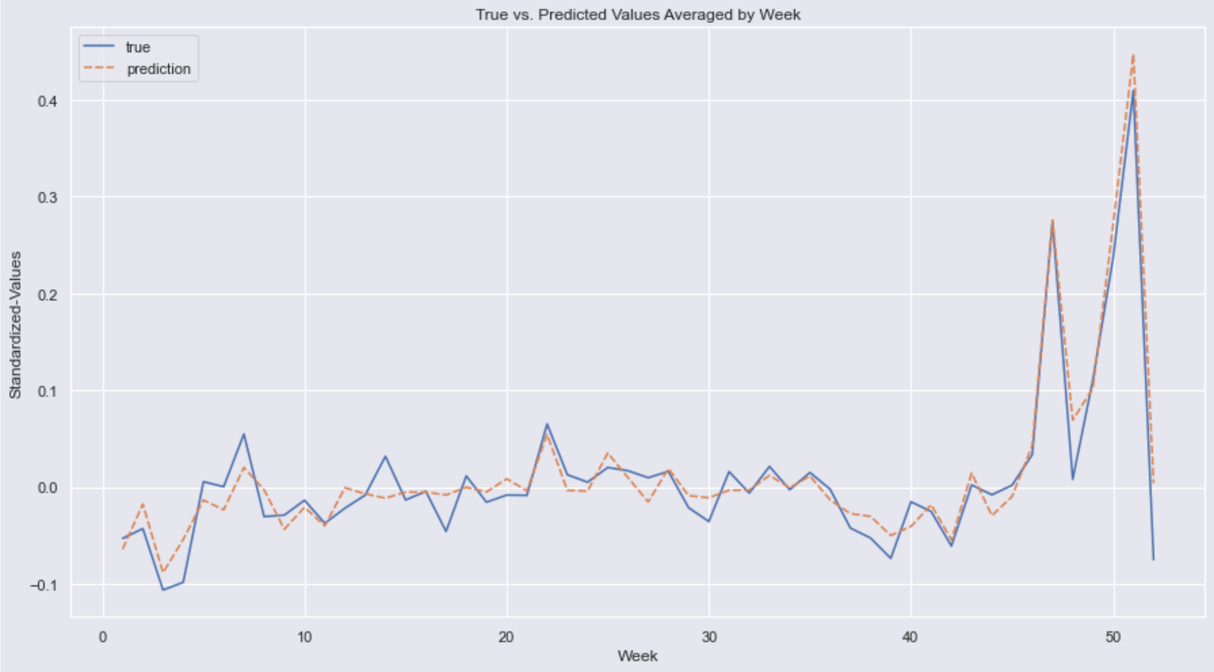
Random-forest achieved the lowest mean 4-fold cross-validated MAE, but only marginally compared to a decision tree; especially when considering the difference in training time.

Because of their mean MAEs, random-forest and decision-tree regressors were chosen to progress and be refined via RSCV (4-fold). Random-forest took 430 minutes to train 12 candidates on my machine, achieving a best MAE of 0.0825. Decision-tree took 18 minutes to train 30 candidates and achieved a best MAE of 0.0973. Due to the high cost of training with only marginal performance improvement, I chose the decision-tree as the final model to refine.

The Final Results

The final refinement RSCV'd over the hyperparameters criterion, max_depth, and min_impurity_decrease, finding the best parameters of 0.1, 119, and friedman_mse respectively after 100 candidates (5-fold CV). On the training data the best MAE was 0.0936. All while only taking 38 minutes on my machine.

Finally, the final model's MAE of predictions against the test set was 0.0863.



Since the testing data spans many stores, departments, years, and is randomized,

this is my best approximation of how to visualize truth vs. predictions