Kaggle Competition: Walmart Weekly Sales Forecasting

William West

April 30, 2014

1 Introduction

Kaggle is a website that hosts data mining competitions. Typically, companies will organize a competition on Kaggle by offering a sample dataset, a well-defined problem, and some incentive for competing, whether it be some monetary prize, internship, or job. Data Scientists can then compete with one another for the best performance, sometimes collaborating with one another and learning new things in the process.

The Walmart Weekly Sales Forecasting competition asks competitors to predict the weekly sales for a set of store/department pairs in 2013, given the weekly sales from 2010-2012. Competitors are not permitted to use any outside data source; only data provided through the competition may be used.

1.1 Dataset Description

The dataset consists of four data files: *features.csv*, *train.csv*, *test.csv*, and *stores.csv*.

- The *train.csv* and *test.csv* files contain the weekly sales for each (store, department, date) triple. Note that the test file contains null entries for the weekly sales column, as expected
- The *features.csv* file contains the temperature, fuel price, markdowns, CPI, unemployment rate, and a holiday indicator for each (store, department, date) triple
- The *stores.csv* file contains the size and type of each store

We combine all files into two distinct datasets–the training set and test set. Each set contains all features and the weekly sales for each (store, department, date) triple. We will assume that these files are combined for the remainder of this report.

1.2 Task

The task for this competition is to predict the weekly sales in 2013 for several stores/departments.

2 Background: Forecasting and Time Series Analysis

A time series is simply a group of measurements which are taken in a non-random order. For example, the set of data captured by a temperature sensor is most likely taken at regular intervals, whether these be every hour, minute, or second. We can assume that each measurement is related to all other measurements by some linear or (most often) non-linear function. Given a measurement of 77 °F, for example, the two surrounding measurements (taken before and after, respectively) have a high probability of being close to 77 °F. This is due to a characteristic of time series data called *seasonality*, which we discuss more below.

2.1 Seasonality

Seasonality is used to describe the way in which a time series goes through cycles as time goes on. For example, consider measurements of monthly water consumption in New Jersey over ten years. While in some states, water consumption may stay relatively constant year over year, New Jersey does not follow the same pattern. Since New Jersey experiences hot summer months, water consumption will be higher during those months, resulting in a time series that exhibits regular peaks during those months every year. Thus, a graph of the time series would show ten regularly-spaced peaks. This is considered the *seasonality* of the time series.

2.2 Trend

Using the same example of water consumption in New Jersey, consider how water consumption may increase or decrease *over time*. That is, over the entire 10 years, what is the *trend* of water consumption? Perhaps there are strict laws put into place 5 years into the time series that gradually limit the amount of water companies are permitted to use in a given year. In that case, we would see a gradual decline in water consumption over the last 5 years of the time series. This is considered the *trend* of the time series.

2.3 Forecasting: Pattern Recognition for Time Series

When dealing with time series data, we often want to

3 Data Exploration

3.1 Features

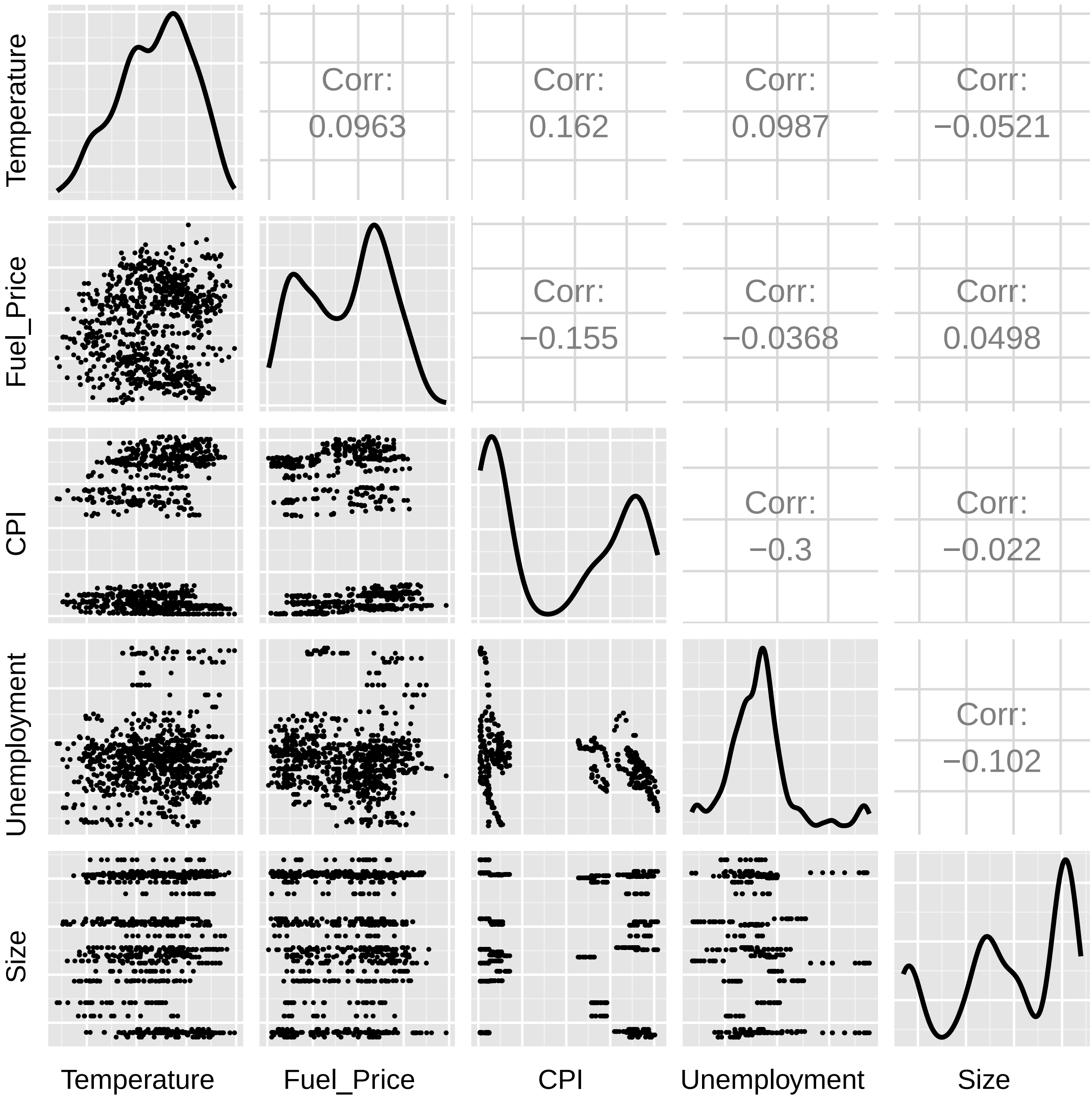

Figure 1: Scatterplot matrix: random sample of feature values. Upper boxes show correlation values between features, diagonal boxes show the distributions, and lower boxes are scatter plots.

3.2 Holidays

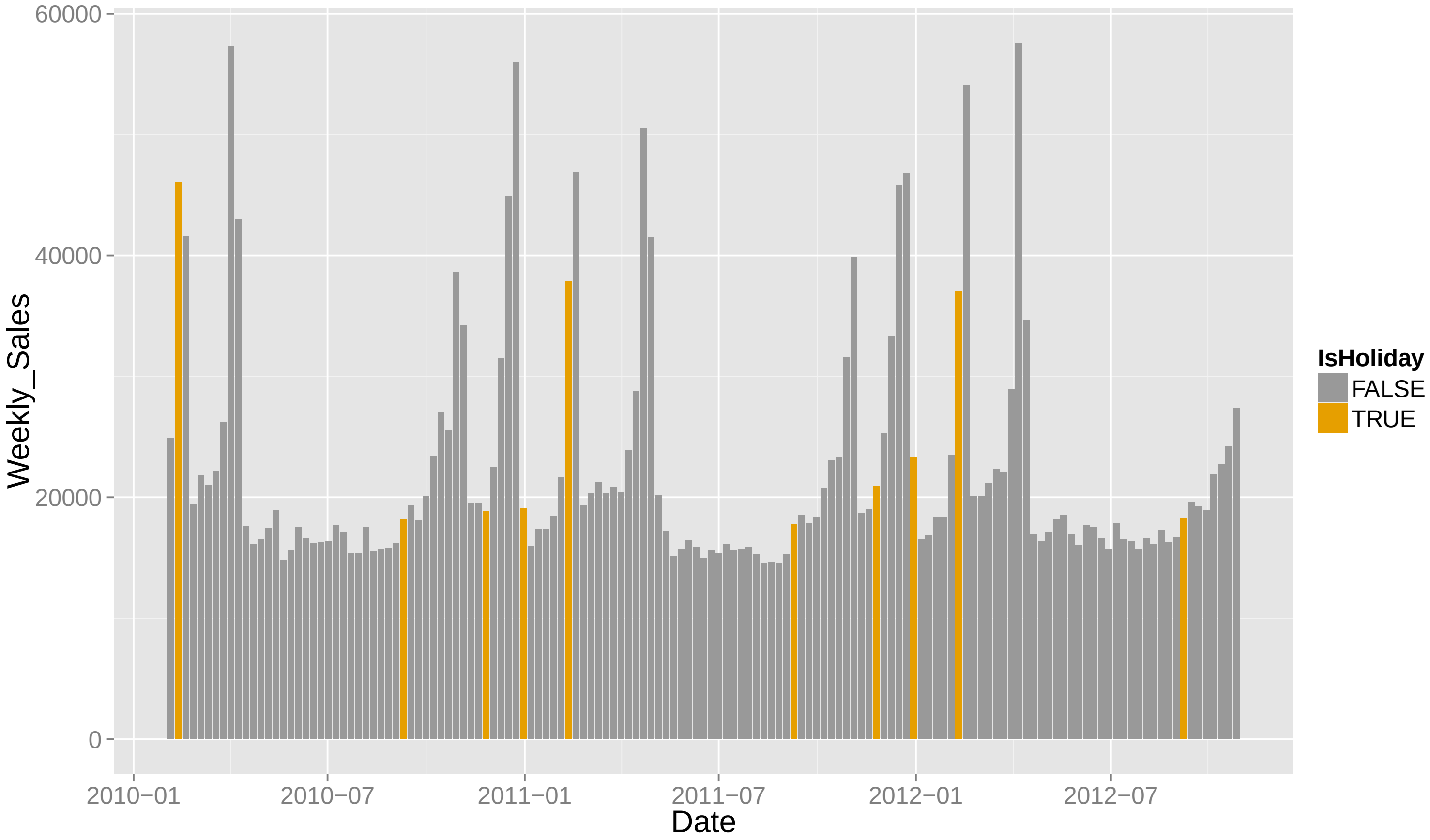

Figure 2: Weekly Sales of store_1, highlighting weeks marked as holidays

3.3 Markdowns

4 Feature Extraction

5 Modeling

6 Evaluation

7 Conclusion

References