TITLE DEMAND FORECASTING COMPARATIVE STUDY USING HADOOP AND SPARK FRAMEWORKS

MOTIVATION/ INTRODUCTION

Forecasting is a useful technique that can help to understand how historical data influences the future. This is done by looking at past data, defining the patterns, and producing short or long-term predictions
Forecasting is used in multiple areas within retail:
Supply chain – Demand forecasting. Have enough inventory available in the store to meet future demand – prevent out of stock, Workforce planning.
Finance – Sales forecasting, Budget planning , goal setting .

There are multiple forecasting techniques from naïve approaches to simple time series models like ARIMA, SARIMAX, and advanced models such as FBProphet, LSTM.

SCOPE OF THE PROJECT

Scope of this project is to explore the sales data using hadoop framework as well as the spark framework where try to give some forecast for the store owners to have the maximum profit .In hadoop we use naïve approach for forecasting and in the spark we use the ar algorithm which is very similar to the linear regression model which uses all the previous sales for predicting the future sales

METHODOLOGY

Naive approach for Hadoop
Naive Method: Forecast the value, for the time period t, to-be equal to the actual value observed in the previous period that is, time period (t – 1)

$$y_t = y_{t-1}$$

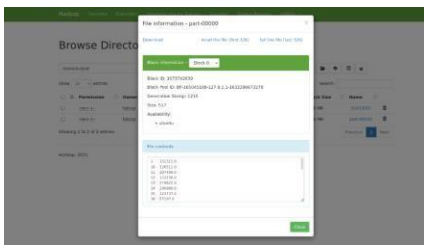
AR model for PYSPARK
Autoregressive models operate under the premise that past values have an effect on current values, which makes the statistical technique popular for analyzing nature, economics, and other processes that vary over time.
We forecast the variable of interest using a linear combination of *past values of the variable*. The term *auto* regression indicates that it is a regression of the variable against itself. Thus, an autoregressive model of order p can be written as

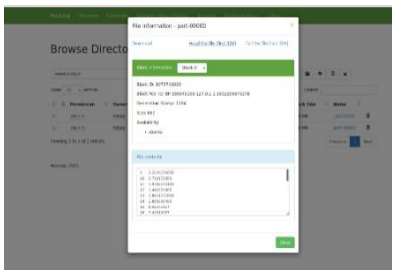
$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \epsilon_t$$

RESULTS OBSERVED

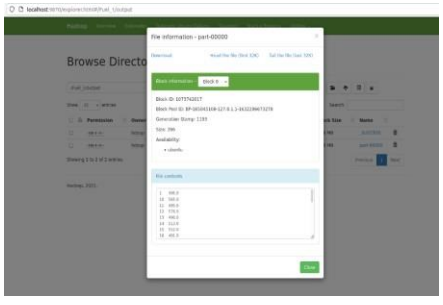
From Hadoop Framework,
In the Hadoop framework we use HDFS where we used naive model for forecasting the fuel price for each store, size of each store and accumulative weekly sales of each store

1. WEEKLY SALES FOR EACH AND EVERY STORE and SIZE OF EACH STORE



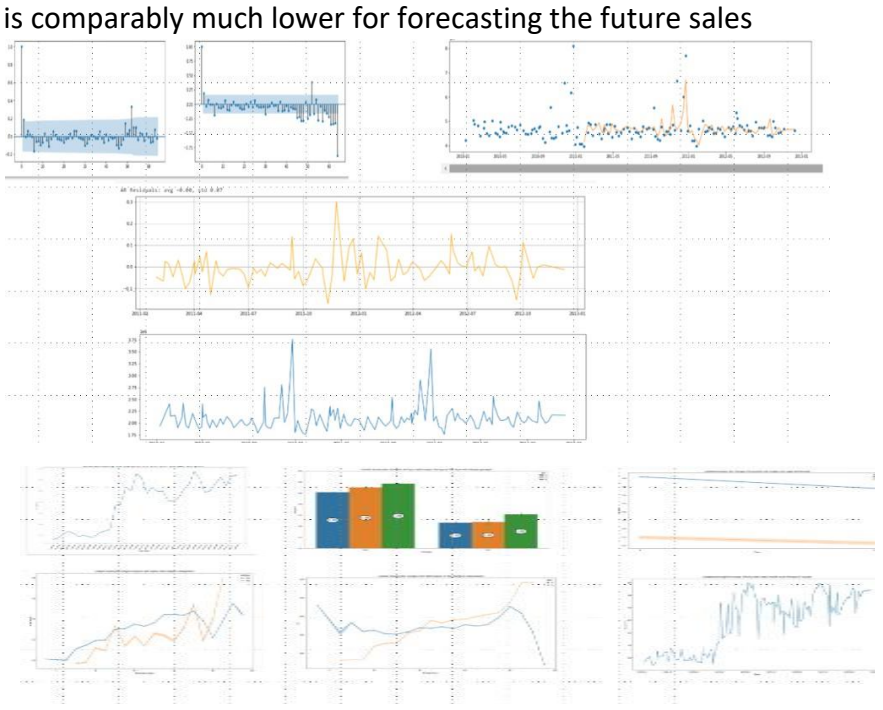


2. CUMULATIVE FUEL PRICE



From Pyspark Framework,

In pyspark we learnt about the data using pyspark SQL and used AR model for forecasting which gave 0.7 as standard deviations which is comparably much lower for forecasting the future sales



From the below table we can observe that the Type A stores has the highest weekly sales among the three stores in which store Number 19 recorded the highest weekly sales among the 40 stores.

	Store	Weekly_Sales
19	20	3.013978e+08
3	4	2.995440e+08
13	14	2.889999e+08

	Type	Weekly_Sales
0	A	4.331015e+09
1	B	2.000701e+09
2	C	4.055035e+08

CONCLUSION

The fuel price increases with increase in temperature steadily during workdays and unevenly during holidays. There is no significant pattern in the data points spread each month in the dataset.
However, one noticeable cue is that no sales data is recorded / happened during the month of September in 2013.There was a peak during the end of the years 2010 and 2011 but not during 2012. This might be due to comparatively very less observations during the last 2 months in 2012.

Inference made using dataset and analytical recommendation can be as follows.

From the basic analytics and forecasting we can suggest the shop owner that they can stock up the high selling products of various season in the low temperature period itself as fuel price tends to increase during the high temperature period. The temperature increase which in turn will reflect in fuel price and the stock prices. Similarly Store number 19 has the highest weekly sales from which all the other stores can take up same trend and cyclic variation in order to increase the sales.