Data Mining Techniques and Machine Learning Model for Walmart Weekly Sales Forecast

José Gil Santaella Colón
Master in Computer Science
Advisor: Dr. Nelliud Torres Batista
Electrical & Computer Engineering and Science Department
Polytechnic University of Puerto Rico

Abstract *— The ability to forecast data accurately is extremely valuable in a vast array of domains such as health, sales, finance, weather or sports. Presented here is the study and implementation of data mining techniques and ensemble regression algorithm employed on sales data, consisting of weekly retail sales numbers from different departments in Walmart retail stores all over the United States of America over the period of 3 years with pre-holiday and holiday data presenting a spike in sales. The model implemented for prediction is Random. The metric to evaluate the model was the Mean Absolute Error (MAE) value. An analysis was performed to evaluate the model and its ability to forecast accurately. It is also notable that artificial neural networks can improve the performance and achieve highly accurate results.*

Key Terms *— Machine Learning, Mean Absolute Error, Neural Networks, Random Forest, Sales Forecasting.*

INTRODUCTION

In a world where immense amounts of data are collected daily, analyzing such data is an important need, For example, large stores, such as Wal-Mart, handle hundreds of millions of transactions per week at thousands of branches around the world [1]. In this modern world where competition is getting greater and thus making business decisions is increasingly difficult. In turn, predicting accurately has become extremely important. Therefore, by using data mining techniques and machine learning algorithms different domains such as retail can make informed decisions based on projections. As mention before, forecasting is of great importance to the retail business, the reason being that sales prediction is a more traditional application of forecasting. Forecasting sales incorrectly can lead to over-estimation and in turn lead to significant losses and costs of inventory holding, in turn, under-estimation of sales in a forecast can lead to loss of business opportunity [2]. By applying data mining techniques, a company can accurately forecast and analyze sales and more importantly how customer behave towards certain products or marketing campaigns. Sales forecasting uses patterns or trends gather from historical data to predicts sales accurately, thus enabling informed decisions to manage efficiently inventory or future production. The problem used in this study is based on a competition from the Kaggle platform on the need to forecast weekly sales for regional stores of the retail organization, Walmart [3]. Exploring and analyzing the results using a machine learning model such as Random Forests [4][5], is a supervised learning algorithm which uses an ensemble learning method for classification, regression and other tasks, that functions by building a large number of decision trees at training time and producing the value that is the mode of the classes (classification) or producing the value that is the mean of the values (regression) of the individual trees.

What Is Data Mining?

Data mining is the process of discovering interesting patterns from massive amounts of data. As a knowledge discovery process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation [1].

What Is Machine Learning?

Machine learning is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available [6].

Dataset

The dataset used for this study comes from the Kaggle Platform which is used for data science projects and competitions. The dataset is made up from an American retail organization, Walmart Inc. It consists of data from 45 Walmart stores centered around their weekly sales. It has 421,517 entries that will be used for training the model used on the study. weekly basis. The dataset has the following attributes: the store (recorded as a number), the corresponding department (each entered as a number), the date of the starting day in that week, departmental weekly sales, the store size, and a boolean value specifying if there is a major holiday in the week. The major holidays being one of Thanksgiving, Labor Day, Christmas or the Super Bowl. Along with the mentioned attributes is a set of features for each entry including Consumer Price Index, unemployment rate, temperature, fuel price, and markdowns. Since there is no test-set provided, we use 20% from the given training data for cross-validation, and final testing

Methods

For this study, we used techniques involve with data mining [1], such as:

- **Data cleaning:** To remove noise and inconsistent data.
- **Data integration:** where multiple data sources may be combined.
- **Data selection:** Where data relevant to the analysis task are retrieved from the database)
- **Data transformation:** Where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data mining:** An essential process where intelligent methods are applied to extract data patterns.
- **Pattern evaluation:** To identify the truly interesting patterns representing knowledge based on interestingness measures.
- **Knowledge presentation:** Where visualization and knowledge representation techniques are used to present mined knowledge to users.

Once we acquired the dataset needed to successfully forecast the weekly sales, we performed the aforementioned techniques or steps. For example, first we explored the data to get a sense of its structure and the values contained within.

Figure 1 shows the variables within the dataset and their correlation. We can see that discounts and holidays are correlated, as is also higher sales with a store with larger size. Therefore, we can see that Also, larger stores generate more sales, discounts generally generate higher sales values and larger unemployment result in a bit fewer sales. There appears to be little relationship between holidays, temperatures or fuel prices with weekly sales. The next step in exploring the data we plot some of the relationships shown in figure 1 to get a clearer image of our dataset.

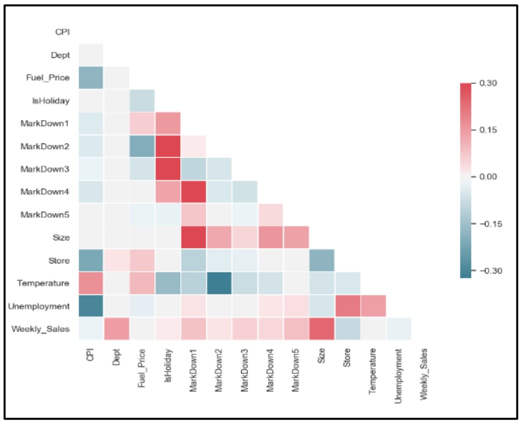


Figure 1
Correlation Heatmap

Figure 2 shows that larger stores generate overall higher sales despite some medium size stores generating a few high weekly sales.

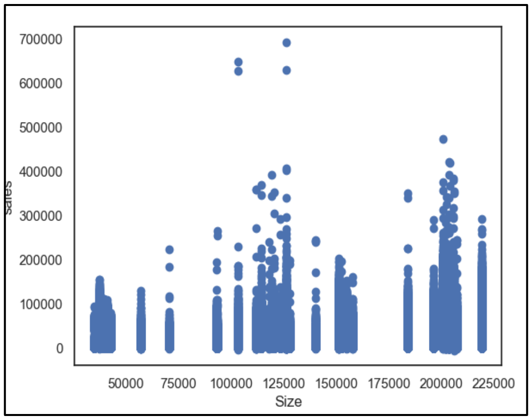


Figure 2
Sales and Store Size Plot

Next we prepared the data for modeling by cleaning the data and removing inconsistent data, then transformed by merging and adding variables useful for the analysis and forecast (figure 3). These operations were implemented on Python 3.7 on the Anaconda distribution using Jupyter Notebooks extension on Visual Studio Code.

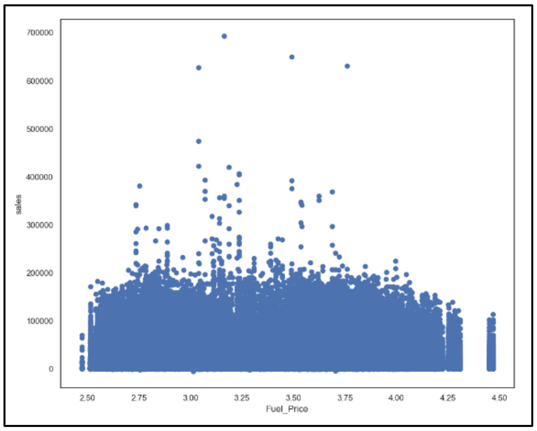


Figure 3
Sales and Fuel Price Plot

One model was constructed for this study on the following algorithm: Random Forest Regressor. Because it is founded to be one of the most accurate learning algorithms available.

Random Forest

The method applied to forecast the weekly sales was the Random Forest Regressor. This machine learning algorithm, as mentioned earlier on the introduction, consists of a large number of individual decision trees that operate as an ensemble. As more trees are constructed, the Random Forest algorithm adds more randomness to the model. It searches for the best feature amidst a random subset of features in place of searching for the most relevant feature while splitting a node. This ensures that the model does not rely too heavily on any individual feature and makes us of all potentially predictive features. Thus, in Random Forest, only a random subset of the features is considered by the algorithm for diverging a node. Adding a further element of randomness that prevents overfitting the model. Its architecture can be best described in figure 4 [4].

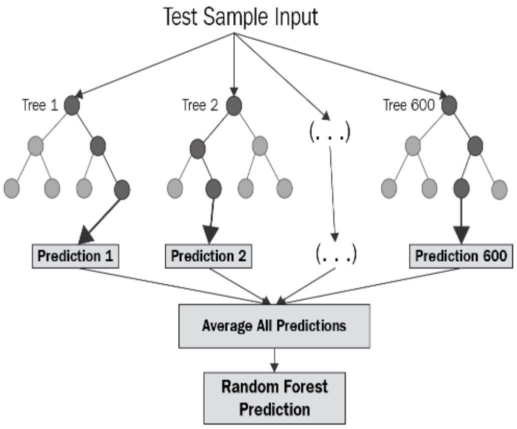


Figure 4
Random Forest Architecture

The features used to train the model are as follows: consumer price index (CPI), fuel price, markdowns, size, store, unemployment, temperature, holiday flag, pre-christmas flag, black Friday flag, lagged sales, sales differences and lagged flag. The algorithm was implemented using Python’s Random Forest Regressor function present in the scikit-learn class. In its implementation, the metric used to evaluate and calculate for the predicted values was the Mean Absolute Error (MAE). The mean absolute error can be defined as [7][8], Prediction Error = Actual Value - Predicted Value. This prediction error is taken for each record after which we convert all error to positive. This is achieved by taking

Absolute value for each error as; Absolute Error → |Prediction Error|. Then we calculate the mean for all recorded absolute errors such as the average sum of all absolute errors. The MAE refers to the results of measuring the difference between two continuous variable which in turn means that the prediction error is the difference between the actual value and the predicted value for that instance.

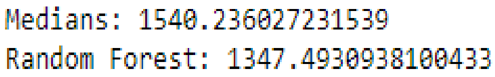
For this case, the variable used to predict the weekly sales was the difference from the median i.e. Difference = Median – Weekly Sales. Then to evaluate the model, the MAE was calculated as follows; Error = Weekly Sales – Prediction.

Results

As mentioned earlier in the study, the problem used is based on a competition from the Kaggle platform. As such, the results of this research was compared against the winning submission for the competition which had a Mean Absolute Error of around 2301 on the private leaderboard meanwhile on the public leaderboard the Mean Absolute Error is around 2237.

The results we first analyzed were when we first constructed the model to observe if it was accurate and ready to forecast. On this model we used 20% of the training set to check the development set. The MAE was found to be around 1347 as seen on table 1.

Table 1
Random Forest Performance on Partial Dataset



```
Medians: 1540.236027231539
Random Forest: 1347.4930938100433
```

These were the results obtained with the n_estimators hyperparameter set at 20, which refers to the number of decision trees that are used for regression. The other hyperparameters were set to their default values. When further evaluating the model, we used a plot to visualize how the prediction fares against the weekly sales as seen on figure 5.

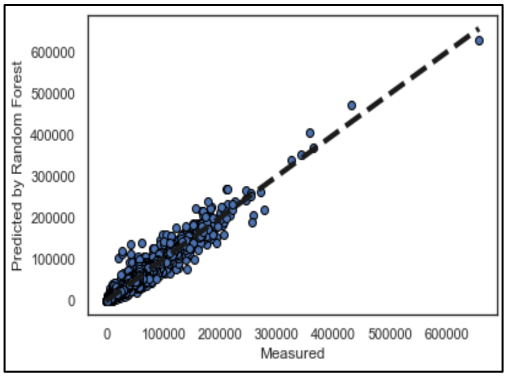


Figure 5
Prediction and Weekly Sales Plot

Figure 5 shows how closely the prediction using 20% of the training set fits the weekly sales. Thus, deciding the model can be used on the full dataset to forecast accurately. The same can be said for figure 6 which presents how closely the difference from the median sales minus weekly sales fits the model.

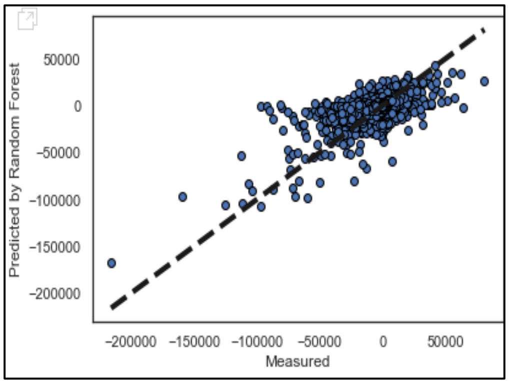
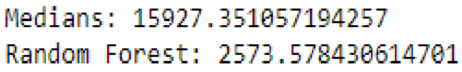


Figure 6
Differences Plot

The model was deemed ready to use on the full dataset to get the maximum amount of information on the model used. On the full dataset the MAE was around 2573, as seen on table 2. This was obtained with the n_estimators hyperparameter set at 80 for faster performance. The other hyperparameters were set to their default value

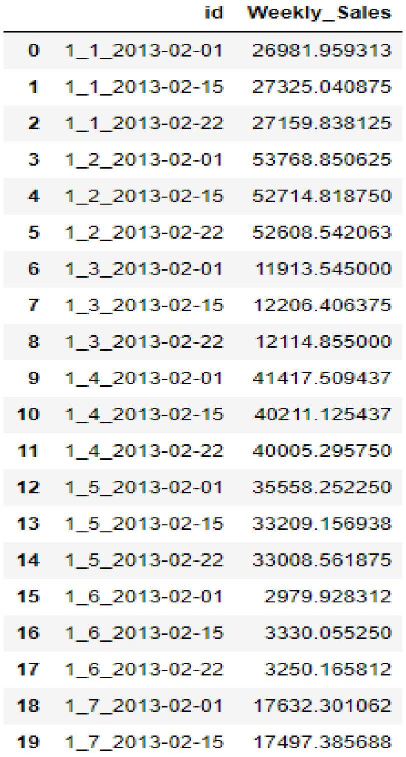
Table 2
Random Forest Performance on Full Dataset



```
Medians: 15927.351057194257
Random Forest: 2573.578430614701
```

Overall, it’s a good result and we obtained a great amount of information for our model with a slightly less accurate weekly sales forecast comparing against the winning submission on the Kaggle competition. With a MAE score of around 2573, it could be in a position on the top 50. On table 3, we can see a sample of the first 20 rows of the weekly sales identified in this order: store id_department_date and on the other column with weekly sales.

Table 3
Weekly Sales



	id	Weekly_Sales
0	1_1_2013-02-01	26981.959313
1	1_1_2013-02-15	27325.040875
2	1_1_2013-02-22	27159.838125
3	1_2_2013-02-01	53768.850625
4	1_2_2013-02-15	52714.818750
5	1_2_2013-02-22	52608.542063
6	1_3_2013-02-01	11913.545000
7	1_3_2013-02-15	12206.406375
8	1_3_2013-02-22	12114.855000
9	1_4_2013-02-01	41417.509437
10	1_4_2013-02-15	40211.125437
11	1_4_2013-02-22	40005.295750
12	1_5_2013-02-01	35558.252250
13	1_5_2013-02-15	33209.156938
14	1_5_2013-02-22	33008.561875
15	1_6_2013-02-01	2979.928312
16	1_6_2013-02-15	3330.055250
17	1_6_2013-02-22	3250.165812
18	1_7_2013-02-01	17632.301062
19	1_7_2013-02-15	17497.385688

The results showed the holiday spikes with the sales estimates almost doubled those of other months. What is interesting is the pre-holiday season effect; despite there only being four holidays in the dataset, the months surrounding those dates see a residual boost in sales. This can probably be attributed to promotions before and after the holiday itself.

Conclusion

This research dealt with the implementation data mining techniques and machine learning algorithms on the Walmart dataset and analysis was made to determine the accuracy of the algorithm, model and forecast. Random Forest is found to be a good model to forecast sales data for its performance on weighted variables such as dollars. Therefore, being a valuable asset to be applied on forecasting sales on the retail industry. It is important to mention that to produce highly accurate predictions with the tiniest of details, more models with larger hyperparameters set could be apply in conjunction with better hardware electronics such as Graphic Procession Units because this is a computationally expensive task and with bigger datasets, it can take hours to train and have a result.

Future work would include on Artificial Neural Networks, which are very powerful machine learning models that are highly flexible universal approximators, needing no prior assumptions during model construction. Neural networks perform end-to-end learning when being trained, determining the intermediate features without any user-feedback [9][10]. Artificial Neural Networks could be further improved by using different functions such as the Rectified Linear Unit (ReLU) activator because a model that uses it is easier to train and often achieves better performance [11]. Combing the Artificial Neural Network, ReLU activator function with the Adam Optimization Algorithm could improve the performance speed for forecasting thus achieving results in the less time possible [12].

This study shows how efficient and impactful is the use of data science and machine learning to forecast sales and how any organization could benefit from valuable insight that leads to an informed and better decision making process.

Acknowledgment

This work would not exist without the guidance of Professor Torres and the faculty of the Department of Electrical & Computer Engineering and Science Department of the PUPR for their teaching and encouragement.

REFERENCES

[1] J. Han, M. Kamber and J. Pei, *Data Mining*, 3rd ed. Amsterdam: Elsevier/Morgan Kaufmann, 2012.

[2] A. Verma et al. (2019). "Optimizing operational spend by predicting product sales," *Final project*, *Indian School of Business*, *Hyderabad*, *India* [Online]. Available: https://www.galitshmueli.com/sites/galitshmueli.com/files/B3_Neeraj%20Nathany.pdf.

[3] Kaggle. (2014). *Walmart Recruiting - Store Sales Forecasting* [Online]. Available: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data. [Accessed: May 1, 2019].

[4] A. Chakure. (2019, June 29). *Random Forest Regression, Towards Data Science* [Online]. Available: https://towardsdatascience.com/randomforest-and-its-implementation-71824ced454f. [Accessed: August 1, 2019].

[5] T. You. (2019, June 12). *Understanding Random Forest," Towards Data Science* [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[6] A. Pant. (2019, January 7). *Introduction to Machine Learning for Beginners, Towards Data Science* [Online]. Available: https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08.

[7] E. Minka. (2018, February 2). *Mean Absolute Error (MAE) ~ Sample Calculation, Medium* [Online]. Available: https://medium.com/@ewuramaminka/mean-absolute-error-mae-sample-calculation-6eed6743838a.

[8] E. Minka. (2018, February 2). *Mean Absolute Error ~ MAE [Machine Learning (ML)], Medium* [Online]. Available: https://medium.com/@ewuramaminka/mean-absolute-error-mae-machine-learning-ml-b9b4afc63077.

[9] J. J. Pao and D. S. Sullivan. (2017). “Time Series Sales Forecasting,” *Final year project, Computer Science, Stanford Univ., Stanford, CA, USA* [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5244336.pdf.

[10] N. S. Elias and S. Singh. (2018). "Forecasting of Walmart Sales Using Machine Learning Algorithms," *Research paper, Dept. of Electronics & Comm. Engineering, BMS Inst. of Technology & Management, Bangalore, India* [Online]. Available: http://www.nikhilelias.com/images/forecasting-walmart-sales.pdf.

[11] J. Brownlee. (2019, January 9). *A Gentle Introduction to the Rectified Linear Unit (ReLU), in Machine Learning Mastery* [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/.

[12] D. Kingma and J. Ba. (2015, May 7-9). *Adam: A method for stochastic optimization*, presented at 3rd ICLR, San Diego, CA, USA [Online]. Available: https://arxiv.org/pdf/1412.6980v8.pdf.