The business problem is to predict the sales of all the stores and departments for the next **13 weeks**. Please go through the datasets as shared in the mail. The evaluation metric will be **MAPE at both store-level and store type level**. The final submission file has to be based on the following format:

|  |  |  |  |
| --- | --- | --- | --- |
| Store | Dept | Date | Weekly\_Sales |
| 1 | 1 | 8/3/2012 | 100 |
| 1 | 1 | 8/10/2012 | 1200 |
| 1 | 1 | 8/17/2012 | 340 |
| 1 | 1 | 8/24/2012 | 4500 |
| 1 | 1 | 8/31/2012 | 5670 |

**Data cleaning and analyses of features.**

- The data has no too much missing values. All columns was checked.

- I choose rows which has higher than 0 weekly sales. Minus values are 0.3% of data. So, I dropped them.

- There are 45 stores and 81 department in data. Departments are not same in all stores.

- Although department 72 has higher weekly sales values, on average department 92 is the best. It shows us, some departments has higher values as seasonal like Thanksgiving. It is consistant when we look at the top 5 sales in data, all of them belongs to 72th department at Thanksgiving holiday time.

- Although stores 10 and 35 have higher weekly sales values sometimes, in general average store 20 and store 4 are on the first and second rank. It means that some areas has higher seasonal sales.

- Stores has 3 types as A, B and C according to their sizes. Almost half of the stores are bigger than 150000 and categorized as A. According to type, sales of the stores are changing.

- As expected, holiday average sales are higher than normal dates.

- Christmas holiday introduces as the last days of the year. But people generally shop at 51th week. So, when we look at the total sales of holidays, Thankgiving has higher sales between them which was assigned by Walmart.

- Year 2010 has higher sales than 2011 and 2012. But, November and December sales are not in the data for 2012. Even without highest sale months, 2012 is not significantly less than 2010, so after adding last two months, it can be first.

- It is obviously seen that week 51 and 47 have higher values and 50-48 weeks follow them. Interestingly, 5th top sales belongs to 22th week of the year. This results show that Christmas, Thankgiving and Black Friday are very important than other weeks for sales and 5th important time is 22th week of the year and it is end of the May, when schools are closed. Most probably, people are preparing for holiday at the end of the May.

- January sales are significantly less than other months. This is the result of November and December high sales. After two high sales month, people prefer to pay less on January.

- CPI, temperature, and fuel price have no pattern on weekly sales. We can do feature engineering to see if there is any effect on the weekly sales.

**Feature engineering**

I removed the outliers and used the KNNImpute for the missing values.

KNNimputer is a scikit-learn class used to fill out or predict the missing values in a dataset. It is a more useful method which works on the basic approach of the KNN algorithm rather than the naive approach of filling all the values with mean or the median. In this approach, we specify a distance from the missing values which is also known as the K parameter. The missing value will be predicted in reference to the mean of the neighbours. It is implemented by the KNNimputer() method which contains the following arguments:

n\_neighbors: number of data points to include closer to the missing value. metric: the distance metric to be used for searching. values – {nan\_euclidean. callable} by default – nan\_euclidean weights: to determine on what basis should the neighboring values be treated values -{uniform , distance, callable} by default- uniform.

The Promos1 and Promos4 are higly correlated and Year and Fuel\_Price are highly correlated. We drop the columns Promo4 and Fuel\_Price.

**Machine learning models and predictions**

I used the sklearn.preprocessing StandardScaler for the features scaling. The machine learning models include KNeighborsRegressor, RandomForestRegressor, LinearRegression,Lasso,Ridge,ElasticNet, LassoCV ,RidgeCV , ElasticNetCV (for the cross-validations), DecisionTreeRegressor, GradientBoostingRegressor, ExtraTreesRegressor.

The randomforest regressors is the best model with max\_features auto with the MAPE 5.59%.

The departments are aggregated in the predictions data. The time series models are not working as auto-arima gave incorrect predictions and exponential smoothing is just the moving average of the past values and might not be too consistent.