# ORIE4741 Midterm Report

## Sales In Stormy Weather

---

**Team members: Faisal Alkaabneh (fma34) and Lawrence Hu (ljh238).**

---

### Overview

We are working solving the challenge of accurately predicting the sales of 111 potentially weather-sensitive products (like umbrellas, bread, and milk) around the time of major weather events at 45 of Walmart's retail locations.

The ultimate goal of our project hence is to help Walmart better predict sales of weather-sensitive products to keep their valued customers out of the rain. The dataset is available through: https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather/data

As far as the collected data is concerned, we will be using the following data to achieve our goal of getting a high accurate prediction model:

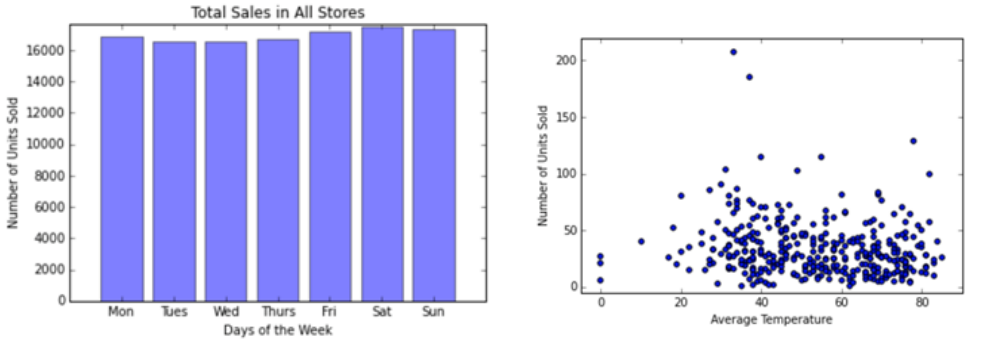
- Weather data following the National Oceanic and Atmospheric Administration (NOAA) standards: a guide to understand the data related to the weather. NOAA weather information is provided for each station and day.
- Sales data for all stores & dates in the training set, the set also include information about each item.

### Dataset

The dataset presents information on the number of units sold for 111 items in 45 Wal-Mart stores as well as information on the weather collected from different weather stations on each day from 2012 to 2014. Due to the nature of this data, there are many different explanatory variables. Weather variables account for 19 of the total number of features and around 20,000 entries in the dataset. Sales data from the different stores contain about 120,000 entries for each of the stores for the years 2012 to 2014. Some information is

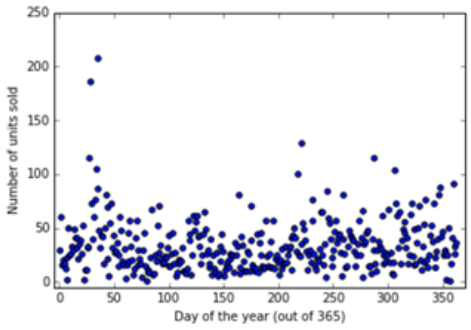
missing from the data collected from the weather stations, so this will have to be accounted for.

Attempts to visualize data were made to gain overall understanding of the data. Only a few were presented in this report:




(a) Total sales vs days of the week. (b) Total sales vs average temperature.

Figure 1: Data Visualization.




(a) Total sales vs days of the week.

Figure 2: Data Visualization.

Based on our experiences, we think that some features contribute to the number of sales more than others. For instance, the day of the week, whether or not it was during the holiday seasons, the amount of precipitation recorded on a given day, the average temperature.

With so many possible explanatory variables, we need to disregard some that are insignificant. Some data collected from the weather station can be discarded since intuitively, they offer very little insight into our observations. For instance, sunrise/sunset times, dew levels, etc.

After assessing all the variables, the ones that can be deemed significant are the following:

- Weekday.
- Is weekend.
- Is holiday.
- Is holiday and weekday.
- Is holiday and weekend.
- Item nbr.
- Store nbr.
- Date.
- Tmin.
- Tmax.
- Tavg.
- Special sales days.
- Preciptotal.

**Modeling And Future Work**

In order to decide the best model among the several suggested ones and validate the several suggested models, a robust methodology has to be followed. To this end, we are planning to have a set of cross validation experiments.

As a first attempt we are in the progress of investigating the efficiency of decision trees. Recall that our data has the sales information for 111 items at 45 Walmart stores locations. Furthermore, since we have no idea about the stores location, it is worth developing several models using decision trees to each set of locations to account for variabilities due to location. The second step in the modeling process is to run Lasso model. A big advantage of lasso model would be the capability of identifying several parameters that might not be a good fit in our model. For instance, by looking at Figure 1, it seems that weekday is not a significant factor. Likewise, several parameters will be omitted due to their insignificance.

The third step after identifying the significant variables to be included in the model would be fitting a linear model. Afterwards, we will investigate more into the nature of explanatory variables (are they linearly or nonlinearly independent?). Therefore, assessing the need to perform regularization or not to guarantee that the model only produces a unique solution.