Table of Contents

Abstract:

One challenge of modeling retail data is the need to make decisions based on limited history. If Christmas comes but once a year, so does the chance to see how strategic decisions impacted the bottom line. Sales forecast is so critical for businesses in inventory as well as demand planning since companies do not want their customers to go home unsatisfied due to unavailability of goods. In this paper, we explore the ways to predict future sales data for 45 Walmart stores located in different regions. The objective is to minimize the prediction error and to predict as accurately as possible. We tried a range of predictive models such as multilevel models with different levels, local regression models for store and type and Random forest and we attempt to demonstrate why multilevel and boosted regression trees work well with this data.

1. Introduction & Problem Description:

The problem at hand is to forecast Walmart stores' sales for each department. The data set we are given is the historical sales data for each department in 45 Walmart stores located in different regions. We are also given the store type and size dataset and features dataset which is different across stores every week.

The challenging part is to consider the effect of promotional markdowns to the sales. Walmart runs several markdowns preceding prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving and Christmas. Getting accurate predictions for weeks including these holidays is more important than non-holiday weeks from a strategic point of view.

2. Data Description:

The dataset has about 421 thousand rows in total with 17 columns. The data is provided in 3 tables - Features, Stores and Train along with a test set on which scoring or predictions are to be done.

Screenshots for the three tables along with their descriptions follow:

2.1 Stores.csv

This file contains anonymous information about the 45 stores, indicating the type and size of store.

Type - indicates the type of the store. The Three types present in the data are A, B, C. It is observed that, Depending on the size of the store, this type variable was assigned to A, B, C.
Size - indicates the size of the store. Size data for all 45 stores were available.

Store	Type	Size
1	A	151315
2	A	202307
3	B	37392
4	A	205863

2.2 Train.csv

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

Store - the store number. 45 different stores were present in the data set.
Dept - the department number. Some stores have more departments than other stores. The highest number of departments available was 99.
Date - the date information for a given record.
Weekly_Sales - sales for the given department in the given store at the given date.
IsHoliday - whether the week is a special holiday week

Store	Dept	Date	Weekly_S	IsHoliday
1	1	2/5/2010	24924.5	FALSE
1	1	2/12/2010	46039.49	TRUE
1	1	2/19/2010	41595.55	FALSE

2.3 Test.csv

This file is identical to train.csv, except the weekly sales are withheld. We need to predict the sales for each triplet of store, department, and date in this file.

2.4 Features.csv

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

Store - the store number
Date - records date information
Temperature - average temperature in the region in a given week
Fuel Price - cost of fuel in the region

MarkDown1-5 - anonymous data related to promotional markdowns that Walmart is running. Markdowns data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.

CPI - the consumer price index
Unemployment - the unemployment rate
IsHoliday - whether the week is a special holiday week

Markdowns time periods

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
1	2/5/2010	42.31	2.572	NA	NA	NA	NA	NA	211.0963582	8.106	FALSE
1	2/12/2010	38.51	2.548	NA	NA	NA	NA	NA	211.2421698	8.106	TRUE
1	2/19/2010	39.93	2.514	NA	NA	NA	NA	NA	211.2891429	8.106	FALSE

3. Preprocessing I: Preliminary

3.1 Merge

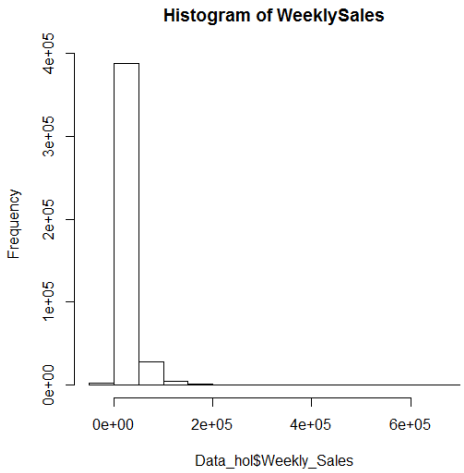
The store.csv, training.csv and features.csv are merged together as a single dataset on which the prediction models were run. Date is split into year, month and week of the month

ID	Store	Date	IsHoliday	Dept	Weekly_S	Type	Size	Temperat	Fuel_Price	MarkDown	MarkDown	MarkDown	MarkDown	MarkDown	CPI	Unemploy	Week	Month	Year
6732	1	########	TRUE	42	7324.21	A	151315	60.14	3.236	410.31	98	55805.51	8	554.92	218.4676	7.866	4	11	2011
6733	1	########	TRUE	28	526.29	A	151315	60.14	3.236	410.31	98	55805.51	8	554.92	218.4676	7.866	4	11	2011
6734	1	########	TRUE	55	30868.94	A	151315	60.14	3.236	410.31	98	55805.51	8	554.92	218.4676	7.866	4	11	2011

3.2 Transformation:

3.2.1 Log transform

The distribution of Weekly Sales looked slightly right skewed as shown in the figure below. We thought about taking a log transform for sales in order to correct for that and also to have the luxury of interpreting the coefficients as elasticities. But, some of the weekly sales were negative, which meant products were returned. Since the log of a negative value is not defined, log transformation of Weekly Sales is not appropriate.



3.2.2 Standardizing the variables:

Standardization of data is important to ensure that the variables measured at different scales contribute equally to the analysis. Though we lose a bit of interpretability, the error does not get inflated due to the scale differences in some of the variables. Since minimizing the prediction error is the primary goal here and not interpretation, we went ahead with standardizing all the variables. In case we are concerned about the interpretation of the coefficients, since it is just a linear transformation we can always multiply by the standard deviation and then add back the mean. The equation shown below is used to transform the dataset to have zero mean and unit variance.

$$x_{new} = \frac{x - \mu}{\sigma}$$

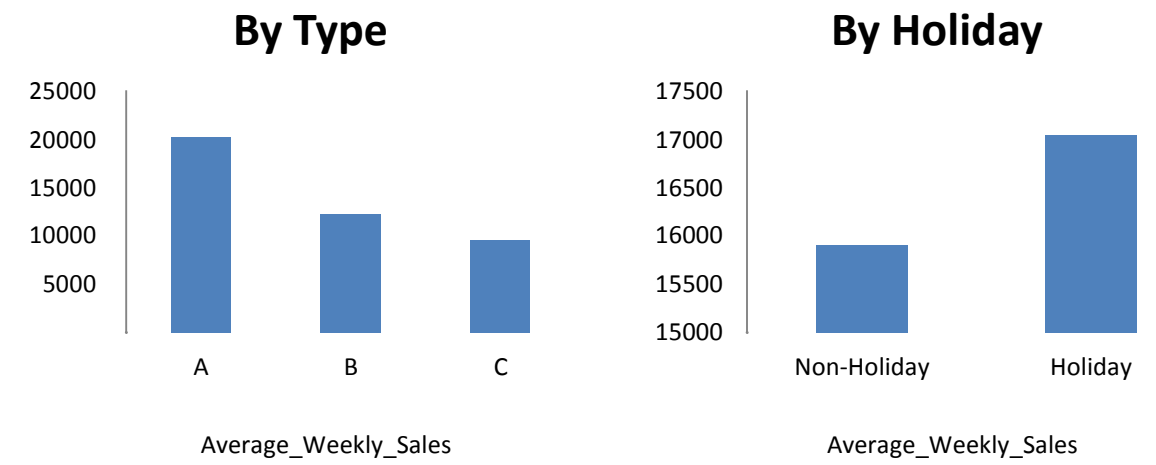
3.2.3 Handling date:

Date here is a very important variable. We separated the date into year, month and week of the month. We thought the month variable would provide significant information because the sales are high during certain months like Thanksgiving, Christmas. In some cases, the week of the month may also have a good impact since people generally tend to buy more during the start of the month as soon as they have salaries in hand and stock up things.

For some models (viz. MLM) , we used the R date function which stores date as a numeric value from a particular point in time. This allows us to clearly see the effect of seasonality in sales over the 3 years.

3.3 Exploratory Analysis & Flat Model:

The average weekly sales for stores according to type and holiday are shown above. The sales are high for store type A followed by B and C. We interpret type A stores as super stores, type B as medium sized and type C as small sized stores. Also, the holiday sales are far more than the non-holiday sales.



3.3.1 Correlations between variables:

Most of the variables in the dataset are uncorrelated. Only 2 significant correlations were observed.

	Date	Dept	Weekly_Sales	Type	Size	mperatu	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	iemployme	year	month	day
Date	1																
Dept	0.0006	1															
Weekly_Sales	-0.0267	0.13713	1														
Type	-0.0195	-0.0172	-0.174545769	1													
Size	0.01976	0.01831	0.212191555	-0.8622	1												
Temperature	0.74277	-0.0017	0.013029329	-0.0557	0.078	1											
Fuel_Price	0.41878	-0.0007	-0.020556682	0.10183	0.005	0.2508	1										
MarkDown1	0.04616	0.0044	0.037340636	-0.1157	0.185	0.0193	0.08430849	1									
MarkDown2	-0.3276	0.00112	0.015829385	-0.0438	0.082	-0.332	-0.2438358	-0.0016151	1								
MarkDown3	-0.2258	0.00347	0.062791754	-0.0185	0.035	-0.088	-0.093533	-0.130754	-0.0512404	1							
MarkDown4	-0.0164	0.00465	0.029471149	-0.0751	0.134	-0.058	-0.0257842	0.8289277	-0.0175171	-0.08021564	1						
MarkDown5	-0.0693	0.00479	0.059649269	-0.1951	0.21	0.0243	-0.1375639	0.1071251	-0.0233904	-0.04337326	0.10111382	1					
CPI	0.02957	-0.0069	-0.016370953	-0.1138	0.02	0.201	-0.3653183	-0.0460047	-0.0369921	-0.02502964	-0.0416622	0.07610638	1				
Unemploymer	-0.0853	0.00281	-0.037735426	0.10073	-0.02	-0.009	0.27656762	0.0644277	0.0183336	0.013967098	0.01881746	0.00333647	-0.2913	1			
year	0.6431	-0.0015	-0.0574983	-0.0036	0.004	0.3907	0.4027707	0.1859573	-0.2264893	-0.32759779	0.16660343	-0.1140067	0.01821	-0.07476	1		
month	0.06716	0.00257	0.050833063	-0.0104	0.011	0.1713	-0.1536852	-0.1767434	-0.0124285	0.212150436	-0.2085726	0.09921791	0.00446	0.01811	-0.7186	1	
day	0.13469	-0.0009	0.004815331	-0.0364	0.032	0.0229	0.13145483	-0.3406514	0.0940223	0.136285682	-0.3479485	-0.1655718	-0.0145	0.018872	-0.07288	0.142084	1

1) High negative correlation between size and type.
2) High positive correlation between Markdown 1 and Markdown 4

We removed the size variable and the Markdown1 variable for subsequent analysis as they don't add any additional information.

To understand more about the data and the variables, we just ran a flat regression model by removing all missing variables. The variables store and department are dummy coded since store 1 and store 45 should mean the same and it does not mean that 45 > 1. Similarly, month, week and year are dummy coded. The data set is split into train and test with 70% of the data going into training set and 30% into the test set by random sampling. The results are given below.

Train R-Square	Test R-Square	Train RMSE	Test RMSE
0.68	0.56	0.56	0.55

4. Preprocessing II: Handling missing values and Sampling

There are missing values for Markdowns from 2/5/2010 to 11/4/2011 (~90 values for each department). Also, in most recent values (5/3/2013 - 7/26/2013), ~ 12 values for each store for CPI and unemployment are also missing.

For imputing the missing value of Markdowns, different techniques are used.

1) Replacing with Global means
2) Replacing with local store level means

4.1 Impute Missing values

4.1.1 Replacing with Global mean

Mean of the variable would be a reasonable estimate for the missing values to start with. So, we imputed all the missing values with global means of the respective variables.

4.1.2 Replacing with local store level mean

As there are different types of stores as discussed above which vary significantly in size and operations, we thought a store level mean would do better than a global mean. So, we imputed all missing values with store level mean.

Imputation Method	Train R-Square	Test R-Square
Global Means	0.65	0.62
Local Means	0.66	0.65

4.1.5 Summary of regression results for different imputations

4.2 Sampling – Representation of Train and Test

4.2.1 Random Sampling:

Random sampling assumes that each observation has equal probability of entering the sample. When 70% of the data is selected at random for the training set, we are not sure whether there is a proper representation of each store in the training set. As a result, the model may not perform well on the test set. We thought we could use the idea of stratified sampling.

4.2.2 Stratified Sampling:

Unlike simple random sampling method, stratified sampling method is advantageous to sample each subpopulation independently. The entire population is divided into strata before sampling takes place. It's a balanced sampling method that improves the representativeness of each stratum and reduces sampling variance.
Here we divided the entire data by considering store as strata and then performed a random 70-30 split in each strata to come up with the test and train datasets. This way, we are sure to have a proper representation of all the stores in the train as well as the test.

But the results did not vary considerably for random and stratified split. When we went back and checked, the random split was performing as good as the stratified split. So, there was not much improvement due to stratified splitting. Hence we continued with random split for all our subsequent analysis

Yet still, it's important to keep in mind that when sub-populations vary considerably, it is advantageous to sample each stratum independently. But when sub-populations do not vary considerably, random sampling does as good a job as stratified sampling.

5. Predictive Models

5.1 Separate Models for Holiday and Non-Holiday:

As we saw in the summary statistics, the sales varied significantly by season. Holiday sales were way higher than the non-Holiday sales. Modeling both of them separately intuitively seemed a better choice.

5.1.1 *Holiday:*

We separated the holiday data and tried two flat regression models with global mean imputation and the local mean imputation (mean of the holiday data).

Imputation Method	Train R-Square	Test R-Square
Global Mean	0.56	0.54
Local Mean	0.63	0.62

The R-square for global mean imputation is less than that of local mean imputation since the markdowns for holiday dates is significantly different from the global mean.

To check for the seasonality the month variable was dummy coded into different holiday seasons. From the regression coefficients, we infer that when compared to super bowl, sales were high during Christmas followed by thanksgiving and labor day.

5.1.2 Non-Holiday:

The results with global and local means imputation of missing values for Non-Holiday data are shown below. There is not much difference between both the imputations as the global and local means for non-holiday data are almost the same.

Imputation Method	Train R-Square	Test R-Square
Global Mean	0.67	0.65
Local Mean	0.66	0.64

As imputing the missing values was not improving the performance, we chose to go ahead by removing all the missing values. Because there was a large dataset at our leverage, even after

removing the missing values store and department combination had a lot of data points. This approach was followed for subsequent modelling techniques as there was not a significant loss of information.

5.3 Local Models:

The plot above shows how a variable CPI varies across different stores. The slope of the variable is steep in some stores while it is not so steep in some other stores. When we run a flat regression, we are assuming that all the variables vary uniformly across each store which is not the case. So, by running local models across levels, we can allow the variables to vary across these levels.

5.3.1 Store Level:

When we run 45 regression models - one for each store and take a mean of R-square and RMSE for all the models, the results are as follows: 13

Train R-Square	Test R-Square
0.88	0.86

Since all the departments are not present in all the stores; one challenge encountered here was that for stores 36 and 44 there were very few data points, so the local model was doing worse. To overcome this problem we replaced the predictions for these stores by the predicted values from local mean imputed flat model which was discussed above.

5.3.2 Type Level:

We thought store was too big a step to take for local level models and wanted to step back and see how local models by type would perform. Since we had 3 types, we ran 3 local models for each type and the results are as follows:

Train R-Square	Test R-Square
0.93	0.92

The results are as good as the store level models with just running 3 individual models instead of 45 models for each. This gives strength to our findings that the 3 types of stores are super stores, medium sized stores and small/regular stores and since they operate very differently across each level, we were able to capture most of the variation in the data by creating type level local models.

5.4 Multilevel/Hierarchical Models:

Multilevel models (also hierarchical linear models, nested models, mixed models) are statistical models of parameters that vary at more than one level. These models can be seen as generalizations of linear models, although they can also extend to non-linear models. These models became much more popular after sufficient computing power and software became available.

Multilevel modeling is a generalization of generalized linear modeling. An MLM models entities at the lowest level but borrows strength from higher levels. When there is very little group level variation, multilevel modeling reduces to classical regression with no group indicators. When group-level coefficients vary greatly, multilevel modeling reduces to classical regression with group indicators. So, it is advantageous when the group level variation is "in between".

In a complete Multi-level analysis, one examines

1) level-1 factors related to the within-group variance
2) group-level factors related to the between-group variation in intercepts
3) group-level factors related to within-group slope differences

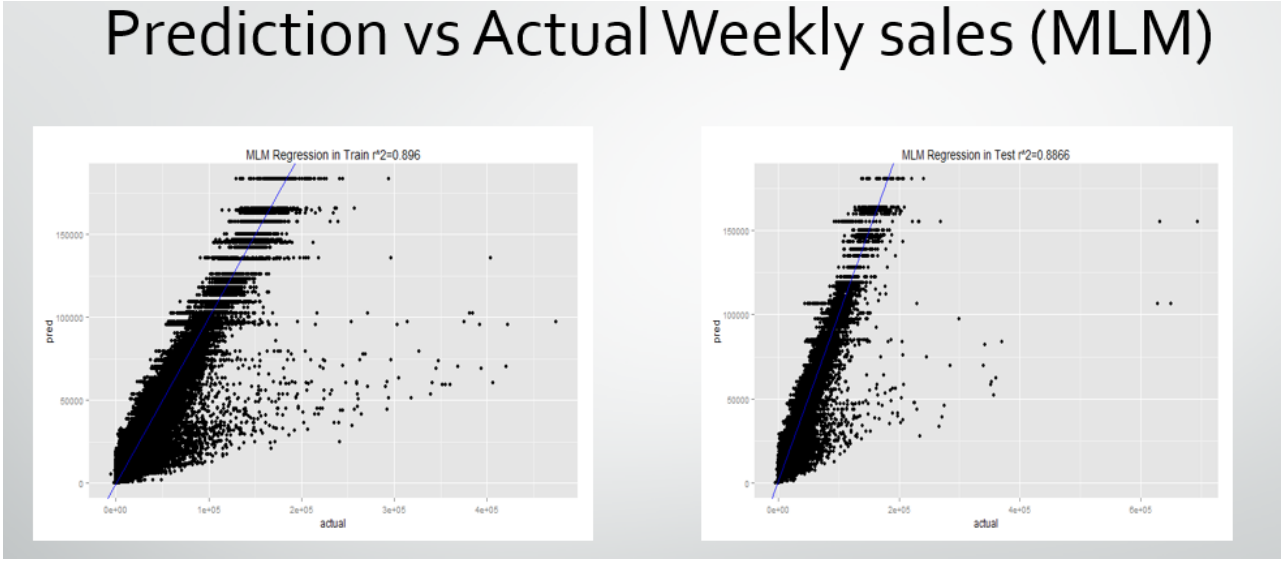
5.5.1 Varying intercepts:

Because multilevel modeling involves predicting variance at different levels, we began the multilevel analysis by determining the levels at which significant variation exists. From our previous analysis, we observed that there is significant variation across Type, Store and Department. The dependent variable is a function of a common intercept, a between group error term and a within group error term. The model essentially states that any Y value can be described in terms of an overall mean plus some error associated with group membership and some individual error.

The results for the multilevel model with level 0 as department, level 1 as store and level 2 as type when only intercepts are allowed to vary among groups are given below:

Model	Train R – Square	Test-Rquare	Train-RMSE	Test-RMSE
MLM-Store/Dept	0.896	0.886	6937.611	7263.428
MLMType/Store/Dept	0.859	0.865	6213.10	6381.43

It is observed that even with two levels Store and Department, we are able to capture almost the same variation as with 3 levels. Increasing complexity does not help here. As mentioned earlier, both type and store capture same amount of variation in the data. Thus either Type/Department or Store/Department will suffice.



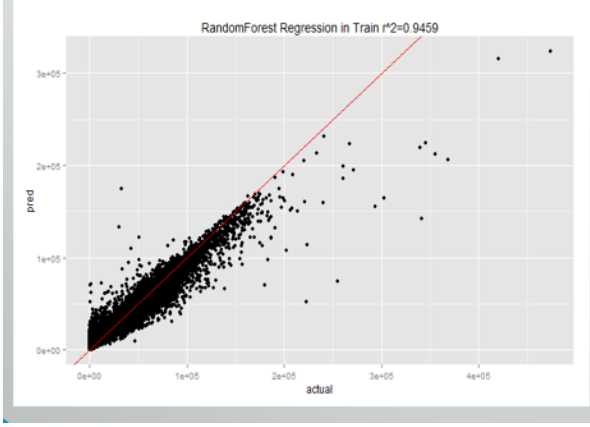
5.5 Random Forest:

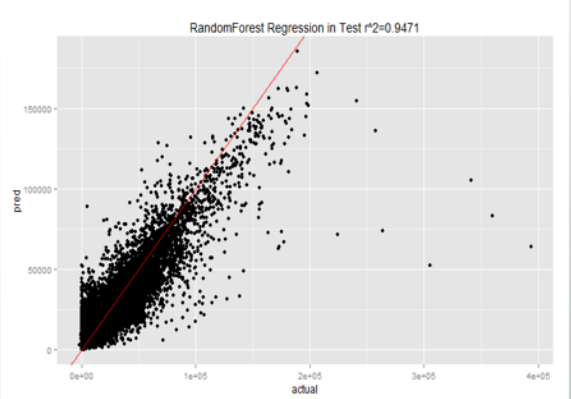
Random Forest is an ensemble learning method that combines bagging and random selection of features. By considering only subset of features at each split, random forest methods effectively reduces the correlation among bagged trees.

Just like previous models, we split our data set into training and test set and ran the random forest model based on training set. We tried ntree = 10 and ntree = 100 to see how the model would improve as the number of trees grows. From the table below, we can see that both of them work decently. Though the RMSE for test set is much larger than that of training set, we don't see any obvious evidence of over-fitting from the graph below where we plotted and compared predicted values and actual sales values.

Model	Train R – Square	Test-Rquare	Train-RMSE	Test-RMSE
MLM-Store/Dept	0.896	0.886	6937.611	7263.428
MLMType/Store/Dept	0.859	0.865	6213.10	6381.43

Prediction vs Actual Weekly sales (RF)





Summary of all the models

Models	Train R-Square	Test R-Square
Flat Model with NA removed	0.61	0.59
Holiday data Global Mean	0.56	0.54
Holiday data Local Mean	0.62	0.62
Non Holiday data Global Mean	0.67	0.65
Non Holiday data Local Mean	0.66	0.64
Store level Local Models	0.81	0.80
Type Level Local Models	0.91	0.92
MLM-Store/Dept	0.896	0.886
MLMType/Store/Dept	0.859	0.865
RF nTree = 10	0.947	0.9471
RF nTree =100	0.954	0.9618

Conclusion:

In conclusion, we believe that multilevel models and Random Forest work best for this kind of data. Multilevel modeling is an increasingly popular approach to modeling hierarchically-structured data, outperforming classical regression in predictive accuracy. This is no surprise, given that multilevel modeling includes least-squares regression as a special case.

References:

1.https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data
2.http://leon.bottou.org/publications/pdf/compstat-2010.pdf
3. http://jaredknowles.com/journal/2013/11/25/getting-started-with-mixed-effect-models-in-r
4. http://faculty.smu.edu/kyler/training/AERA_overheads.pdf
5. http://rt.uits.iu.edu/visualization/analytics/docs/hlm-docs/hlm9.php
6. http://blog.stata.com/tag/multilevel-models/
7. http://www.ats.ucla.edu/stat/examples/msm_goldstein/goldstein.pdf