Walmart Weekly Sales Forecasting

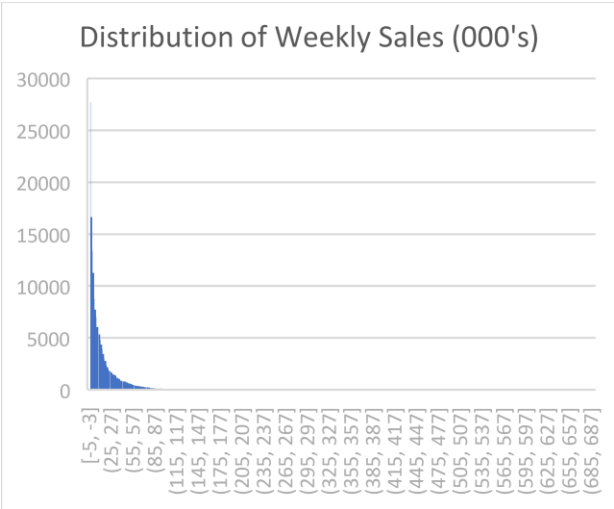
The Walmart Weekly Sales dataset used for this report is downloaded from the CS 598 (Stat 542) Piazza forum under the resources tab. The dataset contains 421,570 rows of weekly sales data from 45 different Walmart stores. Each store further subdivides weekly sales by dept – of which there may be up to 99 different departments within each store. The 421,570 rows are broken down into a train_ini file (covering approximately 13 months beginning in March of 2010 and contains 164,115 rows of weekly sales data) that will be used for training a model, and 10 different test folds (each containing roughly 9 sequential weeks of sales data and over 25,000 rows of weekly sales data) to determine the test accuracy of the model.

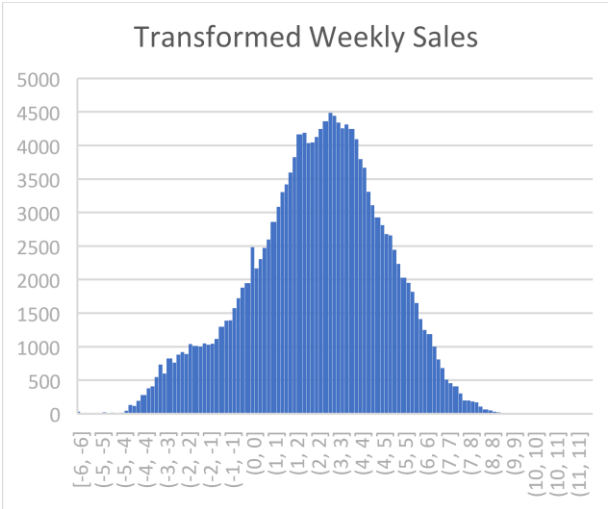
The purpose of this report is to accurately predict future weekly sales volumes by department for each store. The test accuracy measurement used for this report will be a weighted average error (WAE) where the accuracy for holiday weeks (Super Bowl, Labor Day, Thanksgiving, and Christmas) are weighted 5x as much as non-holiday weeks. The mean weighted average error (MWAE) will be reported over all 10 test folds.

Pre-processing of data

The data provided is a clean datatset, there is really no data that needs to be processed. There are some weeks of missing data, but this report does not attempt to replace that data with the mean or with nearby data values.

The `Weekly_Sales` in the dataset can be seen to exhibit a high degree of skew and probably can not be modeled with the a linear model successfully, so this report applies a Box Cox transformation to attempt to make the response more of a normal distribution as can be seen in the below histograms.





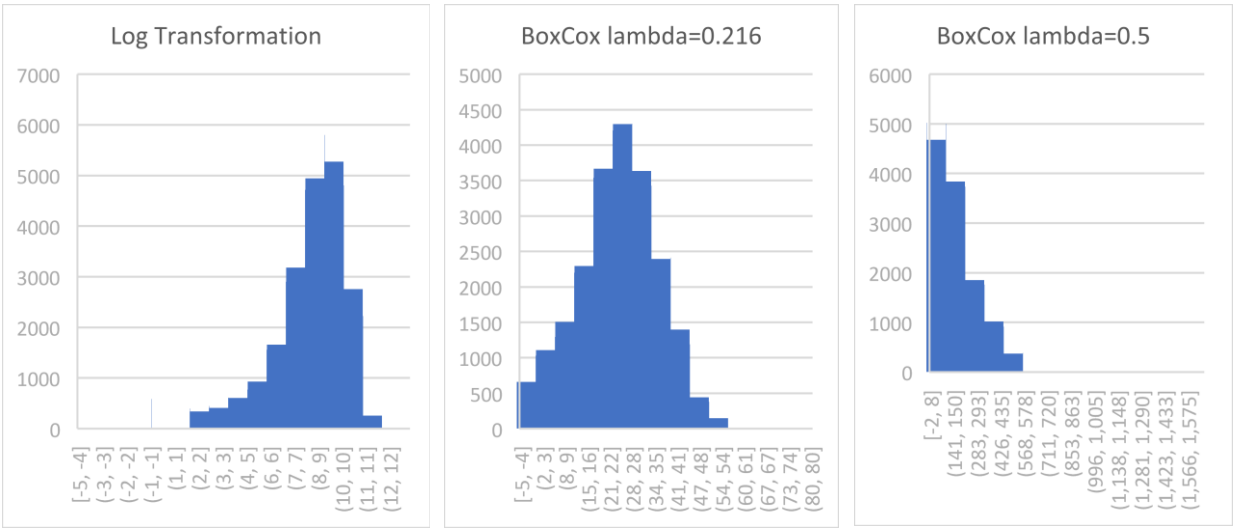
There are some weeks that are showing negative sales numbers. It is not clear how negative sales are generated, but this report made the decision to make any negative sales numbers zero values. The rationale is that negative sales could be a one off event such as opening a new department. Also, negative sales represent a very small subset of the data and are not sustainable in any department – that is the department would be shut down. Negative sales would not add any forecasting value to the report.

Model Implementation

Following the lead of the professor and TA's this report looks at a linear model to predict future sales. The initial thought was that the `Weely_Sales` data needed to be transformed to make it more of a normal distribution. This report looked at a log transformation which created left skewed data and did not increase the accuracy of the forecasting, in fact it significantly decreased the accuracy.

Since log transformation did not work, this report then looked at performing a Box Cox transformation on the `Weekly_Sales` data. The first thought was to apply the Box Cox transformation that would give a normal distribution to this data. The calculated value of lambda that would achieve this was found to be around 0.216. Applying this transformation helped to bring the weighted absolute error down, but still not below the threshold needed for full credit.

The lambda value was increased and decreased to see the impact on the accuracy of the model. Moving the lambda value higher was found to improve the prediction accuracy of the model. A value of 0.5 was close to the optimal value achieving the best mean weighted absolute error. The distribution of the data for the three transformations can be seen below.



Results

The below table shows the error rates for the 10 different folds for the final model.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
WAE	1980.86	1450.16	1423.25	1546.80	2276.24	1631.92	1680.30	1404.24	1426.23	1407.40

MWAE
1622.74

The models were run under the following environment:

Computer Component	Component Characteristic
Processor	**Intel® Core™ I3-3227U CPU @ 1.90 GHz**
Installed RAM	**8.00 GB (7.89 GB usable)**
System type	**64-bit operating system, x-64-based processor**
Operating System	**Windows 10 Home, Version 2004, Build 19041.508**
Running time over all 10 folds	**Approx 6.5 minutes**

Observations

One of the interesting observations for the model used in this report was that the distribution of the transformed response variable was not a normal distribution, it has a good bit of right skew in it. This transformation certainly would not be good for interpretation of the model, but it lends itself well to prediction.

Another thing that was observed was that peak sales did not always occur during the holiday week, sometimes they peaked the week before the holiday period and sometimes they peaked after the holiday season. Some of this variability could be attributed to when the actual holiday occurred within the week – in other words if the holiday was at the beginning of the week, peak sales may have been in the previous week. The model may be able to be improved upon if this is indeed something that is true. This report did not attempt to model this possible behavior.

Fold 5 seemed to be the most difficult fold to get a low MAE with. This could be due to the fact that there were multiple holidays in that period and when the holidays fell compared to the previous year. Fold 1 was also difficult to get a low MAE, although its not very clear why since there are not any holidays in that period – maybe it could be due to differences in weather.