PSL Project 2 Report

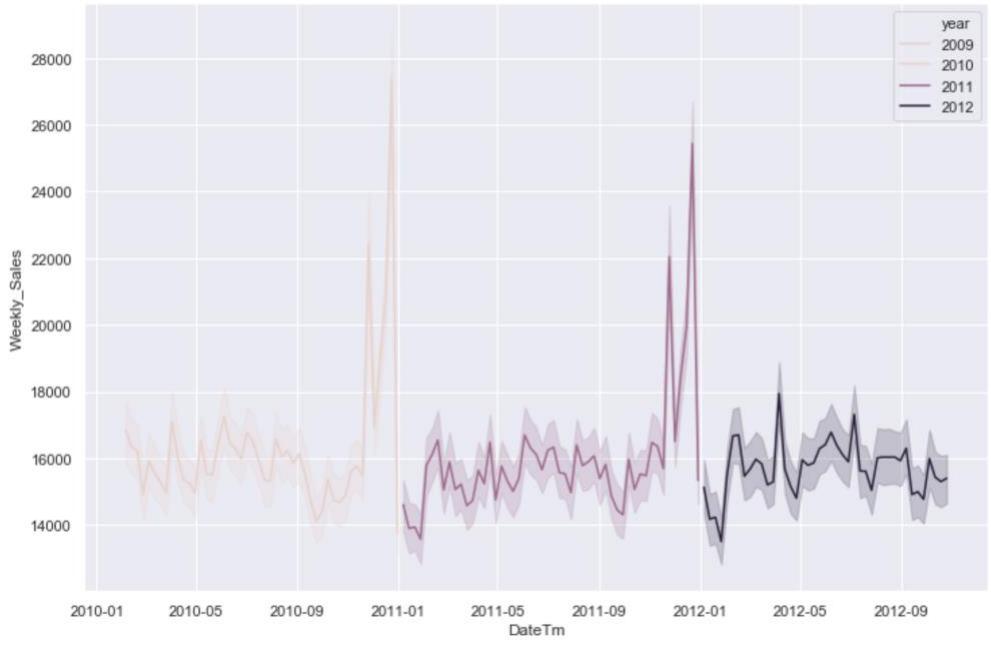
Sushanta Panda (net Id = panda5)

Goal

Goal of the project is to predict the “Weekly Sales” of various Stores and Departments in Walmart based on the history data.

Exploratory Data Analysis (EDA)

- It was observed that, both the year 2010 and 2011, the “weekly_sales” figure looks matching and there is a spike in the sales in the month of dec. However, this pattern is not observed in the year 2012, as the data for 2012 is till “Oct” (“nov” and “dec” data is missing)



***Note**: There is no data for the year = 2009, however the seaborn package shows the “2009” in the legend. Please ignore it. The year “2010” / “2011” and “2012” are only the available year and valid legends.*

Data Processing

As part of the data processing, I have done the following transformation / feature engineering.

- **Label Encoder for "IsHoliday"**– Label Encoded the "IsHoliday" field from "True" / "False" to "0" and "1". Since, it has only 2 categories, not used the pandas's "get_dummies" function to convert it into the column

- **Extraction of useful Year / Month / Week / Day Feature from "Date" Feature** – Extracted the following feature from the "Date" Feature

    - **Day** – Extracted the day from the "Date" . Though it's Weekly, but seems a usefull feature
    - **Week** – Extracted the "Week" number on which the "Weekly Sales" is reported.
    - **Month** – Extracted the "Month" number on which the "Weekly Sales" is reported.
    - **Year** – Extracted the "Year" number on which the "Weekly Sales" is reported.

- **Hyperparameter Tuning** – Used various parameters (**manually**) to see which has the lowest WAE

Fitting Prediction Model

Following are the 2 models created as part of the Project

1. **Linear Regression**
    a. Linear Regression seems not perform well and the WAE is high, where the mean of WAR is **1897**, which is the 2nd highest of the model in the list (Boosting & Random Forest). The reason could be, it's not a linear problem, the decision to predict "Weekly Sales" is kind of the tree structure. The mean WAE comes around

2. **Xgboost (Boosting Model)**
    a. Again, the Xgboost is worst perform and the mean WAE is "8175". Tried to change the various hyperparameter, however the overall WAE doesn't seems comes down

3. **Random Forest:**
    a. Random Forest seems perform well in the data the overall WAE down to "1564".
    b. Initially when explored with "Month" / "Day" / "Week", the WAE seems to be little bit high with *mean(WAE)* comes around ""
    c. However, when added the "Year", the *mean(WAE)* comes down to "1564"
    d. Below are the observations related to the "hyper parameter" tunning
        i. Higher the "max_dept" doesn't help to reduce the overall WAE, though have increase the number of iterations (n_estimators)
        ii. Also, it seems "max_dept" above 7 has the same results
        iii. Observed that , keeping the "max_depth" as blank, seems got good WAE
        iv. Keeping the number of estimators as "2000" is just fine to get a good WAE with blank "max_depth"

Observation / Conclusion

1. “Year” contributes to lower down the WAE by
2. Both Linear Regression and Boosting (Xgboost) seems not working well and Boosting seems the highest WAE with mean(WAE) is around “8175”
3. Random Forest seems perform better than all other model, where the mean(WAE) is below 1564
4. For all the models, it seems the “Fold 5” has highest WAE

Model, WAE, mean(WAE) and Run Time Table

RandomForest is the Best model

Fold	Linear Regression (LR)	Xgboost (Xgb)	Random Forest (RF)
1	2039	7625	1731
2	1685	8074	1424
3	1686	7970	1347
4	1679	7826	1393
5	2700	10473	2633
6	1751	8019	1648
7	2104	7801	1642
8	1849	8121	1342
9	1824	8008	1254
10	1648	7836	1227
Mean (WAE)	1897	8175	**1564**
Total Time Taken	**832** Seconds	**573** Seconds	**953** Seconds

Hardware Configuration

- iMac Pro (2017) 3.2 GHz Intel Xeon W, 32 GB 2666 MHz DDR4
- Jupyter Notebook 6.0.3, Python 3.7.8

References

- **Kaggle** - https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/code