Sales Forecasting with Walmart

K.Ahmed (6601553), P.Dobariya (7099872), R.Solanki (6805329)

Abstract

Consider a large dataset containing a variety of information on factors affecting the Weekly Sales of Walmart such as temperature, fuel price, Holiday Flag, unemployment rate, Date, Store number, and Consumer Price Index (CPI). This dataset (6435 observations) was obtained from the website, Kaggle. Weekly Sales is chosen as the target variable and five of the seven variables have been chosen as response variables to include in linear regression, K-nearest model, Decision Tree model and neural network. The fifth model that was also used to test the dataset was Random Forest Algorithm. Conclusions about the relationship between variables are drawn and the two models are compared, with a conclusion that Linear Regression Model is a better fit for the data.

Models

-> To remain consistent with all the models, the model was trained at 33% and tested it to find the predicted values.

->Linear Regression: This model was comparatively more accurate than the others considering the data and the test was heavily dependent on regression. So multiple linear regression worked out relatively well. And this is what the predictions looked like. The mean absolute error was 2155.5.

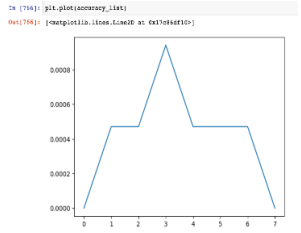
```
[[ 956543.6940401 ]
 [1110992.9186957 ]
 [ 994441.76663751]
 ...
 [1220157.88478776]
 [ 995294.10385453]
 [1049034.58645224]]
```

->The Decision Tree model was little more difficult to build, setting up lab encoding to convert y values to categorical values. Here after, the predicted values, with the MAE and MSE look something like this:

->K-Nearest Model: Predicted value of K

```
[3892  735  611 ...  461 3601 2706]
Mean Absolute Error: 1046850.492212806
Mean Squared Error: 1415263132991.4573
```

using following graph. The mean squared error was 1420573487396.8323.



->Neural Network: It consists of three different kinds of layers: input layer (5), hidden layer(3), and the output layer(1). The output layer then classifies the output within the given classes, according to the regression. The number of epochs chosen were 100. After training the model, the loss and performance looked something like this:

->Random Forest Algorithm: This model works best for the unique dataset because it can handle big data with numerous variables

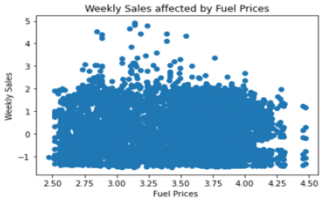
```
performance = model.evaluate(X_test, y_test)
print(performance)
```

```
67/67 [==============================] - 0s 1ms/step - loss: 1040694.8125
[1040694.8125, 0.0]
```

running into thousands. It can automatically balance data sets when a class is more infrequent than other classes in the data. Predicted values: [860546.9905, 844477.085, 1974077.2995,1496068.4895, 449494.2675, 473349.8495]

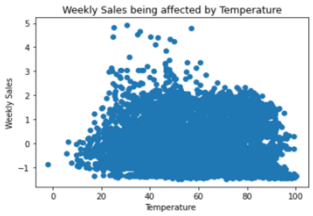
Graphs

This Graph hints towards the negative correlation between fuel process and weekly sales.

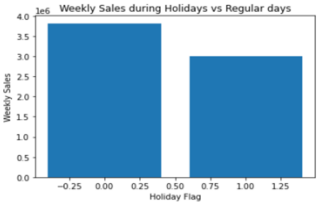


This Graphs shows the negative correlation between the Unemployment rate and the weekly sales.

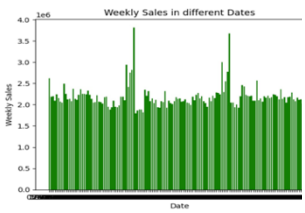
Here's how Temperature influences weekly sales with heavily scattered around 30-80.



The bar graph below describes the weekly sales during a holiday and a regular day. As evident as it is, holiday gives much higher values than regular days.



Below is a graph showing the highest weekly Sales with certain dates throughout the period 2010 to 2012. And two clear major spikes are observed, using max function, it was found that one of them was 24th December in 2010 .




```
col  = 'Weekly_Sales'
max_x = data.loc[df[col].idxmax()]
print(max_x)
```

```
Store                14
Date          24-12-2010
Weekly_Sales   3818686.45
Holiday_Flag            0
Temperature         30.59
Fuel_Price          3.141
CPI                182.54459
Unemployment        8.724
Name: 1905, dtype: object
```

Conclusion

Concluding this analysis on model testing, Linear Regression model worked relatively well with the lowest mean absolute error of 2155.5 in comparison to all the other models, Decision Tree and Neural Network with a 7 figure number for MAE, and KNN with a whopping 13 figure number. It makes sense as this data set required a regression analysis testing, with and linear regression helped evaluate trends and sales estimate, and analyze the impact of price changes. KNN model testing worked poorly on our dataset because it is a distance-based algorithm, so the cost of calculating distance between a new point and each existing point is very high which in turn degrades the performance of the algorithm. Decision Tree and neural network both gave very high errors as well, so naturally these models would not be the best fit for representing this dataset either

Improvements

The data set could have been scaled and transformed into values which would have made it easier to train the model and get good predictions and smaller errors for each model. This is unique dataset consisting of various numbers heavily scattered everywhere, if the dataset was chosen through stratified sampling and then carry out non-parametric regression instead of simple linear regression, the error again, might have been way less than observed in this model testing. Below is the link to the github repository with the codes of this project:
https://github.com/ka18vw/DataScience_GroupProject.git