Walmart Store Sales Forecasting

*ECE225A Probability and Statistic for Data Science leaded by Professor Alon Orlitsky at UC San Diego

1st Yifan Wu
Department of Electrical and Computer Engineering
University of California, San Diego
yiw084@ucsd.edu

***Abstract*—The goal of this project is to investigate how different features affected Walmart stores' weekly sales in order to present an accurate prediction of it. Other than Holidays' discount promotions, Walmart held five promotional markdown events each year. One of the challenges was making predictions based on absence of Markdown data. Various features were evaluated as well and used to predict weekly sales in different regression models.**

I. Introduction

Walmart was seeking a model to predict future weekly sales. Historical sales data of 45 Walmart stores was provided on Kaggle, which included potential features such as Department, IsHoliday, Markdowns, Temperature, Fuel price, CPI, Unemployment, Size and Type.

IsHolidays (Super Bowl, Labor Day, Thanksgiving, and Christmas) were considered as an important categorical feature of weekly sales. According to Walmart, weekly sales during those holidays were weighted five times higher in the evaluation than non-holidays' weeks. Walmart ran several promotional markdowns from 2010 to 2012. These markdowns preceded four prominent holidays. Intuitively, these markdowns would affect weekly sales, but their influence was unknown. Since a lot of markdown data was missing, it would be more challenging to figure out the actual relationship between markdowns and weekly sales.

Other interesting features including fuel price, CPI, unemployment rate, size and type, would be analyzed as well. Correlation between weekly sales and those features would be utilized to help filter out irrelevant features. All the selected features were applied to predict the weekly sales in Linear Regression, Decision Tree Regression, Random Forest Regression, and Extra Trees Regression. The performance of each model would be evaluated by Weighted Mean Absolute Error ($WMAE$), Mean Squared Error(MSE) and R-squared (R^2).

II. Data Visualization and Analysis

A. Sales Trends during each Year and Holiday Period

Figure 1 represented the average weekly sales of each year from 2010 to 2012. During this period, average weekly sales had similar trends. Weekly sales dramatically increased from week 46 and week 51 which were during two important holiday weeks, Thanksgiving and Christmas. (Noticed that sales data in 2012 from November to December was missing which included two important holidays Thanksgiving and Christmas.)

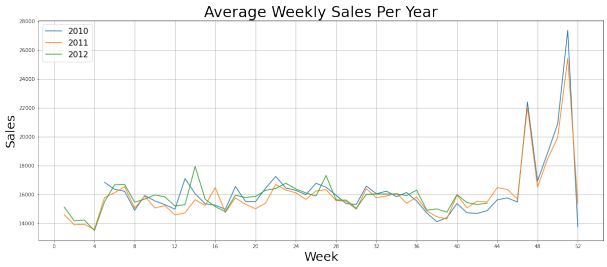

Fig. 1. *Average Weekly Sales Per Year*

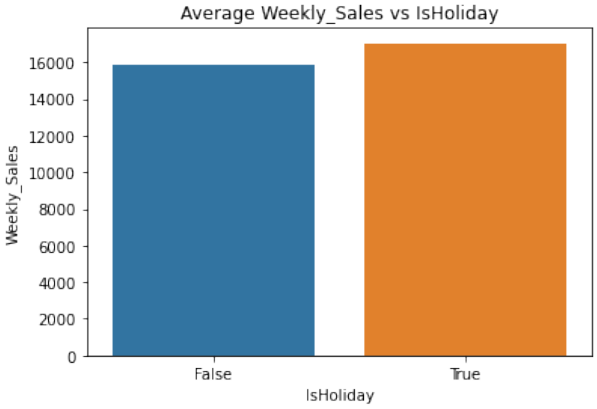

Fig. 2. *Average Weekly Sales vs IsHoliday*

B. Isholiday Feature Analysis

Figure 2 represented the weekly sales of each month from 2010 to 2012. Super Bowl, Labor Day, Thanksgiving, and Christmas were in month 2, 9, 11, 12. During Thanksgiving, weekly sales were much higher than weekly sales in Super Bowl, Labor Day and Christmas. (Noticed that only 4 holidays were defined as Holiday in the data set.) Walmart likely held some markdown in other important Holidays as New Year's Day, Martin Luther King's Day, Independence Day, Veterans Day, etc. Holiday adjustments on data might help improve the accuracy of prediction.

C. Correlation between Type and Size of Store

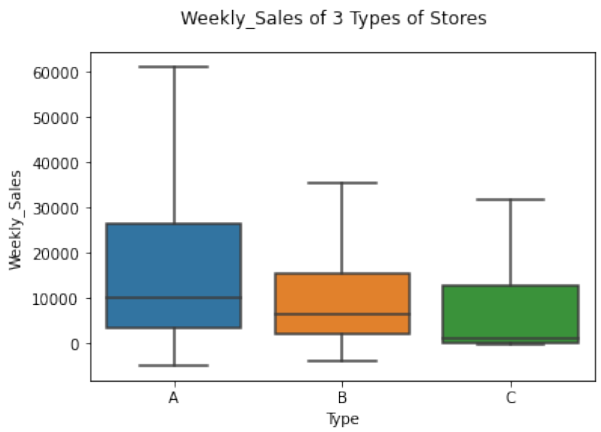

Fig. 3. *Weekly Sales of 3 Types of Stores*

Figure 3 displayed weekly sales of 3 types of stores. It could be inferred that type is relative to size of the store. Large-size stores seem to have more weekly sales. One way to try and understand the data was by looking for correlations between the features (Size and Type) and the target (Weekly Sales).

The strength of the correlation:

- **0.00-0.19** ***"very weak"***
- **0.20-0.39** ***"weak"***
- **0.40-0.59** ***"moderate"***
- **0.60-0.79** ***"strong"***
- **0.80-1.0** ***"very strong***

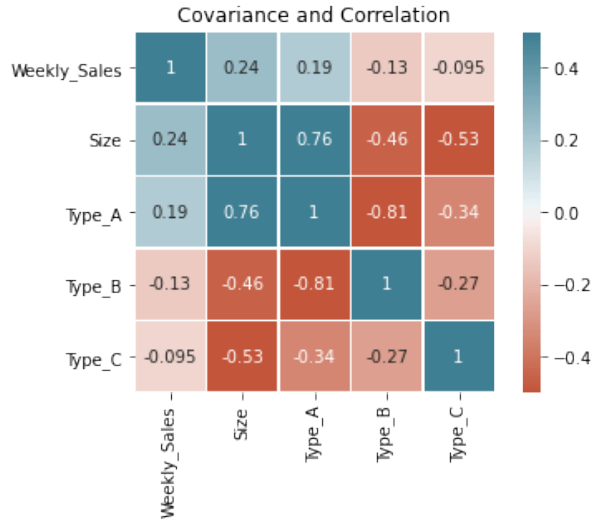

Fig. 4. *Correlation between Size and Type*

From figure 4, it could be concluded that while Type A was strongly positive correlated with size, Type B and C were moderately negative correlated with size. In this case, although type and size (Size and Type) were highly correlated, none of them could be dropped. Type of stores might include information other than size such as geographical location, number of employees, and difference service provided. Both Size and Type A had positive correlations with weekly sales and both Type B and Type C had negative correlations with weekly sales.

D. Feature Correlation Analysis

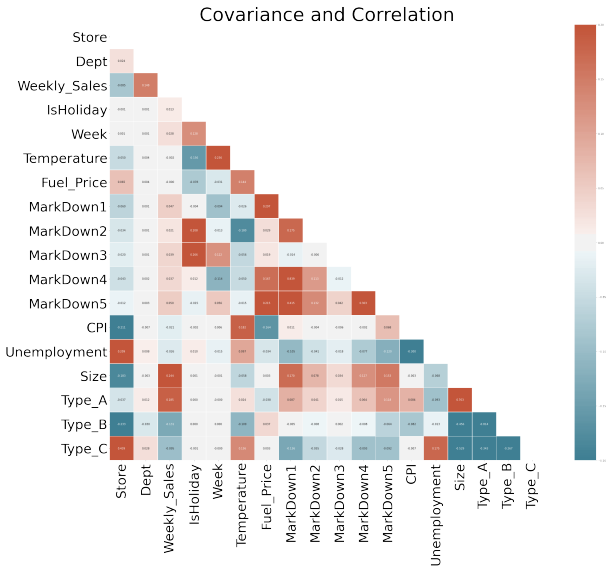

Fig. 5. *Correlation Between Features and Weekly Sales*

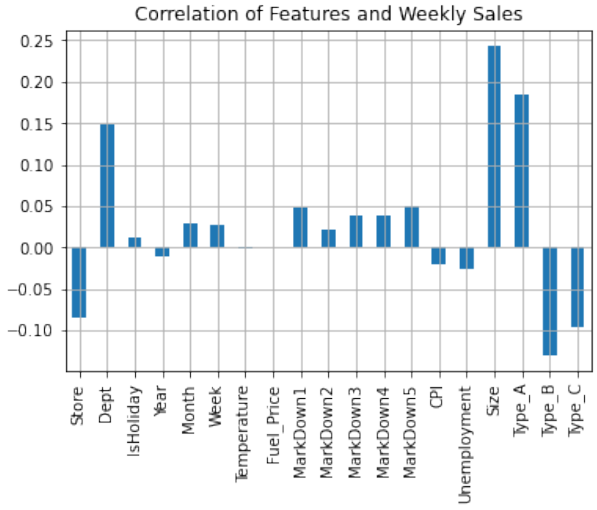

Fig. 6. *Correlation Coefficient of Features and Weekly Sales*

After analyzing the Correlation of all features and weekly sales, it could be concluded that Size, Type A and department were more positively correlated with weekly sales. Store, Type B & C were more negatively correlated with weekly sales. However, correlation was not sufficient to explain the importance of the feature "Store". Store in the training data was an ID. It was a mixture of various information which might affect the weekly sales in distinct ways such as location, size, type and stock.

Markdown1 to 5 had weak positive correlation on weekly sales. The correlation was between 0.02 and 0.05. The correlation of holidays was 0.0128 which was even weaker. CPI and Unemployment had very weak negative correlations. Markdown evens and holidays' promotions indeed stimulated weekly sales, but their influences were not as large as what were expected.

E. Feature Selection and Standardization

Features that were measured at different scales did not contribute equally to the model. It might end up creating a bias. In order to deal with this potential problem, feature-wise standardization (μ = 0, σ = 1) was performed prior to model fitting.

Standardization:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

with mean:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

and Standard Deviation:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

Feature Selection were performed based on the following criteria :

1) Reduces Over-fitting: Less redundant data means less opportunity to make decisions based on noise
2) Improves Accuracy: Less misleading data means modeling accuracy improves
3) Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster

The following feature were dropped:

- Date: It could be converted and represented as a combination of year, month and week.
- Temperature & Fuel Price as shown these two features had nearly zero correlation between weekly sales

III. Models

A. Ordinary least squares Linear Regression

Assuming linear relationship between the dependent variable y and the vector of regressors x, a linear regression model fitted with coefficients β to minimize the residual sum of squares between the targets and predictions.

$$\hat{y} = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon \tag{4}$$

where β_0 is the intercept, $(\beta_1 ... \beta_n)$ are coefficients and error is denoted as ϵ.

B. Decision Tree

Decision Tree algorithms were referred to Classification and Regression Trees, a tree or flowchart-like structure. A training set was broken down into smaller subsets, while an associated decision tree was created. Decision Tree Regression used mean squared error to determine whether to divide a node or not.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5}$$

where:

- n is the number of data points
- $\hat{Y}_i$ is the predicted value
- Y_i is the observed value

C. Random Forest Regression

Random Forest algorithms created forest with number of decision tree. Each tree ran in parallel with square root of number of features. The entire Random Forest outputted mean prediction of the individual tree. Similarly, mean squared error (equation (5)) were used to determine whether to split a node or not. One of advantages of Random Forest was that it can maintain accuracy while a larger proportion of the data are missing. However, performing random forest on regression may not be the best choice, since it did not provide precise continuous nature predictions. Moreover, there was limited control of what the model did.

D. Extra Trees Regression

Extra Trees algorithm was an ensemble method which was much faster than Random Forest yet equally accurate. Samples for every decision tree were selected without replacement. Instead of spending time finding the best splitting point, extra trees randomly picked a point and split based on this point which leaded to more diversified trees and fewer splits waited to be evaluated. Unlike random forest which might over-fitting the noisy data, extra trees outperformed random forest in speed and accuracy.

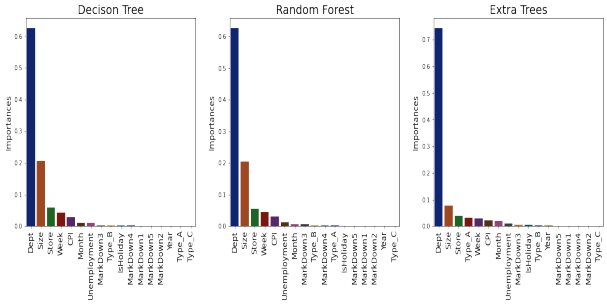

Fig. 7. *Feature importance*

IV. Results and Discussion

A. Feature importance

A nice attribute of decision tree model was that it provided a measurement of feature importance. Feature importance

was calculated as the decrease in node impurity weighted by the probability of reaching that node. From figure 7, top 5 important features weighted were "Department", "Size", "Store", "Week" and "CPI". Weekly Sales was mostly depending on department. The rest of 12 features were less relevant. Backward feature selection was performed with top 8 important features. Accuracy of the model decreased to 76 percent and R-squared drooped as well. Therefore, the experiment would proceed with all selected features.

B. Model Performance

The result was mostly evaluated on the weighted mean absolute error (WMAE):

$$WMAE = \frac{1}{\sum w_i} \sum_{i=1} n w_i |y_i - \hat{y}_i| \tag{6}$$

where:

- n is the number of rows
- $\hat{y}_i$ is the predicted sale
- y_i is the actual sales
- w_i are weights. $w = 5$ if the week is a holiday week, and 1 otherwise

After applying 4 regression models, the cross-validation score for each model was indicated below:

TABLE I
VALIDATION SCORE OF EACH MODEL

Regression Model	Cross Validation Score			
	WMAE	***MSE***	***R2***	***Time***
Linear Regression	12058	342793574	0.069	0.38s
Decision Tree	5054	99790006	0.571	2.7s
Random Forest	4453	82445724	0.772	23s
Extra Tree	4057	61242072	0.807	12.7s

Figure 8 below visualized the prediction error of each regression model. Upper-left plot represented the linear regression model. Feature and weekly sales did not show a linear relationship as the predictions do not converge to the true value (dash line). Hence, linear regression performed poorly, and it was under-fitting.

Decision Tree, Random Forest and Extra Tree fitted the data much better than linear regression. All of them converged to the true value (dash line) with some errors. Extra Tree was the best model with the least WMAE and highest R^2 score. It retained both accuracy and robustness and used only half of the computing time as Random Forest.

Noticed that same two points in each plot were spread out with high value of weekly sales. All of our model failed to predict these two points. Probably, these two points might belong to Thanksgiving's weekly sales which were extremely high. Moreover, data during November 2012 and December 2012 was lost. In the case of regression, Decision Tree, Random Forest and Extra Tree do not predict beyond the range of train data,and they might over-fit the data which were particularly noisy.

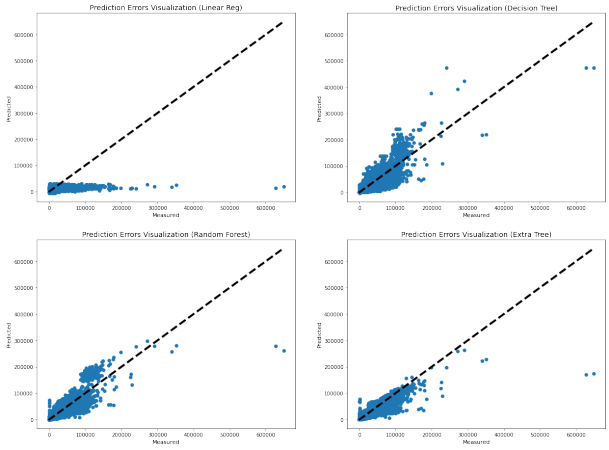

Fig. 8. *True Values vs. Predictions*

V. CONCLUSION

Throughout the data analysis and machine learning, we noticed that there were lots of missing values in our data. Especially, data during November 2012 and December 2012 was not provided. This might be the reason why we had two huge error displayed in Figure 8.

By utilizing Pearson's correlation, more hidden information was revealed by investigating correlation between different features. For instance, type is a proxy of size. Larger-size stores would expect to have larger weekly sales. Therefore, our features were not independent. This helped explain why linear regression model reacted poorly and under-fitting the data. Sometimes, Pearson's correlation might not able to explain relevance of every feature. Feature importance would help reveal more information.

Recalled that Holidays' weekly sales were weighted 5 times important than non-holidays' weekly sales. Data demonstrated that Thanksgiving was the only holiday which had an unusual high weekly sales. The final evaluation based on WMAE for all holidays was not accurate. The solution might be manually adjusted and included other national holidays.

Decision Tree, Random Forest and Extra Tree fitted the data much better than linear regression. All of them provided good indicators of importance. Extra Tree was the best model with the least WMAE and highest R^2 score. It retains, accuracy, robustness and computational efficiency. Even though there might exist other models which have better performance, extra tree is relatively simple, fast and easy to apply.

[1] [2] [3] [4]

REFERENCES

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[2] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[3] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[4] "Walmart recruiting - store sales forecasting." [Online]. Available: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data