## **Walmart Sales Forecast Analysis**

<https://github.com/valarjeya28/Walmart_Sales_Predict>

## Objective – The problem you are trying to solve/predict

 Data from Walmart stores across the US is given, and it is up to us to forecast their weekly sales. The data is already split into a training and a test set, and we want to fit a model to the training data that is able to forecast those weeks sales as accurately as possible.

# Data set – Size of the data, # of features, example features

The train data having **282451 rows × 40 columns**

# Data cleansing/manipulation (if any) – Exploratory analysis, derived variables, missing values, outliers etc.

* The Data cleaning is done using Pandas Data Frame.
* Added the last weeks date as a new variable, we will use this to get sales of previous and compare difference

# Methodology/ML technique used– Please explain the different models that were tried and the reason for choosing the final one

 The base case of our model will be a regression baseline using Random Forest Regression model

Analysis:

* Most of what we see in the correlation table is of little surprise. Discounts are correlated and higher unemployment means lower Consumer Price Index. More interestingly, it appears that higher department numbers have higher sales. Maybe because they are newer? Also, larger stores generate more sales, discounts generally generate higher sales values and larger unemployment result in a bit fewer sales. Unfortunately, there appears to be little relationship between holidays, temperatures or fuel prices with our weekly sales.
* We notice that type C stores have fewer sales in general and holidays clearly show more sales.
* The end of November sees a lot of exceptionally large sales. This special day, better known as Black Friday, causes sales to be on fire, and undoubtedly a dummy variable should be created for this day. Also, Christmas, appears here and there.

### **Missing values**

* We will start with filling in any blank values. There seem to be some missing values in the data.
* Dummy variables for Categorical Data
* We have our entire data frame ready with all the pre-processing.
* However, we do not need all these features. We did all these feature processing for us to understand the data better.
* Now, if we go all these features to the model, some of the features will be repeated. For example, month and date
* The variable to be forecasted to the difference from the median.

### **Split data into training and test sets**

Now we can split train test again and of course remove the trivial weekly sales data from the test set.

## **Model selection:**

### **Random Forest**

## **Forecasting sales**

**After we have created our model, we can predict things with it on the test set**