

Juan Rojas

Jesus Mendoza

Brian Canela

Gian Tolentino

Abraham Vega

Walmart Stores Sale Forecasting

The project we chose involves the company Walmart. Based on the data from 45 different anonymous stores, our goal for this project is to be able to predict the quantity of sales per week based on the data. We are given four data sets. The four data sets are features, stores, train and test. The features data set provides all information of a specific store, the features are Store number, date, temperature, fuel price, 5 columns of markdown, consumer price index, unemployment rate and a Boolean of whether that week contained a holiday. The stores data set contains information of the store itself such as the type and the size of it. The train data set contains information on the store number, the department, date, weekly sales, and whether it is a holiday or not. The test data set contains the same information as the train data set except the weekly sales. Our training data contains 420k + rows; the testing data has 110k + rows.

After finding out what might be the best way to approach this problem, we agreed on how the most effective and efficient way of working together would be. In order to demonstrate our group work structure, we defined this layout:

→ 2 People (P1 , P2) # Where P1 and P2 represent each person in a set

MLA = machine learning algorithm

Chose a machine learning algorithm that can predict a numerical value

- * P1 - Write the code which implements the MLA, meaning

- training, testing, and debugging the code so it works with all features.

- This person will also calculate the accuracy and/or the error of their MLA.

- * P2 - This person uses P1's code, except this person is tasked with improving the MLA by trying to do feature reduction of changing a few variables in MLA.

- This person will also calculate the accuracy and/or the error of their improved MLA.

→ 2 People (P3 , P4)

Chose a different MLA to train and test the data to predict the numerical values desired.

- * P3 - Repeat same procedure as P1.

- * P4 - Repeat same procedure as P2.

→ 1 Person (P5)

- * P5 - This person determines which of the algorithms from each team subset. Meaning that this person must test each of P2 and P4's optimized MLA.
- This person is also in charge of doing all the visualizations for the project. Once the best algorithm is selected, this person will depict how visualizations of how the data, dimensional reduction, etc. contributes to the result of the MLA's accuracy and error.

The above outline describes how the work will be split up amongst our group, specifying what each team member will contribute to the group.

As for what we have so far, the first team subset members (P1 , P3) have started, those who are in charge of implementing the MLA, training, and testing. They are currently working on their part, while their other team subset researches how to implement feature reduction / selection. Once they have finished, they will pass their code to the next team subset members (P2 , P4) to optimize their machine learning algorithm. Lastly, the lone member (P5), is researching how to implement data vacuolation for the best algorithm (while P2 & P4 finish). Once he receives P2 & P5 's enhanced MLA, he will determine which is best, then implement the visualizations to share the significance and improvement in the MLA. In summary, P1 and P3 are currently working on the MLA implementation of the algorithm they selected, while the other team members learn how they will complete their responsibilities (such as feature reduction).