**Sales Prediction (Walmart)**

**PROBLEM STATEMENT:**

We are provided historical sales data for 45 Walmart stores located in different regions. Each store contains a number of departments, and we are tasked with predicting the department-wide sales for each store.

In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data.

The dataset essentially contains five scv files which are descibed as follows:

* stores.csv
  + This file contains anonymized information about the 45 stores, indicating the type and size of store.
* train.csv
  + This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file we will find the following fields:
    - Store - the store number
    - Dept - the department number
    - Date - the week
    - Weekly\_Sales - sales for the given department in the given store
    - IsHoliday - whether the week is a special holiday week
* test.csv
  + This file is identical to train.csv, except we have withheld the weekly sales. We must predict the sales for each triplet of store, department, and date in this file.
* features.csv
  + This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:
  + Store - the store number
  + Date - the week
  + Temperature - average temperature in the region
  + Fuel\_Price - cost of fuel in the region
  + MarkDown1 to MarkDown5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
  + CPI - the consumer price index
  + Unemployment - the unemployment rate
  + IsHoliday - whether the week is a special holiday week
* sampleSubmission.csv
  + We have been provided with dates where we can predict the salesfor each date mentioned in this csv file.

For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):

- Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13

-Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13

- Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13

- Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

**INSIGHTS:**

* Most of the Departments have weekly sales below 200000.
* Weekly sales is highest for Store number 10 and lowest for Store number 44.
* Year 2011 had the highest weekly sales.
* November month witnessed the maximum weekly sales.
* Week 47 of November month had the highest weekly sales.
* Type A store is the largest store and C being the smallest.
* Weekly sales for store type A is more than other store types.
* There is no explicit relation between Department and Weekly Sales.
* We see the sales are higher on holidays than on non-holidays.
* December month had maximum sales.
* In every month Store Type A has the maximum sales.
* The 51st week i.e. third week of December has the highest sales.
* About all sales have happened below 100000. Maximum sales are done at sales value of 1000.
* Most of the Markdowns have been given for Weekly sales of less than 200000.
* Most of the Fuel price value lies in range of 2.5 to 4 for weekly sales of less than 200000.
* Sales of less than 200000 happened for average temperature range of 30 to 80.
* No weekly sales for CPI range of 145 to 180. CPI range (200 - 230) is more when the weekly sales is less than 200000.
* The trend of weekly sales is such that it first decreases then elevates.
* Given walmart dataset is seasonal dataset as it is repeating the same pattern in the month of Nov-Dec.

**DEPLOYED MODELS:**

* Linear Regression
* KNN
* Ridge
* Lasso
* Decision Tree
* Random Forest
* ExtraTrees
* XGBOOST