RUTGERS
UNIVERSITY | NEW BRUNSWICK

Department of Statistics
&
Data Science

Regression and Time Series - Fall 2022

Time Series Forecasting for Business

Anirudh Gaur - 216007356
Manisha Damera - 220002345
Neha Thonta - 219009543
Rohit Macherla - 219008045
Vivek Reddy Chithari- 220002664

supervised by
Prof. Koulik Khamaru

Abstract

Forecasting techniques have wide applications in various domains such as sales, banking, stock-market etc. These forecasting methods vary based on the specific condition of the time series. The time-series dataset has valuable time-related information for prediction and statistical analysis. In our work, we focus on predicting sales for a specific business, i.e. Walmart - a retail corporation. Sales forecasts help businesses make better business decisions. Our present work identifies the trends in the data, and we perform EDA (Exploratory Data Analysis) to find correlations and visualize the data to understand the model's working better.

The proposed work focuses on the Time Series Forecasting model for weekly sales of Walmart, with which we forecast future sales. The SARIMA model is used to attain improved performance regarding prediction accuracy on future sales. The proposed work has examined a few forecasting models and tests, such as The Seasonal Autoregressive Integrated Moving Average (SARIMA) model and the Adfuller test. Further, we analyze the trend and seasonal components of the data and use ACF and PACF plots to identify the best SARIMA model and forecast the sales for the near future. From the forecasting models, it is concluded that the SARIMA fits the best for the future.

Keywords

Time Series Forecasting; Adfuller Test; Seasonality; SARIMA; ACF, PACF

Contents

1 Introduction **4**

2 Problem Formulation **4**

3 Data Set Source **4**

4 Methodology **5**

4.1 Exploration and Findings 5

4.2 Time Series Forecasting 7

4.3 Fitting Seasonal ARIMA 8

4.4 Future Prediction 10

5 Conclusion **10**

1 Introduction

Businesses are always looking for ways to increase profitability. One of the ways to do this is through Business Forecasting using Time Series Forecasting. In Business Forecasting, we aim to forecast future sales using historical Time Series data generated by the business. Sales are considered time series data. We investigate a dataset including historical sales in a large retail store. Hence we perform sales predictions for the Walmart - retail store. We aim to perform reliable forecasts that contribute to efficient planning. As Walmart is one of the eminent supermarket stores, its development is of utmost interest to us. Some major factors have impacts on future sales. These factors can be identified by analyzing the sales patterns of total sales of a retail store. It is worth mentioning how the time period impacts sales differently. We perform data analysis to identify trends and seasonality in the sales pattern that causes complexity in the forecasting process.

Therefore, the main objective of our work is to present a model for sales forecasts which is highly accurate and, at the same time, easily explicable. We perform Exploratory Data Analysis (EDA) on data to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. In this work, a dataset is investigated which has information on retail store sales from February 2010 to October 2012, i.e. a public dataset with the occurrence of seasonal fluctuations is chosen. Walmart sales data suits our requirement to perform Time Series Forecasting. For the data analysis, we target the effect on holidays, i.e. Labor Day, Christmas, Super Bowl and Thanksgiving. We investigate other features of our data set and the impact of those in sales. The most important part of our study is exploring the time series model for weekly sales and predicting future sales. We apply classical time-series forecasting techniques such as SARIMA and adopt the Adfuller test method to help with the stationary version of data. Finally, we test the model by fitting SARIMA, which predicts future sales accurately.

2 Problem Formulation

Sales of a business have been shown to affect the company's profits, and it is becoming a concern to predict the sales and consider taking business decisions accordingly. A prominent retail corporation, Walmart has various features and seasonality trends affecting the business's sales. Thus, our priority is working on modelling data to increase profits by studying sales data. Addressing the issue of predicting sales data and building an accurate fitting model is an excellent job.

Our objective is to work on analyzing past data and trends in sales and train the model using Time Series - SARIMA, and explore other advanced methods to improve accuracy. Accomplishing the most effective time series models to forecast business is our exposition.

3 Data Set Source

As our proposed project is based on business forecasting, sales play a crucial role in business decisions. Thus we obtain a sales dataset from the Kaggle platform.
Walmart is a renowned retail corporation that operates a chain of hypermarkets. Here, we have chosen Walmart data combining 45 stores, including store information and monthly sales. The data is provided on a weekly basis. Walmart tries to find the impact of holidays on the store's sales. It has included four holiday weeks into the dataset: Christmas, Thanksgiving, Super Bowl, and Labor Day. The Walmart sales dataset has date and size as features.

4 Methodology

This report discusses the time series methods for business sales forecasting. We start the data collection process, data collected from an open source retail store sales data, i.e. Walmart. Then we perform data processing, i.e. data preparation, to easily process it for forecasting. We study the forecasting, build models, and perform visualization in python, with essential packages and libraries imported to conduct analysis.

Data Preparation and cleaning: We initially have 'Stores', 'Train' and 'Features data'. Our cleaned final data is the result of combining these based on stores and dates. Data cleaning involves dropping duplicate columns and negative sales values and adding each Holiday boolean column separately. Our date column has continuous values. Since our data has continuity, we can't split it randomly. Thus we split the data manually according to the 70%-30% training testing split which is the thumb rule.

4.1 Exploration and Findings

We perform EDA to study the data for in-depth analysis. We find that there are 45 stores and 81 departments in the data. It shows us that some departments have higher values as seasonal, like Thanksgiving. It is consistent when we look at the top 5 sales in data. All of them belong to the 72nd department during the Thanksgiving holiday time. From the graphs Fig1, it is clear that Labor Day and Christmas do not increase weekly average sales. However, there is a positive effect on sales in the Super Bowl, but the highest difference is observed during Thanksgiving. Furthermore, for all holidays, the Type, A store has the highest sales, and the reason for this is that the size of store A is large and hence justifies the huge sales shown in Fig2. Stores have three types, A, B and C, according to their sizes. Almost half of the stores are bigger than 150000 and categorized as A. As expected, holiday average sales are higher than on regular dates.

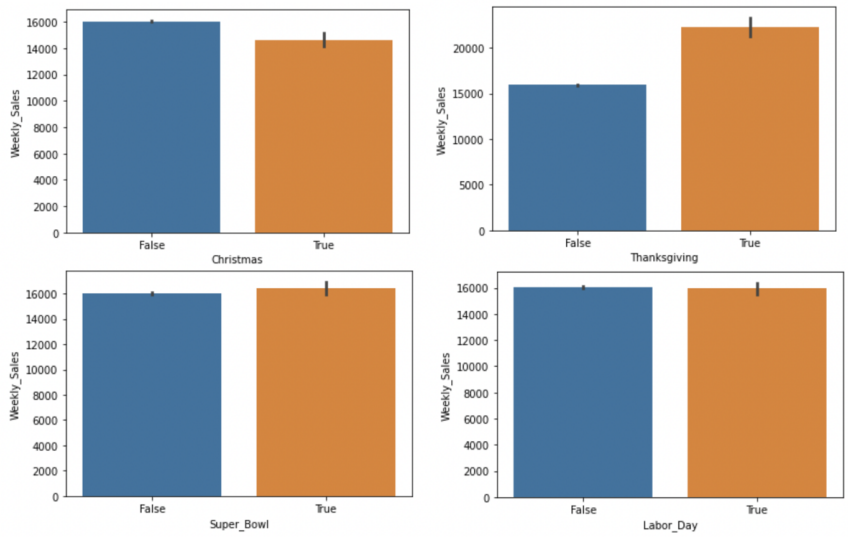

Figure 1: Effect of Holidays Christmas, Thanksgiving, SuperBowl and Labor Day on Weekly Sales

From the graphs in Fig 3, it is seen that 2011 had lower sales than 2010 generally. When we look at the mean sales, it is seen that 2010 has higher values. The year 2010 has higher sales than 2011 and 2012. Nevertheless, November and December sales are outside the data for 2012. Even without the highest sales months, 2012 is not significantly less than 2010.

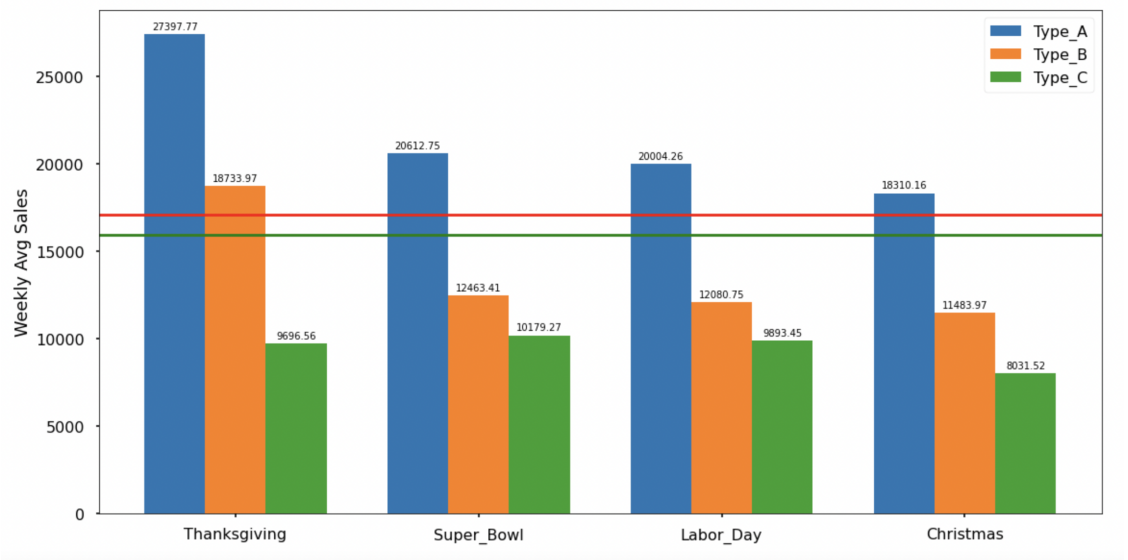

Figure 2: Type of stores effecting the weekly sales on different holildays

Interestingly, the fifth top sales are in the 22nd week of the year. These results show that Christmas, Thanksgiving and Black Friday are more important than other sales weeks. January sales are significantly less than in other months. This is the result of November and December high sales. After two high sales months, people prefer to pay less in January. From the correlation matrix Fig4, we observe that CPI, temperature and unemployment rate have no pattern on weekly sales, i.e. almost no correlation.

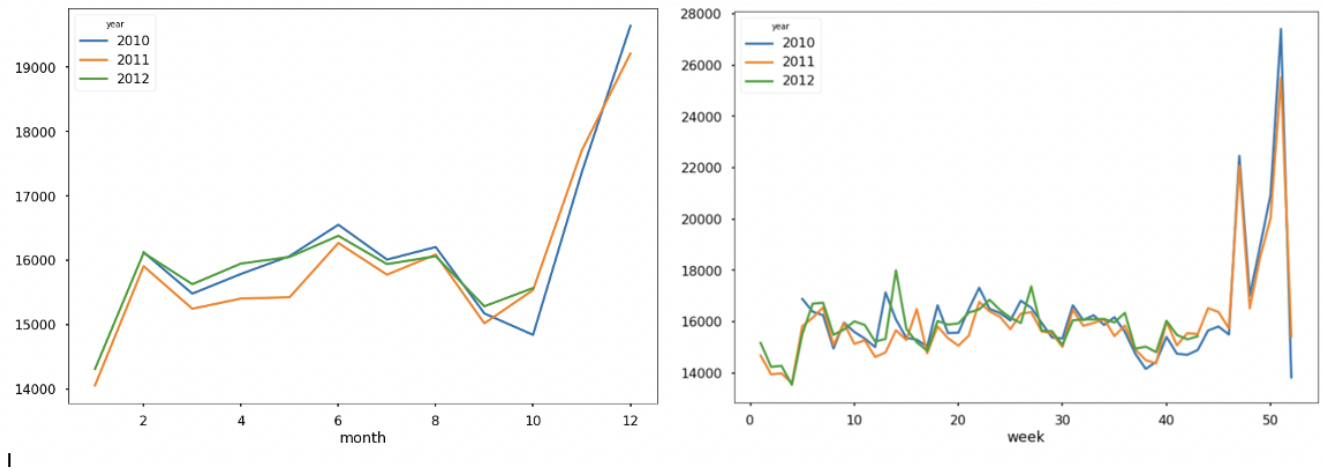

Figure 3: Weekly and Monthly sales of Walmart store in three different years

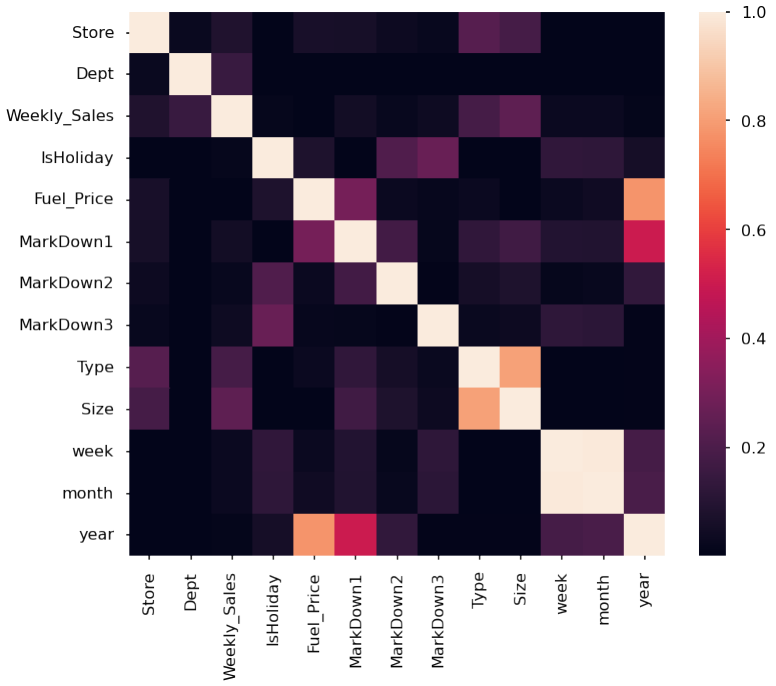

Figure 4: Correlation Matrix: Knowing important features effecting sales

4.2 Time Series Forecasting

There are several different approaches to time series modelling. Traditional statistical models, including moving average, exponential smoothing, and ARIMA, are linear in those predictions of future values that are constrained to be linear functions of past observations. As the data has seasonality (Fig.7), we use Seasonal ARIMA. In this section, we focus on the basic principles and modelling process of the SARIMA. We fit the SARIMA model based on trained data. We see that rather than including date-wise data, we consider weekly data, which gives a pattern. Decomposition of the weekly data finds that seasonality converges to the beginning point. From the plot, we can see the yearly seasonality of increased sales during thanksgiving, which was demonstrated earlier. The above weekly plot is a good time series process. However, the mean of the time series is not constant, i.e. shown in Fig5, and it can be seen that there is a yearly seasonal trend. As the mean is not constant, we take the first difference to get it to a constant mean. Then we split the data into training and testing sets (Fig.6).

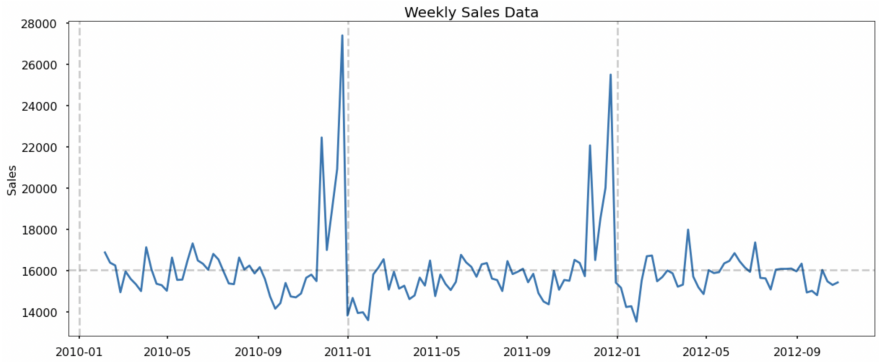

Figure 5: Time Series plot showing weekly sales data

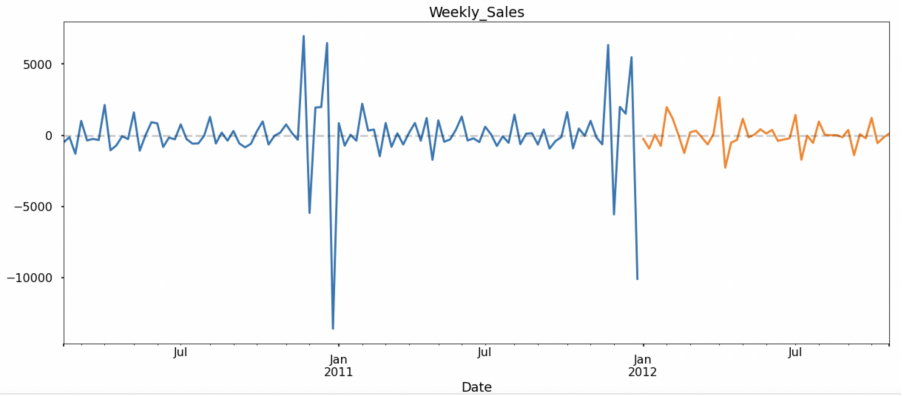

Figure 6: Time series plot with first difference with Train and Test split

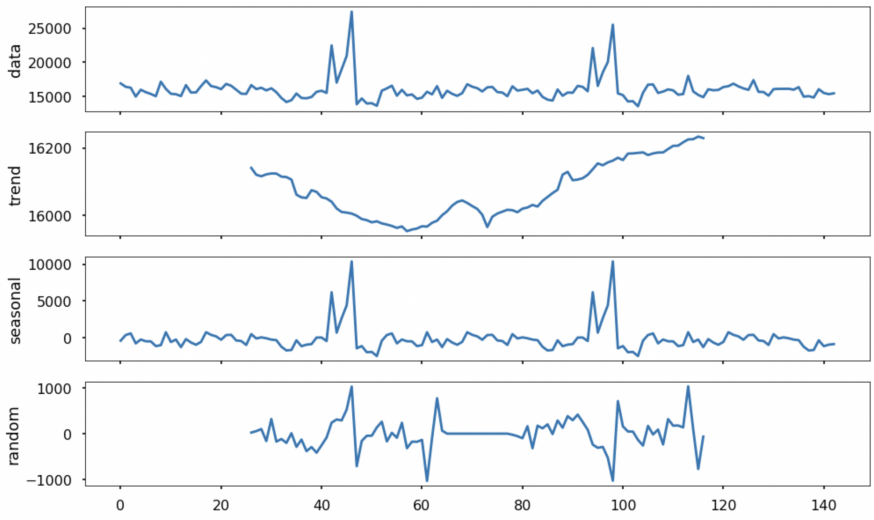

Figure 7: Decomposition graph to check trends

In ARIMA time series forecasting, the first step is determining the number of differences required to make the series stationary. Augmented Dickey-Fuller Test (ADF Test) performs a test hypothesis. It is from the test statistic and the p-value that you can infer whether a given series is stationary or not.

4.3 Fitting Seasonal ARIMA

Box initially presented the seasonal time series ARIMA (SARIMA) model–Jenkins [1], with the advantage of forecasting accurately for a short period. The SARIMA model formulation includes four steps:

i. Identification of the SARIMA(p; d; q)(P; D; Q)s with s periodicity structure: use autocorrelation function (ACF) and partial autocorrelation function (PACF) to develop the rough function.

ii. Estimation of the unknown parameters.

iii. Goodness of fit tests on the estimated residuals.

iv. Forecast future outcomes based on the known data

The difference data has some minus and zero values, so we use additive seasonal. Seasonal periods are chosen from the decomposed graphs above. Tuning the model with iterations takes too much time, so the model for different parameters is tested, and the best parameters are fitted to the model. We first adjust the residuals for our sales data to follow a constant zero mean. We fit SARIMAX(1, 1, 1)x(1, 0, 0, 52)) for the 52 weeks. From the ACF and PACF plots (Fig.9a), we find the orders of the ACF plot's Autoregressive and Moving Average components. From ACF plot, it can be seen that there is MA process and from the PACF plot, it has some AR process. The strongest lag is observed at lag 52 (as it is yearly seasonal). Hence we can fit a seasonal ARIMA model with 52 weeks as the frequency. We initially fit with seasonal order(1,0,1,52), not taking the difference, and realized that MA process for the seasonal component has more p-value (Fig.8) indicating that MA is not suitable. Hence from the AR and MA behaviour, we conclude that the SARIMA (1,1,1)x(1,0,0,52) is the best fit.

SARIMAX Results

Dep. Variable:	Weekly_Sales		**No. Observations:**	99	
Model:	SARIMAX(1, 1, 1)x(1, 0, [], 52)		**Log Likelihood**	-845.825	
Date:	Mon, 12 Dec 2022		**AIC**	1699.650	
Time:	05:33:17		**BIC**	1709.990	
Sample:	02-12-2010		**HQIC**	1703.833	
	- 12-30-2011				
	coef	**std err**	**z**	**P>|z|**	**[0.025 0.975]**
ar.L1	-0.4009	0.053	-7.564	0.000	-0.505 -0.297
ma.L1	-1.0000	0.626	-1.598	0.110	-2.226 0.226
ar.S.L52	0.8916	0.103	8.656	0.000	0.690 1.093
ma.S.L52	0.3882	1.112	0.349	0.727	-1.791 2.567

SARIMAX Results

Dep. Variable:	Weekly_Sales		**No. Observations:**	99	
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 52)		**Log Likelihood**	-845.832	
Date:	Mon, 12 Dec 2022		**AIC**	1701.664	
Time:	05:14:50		**BIC**	1714.589	
Sample:	02-12-2010		**HQIC**	1706.892	
	- 12-30-2011				
	coef	**std err**	**z**	**P>|z|**	**[0.025 0.975]**
ar.L1	-0.4017	0.053	-7.584	0.000	-0.505 -0.298
ma.L1	-1.0000	0.129	-7.772	0.000	-1.252 -0.748
ar.S.L52	0.9348	0.009	103.345	0.000	0.917 0.953

Figure 8: Summary tables for SARIMA model with different orders

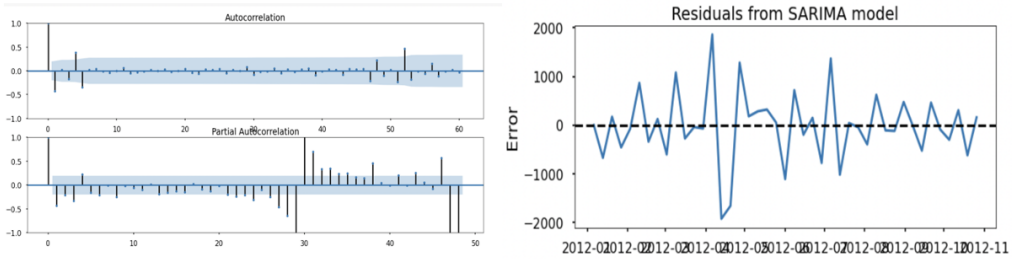

Figure 9: a) ACF and PACF plots b) Residual plot for SARIMA model

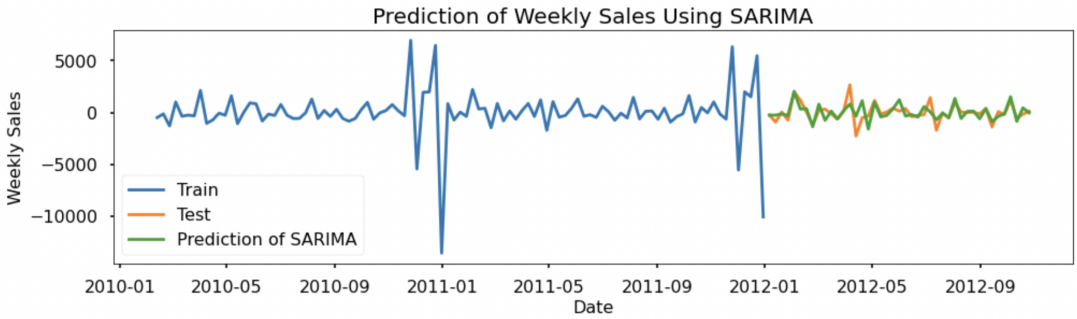

Figure 10: Fitting SARIMA model to predict weekly sales

4.4 Future Prediction

We use the seasonal ARIMA model to forecast sales which we aim to do. For this, the future dates, i.e. next 24 weeks' data, are generated and predict the sales values by applying SARIMA. Fig 12 shows the predictions, and we find that the seasonal trend of having more sales around thanksgiving week is very much evident, and the model performed well.

```
my_order = (1,1,1)
my_seasonal_order = (1,0,0,52)
second_model = SARIMAX(df_week_diff, order=my_order, seasonal_order=my_seasonal_order)
second_model = second_model.fit()
```

Figure 11: SARIMA model fit for the future data

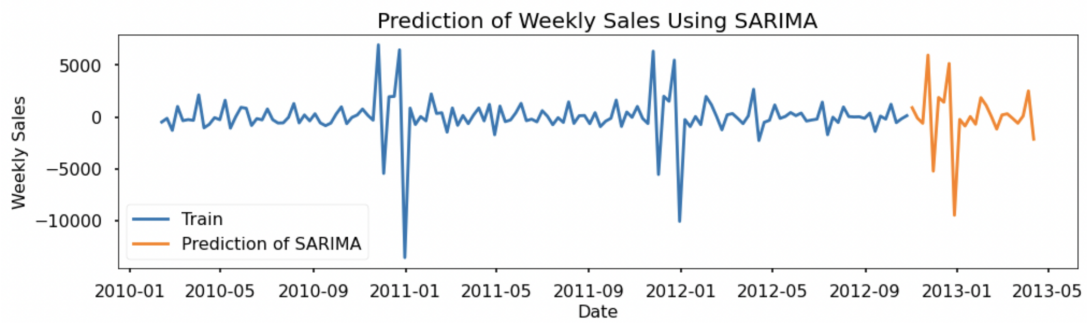

Figure 12: Future prediction of weekly sales using Seasonal ARIMA

5 Conclusion

Nowadays, sale forecasting is an inseparable part of every industry, especially businesses that work with seasonal items. In our study, we considered different model approaches for time series forecasting. Sales forecasts help make informed decisions about everything from staffing and inventory to new product lines and potential marketing efforts, which are crucial for business. Using time series methods for sales data helps make valuable predictions. The accuracy of the test set is an essential indicator for choosing the best-fitting model. The seasonal ARIMA we selected to fit make sales predictions based on historical data for specific sales time series.

The results of the model indicated the superiority of SARIMA in forecasting time series data, i.e. residual plot for the SARIMA model does not have any pattern and has a mean zero indicating good predictions (Fig 9b). The predictions graph shows that the predicted values closely follow the original (test) values, indicating that the model we fitted is a good model(Fig 10). When we implement this model, we get a precise graph following a similar trend. We find that the prediction capability of the The SARIMA model is encouraged to work on time series data. Thus, using seasonal ARIMA and more advanced time series models to estimate the future accurately can help businesses boost their work.

References

[1] F.-M. Tseng and G.-H. Tzeng, "A fuzzy seasonal arimamodel for forecasting," *ELSEVIER*, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165011401000471

[2] D. J. Dalrymple, "Sales forecasting practices: Results from a united states survey," *ELSEVIER*, vol. 3, pp. 379–391, 1975. [Online]. Available: https://doi.org/10.1016/0169-2070(87)90031-8

[3] R. F. P. H. F. By Philip Hans Franses, in *Time Series Models for Business and Economic Forecasting*, Cambridge, pp. 251–260. [Online]. Available: https://books.google.com

[4] N. P. . Patimaporn Udom1, a, "A comparison study between time series model and arima model for sales forecasting of distributor in plastic industry," IOSR Journal of Engineering (IOSRJEN).

[5] M. Navratil and A. Kolkova, "Decomposition and forecasting time series in business economy using prophet forecasting model," 2019. [Online]. Available: http://cebr.vse.cz/pdfs/cbr/2019/04/02.pdf