Predicting Walmart Sales

Problem Identification

Walmart is known worldwide as a multinational retail corporation characterized for the competitive prices on their products. Walmart is formed by different departments, and stores. The main idea of this project is trying to predict the amount of weekly sales a store could have as a function of variables such as store, and department number, unemployment rate of the area, CPI, size and type of store, and many others.

To solve this problem, I created a model that predicts the number of weekly sales a Walmart store could have using a dataset Walmart from Kaggle.

Data

To give a brief explanation of the dataset, we had 4 different datasets, each one with information we needed to make our model predictions.

The first data set called "Features" consisted on 12 columns and 8190 rows. The features included in this dataset were: Date, Store, Temperature, Fuel Price, 5 Markdowns, CPI, Unemployment and if it is Holiday or not.

The second dataset had 3 columns and 45 entries. The column names were: Store, its type and Size.

The third data set included 5 columns, and 421,570 rows. The features were: Store, Department, Date, Weekly Sales and if it is Holiday or not.

The last dataset we decided to include as additional work to see if the stock price has whether or not influence on the weekly sales.

We had to follow different steps in order to create our model. The steps were Data Wrangling, Exploratory Data Analysis, and finally our Modeling part. The description of each step is described below.

Data Wrangling

This step consisted on converting the messy dataset we extracted from Kaggle into a more organized data frame. This step is crucial to make the modeling process so much easier and faster.

We completed this process thanks to the Python's library Pandas.

Firstly, we needed somehow to merge all of the 4 datasets we initially had into just one. This would make easier to define the dependent and independent variables in the modelling part. We have the

pandas function merge to put together all of the data frames in just one and to avoid repetitiveness of columns, we specified the matching columns. Since we had 4 different datasets, we needed to make 3 inner merges to complete this process. The final result of this process was a data frame with 17 columns and 409,727 rows.

Secondly, we handled the missing values of our dataset. While we were doing the data wrangling, we realized the big amount of no-values. We were on the obligation to fill or remove those values using the methods we learned during the journey at springboard. We established a threshold of 30%. This means the following: If a column has more than 30% of no-values, we simply removed this column and row for the analysis because we did not have enough data to make an acceptable analysis and mostly this would be supported from assumptions. After doing so, the Markdown columns were removed.
Now it is time to fill the No-values for the missing columns. For stock price we decided to use the back-fill method (Stocks are not traded during holidays, so we assumed Walmart Stock's price did not change during all the weekend). Finally, for unemployment and CPI, we filled the no-values using the mean.

Lastly, we needed to convert the departments, stores, and type of stores as categorical features for our analysis. To do so, we used the get dummies function available in pandas. We got a final data frame with 409,727 rows and 138 columns. The dataset is ready to use for exploratory data analysis and modeling.

Exploratory Data Analysis

At this step, we figured out the relation among the variables with weekly sales Walmart had. To do so, we made correlation heatmaps, and to see how some variables were distributed, we took advantage of using histograms and boxplots.

Before showing some plots, we removed the outliers. Outliers have negative influence on the plots and statistical inference. Removing outliers could be a subjective matter, but in order to do this process, we used a statistical agreement that is considered outlier if the point is further the mean plus or less than 1.5 times the Interquartile Range. Special mention that the Interquartile Range is the third quantile minus the first quantile. The equation is below:

$$Outlier\ if\ x > mean + 1.5 * IQR\ or\ x < mean - 1.5 * IQR$$

After removing the outliers, we wanted to check how the weekly sales variable was distributed. We found the boxplot as a perfect way to see its distribution. Below is the boxplot we got.

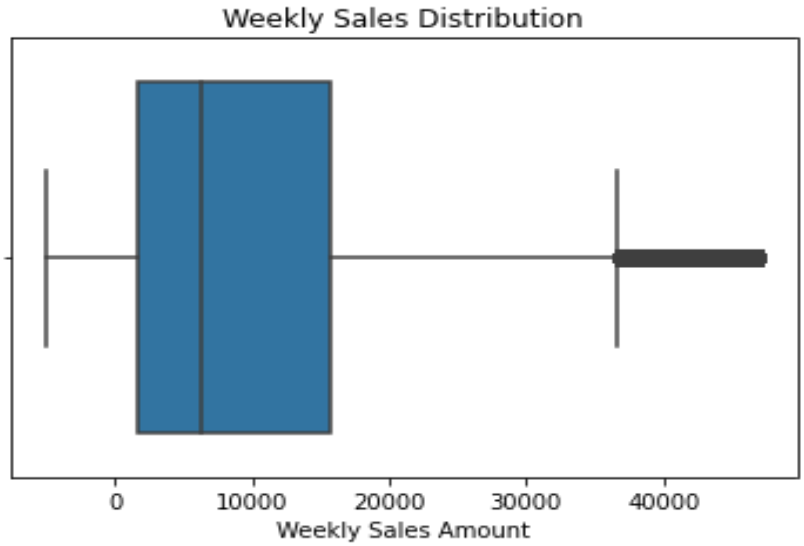

Exhibit 1. Weekly Sales Distribution.

From the figure 1 I was able to draw some conclusions. The weekly sales variable is not normal distributed, and it has positive skewed (or right skewed).

After, I plotted a correlation heatmap of the numerical features (I decided not to take the categorical features because of the big amount of data would cause the heatmap very difficult to read). The heatmap is shown below on figure 2:

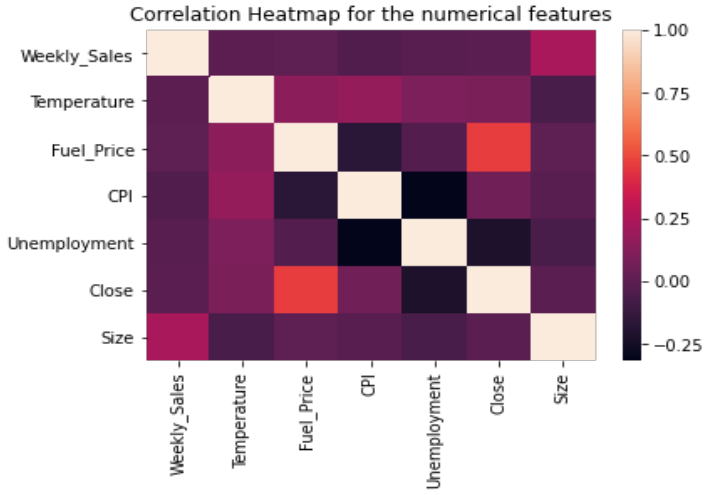

Exhibit 2. Correlation Heatmap for the numerical features

From the correlation heatmap, we did not see variables highly correlated between them. Weekly Sales is more correlated with the size of store than any other variable and there is no other variable that has some kind of correlation with weekly sales. We can reduce the number of features for the

final model and by doing so, we would reduce the dimensionality problems this specific problem could have.

We also did a correlation heatmap for the categorical features. It creates some complications to read the heatmap for all the departments and stores, so we did an extra step before. We reduced our analysis to the 10 departments and stores with higher amount of weekly sales (by taking the mean), so this can tell us how the other departments may behave (and the results were close enough). The departments selected were 8, 13, 4, 79, 46, 23, 10, 40, 7, and 1 and the stores selected were 23, 6, 11, 27, 2, 13, 20, 28, 19, and 1.
The boxplots for the department and stores distribution of weekly sales are shown below:

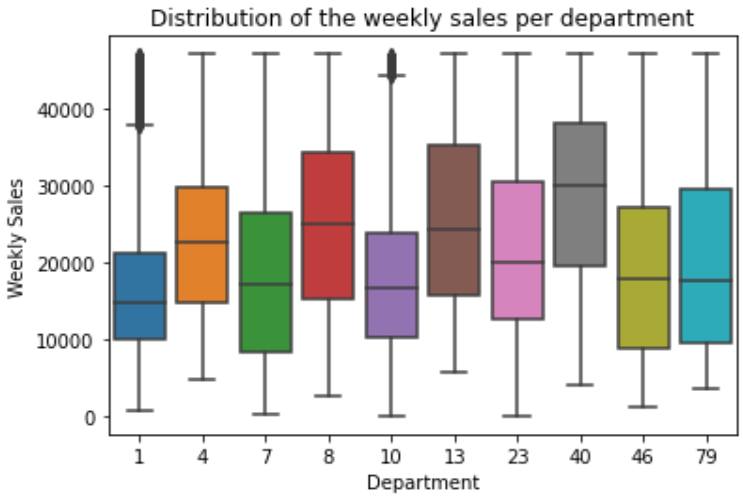

Exhibit 3. Department's

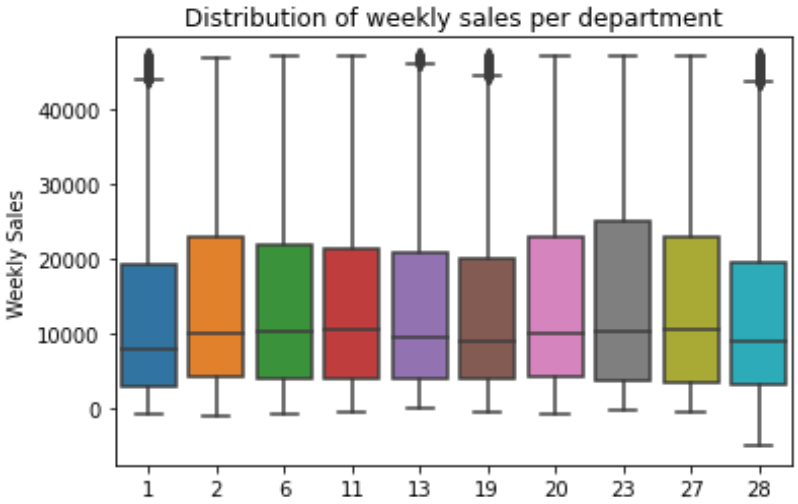

Exhibit 4. Store's distribution.

There is not a clear pattern of how the weekly sales is distributed based on the department, but on the other hand, all the weekly sales distribution using the stores as variable are skewed to the right (same as the weekly sales distribution).

Now, let's take a look to the categorial features heatmap.

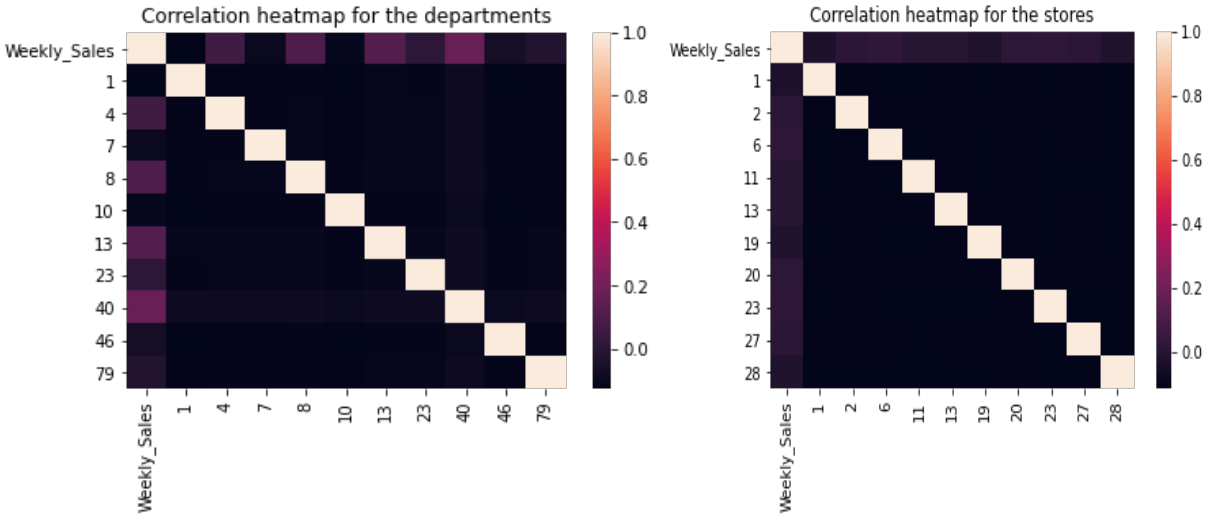

Exhibit 5. Correlation heatmap for departments (on the left) and stores (on the right)

From figure 5, we did not see any significant correlation between any department or store and the weekly sales variable.

Hypothesis Testing

We wanted to repeat the experiment over and over again to see if this dataset had those numbers by chance. To complete this step, we used the bootstrapping method. Basically, repeat the experiment over and over several times, and finally see if the department 8 (that was the department with higher amount of sales) would tend to have higher mean than the mean of the other nine departments. The bootstrapping results are shown below.

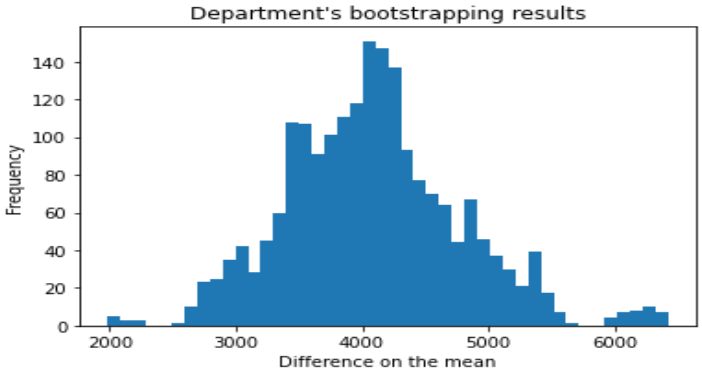

Exhibit 6. Bootstrapping results

From figure 6, we can draw the conclusion that the department 8 would have higher sales than the mean of the other departments, which basically means department does matter to calculate the amount of sales a store could have.

Now, for stores, we got the following results:

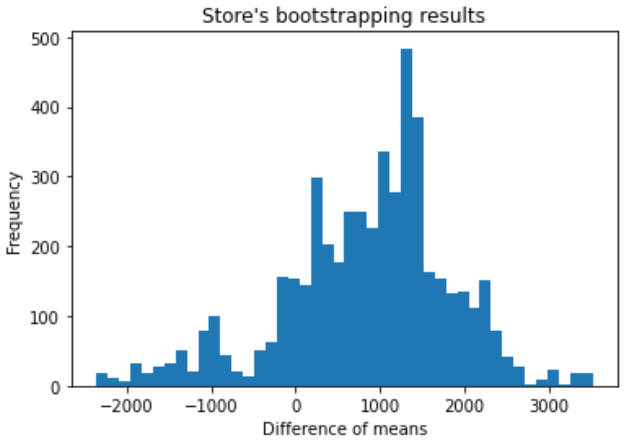

Exhibit 7. Store's bootstrapping results

For this bootstrapping, we got a p-value of 0.1718 that shows no statistical significance, and therefore the store that had higher amount of weekly sales would not necessary tend to have higher amount of weekly sales than the other stores.

Modeling

For the modeling part, and by taking into consideration the wide dataset, I used the correlation heatmap to see how we can reduce the dimensionality of our dataset, and after that use scikit learn library to start the modeling and creation of our model. We also counted the values for department and stores, so I was able to remove some variables when I considered the data obtained was not enough. By looking the correlation heatmap, all departments had week correlation between the department and the weekly sales, so we defined a threshold of 0.10. Any department with higher correlation than 0.1 was included in the modeling part.

We got a model with 17 departments, type of store (3 different types in total) and size of the store. A total of 21 independent variables and 1 dependent variable (weekly sales).

We applied standard linear regression, Lasso Regression, Ridge Regression, Random Forest Regression, Gradient Boosting Regressor, and Decision Tree.

The model we saw good accuracy (0.9284 on the train data and 0.9283 on the test data) was on Gradient Boosting, also because we did not feel we were overfitting the model and be influenced by the noise. Comparison of the results are shown below:

	Actual	Prediction
8927	33663.65	32754.031722
2993	28477.46	28937.577945
4134	35973.05	33451.138653
10271	25363.63	28034.463433
1295	39196.18	23339.395267
12806	47329.06	41958.640633
2806	495.60	1698.934306
7459	288.00	1620.273834
9758	7.98	1030.939735
13259	32767.40	32898.747156
4492	42723.81	34938.940346
7908	25629.14	32017.974920
1733	36515.52	34021.138511
14599	20405.78	23166.250122
5856	16771.51	11227.158383

Exhibit 8. Actual and Predicted values of the model with selected departments

I also decided to do a model by taking into consideration all the variables and not just the variables with the threshold of 0.1 and we also got decent results. Some comparisons between the actual input and the predicted value are shown below.

	Actual	Predicted
13386	213.80	1334.081225
14233	42534.06	37534.619609
9116	19.39	1116.285578
15951	27400.48	28678.358873
5374	34.88	1092.740586
17299	17290.67	23264.942656
8757	115.56	992.779161
2146	234.97	1352.929939
16798	31893.42	26633.947318
17651	459.00	1384.517649
18150	30853.03	28623.481813
11690	46078.83	38178.134872
16330	3409.81	4216.630486
15326	3437.23	3449.035980
3862	10633.94	11687.697497

Exhibit 9. Actual and Predicted values of the model.

The accuracy was of 0.94.29 on the train data and 0.934 on the test data.

For the feature importance, we got the following graphic:

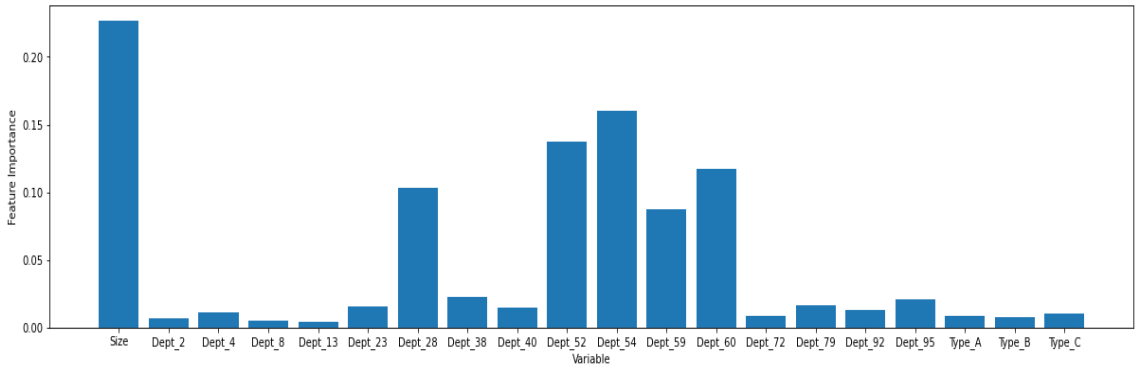

Exhibit 10. Feature importance graphic

From the graphic, we can conclude that the most important variables are: Size, and departments 54, 52, 60, 59 and 28.

I concluded based on the comparison between actual and predicted model for all the variables that our model is not heavily influenced by the noise of the data and moreover, since we included a learning rate of 0.1 as hyperparameter, we expect this model has low variance and low bias as well, that is something we look when we are building a model.

As an important mention, we also created a data frame with the feature importance for the Gradient Boosting Regressor. From the following exhibit, we can know the most important features.
Some of the most important coefficients are shown below.

	Feature Importance
Size	0.179931
Dept_54	0.154807
Dept_52	0.132801
Dept_60	0.113299
Dept_28	0.100208
Dept_59	0.084867
Dept_38	0.024187
Dept_95	0.021854
Dept_23	0.018439
Dept_79	0.017046
Dept_40	0.016444

Exhibit 11. Most important variables based on the Ridge Regression.

It is important to mention that the most important features of exhibit 10 are the same than the ones reflected on the exhibit 11, nevertheless some variables are in different order.

Key Insights

- Department, store and store's size are the variables that has more influence on the amount of weekly sales and should be studied before any decision is made in terms of opening a new retail.
- The correlation heatmaps and the statistical inference also supported what is mentioned above. Though, the number of department has more influence than the store number.
- Variables such as unemployment, CPI, Fuel Price, and any other variable but department, store and size of the store does not have too much influence on the outcome of the weekly sales. Basically, this is because Walmart is an international retailer that competes with lower prices and buy in-bulk to the suppliers.