***The Walmart sales analysis***

Team Members

Anshu Raina

Gopalakrishnan Kailasanathan

Sarath Cheriyan Joseph

Sneha Chitturi

Date

05/08/2014

ISTE 600 Analytical Thinking

**1. Introduction**

* 1. **Problem Description**

Walmart is the one of the biggest retail companies that comprises many departments within wide spread stores in the US. Modeling retail data can be challenging because decision making has to come of off limited history. Our main goal is to project sales for stores and departments for the year 2013. While doing so, we have to consider historical markdown events in predicting the sales.

Markdown events could be anything from clearance sale, rollbacks, special buys, damaged items which amounts to the reduction in the selling price of an item. For example, holidays like Thanksgiving, Super bowl, Labor Day, Christmas may have significant number of markdown events. These events usually take place around important holidays. The days preceding these events could be of particular significance in our analysis. These markdown events can create a big impact on the sales. Using the historical data, we have to predict which departments would be affected and the extent of this impact along with the sales for each department.

* 1. **Motivation**

The main motivation of our Walmart sales analysis is to help Walmart management to analyze their sales statistics and make necessary decisions to improve their sales. This can be facilitated by evaluating the performance of each store and department in the future particularly during holiday weeks which may include key events like markdowns.

* 1. **Report Organization**

Analysis of the problem domain revealed it is a statistical problem and that we would be using regression algorithms. Before which, we analyzed the data and found that it needed preprocessing. We performed a few preprocessing steps to clean the data and handled missing values. We proceeded with testing out different regression algorithms and evaluated them. SMOreg seemed to be the best fit and we proceeded predicting weekly sales for year 2013 using SMOreg.

**2. Data Exploration (Checkpoint 2 summary)**

We found data sets from Kaggle website. Kaggle provided 4 data files which are test, train, stores and features. These files are in .csv format.[2]

**1. Feature data set**

This data set describes data related to Walmart store. In this data file we have 8190 records and 12 attributes which are store, date, temperature, fuel\_price, markdown1, markdown2, markdown3, markdown4, markdown5, CPI, unemployment and holiday. [2]

**2. Store data set**

This data set provides information about 45 Walmart stores spread across the country. In this file we have 45 records and 3 attributes namely store, type and size. [2]

**3. Train data set (test.csv)**

This data set contains historical information from May 2nd 2010 to January 11th 2012 about the Walmart stores. In this file we have 421570 records and 5 attributes namely store, dept, date, weekly\_sales and isHoliday. We will build our model using train data set. [2]

**4. Test data (test.csv)**

This data set is very much similar to train data set except it doesn’t contain weekly\_sales attribute. The model built with train data set would be used here to predict weekly sales of Walmart store. [2]

**3. Methodology**

**3.1 Data Preprocessing**

We have missing values in markdowns 1-5. All the missing values are represented as “NA” in the data set. Weka does not identify NA as missing value. We replaced NA with a “?” so that WEKA will consider those as missing values. We found that 67% of the data in markdowns is missing. Missing values have been handled by using a filter called Replace Missing Values, which will replace all the missing values for numeric attributes in a dataset with the modes and means from the training data.

**3.2 Mining the data**

We have tried mining the data with classification tab and time forecast tab in weka 3.7.10.

In linear regression, under classification tab, we initially thought of taking the formula from the linear regression model applied to the merged train data and predict the sales for the test data with the intention of getting accuracy.

**Parameterization**: We preprocessed the train data first and we chose linear regression under classify tab.

**Conclusion**: The relative absolute error rate was so high 99%. So we dropped that technique.

Under **Time Forecasting**, we couldn’t load entire merged train data, so we analyzed the train data and took 5 stores and 3 departments each for year 2010 and 2011 and predicted for 2012.

Weka provides forecast option, where we can predict the sales with the help of some parameters. Moreover, when we tried this approach we found that the predictions are far better compared to the previous approach.

We applied the following algorithms to our data set.

**Linear Regression**: It is one of the popular regression algorithms. It helps to determine the relationship between dependent variable and independent variable in a linear manner (Straight line). It is more applicable for predictions whose coefficients are either increasing or decreasing.[3]

**SMOreg**: It is a nonlinear regression approach. Since our independent variables vary a lot, this algorithm finds a suitable regression curve to provide a close fit.[3]

**Multilayer Perceptron**: Here each and every independent variable is passed to the hidden layer where the weights for each of the independent variables will be calculated. The output is the sum of all the inputs. [3]

**Parameterization**: Under Basic configuration tab of Forecast weka 3.7, we set the number of units to forecast depending on the train data set input to weka, changed the periodicity to weekly and checked perform evaluation. Under Advanced tab, we selected linear regression, multilayer perceptron and SMOreg, and observed the prediction for weekly sales.

**Conclusion:** While analyzing the predictions from three regression techniques, we came to know that SMOreg gave the best predictions. To confirm once again, we took the test data which contains Nov, Dec 2012 and Jan-July 2013 for stores 3, 5,14,18,20 and departments 11,12,13,92.

We took the stores and departments which has high, medium and low weekly sales. Here is the order of the stores and departments from high to low.

Store : 20,14,18,3,5 Dept: 92, 13, 11, 12

So we took SMOreg as our core algorithm.

**3.3 Models Performance**

We used Mean Absolute Error (MAE) to check the performance of the algorithm in predicting the future sales. Mean absolute error is a measure used to determine how close the predictions are with respect to the reality. Since our project is a forecasting problem, we concentrate on identifying the error between actual and predicted value. Moreover, the MAE is simplest and more popular. We estimated the performance measure by taking absolute difference between actual and the predicted values and dividing by the total number of predictions for linear and SMOreg. Below is the mean absolute error for stores 3, 5, 14, 18, 20.

|  |  |  |
| --- | --- | --- |
| **Store** | **Linear MAE** | **SMOreg MAE** |
| 3 | 13037.41436 | 3835.709204 |
| 5 | 1512.015854 | 1095.24153 |
| 14 | 12954.22138 | 10337.50745 |
| 18 | 4818.274806 | 4407.581072 |
| 20 | 12557.98397 | 5991.152678 |

**4. Logic of the problem:**

The elements of logic of problem gave us the head start we needed for our project in addition to clarifying our stand on the project with each subsequent checkpoint. In reasoning through each element, we addressed some of the key questions required for our project analysis.

**Checkpoint 1:**

We understood and established the information and data we needed to work with for the project.

An initial perusal of the data led us to conclude that sales varied across departments within stores, markdowns were not constant across departments and that there were variations based on the day being a holiday or not. We also established the purpose of our reasoning through the issue which was to help the Walmart Management dissect their sales statistics to improve and better their decision making.

**Checkpoint 2:**

During the second checkpoint we zeroed in on the missing values and kicked off data pre-processing. We found outliers in our data. We used the WEKA tool to identify them. The logic of the problem was updated about the state of data exploration. This exercise helped us know the input attributes better.

**Checkpoint 3:**

Continuing with preprocessing steps and algorithm selection processes, we updated the Information and concept steps in the logic of the problem. After comparing different regression algorithms, we found SMO to be the best fit.

**Advantages of Logic of problem**

1. The logic of program tool gave us an idea on how to analytically think about a problem and what factors need to be taken into account to begin with.
2. The eight elements of thinking namely ‘Purpose’, ‘Question’, ‘Information’ , ‘Interpretation and Inference’, ‘Concepts’ , ‘Assumptions’ , ‘Implications & consequences’ and ‘Point of view’  helped us in understanding the core concepts and ideology behind the project.
3. It gives the analyst a thoughtful and disciplinary approach in converging on a solution to a problem
4. Given time constraints, a tool like this helps streamline the thought process for quick analysis and progress.

**Disadvantages of Logic of problem:**

1. The problems with logic of problem would be constant need to update the model.
2. There should be way to keep track of changes done in previous time.
3. The user interface is not friendly on navigating to a particular element, it doesn’t stay there.

**Recommendation:**

We would like to recommend a change where the logic of tool can be updated to keep the track of changes so that its gives flow from where we started and how it changed the solution building and the final result.

**5. Conclusions:**

**Comparison:**

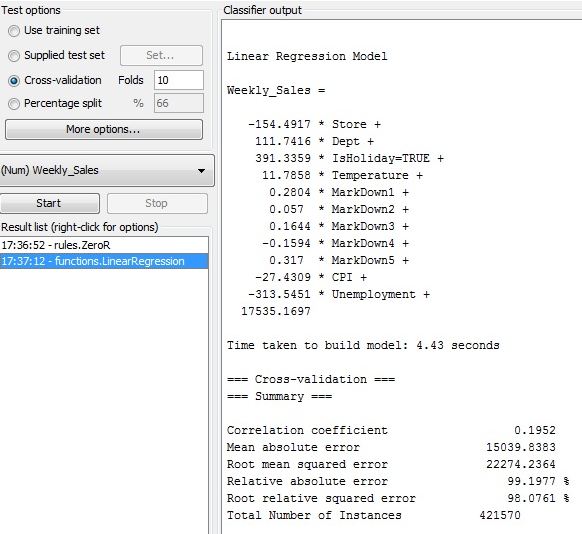
* After analyzing the results through graph, linear predicts lots of negative values when compared with SMOreg.
* Moreover, as we mentioned earlier, the predictions for dept 92 were high, which again confirms that the algorithm works better.
* SMOreg gave closer prediction for selected stores and departments. Linear gave huge variations in predicted weekly sales.

**Summary**: We took a sample test data from store no 3, 5, 14, 18, 20 with each having departments 11,12,13,92. We chose the store & department with respect to highest weekly sales -medium and lowest. From checkpoint 3 ,we found that SMOreg works better, hence we predicted the test data (future date) using SMOreg. We compared the prediction with 2011 Nov, Dec and 2012 Jan-Jul data. We also compute mean absolute error and it is confirmed that SMOreg gives a closer prediction. So we conclude that SMOreg suits best for our project.

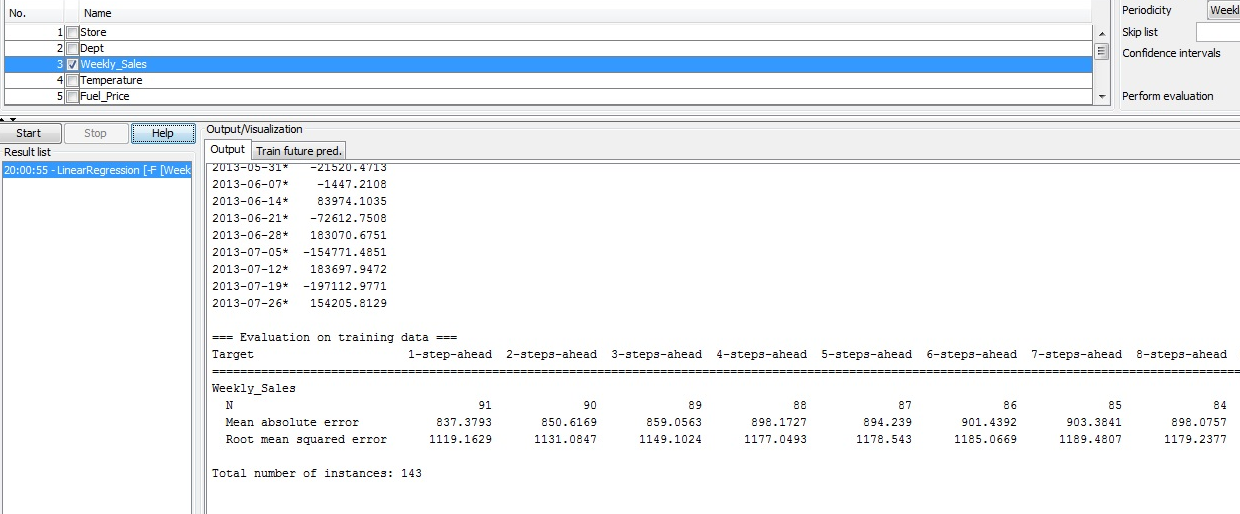
**Future work:** Since the majority of data consisted of null values for markdowns (1-5), given time and extension, we would like to analyze patterns in the markdowns & CPI and predict all the null values in the markdowns listed.

**6. Appendix**

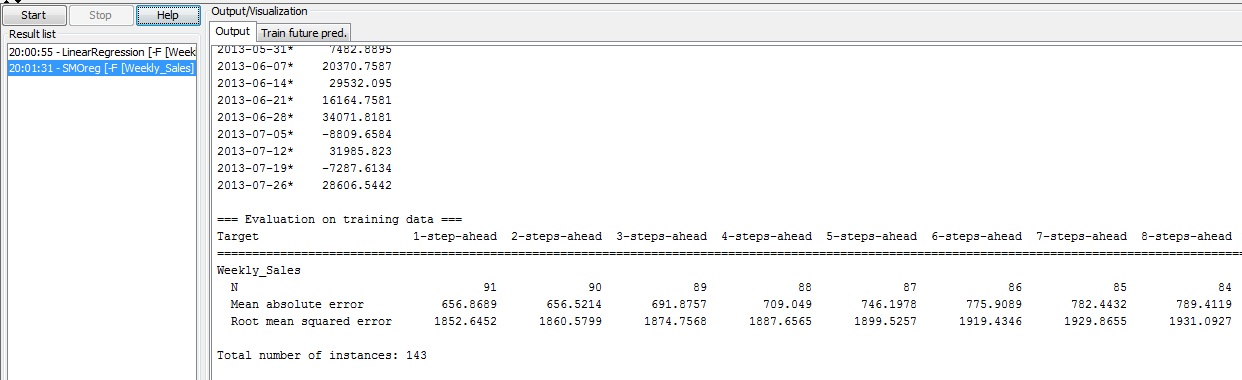
1. Weka output for Linear under classification tab which showed error rate close to 99%.



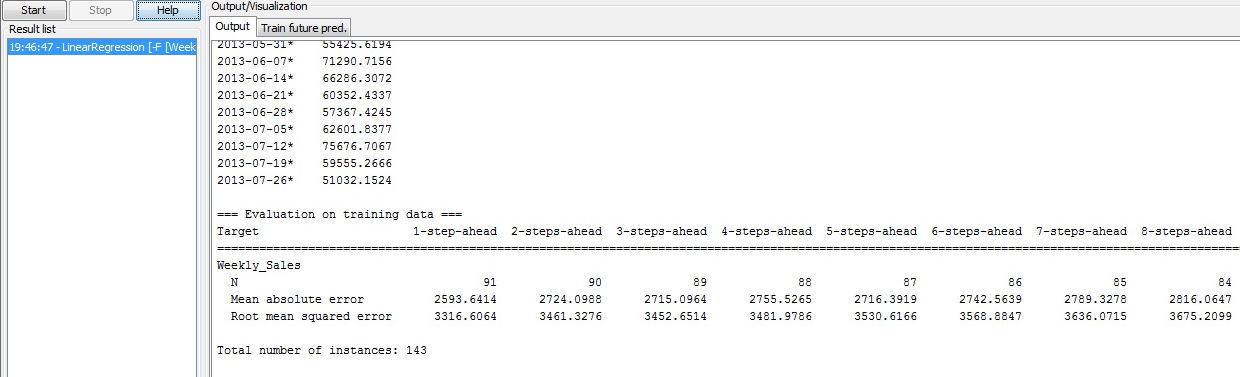
1. Weka output for Store 3 dept 11 forecasted using linear regression



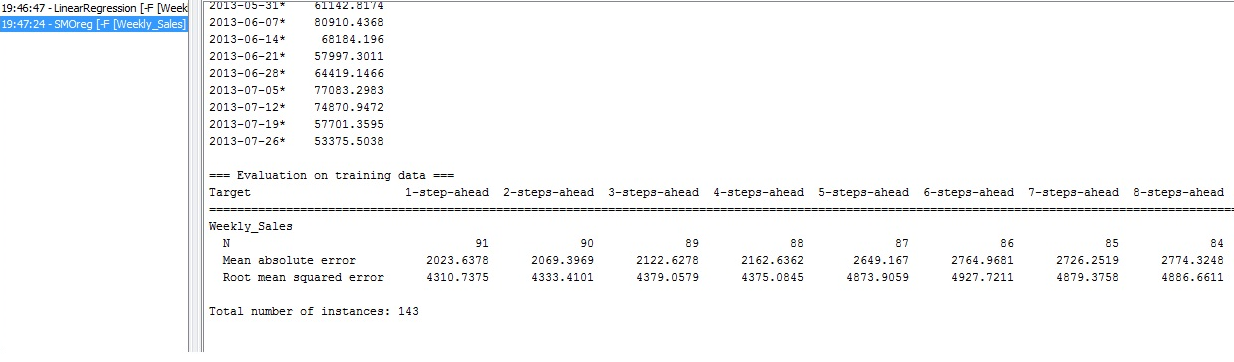
1. Weka output for Store 3 dept 11 forecasted using SMO regression



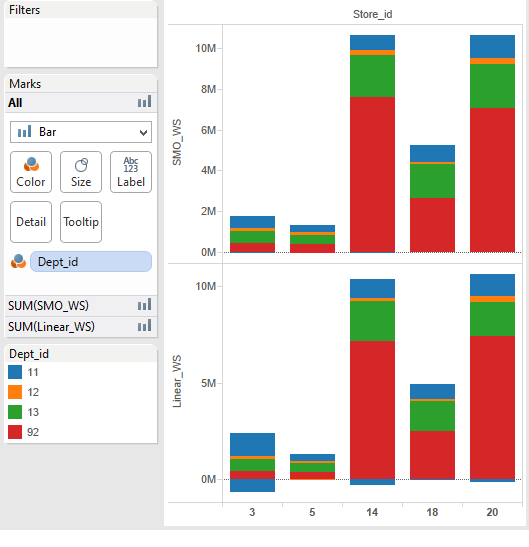
1. Weka output for Store 18 dept 92 forecasted using linear regression



1. Weka output for Store 18 dept 92 forecasted using SMO regression



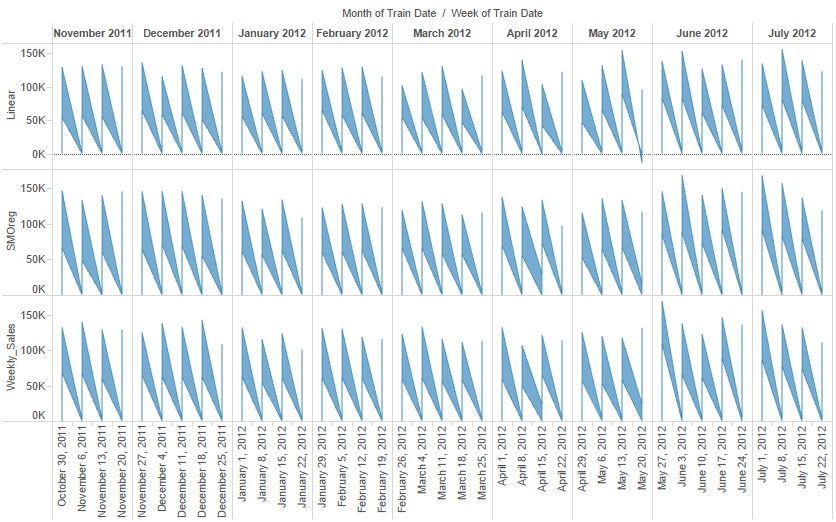
1. SMOreg and linear comparison using tableau for stores 3,5,14,18,20 and dept 11,12,13,92



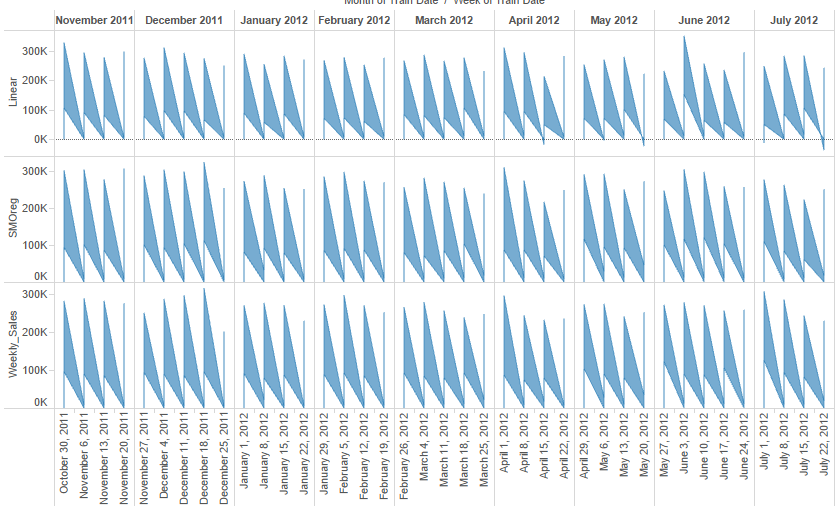
1. Store 5 comparison with respect to SMOreg, linear and actual weekly sales[4].



1. Store 18 comparison with respect to SMOreg, linear and actual weekly sales[4].



1. Store 20 comparison with respect to SMOreg, linear and actual weekly sales [4]



**7. References:**

[1] Retrieved from • onlinesvr.altervista.org/

[2] Retrieved from • www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[3] Ezekiel, M., & Fox, K. A. (1959). Methods of Correlation and Regression Analysis. London: John Wiley & Sons.

[4] Murray, D. (2013). Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software. Hoboken: Wiley.

[5] Provost, F., & Fawcett, T. (2013). Data science for business: [what you need to know about data mining and data-analytic thinking]. Sebastopol, Calif: O'Reilly.