Project Report Dec. 4, 2017

Final Report

Author: *ORIE4741*

Sales In Stormy Weather

Name	netID
Faisal Alkaabneh	fma34
Lawrence Hu	ljh238

Abstract

Walmart operates 11,450 stores in 27 countries, managing inventory across varying climates and cultures. Extreme weather events, like hurricanes, blizzards, and floods, can have a huge impact on sales at the store and product level. In this project, we are developing models to predict the sales of 111 potentially weather-sensitive products at 2 of Walmart retail locations.

Lasso and linear regression models were explored. The results show that Lasso model performs better and predicts sales with MSE 43.0062. Lastly, we provide some recommendations and limitations of our modeling procedure.

1 Introduction

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:

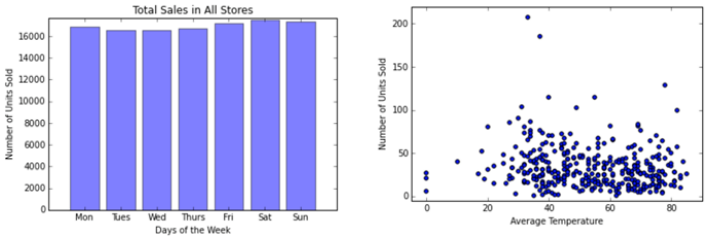
- Weather data following the National Oceanic and Atmospheric Administration (NOAA) standards: a guide to understand the data related to the weather. NOAA weather information is provided for each station and day.
- Sales data for all stores & dates in the training set, the set also include information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

1.1 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.

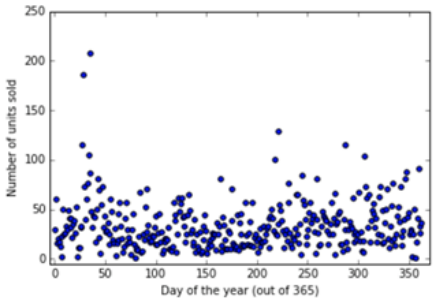
Attempts to visualize data were made to gain overall understanding of the data. Only a few were presented in this report:



(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.

1.2 Data Cleaning and missing data points

The data provided on the website is missing some critical information on the sales of products that are out of stock. In several data points, we have demand of zero implying that no units are sold of this unit. Nonetheless, per the guidelines provided in the competition: "The sales data does not capture the difference between the stock and the demand. In other words, sales number 0 doesn't necessarily mean there was no demand for this product." Hence if the model predicts a stocking policy of zero for that product, it may lead to loss in sales and hence poor prediction. To circumvent this deficiency, we will take off these values to provide better training. Nonetheless, in the last section we provide some recommendation of how to deal with that in real-world settings.

The other difficulty we ran into is the location independent variable. Having location data in our dataset adds more challenges to the model which is already

complex. One way to handle the location issue is to use dummy variables; however, given the large number of locations, such trick may result in bad prediction. Another idea is to extract some information about the locations and try to group the locations based on some properties. But since the location data was censored by Walmart for the competition purposes, we can't use such technique. What we did hence is to only use the sales of two stores and then build the models based on that.

1.3 Data Visualization

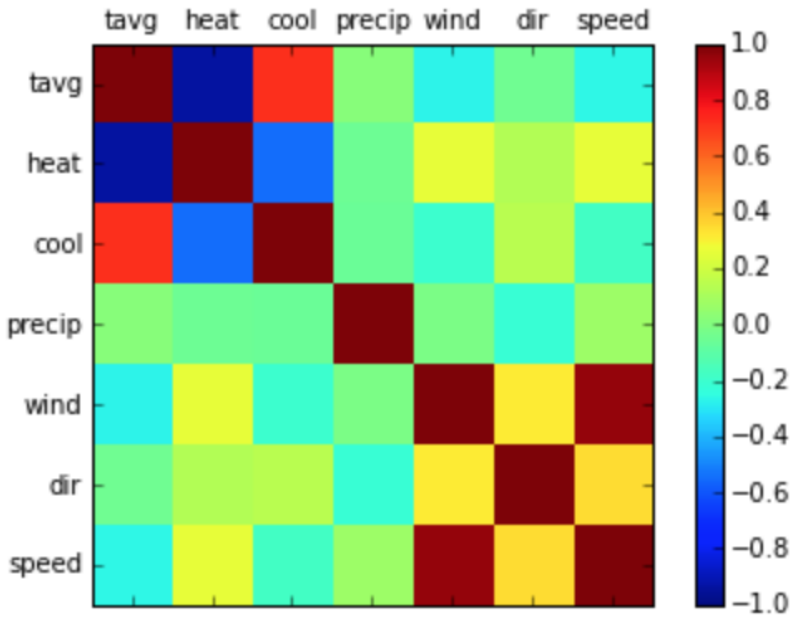


Figure 3: Correlation Visualization.

2 Models

2.1 Linear regression

To start off. we investigate the linear regression to get a better idea of how parameters affect the results on the dataset.

2.2 Lasso Model

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces [1].

lasso optimization problem can be described as:

$$\begin{aligned} \underset{w}{max} \quad & \sum_{i \in \mathcal{N}} (y_i - w^T x_i)^2 + \lambda \sum_{i \in \mathcal{N}} |w_i\| \\ & w_i \in \mathbb{R} \quad \forall \quad i \in \mathcal{N}. \end{aligned}$$

The intuition behind using Lasso is due to the following:

- It can provide greater prediction accuracy since we have a relatively small number of observations and a large number of predictors, then the variance of the OLS perimeter estimates will be higher. Nonetheless, Lasso Regression is useful because shrinking the regression coefficient can reduce variance without a substantial increase in bias.
- Second, Lasso Regression can increase model interpretability. With Lasso Regression, the regression coefficients for unimportant variables are reduced to zero which effectively removes them from the model and produces a simpler model that selects only the most important predictors.

3 Results and discussion

For the first model, we initially proceeded with linear regression without any regularization against the explanatory variables that we believed would contribute the most to the sales of these weather-sensitive items. We claimed that the average temperature and the amount of precipitation would have affected sales the most. After fitting the model to our data, we found an average MSE of about 70.0018.

To better visualize which of these weather variables had the biggest effect on store sales, we plotted each of our coefficients, see Figure 4.

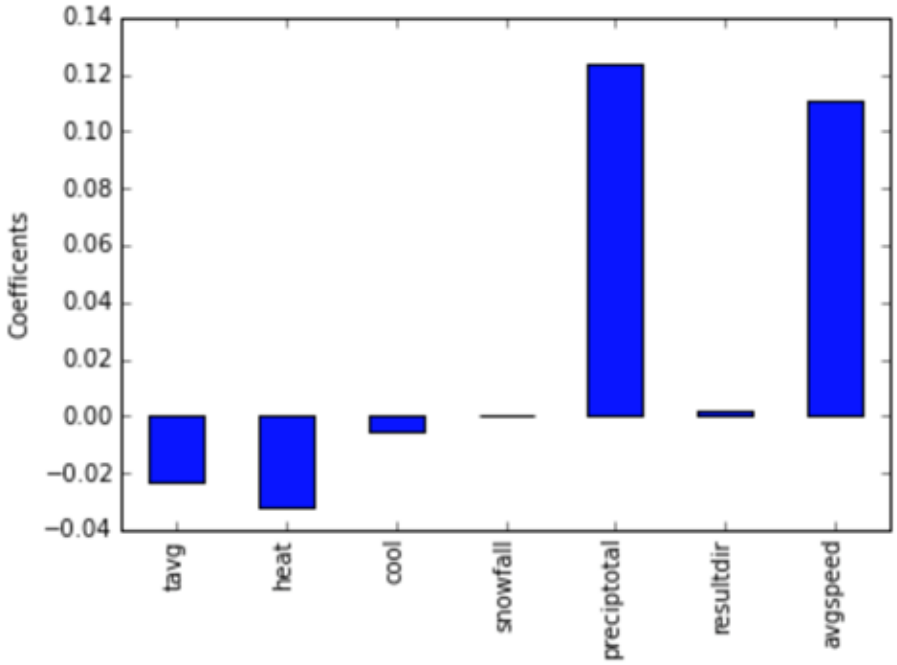


Figure 4: Coefficients of independent parameters.

From the plot, we can see that the values for preciptotal and avgspeed are much higher than the rest, indicating that the total sales for these stores are more driven by these two features.

Our second model involved using Lasso regression to help confirm the results we saw in plotting our coefficients. Fitting our Lasso model yielded an MSE of 43.0062, which shows an improvement over our previous model without the regularization.

By looking at the results and conclusinos provided by the people who participated in the competition at Kaggle, we found that they arrived at similar conclusion. Basically all the parameters are not very significant implying that people are going to buy the products regardless of the stormy weather.

We were hoping to get more insightful results, unfortunately the results we got at the end are not very encouraging. Nonetheless, we can provide some conclusions and recommendations that are discussed in the next section.

4 Model Limitations and recommendation

As stated in the dataset provided at Kaggle: "The sales data does not capture the difference between the stock and the demand. In other words, sales number 0 doesn't necessarily mean there was no demand for this product." Such missed information affects the prediction capabilities of our model and hence for more accurate estimations it would be better to observe the demand. We are aware of the fact that in real-world it is very challenging to get accurate estimate of loss of sales data. Hence, as a recommendation for the company, try to use customer choice modeling techniques to find a good estimate of the loss of sales to get better idea of how much loss of sales there is. Once customer choice models are developed to estimate the sales-loss, the model can use this data to provide better estimation of the sales.

Regarding the location data. The location data for the sake of keeping Walmart Sales confidential, location data where coded as a set of locations with no other data provided. Generally speaking, location data can be very important part of the model and can be used in a better way if available. Unfortunately, in our models we did not utilize the location factor for the reasons discussed earlier hence our results may not work if tested on different locations of Walmart stores.

Lastly, more data is needed to see if there are correlation between the products or if there is any kind of substitutional effects. For instance, if someone wanted to buy milk and milk was out of order, what other products the customer might have

looked at to substitute for the milk and hence more investigation is needed to obtain that data too to get better estimate.

References

[1] https://en.wikipedia.org/wiki/Lasso˙(statistics)