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

Author: José Gil Santaella Colón
Advisor: Dr. Nelliud Torres Batista
Electrical & Computer Engineering and Science Department

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.

Introduction

In a world where immense amounts of data are collected daily, analyzing such data is an important need, Therefore, by using data mining techniques and machine learning algorithms on different domains such as retail, one can make informed decisions based on projections. 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.

Problem

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. The objectives being to fit a model to the training data able to forecast the weekly sales as accurately as possible. The success of our model being defined by the metric used, Mean Absolute Error. The dataset has the following attributes: the store, the corresponding department, 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.

Background

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.

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.

Methodology

For this study, we used techniques involve with data mining, such as: **Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation** and **Knowledge presentation.**

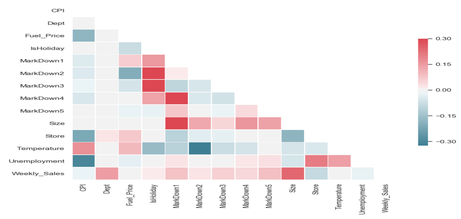

Figure 1. Correlation Heatmap

The method applied to forecast the weekly sales was the Random Forest Regressor. This machine learning algorithm that consists of a large number of individual decision trees that operate as an ensemble and producing the value that is the mean of the values (regression) of the individual trees.

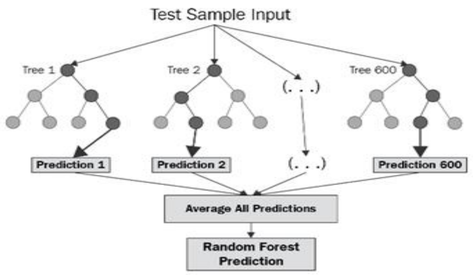

Figure 2. Random Forest Architecture

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.
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**: Prediction Error = Actual Value - Predicted Value**.

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.**

Table 1. Random Forest Performance on partial dataset

Medians	Random Forest
1540.23602	1347.49309

Results and Discussion

The results of this research were 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.

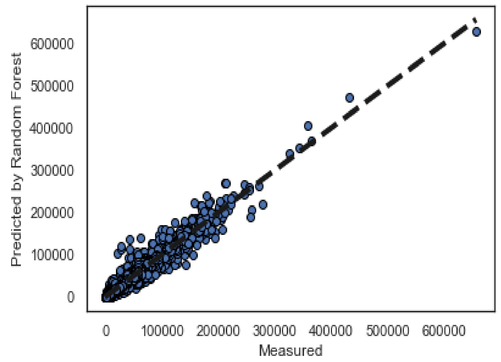

Figure 3. Prediction and Weekly Sales Plot

Figure 3 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 4 which presents how closely the difference from the median sales minus weekly sales fits the model.

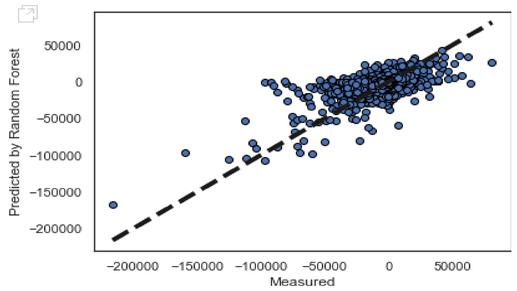

Figure 4. Differences plot

Table 2. Random Forest Performance on full dataset

Medians	Random Forest
15927.351057	2573.578430

The results showed that the holiday spikes with the sales estimates almost doubling 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.

Conclusions

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. 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 a informed and better decision making process.

Future Work

Future work would include 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. 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. 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.

Acknowledgements

This study would not existed 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] "Optimizing Operational Spend by Predicting Product Sales", Galitshmueli.com, 2019. [Online]. Available: https://www.galitshmueli.com/sites/galitshmueli.com/files/B3_Neeraj%20Nathany.pdf. [Accessed: 01- Aug- 2019].
[3] "Walmart Recruiting - Store Sales Forecasting | Kaggle", Kaggle.com, 2019. [Online]. Available: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data. [Accessed: 01- May- 2019].
[4] A. Chakure, "Random Forest Regression", Medium, 2019. [Online]. Available: https://towardsdatascience.com/random forest-and-its-implementation-71824ced454f. [Accessed: 01- Aug- 2019].
[5] T. Yiu, "Understanding Random Forest", Medium, 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest 58381e0602d2. [Accessed: 02- Oct- 2019].
[6] A. Pant, "Introduction to Machine Learning for Beginners", *Medium*, 2019. [Online]. Available:https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08. [Accessed: 01- Aug- 2019].
[11] J. Brownlee, "A Gentle Introduction to the Rectified Linear Unit (ReLU)", Machine Learning Mastery, 2019. [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. [Accessed: 08- Oct- 2019].
[12] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", arXiv.org, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980v8. [Accessed: 02- Aug- 2019].