PROJECT 2

stat4squirrels

Xiaoqian Zhang
Xinyi Liu
Shan He
Rachneet Kaur

1. Introduction

The Walmart data we use is downloaded from Kaggle. There are 45 stores and 81 departments in this data set. Our goal is to predict the weekly sales in each store under each department. First, we fill missing values in our data set and plot our data to find their patterns. Then, we applied three models including Seasonal Naïve Method, Product Method and ARIMA Model to do prediction and evaluate their performance using WMAE.

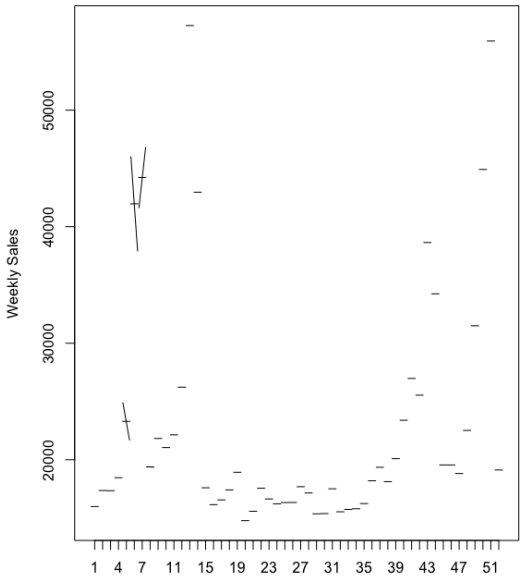
2. Preprocessing

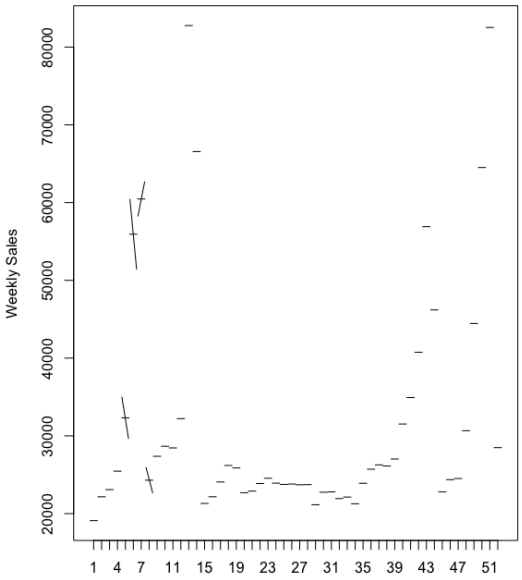
2.1 Formatting the Data

For the given training data set, we firstly need to preprocess it so that it will be convenient for us to make prediction. First, we separate the data by each department. For stores in the same department, we transform the data into a matrix. Certain matrix has the same number of rows as the number of unique dates in the trainset, while the columns are the date and data for each store in certain department. Thus, the number of such matrices is equal to the number of departments. And the matrices are stored in a list.

2.2 Description of the Data

Due to the large number of stores multiplied by departments, we only choose the first two stores in the first department as examples to see the seasonal fluctuation that is shown in the data. (Weekly Plot)





2.3 Filling Missing Values

For Seasonal Naïve and product method, we fill the missing values of each column by its column mean. For the columns whose values are all missing, we can fill the missing values with 0.

For ARIMA model, if the number of missing values in each column is more than 1/3

of the column, we fill the missing values with mean sale values in the last month of the store and the next store.

3. Model Selection

We treat the weekly sales in each store under each department as time series data and predict them separately. When predict each time series data, we try the following three models.

3.1 Seasonal Naïve Method

Firstly, we can apply a simple method called seasonal naïve method. For each observation in the test data set, we can predict the sale values as the sale values of the same week in the last year.

3.2 Product Method

Secondly, we can apply the product method. That is, for the observations in the same department, we can compute the mean of the sale values by store and by date. Also, we can compute the overall mean of the observations from the same department. Then the predicted value for a specific date d and store i in the same department j is mean_by_store_i * mean_by_date_d / overall_mean_j.

3.3 ARIMA

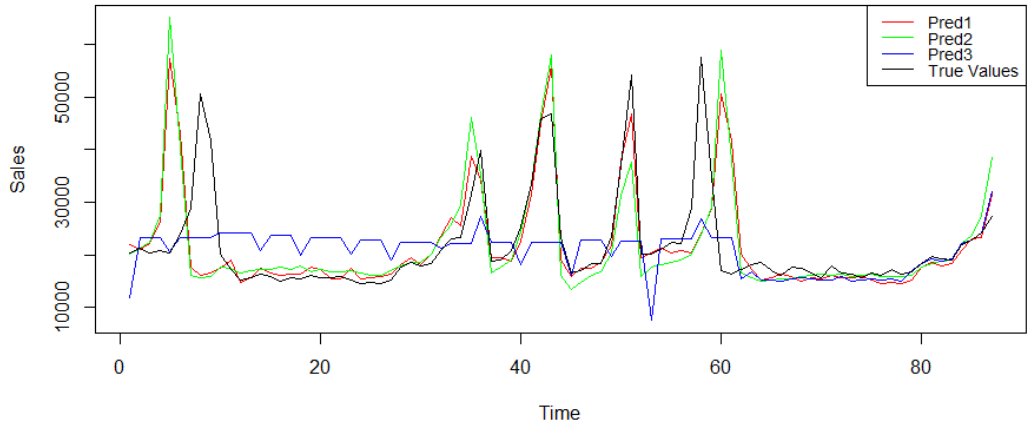
Thirdly, we apply ARIMA model. We use auto.arima function, in which we set BIC to automatically select parameters in ARIMA and apply seasonal test to test whether there is seasonal effect in our time series data.

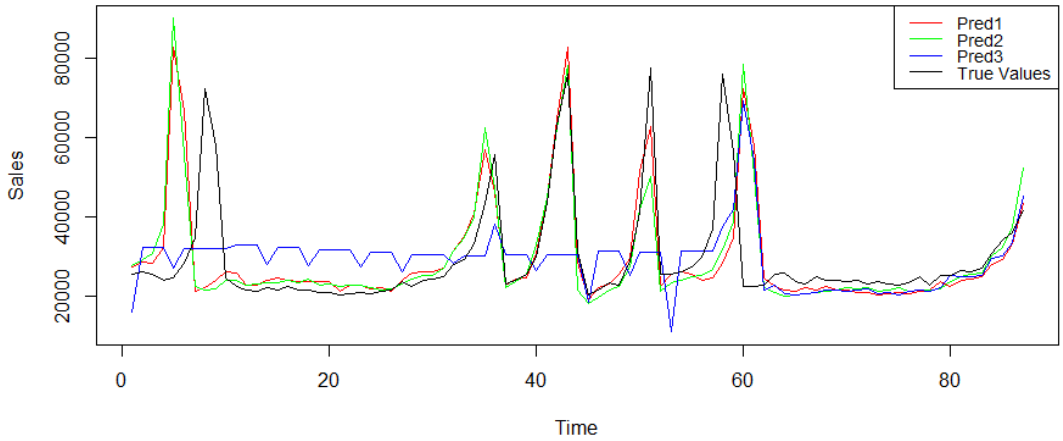
4. Result

t	Seasonal Naive	Product	ARIMA
1	1884	1729.4	1734.4
2	2543	2262.4	1759.3
3	1815.7	1608	1889
4	1816.1	1687	1909.5
5	1811.7	1680.1	1840.9
6	1787.0	1793.6	1926.5
7	1662.7	1667.0	2043.9
8	1838.4	1607.7	1729.5
9	2138.8	2425.9	3132
10	2628.2	2609.5	2693
11	1497.7	1430.2	2637
12	1789.6	1777.1	1930.1
13	1730.6	1607.5	2359.2
14	2529.2	2338.1	2683
15	1823.6	1673.2	2204.6

16	1715.7	1631.5	2252.5
17	1706.8	1635.6	2168.7
18	1782.3	1689.6	2217.3
19	1689.6	1592.9	2174.6
20	1659.6	1508.9	1905.3
Average	**1893**	**1798**	**2160**
Time	**11016.87s**		

We still take the first two stores in the first department as examples. Plot the true values as well as the predicted values from the three models as below.





From the results above, the product model has the best performance among all the models.