DS502 Final Proposal

a. Members' names

Enbo Tian, Xiuhan Li, Ziyang Xu

b. Description of the problem

In this problem, we want to use the dataset of Walmart from 2010-02-05 to 2012-11-01 to analyze several tasks.

1. Which store has maximum sales
2. Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation
3. Which store/s has good quarterly growth rate in Q3' 2012
4. Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together
5. Provide a monthly and semester view of sales in units and give insights

The main goal is to build prediction models to forecast demand

c. Description of the dataset (dimensions, names of variables with their description)

The dataset contains 8 dimensions.

Store - the store number

Date - the week of sales

Weekly_Sales - sales for the given store

Holiday_Flag - whether the week is a special holiday week 1 - Holiday week 0 - Non-holiday week

Temperature - Temperature on the day of sale

Fuel_Price - Cost of fuel in the region

CPI - Prevailing consumer price index

Unemployment - Prevailing unemployment rate

d. Regression or classification?

Regression

e. The methods you plan to try.

Linear Regression

Tree Basic method

f. The error metrics you plan to use and the algorithms for assessing them.

We tend to use MSE, MAE and RMSE as the error metrics.

g. Comments and/ or concerns?

For Store 1 - Build prediction models to forecast demand

Linear Regression - Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.

Change dates into days by creating new variables.

Select the model which gives best accuracy