Forecasting sales and number of people (Traffic) in-store using ARIMA Model

Feziwe M Shongwe
melvin.shongwe@gmail.com

Abstract

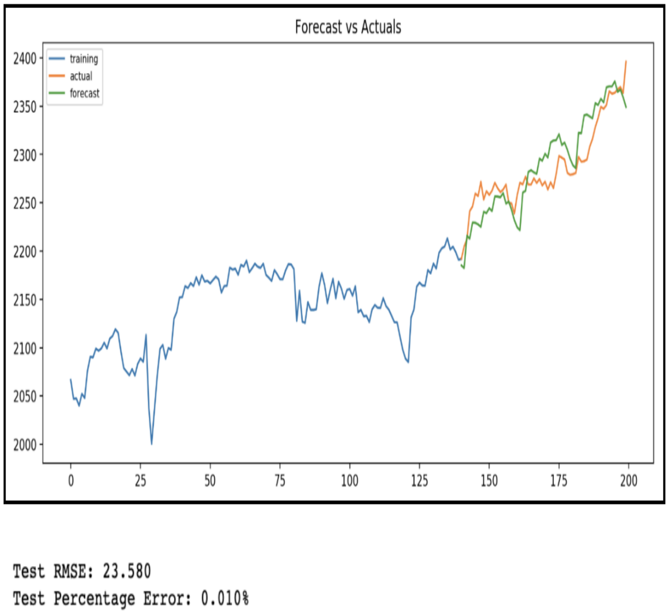
Forecasting total sales and customer numbers in-store assists in estimating a store's profits and losses for a given period. Estimating manually may be difficult because some days may be different due to weather conditions and other factors we may consider. An Auto-Regressive Integrated Moving Average (ARIMA) model is used in this paper to forecast future sales (in-store sales in cents) and traffic (number of people in-store) because its main purpose is to predict future values based on past values.

I. Introduction

[5] discusses sales forecast as the foundation of the financial story that you are creating for your business. Once you have your sales forecast complete, you'll be able to easily create your profit and loss statement, cash flow statement, and balance sheet. Similarly, to the number of people in the store, if we can predict how many people were in the store for a specific hour and how much money was made in sales, we could also be able to compute the rate of people purchasing in the store this way.

On the other hand, counting the number of people in the store could also assist in optimizing costs by the management team deciding how many workers must occupy the shop for specific days/hours. [6] discusses the benefits of counting the number of people in the store, which include assisting your business in planning ahead of time, understanding factors that impact your business, improving energy efficiency, and so on.

In this project, I will be implementing a model to forecast sales and traffic in-store during the store's operating hours. When forecasting any time-based quantity, using previous data to estimate the future value is useful. The Auto-Regressive Integrated Moving Average (ARIMA) model is used in this paper because it is a time-based statistical model for forecasting future events using previous data such as averages and other information. Forecasting models have been shown in studies to be useful in a variety of industries, including the stock market, manufacturing, and retail. Figure 1 depicts an example of an ARIMA model forecasting the price of the S&P 500 from June 2016 to March 2017. [2] discovered that models used to forecast manufacturing are also used to forecast retail sales.



```
Test RMSE: 23.580
Test Percentage Error: 0.010%
```

Fig. 1. [1] ARIMA model implemented for forecasting S$P500 price

II. Model building Phases

In this section, I will discuss the methods used in this project when developing the ARIMA model. These methods are divided into four phases that we followed in order to develop an optimal method for forecasting sales and traffic of a store. These phases are as follows: data preparation and exploration, explanatory variable reduction, model building and selection, and model validation.

A. Data Preparation and Exploration

Conversion of intervals

The datasets' interval is 15 minutes, and the goal is to forecast sales and traffic per hour for the following month. I converted the timestamps to 60-minute/hour intervals before cleaning the datasets to make analysis easier, as the main goal of this exercise is to forecast the store's sales and traffic for the next month.

Based on the results obtained (before converting the intervals), the Initial datasets didn't appear to have missing values; however, after converting the intervals, some hours have nulls after resampling the data using an hour interval because the store may not have sales for an hour at times or no customer entered for traffic counting, and other null values represent when the store was closed.

Handling missing values

The dataset is a time series, we considered the business hours of the United States, and adding a day of the week column in the data will help to identify the times when the store was closed and when the store was not making sales, because removing all the null values implies that the store does not experience hours where there are no sales, which is likely false. Removing the hours when the store is closed may appear to be removing information, but domain knowledge also aids in data exploration. Furthermore, the missing values during operating hours can be treated as a case of *Missing Not at Random (MNAR)* whilst the ones where the store was not operating, setting them to zero is illogical because it will be assumed that their store did not make any sales or did not have customers while it was closed. [4] provides the following information about business hours in the United States:

- Shopping Malls
  - Monday-Saturday: 10 a.m.-9 p.m.
  - Sunday: 11 a.m.-6 p.m.
- Supermarket
  - Daily: 8 a.m.-8 p.m.
  - Some open 24 hours

Outliers

I used a scatter plot to highlight obvious inconsistencies that could indicate that the value was incorrectly entered because it differed from the other data points in the datasets. There do not appear to be any outliers in the traffic values. For sales in the original dataset, the store's sales in 2017-07-09 19:15:00 appear to be an outlier when compared to other points, as sales typically range from $0 - 3000$, and $> 6000 = 6435$ appears to be an outlier when compared to other points.

The outlier from the original dataset was automatically adjusted after the intervals were converted from 15 minutes to an hour, and it is no longer an outlier. This clearly indicates that the data point was an outlier in the 15-minute interval; it is possible that the store made a lot of sales on July 9, 2017, between 19:15:00 and 19:15:00. Figures 2 and 3 below show the scatter plot before and after conversion.

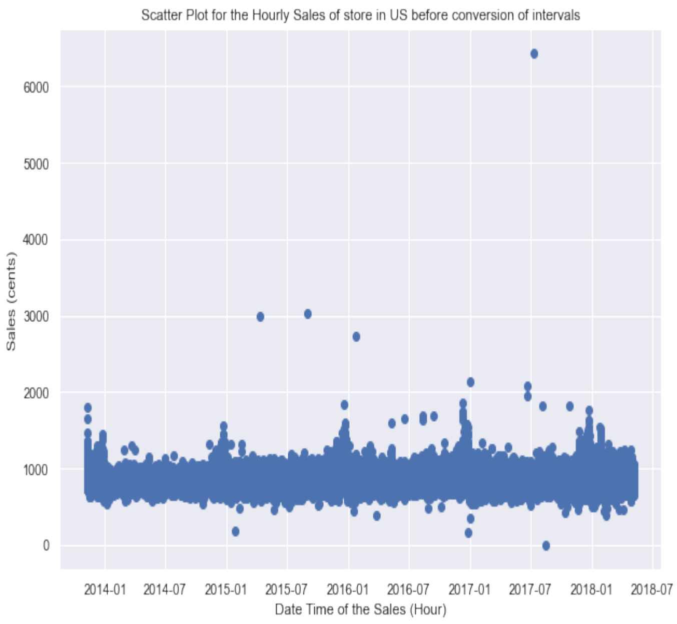


Fig. 2. Scatter Plot for the Hourly Sales of the US store before conversion of intervals

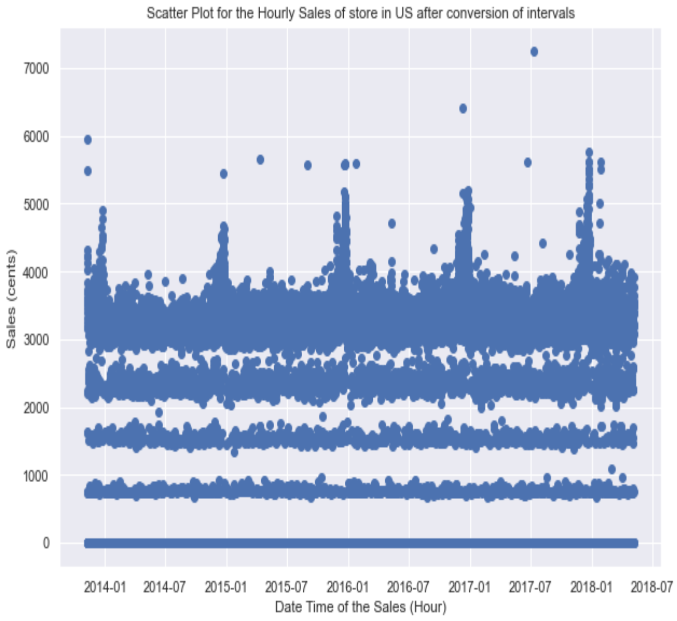


Fig. 3. Scatter Plot for the Hourly Sales of the US store after conversion of intervals

Handling Values missing completely random(MCR)

The datasets are not equal in terms of the number of data points; the traffic dataset has 17278 data points, while the sales dataset has 23215 data points; the difference between the two is approximately 25% of data points, so it is reasonable not to remove so much data. There are several approaches I could take to make these two equals. Removing data that is not common on the datasets but removing these data points will result in the loss of important information about the dataset because this data is a time series. Because the data is missing from 2013-Nov-2013 to 2015-01-01, keep in mind that these datasets are Time Series, so such information will not be useful when modelling with other supervised algorithms. I will impute the data using an average because it makes more sense to impute using an average in a time series dataset.

keep in mind that the Average size must be appropriate for this problem. Calculating the annual average traffic for the store makes no sense because it is a large sample to estimate the actual traffic.

Descriptive statistics

Descriptive statistics provide me with measures of central tendency, a summary metric that attempts to describe an entire set of data by assigning a single value to the middle or center of its distribution. Furthermore, this descriptive shows whether the dataset's overall mean is trending or not, and the mean and standard deviation of the traffic edited dataset are close, implying that the edited dataset may have similar metrics to the original dataset. We can continue to use the original dataset despite the fact that they are not equal in this way.

B. Reduction of explanatory variables

After performing some diagnostics for missing data points and outliers, I applied remedies to the exploratory variable(s). I only reduced the variables used to clean/analyse the data in this study, such as the hour, weekday, and year for each data point. As a result, because the variables appear to be significant, this phase of model development is not applicable in this study.

C. Model Building

Identifying trend and Stationarity

Before computing the model, I plotted the data to observe whether the datasets show trends and seasonality.

Moving averages are based on the concept of windowing because the datasets are not too complex to implement the simple average for smoothing. Furthermore, Simple Moving Average is calculated using the following algorithm: each subsequent value is the mean of the previous 92 (week) observations. Remember that a Window length should be greater than the frequency of the time series, which is an hour in this case, and is typically defined as a minute, hour, day, week, or month.

Because the shop is not open 24 hours a day, my window for a month will not be 30/31 data points. Similarly, my window for a week will not be 7 data points because the data point interval is an hour and the hours for each day are not equal. To recap, the shop is open 14 hours during the week, 13 hours on Saturdays, and 9 hours on Sundays. Each data point corresponds to one hour. Thus, in this case, I will use a weekly average, which implies that the window will be $(14 * 5 + 13 + 9 = 92)$, and I will use a monthly average to identify the trend.

Choosing a window size is frequently an iterative process; smaller windows will still smooth out the data, but not to the same extent, whereas larger windows will result in too much information loss. As a result, one week is sufficient because a month would cause me to lose a lot of information.

The plots (rolling window) show a slight trend in the sales dataset in the averaged plot, but it is difficult to tell whether there is a pattern or not in the original plot. Similarly, the traffic dataset shows a trend with an annual peak (December-January). Interestingly, the plots show an increase in sales over time while traffic remains constant, which makes sense given that inflation causes prices to rise in the store. People, on the other hand, continue to buy goods from the store whether prices rise or fall. see figure 4 and 5.

The monthly rolling window appears to show the trend better than the weekly rolling window, but that does not mean I can use it because it is too broad and will not help me forecast sales and traffic because the plot shows that it does not spike closer to the actual value.

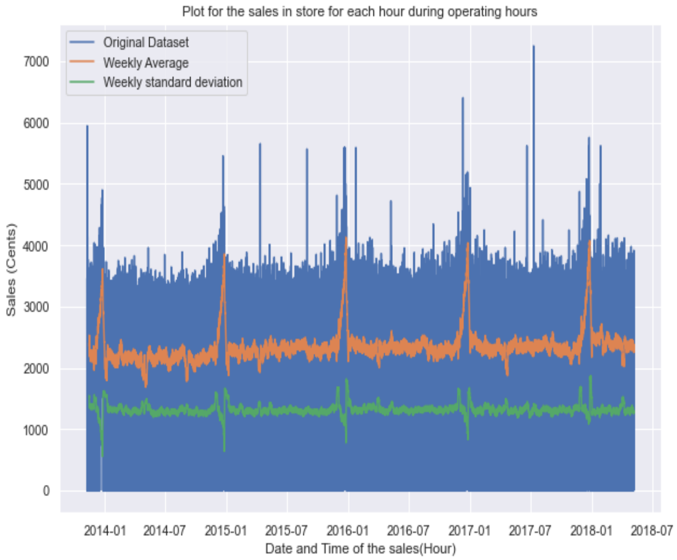


Fig. 4. Plot for the sales in store for each hour during operating hours

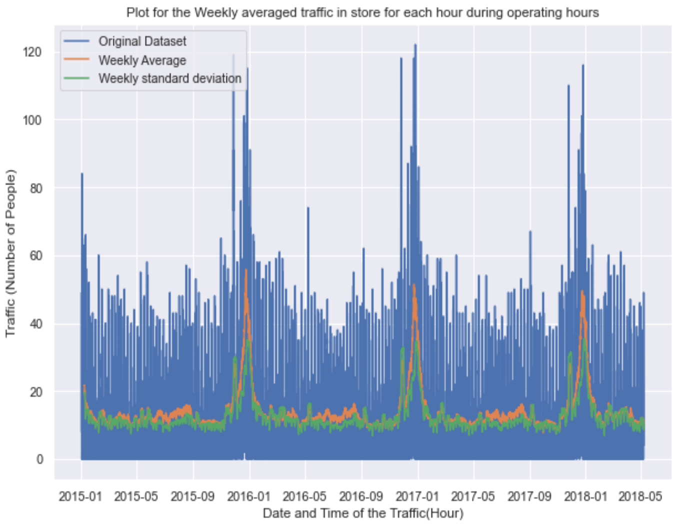


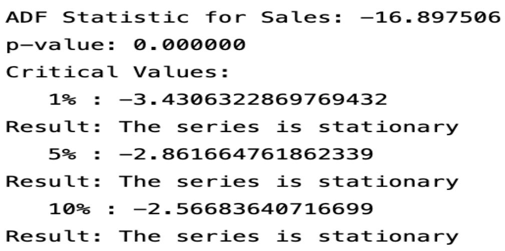
Fig. 5. Plot for the sales in store for each hour during operating hours

Statistical tests to check stationarity

I used the Augmented Dickey Fuller(ADF) test.

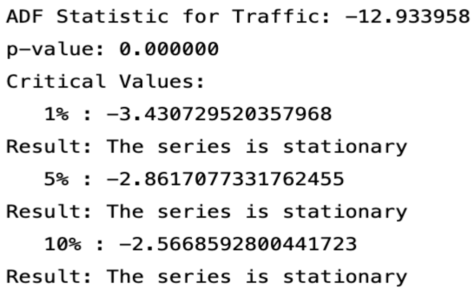
$p - value > 0.05$: Fail to reject the null hypothesis, the data has a unit root and is non-stationary. $p - value <= 0.05$: Reject the null hypothesis, the data does not have a unit root and is stationary.

The results show that the datasets (Sales and Traffic) are stationary, so I will not transform them to make them stationary. Figure 6 and 7 show the test results obtained.



```
ADF Statistic for Sales: -16.897506
p-value: 0.000000
Critical Values:
   1% : -3.4306322869769432
Result: The series is stationary
   5% : -2.861664761862339
Result: The series is stationary
   10% : -2.56683640716699
Result: The series is stationary
```

Fig. 6. ADF results for the Sales dataset



```
ADF Statistic for Traffic: -12.933958
p-value: 0.000000
Critical Values:
   1% : -3.430729520357968
Result: The series is stationary
   5% : -2.8617077331762455
Result: The series is stationary
   10% : -2.5668592800441723
Result: The series is stationary
```

Fig. 7. ADF results for Traffic dataset

Modelling and fitting the model

Model configuration and building

[3] discusses the procedure for determining the correct ARIMA(p, d, q) values, where p represents the number of significant terms in partial autocorrelation factors (PACF), d the order differencing, and q the number of significant terms in autocorrelation factors(ACF).

Before I could build the model for forecasting future sales and traffic in the store, I determined the model's hyperparameters using the autocorrelation and partial autocorrelation factors. Due to the errors and other limitations of this procedure for determining the order of the model, I implemented a stepwise approach that is used to search multiple combinations of the order (p, d, q) parameters and select the best model with the lowest Akaike information criterion(AIC) as it helps to choose which model fits the data well from the generated models.

After performing the stepwise approach for selecting the optimal order for the models, the order of the two datasets is the same $(p, d, q) \rightarrow (3, 0, 2)$. The results obtained using the stepwise approach are the same as the ones obtained using autocorrelation.

Model fitting and Visualisation

The fitted model values are within the range of the original dataset after building the ARIMA model as the results in the previous section indicate that the datasets are stationary. Thus, the will be no need to perform differently in this problem. To keep the model's predicted values within the range of the actual sales and traffic values, I use inverse differencing, which was previously used to make the datasets stationary.

Plotting the actual and predicted values for all data points yields ambiguous results, despite the fact that the plots are correct for visualisation purposes. The notebook depicts various time periods. Figures 8 and 9 visualises the plots for all actual values and fitted values(unclear) and the first two weeks of each dataset robustness.

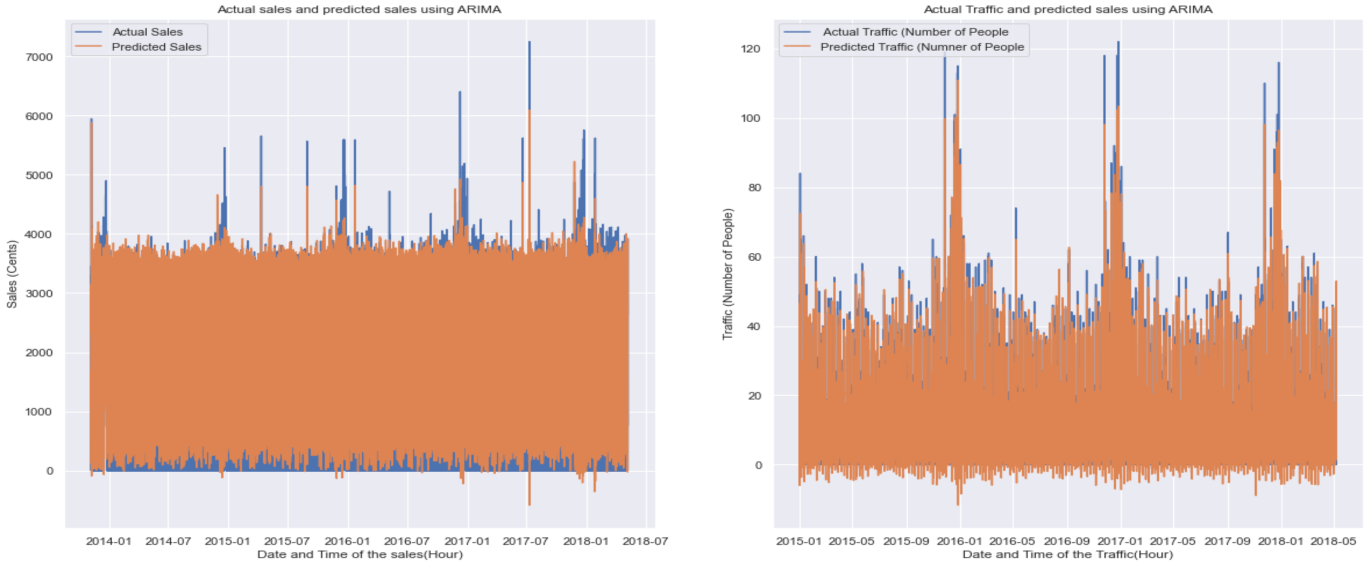


Fig. 8. Actual values(Blue) and Predicted values(Orange) for a store's sales and traffic

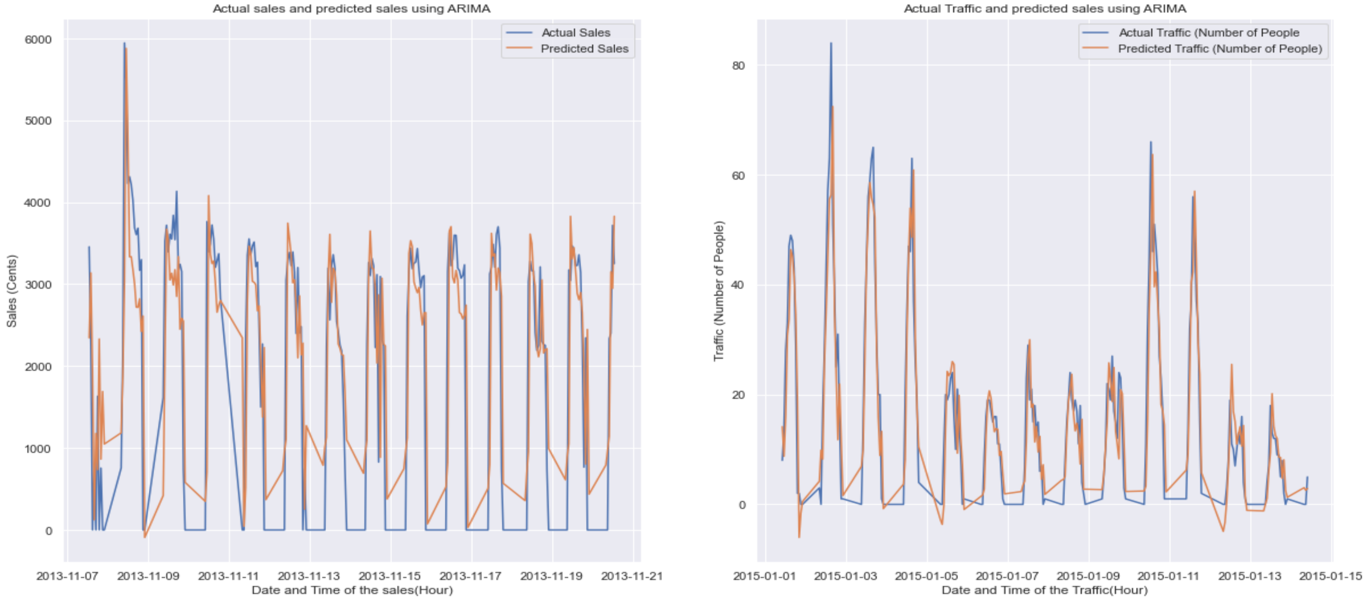


Fig. 9. Actual values(Blue) and Predicted values(Orange) for a store's sales and traffic for the first two week

III. Forecasting and Discussion

In this problem, generating the next month's timestamps will be the first task to complete before forecasting because the intervals are not the same throughout as non-operation hours were removed for each day because the store operates at different times on weekdays and weekends. As a result, I'll repeat the procedure I used to remove the missing values in Data Preparation and Exploration.

There are two different methods to use for this problem *Prediction* and *Forecast* where *Prediction* is an in-sample forecast whereas *forecasting* is an out-of-sample forecast.

What do I mean by that, if I wanted to focus sales or traffic for a range that lies in the given data, that would be called in-sample forecasting on the other hand if I want to forecast sales or traffic for a time period which is not in the data then that is called out-of-sampling forecast.

Figure 10 shows the forecasted sales and traffic for the next month, starting with the most recent data point in each dataset. The plot displays the actual value from the last month of the dataset rather than the forecasted value because the plot only shows a month period.

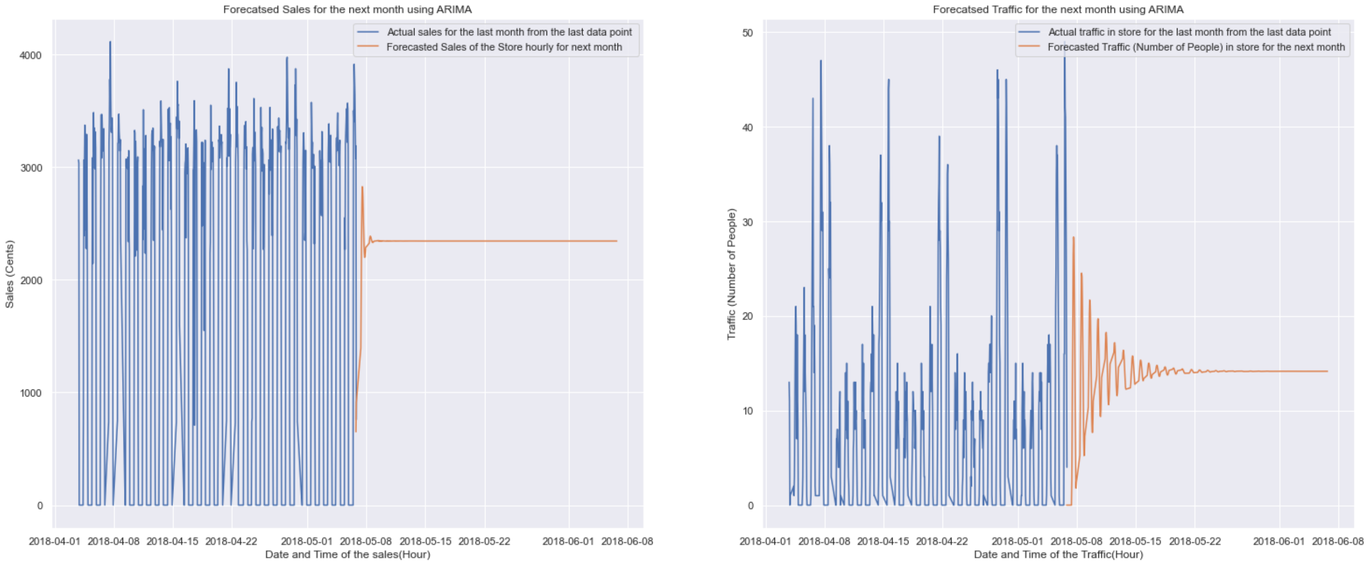


Fig. 10. Actual values(Blue) and Predicted values(Orange) for a store's sales and traffic for the first two week

Based on the nature of the ARIMA equations, out-of-sample forecasts tend to converge to the sample mean for long forecasting periods. Thus the plots show the noise at first and converges after some period.

IV. Conclusion

The model was able to forecast the hourly sales and the traffic of the store for the next month using ARIMA. I can conclude that using machine algorithms and statistics models in the retail industry would help in decision making as stated that knowing an estimation of sales and number of customers would help when allocating stuff and in other different scenarios. By that, retailers will no longer guess what their customers want; instead, they will use future-ready demand forecasting tools and machine learning algorithms to more accurately predict customer behavior.The main focus of this study was on ARIMA because the datasets were stationary as the ADF showed, because if the datasets had some seasonality, I would have implemented SARIMA because it performs well in seasonal data.

References

[1] Anthony Yu Abhinav Bhaskar. Autoregressive integrated moving average (arima) models, Sep 2019.

[2] Michael D. Geurts and J. Patrick Kelly. Forecasting retail sales using alternative models. *International Journal of Forecasting*, 2(3):261–272, 1986.

[3] Sanjay Nandakumar. Time series analysis complete tutorial for beginners (part 3), Mar 2022.

[4] The official travel site of the USA. Time & business hours.

[5] Noah Parsons. How to do a sales forecast for your business the right way-2021 guide, Mar 2016.

[6] Chris Wadsworth. Measuring retail store traffic: How people counting works, Apr 2019.