Donia Zaheri (DoniaZ2)
Abishek Samuel (asamuel4)
Praveen Bhushan (bhushan6)

PROJECT 2

Walmart Store Sales Forecasting

Introduction and Goal

For this project, we are provided with the historical sales data from 45 Walmart stores, located across different regions. Each store could have multiple departments and the project goal is to predict future weekly sales of these departments, based on the historical data. The prediction model will be evaluated based on the Weighted Mean Absolute Error (WMAE/WAE) and its value should be less than the specified target of 1630. Three models were implemented for these purposes: A Naïve approach, Seasonal Naïve and a Time Series Linear Model (TSLM).

Preprocessing the Data

The data file train.csv consists of 421,570 observations from February 2010 to October 2012. Each row represents a weekly sales record for a department in the Walmart store & contains the following attributes:

- Store - Unique Store Identification Number
- Dept - Unique Department Identification Number
- Date - Date specifying the last day of the week (Friday)
- Weekly_Sales - Sales of the week
- IsHoliday - Boolean to specify a holiday in the week

Pre-processing 1

The attribute 'Date' is converted into a 'datetime' data type from strings in the data set. We then extracted year, month, week and day information from the 'Date' attribute to be able to plot figures and better visualize the data.

Pre-processing 2

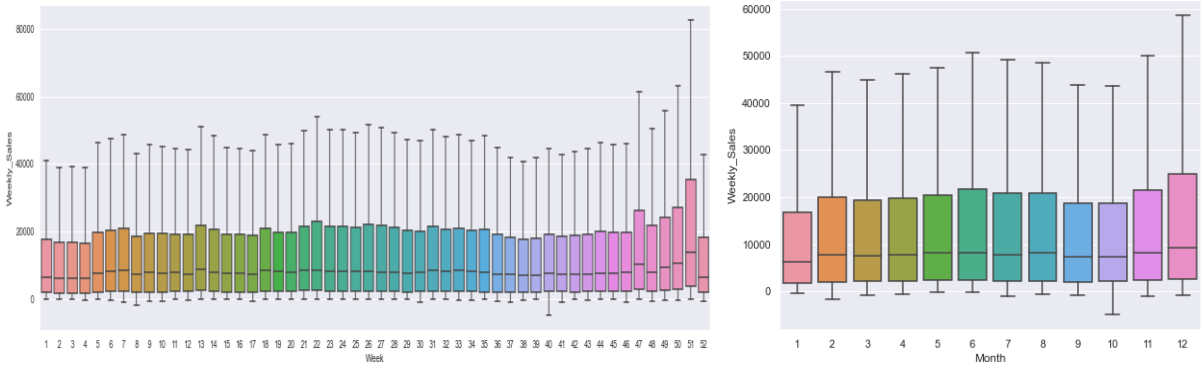
The attribute 'Date' is converted into a 'datetime' data type from strings in the data set. We then extracted year, month, week and day information from the 'Date' attribute to be able to plot figures and better visualize the data.

Pre-processing 3

Any missing sales records were assigned '0' value in the training data. Unique test dates, stores & departments within each store were computed for prediction. Test data of stores that weren't part of the original training sample, were removed before predictions were performed.

Data Analysis

The seasonality can be seen by plotting the weekly sales data.



The plot on the right illustrates the months of June, November and December of having the highest weekly sales. These months contain the main U.S holiday shopping seasons as well (Christmas and Black Friday/Thanksgiving).

Sales Prediction Models

This dataset is a time series dataset. So, we tried three different time series models. Naïve, Seasonal Naïve and Time-Series linear model.

Naïve Model

This was the first model we implemented. This model as the name suggests, makes a naïve assumption by predicting the 'Weekly Sales' for the test set length (two months in our case) based on the selected number of the last data points in the training set (two last points in our case). In this approach we went over each store in each department in each fold. We used only 'Weekly_Sale' (target feature) as our 1-D training set given to the model. The number of observations that we want the model to use to make the prediction were passed to the model with 'n_lags'. Naïve is a simple model; it worked very well on the training data, but not well enough on the test data. Individual WAEs and the mean WAE for the naïve model is provided in the performance section.

Seasonal Naïve Model

In the second approach, we improve the previous model by considering the seasonality characteristic of the data. This model is the same as the naïve model, with the only exception that it takes one more parameter called 'period' that considers the period of the data. For our dataset the period is 52 weeks (one year). Passing 52 as the period to the seasonal naïve model, will make the model predict the 'Weekly_Sales' based on the same week in the previous year. This makes sense as we expect similar sales trends around the same weeks' year-to-year (e.g. holidays). So, we gave the model the same input as the simple naïve model and specified the period as 52 (number of weeks in a year). Using Seasonal Naïve, as seen in the table below, improved our test results significantly.

Time Series Linear Model

TSLM predicts using linear regression and by adding trend and seasonality factors to the time series. The year is added for trend and the weeks are implemented as categorical features. Performing Single Value Decomposition

(SVD) on the training data resulted in a much lower WAE. The WAE value of ~1616 was within the benchmark (1630) by applying the above steps. No further data pre-processing was necessary.

Performance

Fold #	TSLM (with SVD)	TSLM (sans SVD)	Seasonal Naïve	Naïve
1	1969.15	2042.40	1891.48	1898.38
2	1379.03	1440.11	1975.60	2138.69
3	1397.96	1434.68	1527.35	1628.04
4	1550.52	1596.98	1945.29	1610.92
5	2308.69	2327.53	2221.03	2837.49
6	1637.73	1673.30	3922.89	3969.96
7	1689.53	1719.39	1767.19	1838.19
8	1391.87	1421.05	1524.71	1561.47
9	1413.10	1431.87	1467.54	1491.01
10	1425.51	1447.09	1463.79	1536.46
Mean (WMAE)	1616.31	1653.44	1970.69	2051.06
Runtime (s)	367.54	347.39	307.21	269.95

Conclusion

The project provided profound insights into the time series models. We preprocessed the data and made predictions using the 3 models - Naïve, Seasonal Naïve & TSLM. The linear regression model is adequate for the time-series data with small modifications. TLSM (with SVD) worked the best of all the 3 models and the corresponding WAE value was within the stated benchmark (1630). Simple noise reduction with SVD improved the prediction accuracy without adding too much to the computational costs involved.

System and Software Packages

System Specs: Operating System - Windows 10 (64 Bit), Processor - Intel i7 @ 2.60Hz, RAM - 32 GB

Software Packages: Python: 3.8.3; Scipy: 1.5.0; scikit-learn: 0.32.1; numpy: 1.18.5; pandas: 1.0.5; statsmodels.tsa: 0.12.1

Individual Contributions

Donia Zaheri (DoniaZ2) - Worked on data wrangling, Naïve model, Seasonal Naïve model and the two modeling parts of the project report.

Abishek Samuel (asamuel4) - Worked on Data pre-processing, and the two TSLM model implementations.

Praveen Bhushan (bhushan6) - Worked on the project report, data visualization and data pre-processing.

Reference

[1]Original Project - https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[2] Certain coding and model implementation practices were provided from the UIUC STAT 542 course by Prof. Feng Liang and the course's TA's.