Capstone Project

Data Scientist Nanodegree

Abdullah AlKanhal
Feb 19th, 2020

Definition

Project Overview

One of the main challenges that face all business all around the world is to predict their yearly demand, that will help the business to plan their stocks, workforce, and to budget based on that. As machine learning is developing, now we can utilize it in predicting sales of the stores, based on the data we have we can go as predicting by store per department or by item.

In this project, we analyze the sales data for Walmart all over the states and try to predict the sales per week per store per department for a sample of future dates. The project was inspired by a competition on Kaggle by Walmart. You can find it here.

Problem Statement

The goal is to predict the sales of different stores and departments per week; the tasks involved are the following:

1. Download and preprocess the competition data
2. Train multiple ML models
3. Find the best model.
4. Predict the sales of the specified stores

Metrics

This competition is evaluated on the weighted mean absolute error (WMAE):

$$WMAE = \frac{1}{\sum w_i} \sum_{i-1}^{n} w_i \, |y_i - \hat{y}_i|$$

where

n is the number of rows

$\hat{y}_i$ is the predicted sales

y_i is the actual sales

w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

In WMAE, the error is calculated as an average weight of the weighted absolute differences between the target values and the predictions.

What is important about this metric is that it penalizes huge errors that not as that badly as MSE does. Thus, it’s not that sensitive to outliers as a mean square error.

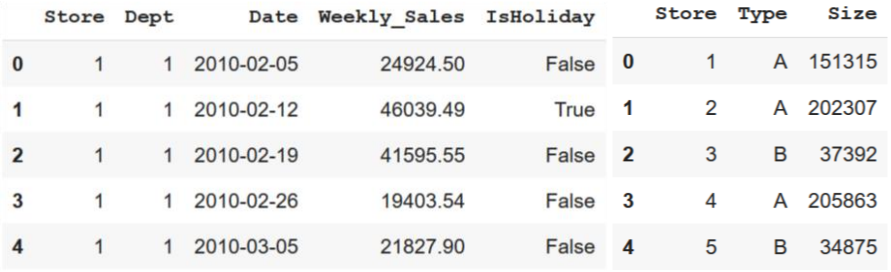
WMAE is widely used in finance, where $10 error is usually exactly two times worse than $5 error. Also, it takes consideration of the importance of one point to the other based on the weight injected by the user.

https://medium.com/@george.drakos62/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0

Analysis

Data Exploration

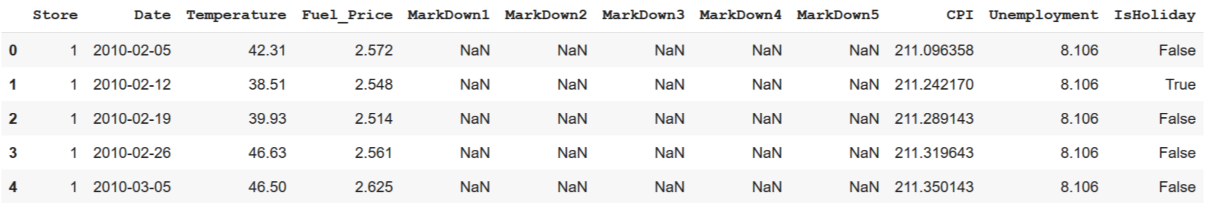
The data set provided in the competition consists of 5 files (Store Data, Training Data, Testing Data, and features of the weeks included in the analysis. The stores' data consists of info about the stores in terms of store type and size. The second file (Training Data) consists of five columns as follows, store number, dept number, week start date, weekly sales, is this week a holiday week. Also, the file has 421570 rows. The third file is a set of 115056 rows of the store, dept, week date, and wither it's a holiday or not that we need to predict the weekly sales of them. Finally, the features file, where we have extra info about the special factors like CPI, Temperature, unemployment, fule price, and markdowns. The features data set contains a high number of nulls in the markdowns, CPI, and unemployment columns.



	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

Figure 1: Training Data and Stores Data sets



	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

Figure 2: Features data set

Exploratory Visualization

The plot below shows the relations between our values (weekly sales) and the other main factor which is week number, and we see that sales have a certain habit during the year, they tend to be low during the begging of the year then they increase during Summer during schools holidays, and then they acclimate during the holidays season by the end of the year, we can see that effect clearer by plotting the sales vs holiday weeks as showen in the second graph.

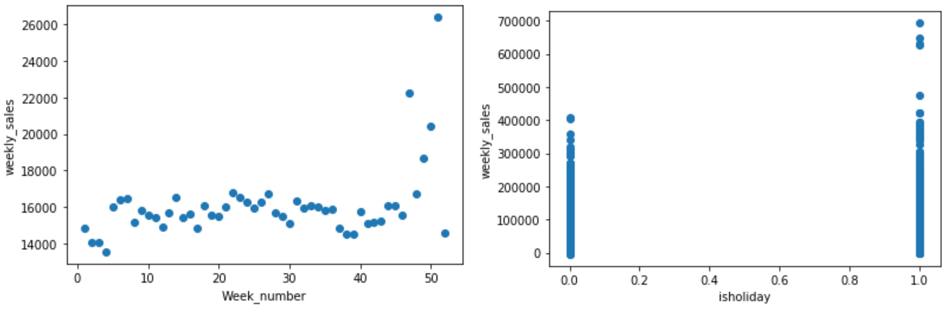


Figure 3: Weekly Sales Vs Week Number and Weekly Sales Vs Holiday Weeks

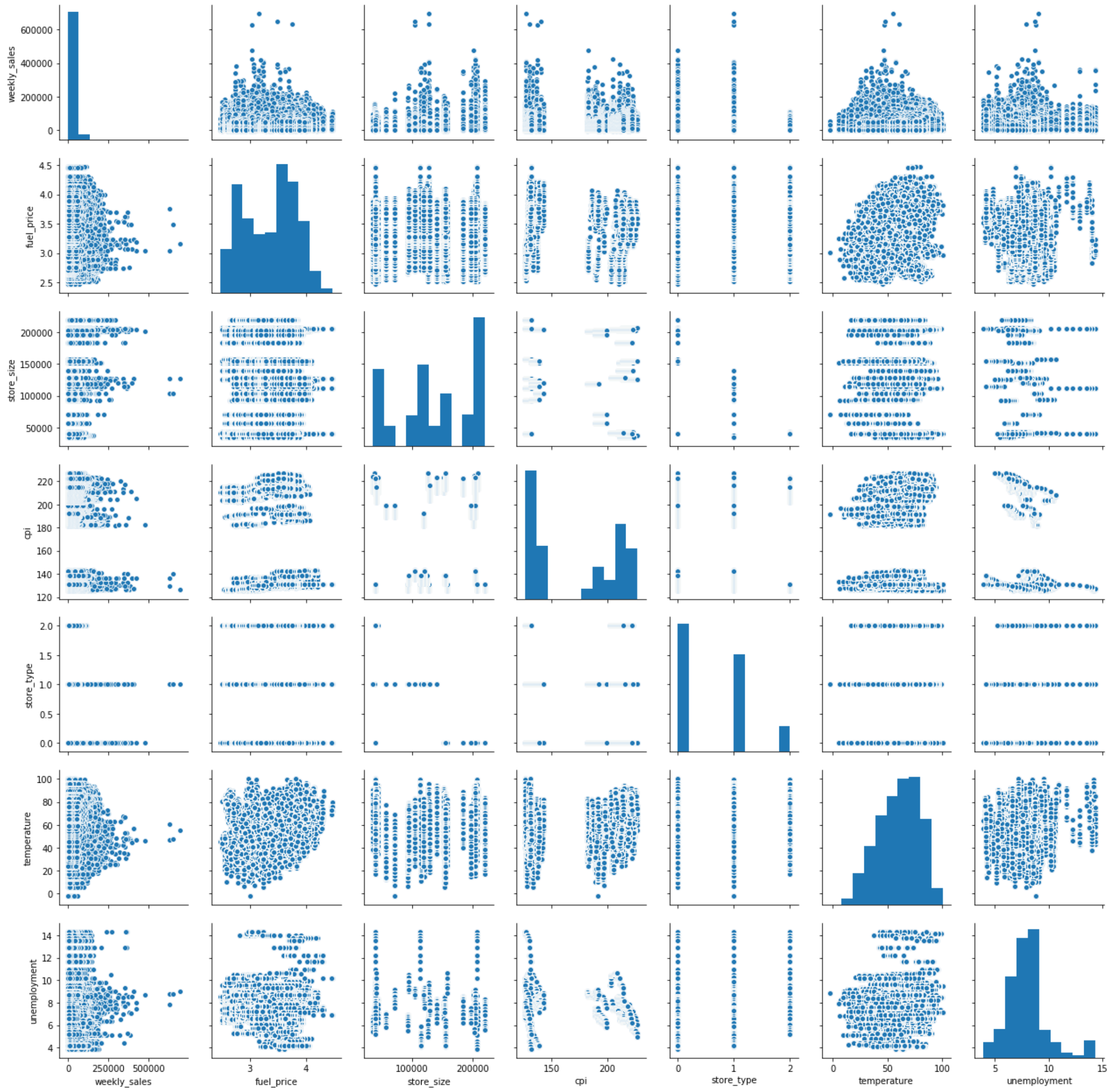


Figure 4: Features relations between eachother

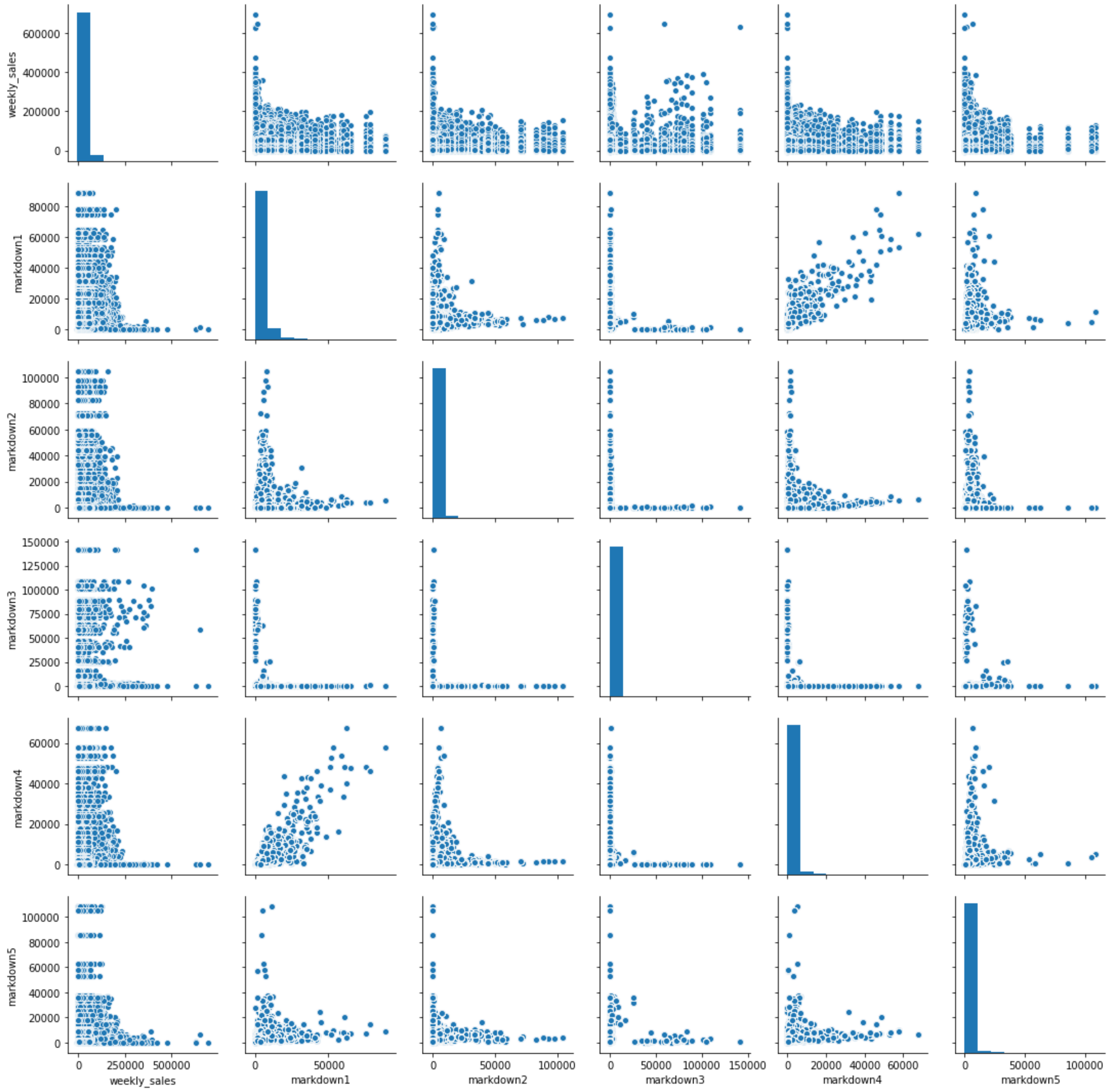


Figure 5: realtion between Markdowns and Sales

Algorithms and Techniques

In this problem we tried to test multiable models, then choose the model that has the lowest WMAE. The models considered are RandomForestRegressor, KNeighborsRegressor, MLPRegressor, and SGDRegressor. The models are chosen based on the following:

- RandomForestRegressor: the model has a proven record of giving a high accuracy over data, also it has the power to handle large data sets.
    - Main Training parameters:
        - **n_estimators** (number of trees)
        - **criterion** (maximum depth)
        - **Max_features** (The number of features to consider when looking for the best split)
        - **Verbose** (Controls the verbosity when fitting and predicting.)
- KNeighborsRegressor: K nearest neighbors is a simple algorithm that has been used in of 1970's as a non-parametric technique.
    - Main Training parameters:
        - **n_neighbors** (Number of neighbors to use by default for kneighbors queries)
        - **criterion** (maximum depth)
        - **min_samples_split (minimum number of samples in the node to split)**
- SGDRegressor: In simple regression, we try to minimize the error rate. While in SVR we try to fit the error within a certain threshold.
- MLPRegressor: we used this model as an example of neural network model example.
    - Main Training parameters:
        - **hidden_layer_sizes** (The ith element represents the number of neurons in the ith hidden layer.)

to help in the selection of the best parameters and to find the best model, 2 methodologies were used a grid search function a cross-validation model was used.

Methodology

Data Preprocessing

The preprocessing done in the “Prepare data” notebook consists of the following steps:

1. the different data files are joined together to link the sales with all the possible features.
2. Null data were replaced by the median in CPI and unemployment and replaced by zeros in markdown columns.
3. Categorical data were replaced by numbers.
4. The data were interpreted graphically to see if any of the features don’t correlate with each other or with weekly sales.
5. Markdown 2 and 5 were dropped as they seemed to have no relation with other data points, and seems that they were acting as noise.
6. Features and targets were separated.
7. Data then scaled using a standard scaler.
8. In the case of SGD_Regressor we need to preprocess the data more using Nystroem kernel.

Results

Model Evaluation and Validation

During development, a validation set was used to evaluate the model.
The final architecture and hyperparameters were chosen because they performed the best among the tried combinations.
based on the grid search conducted and the K-Fold model selection methodologies,

the best model found to be is SDGRegressor

with the following parameters, max_iter=1000.

Conclusion

Future prediction models can be very tricky because the future can be changed dramatically and totally new factors can be introduced, for example, the data here does not take in factor the online shopping effect on store sales. In problems like this, we always try to keep very sure to not overfit the data and to review the models with updated data to get a better picture.

Improvement

To achieve the optimal model, using more capable hardware and trying more models and parameters can be done to get a better estimate while ensuring not to overfit the data.
Different styles of cleaning can be experimented with in imputation.