**WALMART SALES FORECASTING**

**PROBLEM STATEMENT:**

Unable to manage the inventories as the retailers are not sure of the demand for the products that they sell.

**Project Objective:**

Objective is to create a model that can be able to predict the demand for the input features provided.

**Data Description:**

The dataset has 7 features in total as below:

* Store – Store ID.
* Date – Date of Sales.
* Holiday\_Flag – 0 – No; 1 – Yes
* Temperature – Temperature during sales.
* CPI – Customer Purchasing Index.
* Unemployment – Unemployment rate.

**Data Preprocessing Steps and Inspiration:**

* There are no null values/duplicate rows observed in the dataset.
* Convert Date Feature from int type to datetime type.
* The dataset comprises of sales that happened only on Fridays.
* Outliers are observed in Weekly\_Sales, Temperature and Unemployment features.
* Functions written to remove the outliers from the above features.
* Checked the sales on the day when maximum/minimum temperatures are observed.
* Checked for each feature and their impact on Weekly\_Sales.
* Perform model training using ARIMA with univariate feature as Weekly\_Sales and Date as index.

**Choosing the Algorithm for the Project:**

As it is a forecasting based use case, we would employ Time Series based models like ARIMA.

**Motivation and Reasons for Choosing the Algorithm:**

ARIMA is a statistical based model which helps us in forecasting the future sales/trends based on the historical data provided. The name ARIMA stands for Auto Regressive Integrate Moving Average which uses Auto regressive part as p which is obtained from Partial Auto Correlation Plot, Integrated as d which is obtained from differencing and Moving Average as q which is obtained from Auto Correlation Plot.

**Assumptions:**

One of the key assumptions that ARIMA model expects is that data should be stationary. This stationarity can be identified using Augmented Dickey Fuller Test (ADF Test) where the stationarity of the data can be confirmed when p < 0.05 and T-stats < Critical values.

**Model Evaluation and Techniques:**

Below metrics are used for evaluation of model’s performance:

* Mean Absolute Error.
* Mean Squared Error.
* Root Mean Squared Error.

**Inferences from the Same:**

* Store 20 has the highest sales whereas store 38 has the lowest sales.
* Highest Sales has happened on Sep 12,2011 whereas lowest sales has happened on Jan 27,2012.
* Most of the sales are ranging between 40000000 - 50000000.
* Sales has happened the highest during non-holidays when compared to holidays.
* Highest sales have happened when the temp was between 51-75 and the lowest sales have happened when the temperature was between 0-25. This is obvious as most people prefer shopping when the temp is moderate and not when temp is too low/high.
* Sales happened to be high when Fuel price is between 3.6 - 4.0 whereas it is low when fuel price is between 2 - 2.5
* Sales are high when the price of fuel is moderate (3.6-4.0). However, when the prices were low, there was not much of sales happened. May be due to rise in temperature or any other factors that would've affected the sales.
* Higher Sales happened when CPI is between 125 - 150. In contrast, lower sales happened when CPI is between 226-250. This is because when the prices of goods increases, their sales will automatically decrease.
* Sales has happened high when unemployment rate is moderate (6.1 - 8). However, least sales has happened when unemployment rate is between 10.1-12 which is obvious as most people won't have much spending power as they are unemployed.

**Future Possibilities of the Project:**

The model can be helpful to forecast the weekly sales for any given week.