Tarik Koric netid: koric1
Pranav Velamakanni netid: pranavv2

Introduction and Explanation of dataset

The goal of the project is to build a model that can predict the weekly sales of a Walmart store using variables like the department and weeks marked with holidays.

The dataset is provided by Walmart and contains data from 45 stores located across the United States. The following fields are provided in the dataset:

- Store - the Walmart store number
- Dept - the Walmart department number
- Date - time stamp of the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - boolean parameter providing information whether the week is a special holiday week

Pre-processing procedure

The dataset is provided as a CSV file. The first pre-processing was performed using R and it involved the following steps:

- Generate train_init.csv which includes all fields of the dataset mentioned above from 2010-02 to 2011-02.
- Generate test.csv which includes all fields of the dataset mentioned above except Weekly_Sales from 2011-03 to 2012-10.
- Splitting the data into 10 folds, where each fold contains data for 2 months starting from 2011-03 to 2012-10.

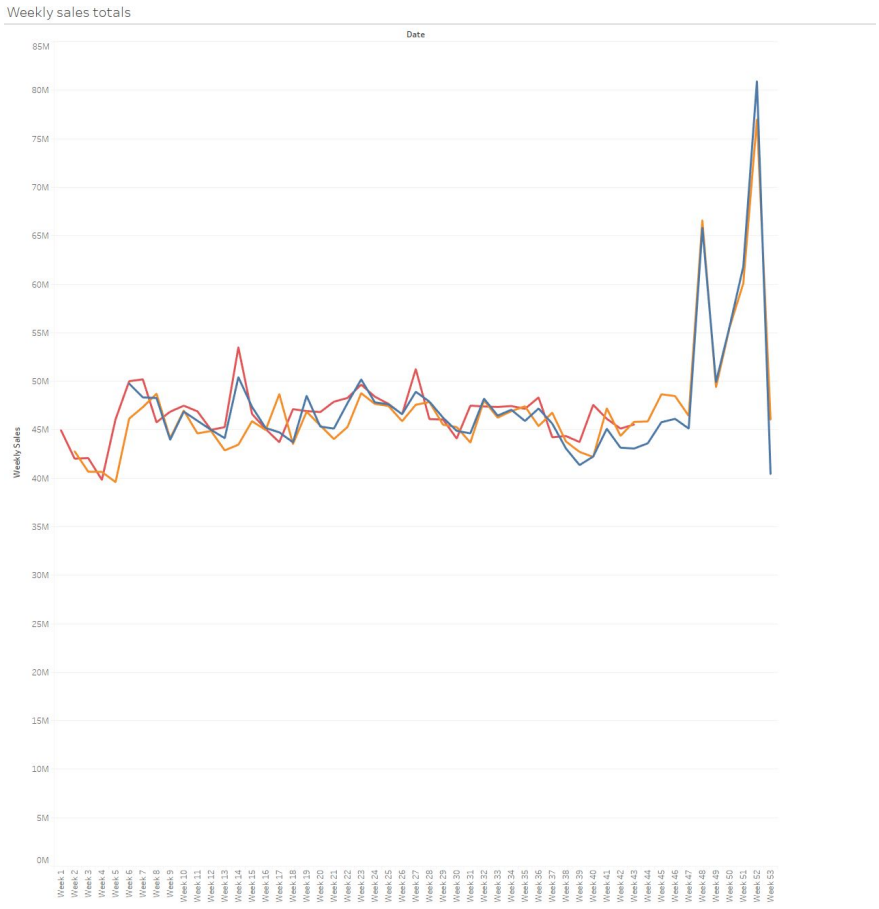
Further pre-processing steps were performed using Python and it involved the following steps:

- The training set (train_init.csv) is concatenated with data from each of the folds using an inner join
- A new column named month is created which contains an integer representing the month
- A new column named year is created which contains an integer representing the year
- A new column named week is created which contains an integer representing the week
- A new column named holiday is created which maps the boolean data from the IsHoliday column to 1 and 0

The above pre-processing steps were repeated for the test data as well. The Store, Dept, week, holiday and year columns are the fields that are used to predict the weekly sales.

Models and Results

Looking at the data we were able to determine what features to use in our calculation. The graph below shows all total sales of each year by week. The correlation is very high between the week number and the sale price for each.



Knowing this we selected the features that we wanted for our model along with the ones already given to us:

- Store number
- Department number
- Week number
- Holiday yes/no (1,0)
- Year

We first thought that a simple linear model would work to be under the given error. However this model was way off almost by a factor of 10 of the target error. We then used a Random Forest Regression with the default version in python with the number of estimators being 100 and setting the random state variable to 0 to be able to get the same results each time. Using this we got the results we wanted and are shown below.

Final Data

t	score
1	1748.25
2	1455.57
3	1371.18
4	1472.1
5	2652.64
6	1641.05
7	1695.94
8	1362.14
9	1296.25
10	1366.23
average	1606.13

Computer specifications
intel(R) Core(TM) i7-6500U CPU @2.5 GHz 2.6GHz 16.0 GB Ram

Program Run Time
947.18 seconds

Group Duties

Tarik:

- Random Forest Model
- Cleaning Data
- Model and Results
- Feature engineering

Pranav:

- Linear Model
- Cleaning data
- Introduction and Explanation of dataset
- Pre-processing procedure