Walmart's Sales Forecasting Final Report

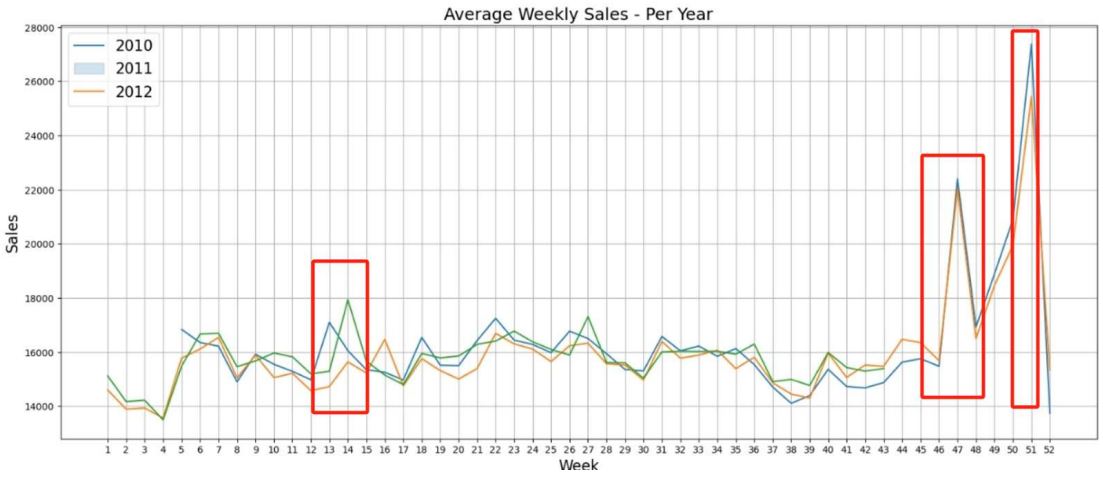
By Ye Guo

1. Introduction

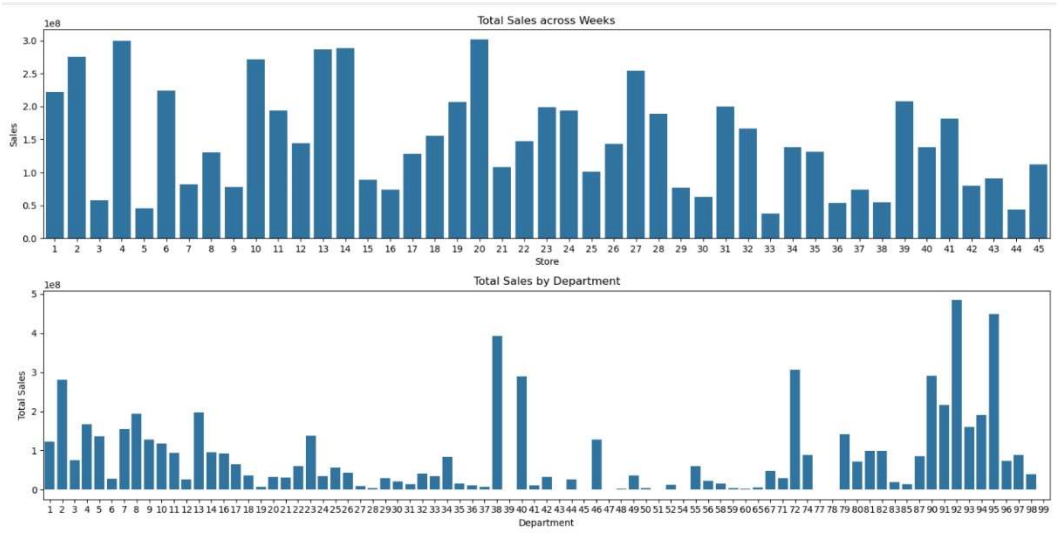
This is a brief report about some business insights based on the predictive analysis of data of Walmart's sales performance. I will first conduct and get insight from the EDA, then provide three main suggestions for Walmart's to improve the sales performance.

2. EDA analysis

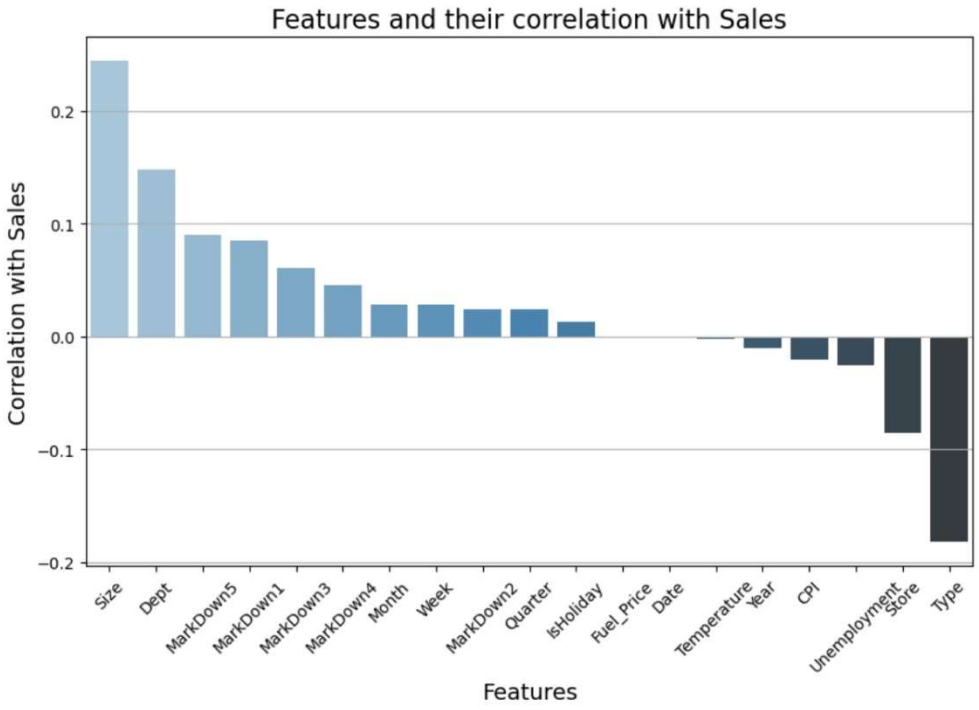
This line chart shows the sales change by time. We can see that sales remains a stable fluctuation in most of the period of year. However, a notable decrease around week 42 and a subsequent resurgence during the holiday season. It is because two big holidays – Thanksgiving and Christmas, are around those weeks.



This line char shows the average sales per week by different years and weeks. We can find that there is so much significant difference change of sales among 2010-2012. That means that we can assume all the features and factors that influence the sales can have the same impact on the following year 2013 and 2014. The three peak of average sales is Super Bowl, Thanksgiving and Christmas week.



These two bar charts show the sales of different stores and departments. We can see that the sales of different stores and departments is uneven. Walmart can focus on the stores generating low sales and gather more information about these stores and departments. More customized actions should be done to different stores. Plus, for department perspective, some departments generate nearly 0 sales during the year 2010-2012. Considering there is a huge number of departments for each stores, Walmart can consider to cut some departments and reallocate the resources to prime departments. More research and field experience should be undertaken before making decisions.



This bar chart shows the linear correlation between different features, and how they are important to boost sales in the perspective of linear relationship. We can see that the size and department is two of the most important factors to make impact on sales. Markdown 1-5 is also very important. However, below conclusions are only eligible in the situation of linear relationship. For more complex predictive models and machine learning, we need to do more features importance check and evaluation.

3. Modelling

3.1 Modelling evaluation and choosing

Below is performance of two different models:

Model	R-squared	RMSE	WMAE	OOB score	Score on train	Score on test	Training time
Random Forest	0.9746	3638.22	1615.62	0.975	0.997	0.975	425.88s
XGBoost	0.9470	5257.83	3118.93	/	0.952	0.947	0.68s
Randomized Search for XGBoost	0.6373	13752.48	8780.22	/	0.637	0.637	1.61s

We can see that while XGBoost have a sharply decrease of the training time, random forest has a better performance not only WMAE, but also in RMSE, using the default parameters set. Although using randomized search to find the best parameters of the modelling, I find it still does not overperform the random forest. Therefore, the final model should be a trade-off. In business reality, longer training time means more money to invest in maintaining fees and computing power. The costs is tend to increase exponentially. If XGBoost's RMSE and WMSE falls in the expected range of Walmart, then XGBoost is

the best choice. Otherwise, random forest should be accepted. It should be case by case.

3.2 Features Importance

Features Importance

Dept	0.625071
Size	0.193173
Store	0.056783
CPI	0.025782
Thanksgiving	0.017631
Type	0.014652
Unemployment	0.010805
Temperature	0.009938
Week	0.009820
Days to Thanksgiving	0.006369
Days to LaborDay	0.006273
Days to Christmas	0.006186
Fuel_Price	0.004353
Days to SuperBowlWeek	0.003325
MarkDown3	0.002543
MarkDown4	0.001476
MarkDown5	0.001146
MarkDown1	0.001019
MarkDown2	0.000955
Month	0.000764
Christmas	0.000727
IsHoliday	0.000568
Year	0.000217
SuperBowlWeek	0.000159
LaborDay	0.000156
Quarter	0.000109

Assuming to choose random forest model, this importance table shows that Department and size is the two most important features that decide the model performance. Out of my expectation, the markdowns are not so important to influence the sales performance. If the model is good enough, then it means that maybe the result of all kinds of markdown activities is not well enough, so more investigation and improvement should be taken to these markdown activities.