Walmart Sales Forecast

Marcus Tan

May 13, 2021

1 Pre-processing and data exploration

This project was implemented with the Python packages `Pandas`, `numpy`, `sklearn`, , `datetime`, `matplotlib`. A MacBook Pro with 2.2 GHz Quad-Core Intel Core i7 CPU and 16 GB 1600 MHz DDR3 memory was used to run the code.

The data consist of the weekly sales for stores labelled from 1 to 45 with each store having department labelled from 1 to 99. Some stores or departments may be not present at the earliest date of the data. The labels are no necessarily consecutive. No invalid value is found in the data. The years and weeks are extracted from the dates and used as predictors to exploit the seasonality of the data as shown in the weekly sales vs date in Fig. 1. It is observed that if a spike occurs for a department, it does so around the same week every year. Moreover, the same department of different stores peaks at around the same time of the year.

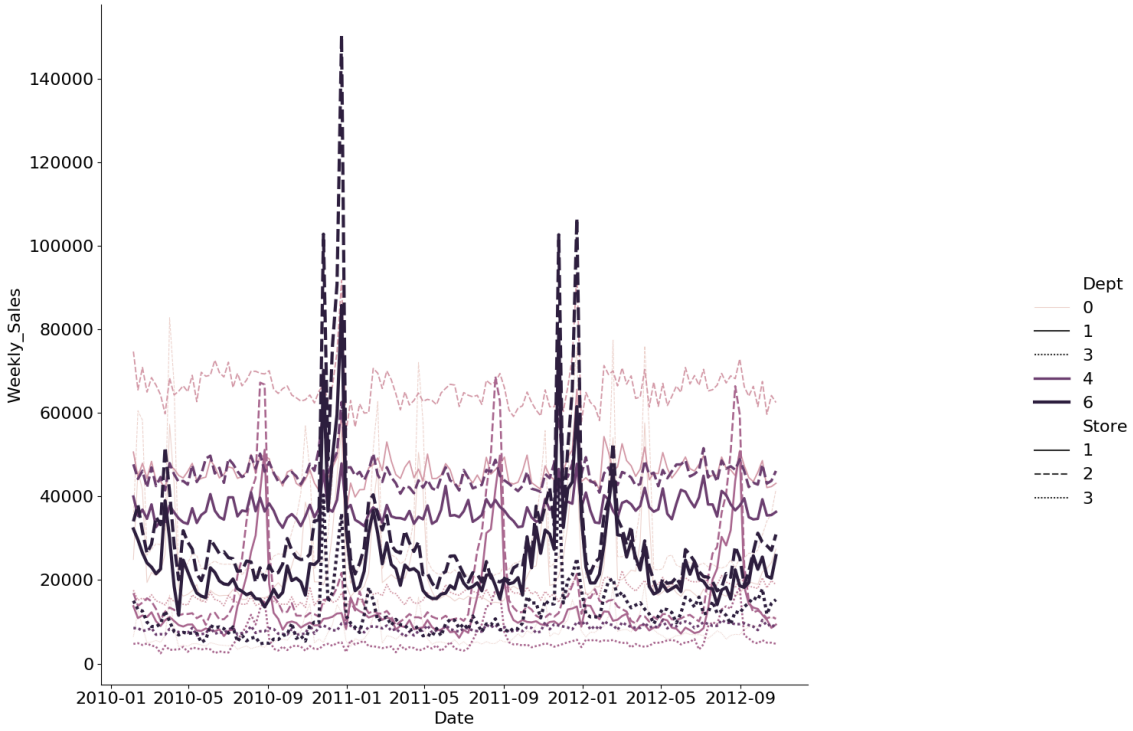


Figure 1: Weekly sales for a subset of stores and departments.

2 Models

Two models from `sklearn` were used in this work: (1) k-nearest neighbor (KNN) and (2) linear ridge regression. The KNN model uses two predictors — year and week — as continuous variables. For the ridge model, the week numbers are one-hot-encoded with `OneHotEncoder` from `sklearn`, which will ensure consistent encoding of the train and test data sets even when some week numbers are missing. Ridge regression is selected because (i) the

number of data points is often smaller than the number of predictors and (ii) encoded week columns for each department and store may contain only a single value (0) due to the lack of sales data for all weeks. If regular linear regression were used, very large coefficients can result due to the ill-conditioned matrix leading to predictions with large magnitudes.

Negative and invalid predictions are set to 0. Negative predictions can occur in regression while invalid predictions occur when no prior data exist for a store and department combination.

3 Results

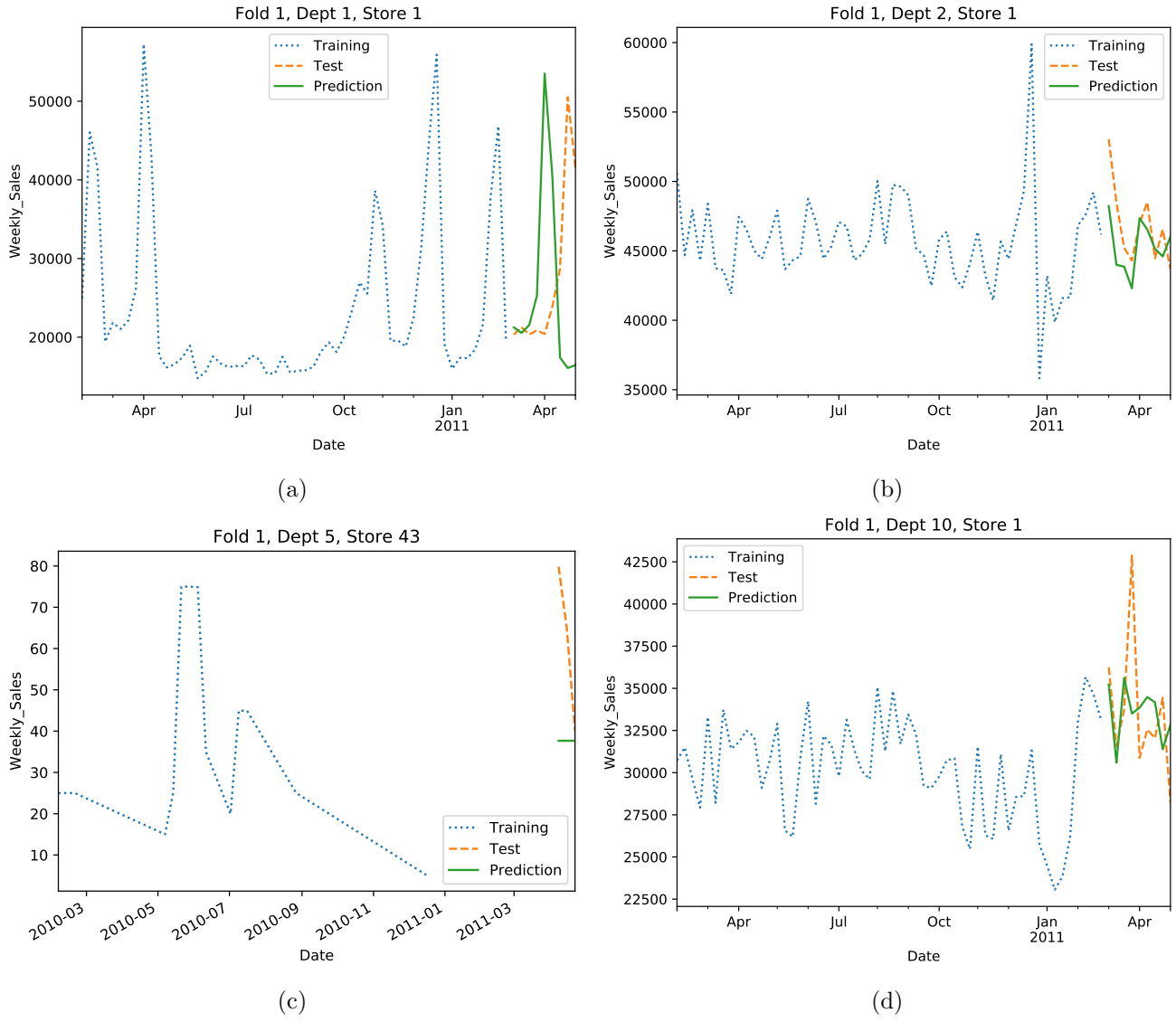
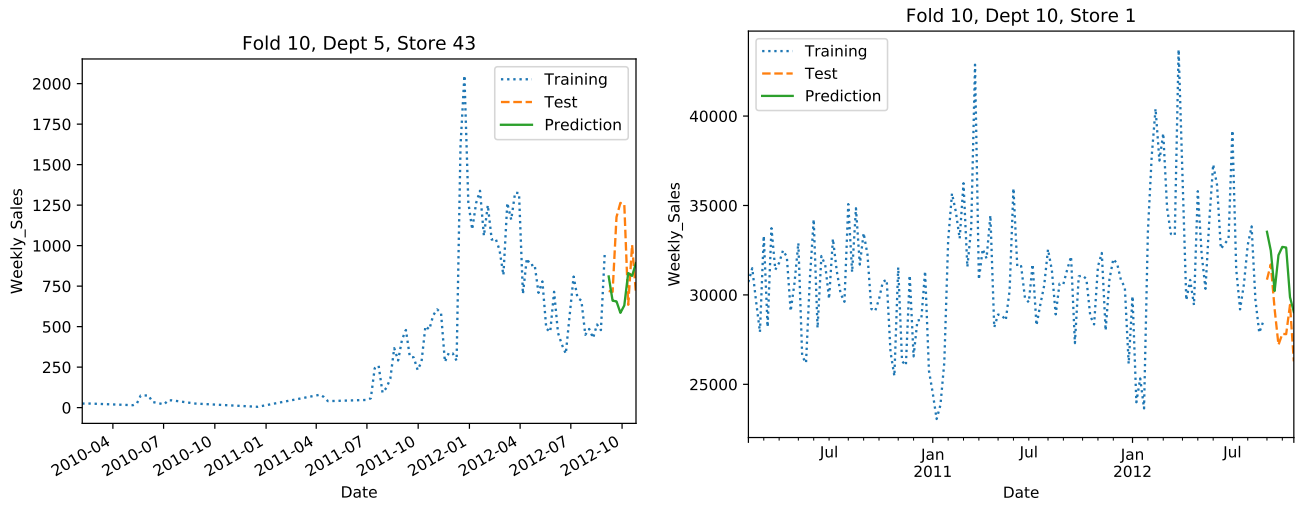


Figure 2: Ridge prediction for fold 1 (from 2011-03-01 to 2011-05-01) compared with actual sales (test data) using training data from earlier dates

Fig. 2 and 3 show the predicted sales obtained from the ridge model compared with the actual sales date. One can see that the ridge model can capture the spikes in sales well with some offset in the date where the spikes occur (see for example Fig. 2a). The offset is due the same holiday falling in slightly different weeks in different years. In the case of intermediate weeks with missing data, the ridge model is able to produce a reasonable prediction as apparent from Fig. 2. As a sanity check, the final prediction is also shown in Fig. 3.

Since KNN is similar to the method of using the sales of the same week in previous years to predict subsequent years, its weighted average errors (WAE) as shown in Table 1 is similar to that shown in Piazza Note 355 "Project 2: What we have tried (II)". There is a small improvement in accuracy when the nearest neighbors is increased



(a) (b)

Figure 3: Ridge prediction for fold 10 (from 2012-09-01 to 2012-11-01) compared with actual sales (test data) using training data from earlier dates

Table 1: Weighted average errors associated with the next-two-month sales predictions of the k-nearest neighbor model with 2 nearest neighbors and ridge linear regression model with $\alpha = 0.1$ fitted to the data from earlier dates

Fold	1	2	3	4	5	6	7	8	9	10	Mean
KNN	2251.0	1781.6	1778.3	1714.7	2397.9	1695.6	2086.0	1748.6	1719.0	1679.2	1885.2
Ridge	1825.7	1393.4	1399.6	1499.0	2226.9	1576.7	1649.9	1360.4	1390.7	1375.0	1569.7

to 2. The ridge model is significantly more accurate than the KNN model for all folds since it can exploit both short term trend over weeks and long term trend over years. $\alpha = 0.1$ is obtained after trying various α on the 5th fold test data. The computational time per fold is roughly 30s, where the cost comes from fitting the model to a large number of store and department combinations in the training data.

4 Conclusion

A KNN and ridge regression models were used to forecast future weekly sales by taking advantage of the seasonality of the data, which is built into the model by using two predictors – year and week. The ridge regression model with week taken as one-hot-encoded categorical variable and carefully selected α is by far the more accurate model.