FORECASTING ROSSMANN STORE SALES

Rhys Kilian

November, 2017

Executive Summary

This report outlines the methodology to determine an appropriate time-series model for forecasting 6-weeks of sales data for Rossmann Store 1. Rossmann is a retail store chain which requires accurate forecasts of future sales to determine stock levels, staffing requirements and the possible impacts of any promotional sales. In order to generate these forecasts the past sales of various stores were processed which involved filtering and cleaning prior to any model selection. However, prior to this, an exploratory data analysis was completed in order to find any trends within the data. This information was used in the model selection process.

The candidate models discussed in this report are: simple exponential smoothing (SES), Holt-Winters Exponential Smoothing, ARIMA and Seasonal ARIMA models. These models were evaluated based on their performance on the out-of-sample data for various forecast horizons, including 2 weeks, 6 weeks and 26 weeks. The two models selected for further analysis were based on the the mean squared error (MSE) and the mean absolute percentage error (MAPE) for the 6 week validation test based on the client requirements. This further analysis involved checking the assumptions of the models using diagnostic plots. Finally, a forecasting exercise was completed on the Seasonal ARIMA (2,1,0)(0,1,1) model which was found to be the best performing model. This involved producing fan charts to show the various levels of confidence for sales forecasts over different horizons, and point forecasts for a week. It was concluded that the Seasonal ARIMA model would allow the manager's at Rossmann Store 1 make more informed decisions regarding stock and staff levels, with the 6 weeks of sales forecasts.

Contents

1 Introduction and Business Context

For any retail store, being able to forecast future sales is critical to determining resource requirements, including, but not limited to restocking of merchandise, staffing of stores and opening hours. Rossmann is a Germain drug store chain which operates in 7 different European countries. With over 3000 stores, understanding future sales at each location, with its unique operating conditions is a challenging task. This report focuses on forecasting methods which can be used to predict sales for 6 weeks in advance at a single location using simple forecasting methods.

Time series data from 1,115 stores, over 2 and a half years (January 2013 - July 2017) is used for this analysis. The data provide contains information not only about the date (year, month and day) and sales, but also provides cross-sectional data of the day of the week, the number of customers, if the store was open, if a promotion was occurring and if it was a state or school holiday. Whilst all these factors have an influence on the sales, it is possible to use only the time series data to construct a forecasting model.

2 Data Processing

The original dataset presented to the team contained the combined data of the Rossman stores in terms of the date, sales, number of customers, day of the week, whether the store was opened on that particular day, whether a promotion was active and what type (if any) state holiday was valid.

Firstly, the data was arranged and indexed by date, starting from oldest to newest for simplicity. Secondly, the entire dataset was cleaned appropriately. This involved checking for any missing values, of which none were found, and sales data points less than or equal to zero were removed. This is because negative sales do not make sense in the context of the business problem (unless refunds are considered negative sales). Additionally, any days in which the store was not open (i.e. Sunday) was removed in this step as well. These days were not considered in the forecast as they would have an effect of pulling the forecasts towards zero. Therefore, the working week would be six days, and any forecasts involving Sunday, would be forecasted to have sales of $0.00 under this assumption.

Furthermore, the holiday variables were converted into k-1 dummy variables. These additional predictors could be used to predict future sales using a combined ensemble and time-series model. Finally, as only Store 1 was being considered for this analysis the data was filtered for this. Therefore, all analysis presented in this report is in relation to Rossman Store 1.

3 Exploratory Data Analysis

For time series data, understanding the behaviour of the data, particularly changes in the data overtime is essential to ensuring accurate foreecasting models are applied. Exploratory data analysis on time series data establishes if there are trends over time (long-term increase or decrease) or seasonality (repeated patterns) in the data. This analysis also allows for an exploration into the other factors which might be contributing the response (sales).

3.1 Time Series Data

The easiest way to visulize the longterm trends in the data is to simply plot the time series data. Figure 1 shows the full time series data for store 1. Whilst in Figure 2, days with zero sales are removed, and the time series is further decomposed into yearly plots allowing more detailed analysis. The decomposed year plots highlights the spike in sales each December/January which can be attributed to holiday shopping trends. Furthermore there appears to be a degree of seasonality with fairly constant spikes in sales throughout the year. Interestingly there appears to be no longterm increase or decrease in values, suggesting to no trends in the data are present.

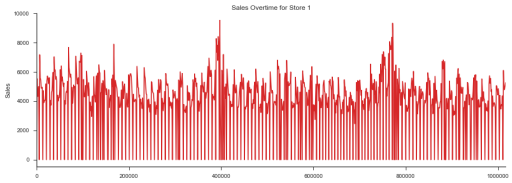

Figure 1: Time Series of Store 1 Sales

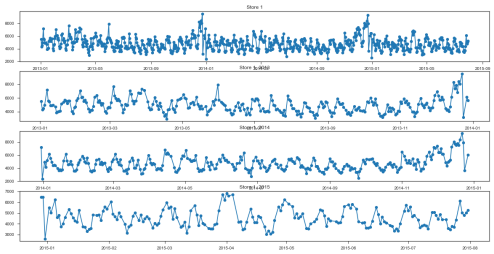

Figure 2: Time Series of Store 1 Sales, Yearly Decomposed

Further analysis of the data shows that the seasonality in the data appears to occur at a period of 12 days (or 2 weeks taking into account the days when the store is closed). Figure 3 shows a decomposition of the time series data into trend, seasonal and residual components. As was previously suspected there is no distinct trend, with the trend plot showing a rough trend which accounts for the yearly spike in sales around January. The seasonal plot is interesting, showing and accounting for the seasonal variation which appears at a 2 week interval. However as can be seen from the residual plot, there is a lot of noise in this decomposition, suggesting the data does not have a strict seasonal period.

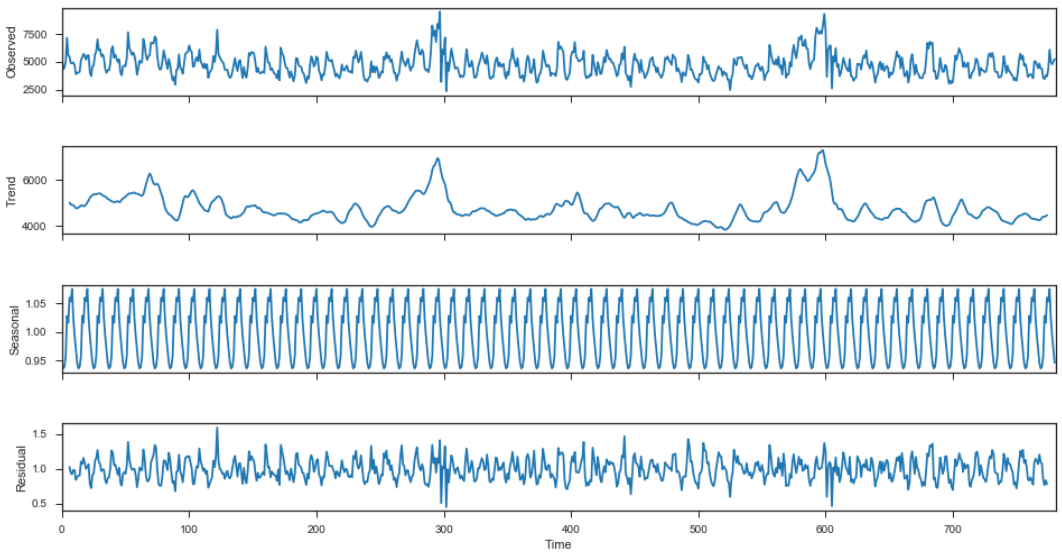

Figure 3: Time Series Decomposition

Critically, as the seasonal variation does not seem to be highly proportional to the long term trend (which has significant variability) it can be hypothesized that an additive model will be more suitable to the data. However, if the trend is considered to be significant (despite the noise present) a multiplicative model might be suitable.

3.2 Cross Sectional Data

Given the cross-section features which are also given in this dataset, it is of some interest to investigate the impact of these predictors on the sales. In particular these variables can help explain some of the patterns present in the time series data.

As seen in Figure 4, there is a distinct relationship between a promotion running, and higher sales, with both the mean and interquartile ranges for sales being higher than for the cases where there is no promotion. Further investigation into the data reveals that store 1 has a 5 day promotional period approximately every 6-7 days, accounting for the approximately fortnightly seasonality previously observed.

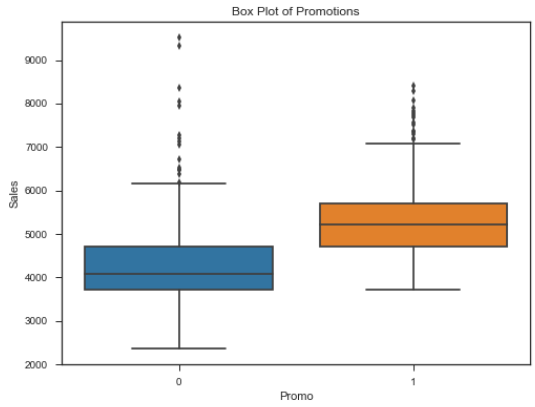

Figure 4: Box Plot for Promototions

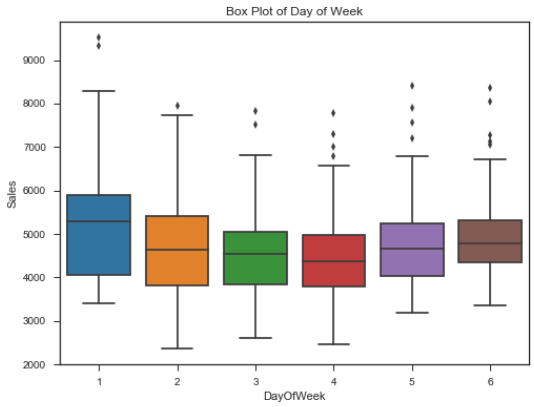

Figure 5: Box Plot for Day of Week

Looking at the relationship between sales and day of the week as seen in Figure 5 shows that whilst there is a dip in sales during the middle of the week, the difference between the interquartile ranges and means of sales is slight compared to the impact of promotions. However this information is useful to consider, in particular the fact that store 1 is always closed on Sunday (day 7).

Further investigation of other cross section data reveals that the store is always closed on state holidays, meaning there is no discernible impact of these on sales. These additional predictors thus help reveal important seasonal (promotion) and cyclic (day of the week) trends in the data which can thus be accounted for in a time series model. Though outside the scope of this report, developing a dynamic regression model incorporating these predictors could result in a more robust forecast, although given the clear relationship with these predictors and time series behavior suggests a time series model will be sufficient.

4 Modeling

There are several different forecasting methods which can be used to forecast future data. The models considered include random walk, exponential smoothing and ARIMA as well

as variations within these models. These models were selected and fit based off the exploratory data analysis and through the use of diagnostic plots.

4.1 Model Selection Criterion

To select the models the diagnostic and identification images, AIC and the validation method were used. For the purposes of this report, only the image and validation methods were considered to avoid relying on the theoretical justifications and assumptions required for the AIC. Furthermore, it also requires a constant dataset and prevents dissimilar models, such a additive and log-additive, from being directly compared.

4.2 Justification for Models Considered

The models considered in this analysis are all time series forecasting methods. Time series forecasting allows for changes in the response over time, such as long term increases/decreases (trends) or periodic fluctuations (seasons) to be taken into account. In comparison cross-sectional methods only focus on information which has been observed at a single point in time. The sales of a store, which whilst dependent on many factors are expected to have a time-varying behavior which is important to the prediction of future sales. The exploratory data analysis performed in Section 3 confirmed such time varying behavior.

Whilst predictor variables can be included in time series forecasting, such as dynamic regression models, panel data models or linear system models, these models have significant difficulties with forecasting and required significant knowledge about the variables of interest. Given time series models have the potential to outperform explanatory or mixed models, it is seen to be sufficient to focus on these time series forecasting models.

4.2.1 Benchmark - Random Walk Model

Random walk models use the value of the last available observation to forecast the series. This model assumes that the future values are equal to the last observation, and as thus is known as the naive method.

$$Y_{t+h} = Y_t + \Sigma_{i=1}^{h} \epsilon_{t+i} \tag{1}$$

Assuming that the errors are normally distributed according to $\epsilon_t \sim N(0, \sigma^2)$, it is possible to generate a forecast interval of the data, to estimate the possible bounds of future predictions.

$$y_t \pm z_{\alpha/2} \times \sqrt{h\hat{\sigma}^2} \tag{2}$$

$$\hat{\sigma}^2 = \frac{\sigma_{t=2}^{T}(y_t - y_{t-1})^2}{T-1} \tag{3}$$

The random walk model will be used as a baseline for comparison of further forecasting models due to its simplicity and relative straight-forward implementation. Notably it does not consider any long term trends or seasonality in the data and thus represents a simple forecast estimate.

Variations on Random Walk

However, if the errors are non-Gaussian, other methods such as a bootstrap algorithm should be used. A bootstrap forecast interval samples residuals with replacement to obtain bootstrap residuals, to then estimate $y_{s,t+h}^{*} = y_t + \Sigma_{i=1}^{h} e_{s,i}^{*}$. These values are used

to compute the desired quantiles. However even this method the errors are assumed to be independent and identically distributed.

Given the data does not feature highly seasonal patterns (although some seasonality is present), the seasonal random walk model was not considered. This model considers the forecast to be equal to the last observed value from the same season of the year.

4.2.2 Exponential Smoothing Models

Exponential smoothing methods use weighted averages of past observations, thus taking into account how time series components change over time. The weights of the past observations exponentially decay the further in the past these observations are.

Simple Exponential Smoothing

The simple exponential smoothing (or exponentially weighted moving average EWMA) is the simplest of the exponential smoothing methods, and is useful for time series with changing levels.

$$\hat{y}_{t+1} = \ell_t \qquad \text{Forecast Equation} \tag{4}$$

$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1} \qquad \text{Smoothing Equation} \tag{5}$$

For higher values of α, the forecasts will reflect recent changes in the series as larger weight is given to recent observations. Conversely, for lower vales of α the forecast will tend to be smoother, larger weight being given to past observations. α can e estimated by least squares $\hat{\alpha} = argmin\Sigma_{t=1}^{N}(y_t - \ell_{t-1})^2$.

Again assumed the errors are distributed normally according to $\epsilon_t \sim N(0, \sigma^2)$, is it possible to find an interval forecast. Similarly if errors are not normally distributed another method (such as a Bootstrap) should be used.

$$\hat{\ell}_t \pm z_{crit} \times \sqrt{\hat{\sigma}^2[1 + (h-1)\hat{\alpha}^2]} \tag{6}$$

$$\hat{\sigma}^2 = \frac{\sigma_{t=2}^{n}(y_t - \ell_{t-1})^2}{N-1} \tag{7}$$

Holt-Winters Exponential Smoothing

Holt-Winters exponential smoothing, or seasonal method extends the Holt model to account for seasonal data. There are two distinct variations of this method; the additive method is used with roughly constant seasonal variations, whilst the multiplicative method is used when seasonal variations change in proportion with the level of the series. Given the data under consideration a slight link between seasonal variation and the trend was found, therefore, the multiplicative model will be used. This is verified by the results of the validation test in the next section.

$$\hat{y}_{t+1} = \ell_t + b_t + S_{t+1-L} \qquad \text{Forecast Equation} \tag{8}$$

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \qquad \text{Level Equation} \tag{9}$$

$$b_t = \beta(\ell_t - \ell_{t-1} + (1-\beta)b_{t-1} \qquad \text{Trend Equation} \tag{10}$$

$$S_t = \delta(y_t - \ell_t) + (1-\delta)S_{t-L} \qquad \text{Seasonal Indices} \tag{11}$$

4.2.3 ARIMA Model

ARIMA models are another commonly used model for time series forecasting based on autocorrelations, that is the relationship between time series components in the data.

ARIMA models rely on first finding a stationary transformation of the data. Stationary data does no depend on the time which it is observed. By dentition time series data with a clear trend or seasonality is not stationary.

Common data transformations include log transformations, Box-Cox transformations and differencing. Differencing looks at the change between consecutive points in time ($\Delta Y_t = Y_t - Y_{t-1}$). Second order differencing differences the series a second time, which is occasionally necessary to make the series stationary ($\Delta^2 Y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$. For data that is non-stationary due to seasonality, a seasonal differencing can be used to transform the data.

Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots are used to assess if the data is stationary. ACF plots of stationary data should show a quick drop to zero.

Next the autocorrelations are modeled and used to build the forecast.Non-seasonal ARIMA models are usually specified by three numbers as an ARIMA(p,d,q) model. Here, p refers to the number of autoregressive terms (lags of stationarized series), d the number of nonseasonal differences and q the number of moving-average terms (lags of forecast errors). The general ARIMA(p,d,q) model is given by Equation 12.

$$(1 - \Sigma_{i=1}^{p} \phi_i B^i)(1 - B)^d) Y_t = c + (1 + \Sigma_{i=1}^{q} \theta_i B^i)\epsilon_t \tag{12}$$

Autoregressive terms are usually associated with time series behavior which has a restoring force, pulling values towards the mean. The AR coefficient/s determines the speed of this pull towards the mean. Meanwhile moving average terms are associated with time series with random shocks felt in two or more consecutive periods. However the dependence on autoregressive or moving average terms depends on the extent of differencing.

The order of p and q are selected using a combination of visual identification, AIC and model validation. Through these techniques, ARIMA(0,1,1) which is related to exponential smoothing was seen to be the most suitable non-seasonal ARIMA model for this dataset.

Seasonal ARIMA

Seasonal ARIMA(p,d,q)(P,D,Q)$_m$ models are a variation on ARIMA models which takes into account seasonal patterns, where D is the order of seasonal differencing, P and Q are the seasonal autoregressive and moving average components respectively and m is the number of seasons. If a strong seasonal trend is present, seasonal ARIMA should be used otherwise the trend will die out in long-term forecasts.

Again the order of p, q, P and Q is done through visual identification, AIC and model validation. The best seasonal ARIMA model is thus found to be ARIMA(2,1,0)(0,1,1) as shown in Section 5.

Whilst ARIMA models lead to stable estimations of time-varying trends (and seasonal patterns) with the use of relatively few parameters, there are some drawbacks to this model. Specifically the interpretability of the coefficients is limited and it is difficult to explain the underlying mechanisms of the model. Furthermore there is significant danger of over-fitting and mis-identification.

4.3 Model Selection Identification and Diagnostics

4.3.1 Multiplicative Holt-Winters Diagnostics

The fit of the multiplicative Holt-Winters model to the training data is shown in 6. The residual diagnostics helps determine the appropriateness of this model. Figure 7 shows

the residual autocorrelation plot, which highlights any autocorrelation in the residuals. If autocorrelation is present in the residuals it suggests there is valuable information not accounted for by the model. Whilst there are some spikes in the ACR plot suggesting some autocorrelation, overall the model is assumed to capture most of the information.

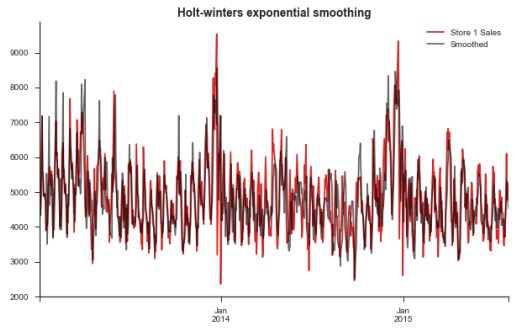

Figure 6: Fit of MHW

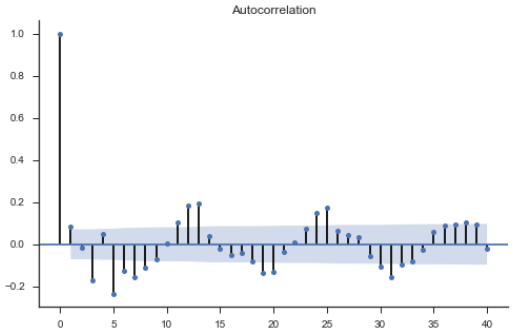

Figure 7: Autocorrelation Plot of MHW

Looking at the residual plot in Figure 8 there are two distinct points which appear to be outliers. Notably these occur around December/January which is seen to have a significant spike. Thus whilst these are a potential outlier (and accounting for some of the autocorrelation previously identified), given the forecasting horizon of 6 weeks the potential effect is seen to be minimal. However given the perceived seasonality of this spike, further modeling could be done to try and capture this information.

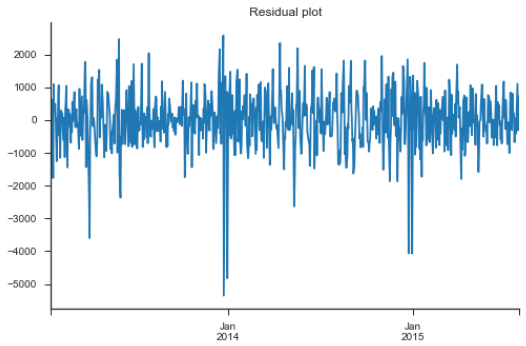

Figure 8: Residuals of MHW

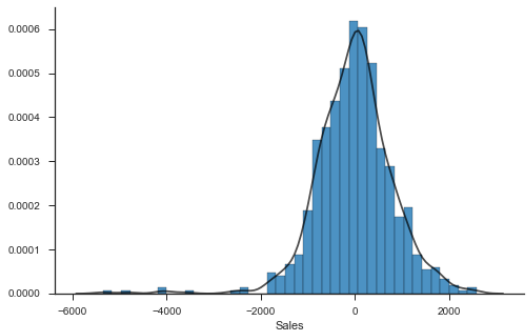

Figure 9: Residual Distribution of MHW

The residual distribution in Figure 9 has a significant left tail, and a slight positive skew. However part of this can be attributed to the December/January spikes previously discussed. Aside from this tail, the residuals are quite symmetric with near Gaussian distribution. Thus whilst questionable, there is reasonable evidence to support the assumption of normally distributed residuals.

4.3.2 Seasonal ARIMA(2,1,0)(0,1,1)

ARIMA specifications are best identified using autocorrelation and partial autocorrelation plots to identify the impact of differencing/transformations on the data. Further details and ACF/PACF plots demonstrating the steps taken to find the correct ARIMA specification can be found in Appendix B. The ARIMA(2,1,0)(0,1,1) model was identified by these autocorrelation plots as the most suitable model for the dataset in question.

The autocorrelation and partial autocorrelation plots shown in Figure 10 show there is minimal autocorrelation between residuals are there is no distinct pattern in the plot.

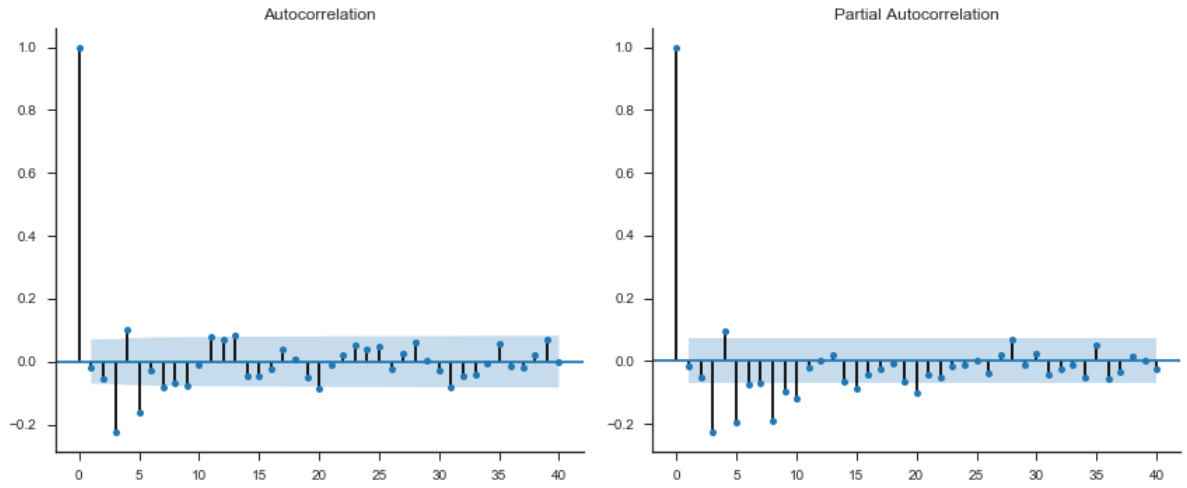

Figure 10: ACF and PCF of ARIMA

The residuals and distribution of residuals as shown in Figures 11 and 12 are similar to that of the multiplicative Holt-Winters model considered. There are two distinct spikes during the December/January period representing the high sales during this period. The residual distribution is likewise somewhat tailed, however otherwise has a normal distribution, which a high degree of symmetry, meaning the normal assumption is not severely violated. Although further transformations could be conducted to remove this tail.

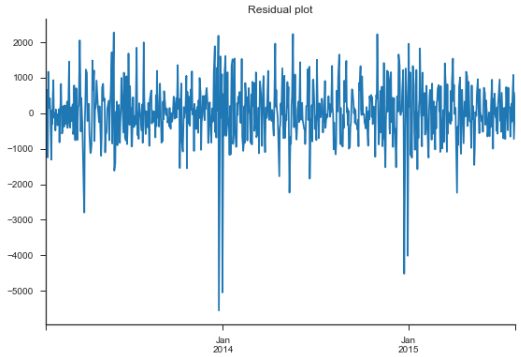

Figure 11: Residuals of ARIMA

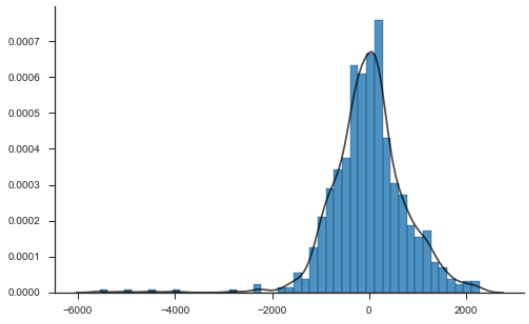

Figure 12: Residual Distribution of ARIMA

5 Model Validation

Following the completion of the model selection phase, the model validation stage was conducted as part of the forecasting process. During this phase a validation sample was set aside for estimating and comparing the performance of the different models outlined in the previous section of the report. It is important to mention that this is the final stage of the forecasting process as there is no test set. This is due to the dynamic nature of the process.

From the overall dataset, the last part of the part was allocated to the validation set. Various sample sizes were used to validate the performance of the model at different forecast horizon. In particular, a real time forecasting was used for validation. Such that, at every period t, all the available data was used to estimate the model and predict the future value of the time series. The horizons considered and the reason for this are outlined in Table 1:

Table 1: Validation Data Sets

Horizon	**Dates**	**Reason**
2 weeks	18/07/2015- 31/07/2015	Approximate average sales period
6 weeks	20/06/2015 - 31/07/2015	Based on the client requirements
26 weeks	24/01/2015 - 31/07/2015	Two sales quarters of data

However, based solely on the client requirements alone the model which exhibits the lowest validation score for the 6 week horizon will be the model selected for further forecasting.

5.1 Validation Metrics

To measure the forecasting accuracy of the proposed models several validation metrics were considered. Firstly, the out-of-sample mean squared error (MSE) was used as the primary metric to measure the performance of the model. Models which produce more accurate forecasts will result in lower out-of-sample MSE, and the model which has the lowest MSE for the 6 week validation sample will be the final model selected.

Nonetheless, one other popular metric was also considered to provide a comparison to the out-of-sample MSE. The mean absolute percentage error (MAPE) was calculated for each model on the validation data and has the following form:

$$MAPE = mean(|p_t|) \tag{13}$$

Where,

$$p_t = 100 \times ((y_t - \hat{y}_t)/y_t) \tag{14}$$

The MAPE is a useful metric to validate the performance of models since it is scale-independent and is simple to understand. However, this metric does have several drawbacks which must be understood when interpreting the results. Namely, it cannot be used if there are any zero values in the data, therefore, any closures of the store due to public holidays must first be cleaned. Moreover, for forecasts which are low the percentage error cannot exceed 100%. However, there is no upper limit to MAPE for forecasts which are too high.

To summarise, the model which produces the smallest out-of-sample MSE for the 6-week period will be the model selected for final forecasting.

5.2 Validation Results

Based on the client requirements, the validation results for the forecast horizon of 6 weeks is shown in table 2. Furthermore, additional validation scores were also considered for horizons spanning 2 and 26 weeks. These are contained within Appendix C in Tables 4 and 5 respectively.

Table 2: 6 Week Validation Results

Model	**MSE**	**MAPE**
Random Walk (Baseline)	601.348	9.650
SES	829.934	13.419
Additive Holt-Winters	779.54	15.773
Multiplicative Holt-Winters	731.47	14.810
Log Additive Holt-Winters	782.68	15.302
ARIMA (0,1,1)	608.140	9.703
Seasonal ARIMA (2,1,0)(0,1,1)	493.680	9.407

As can be seen in Tables 2,4, and 5 the best performing model in terms of both MSE and MAPE, other than the baseline, was the seasonal ARIMA (2,1,0)(0,1,1) model. However, based on the client requirements, the purpose of this analysis was to determine a model which could accurately forecast the sales of a Rossman store 6 weeks in advance. Interestingly, this was the baseline random walk model. However, when this analysis is extended to consider periods of 2 and 26 weeks, as well as the MAPE metric, in fact the seasonal ARIMA model performs better. Therefore, while the selection criteria outlined in Section 5.1 is still valid, based on this information a decision was made to expand the criteria to take into consideration forecasts of different horizons and using other relevant metrics. On average, the seasonal ARIMA model outperforms the random walk model. In short, this is the model selected for further forecasting in the next section.

The validation performance of the seasonal ARIMA model does in fact suggest that there is seasonality present in the data. This is support of the trends found in Section 3. The apparent performance of the seasonal ARIMA model may be due to the solid underlying theory of which underpins the model, and the stability of estimating time-varying trends and seasonal patterns contained within the data (Nau, 2014).

However, this model is susceptible to over-fitting or mis-identification if not used properly (Nau, 2014). Indeed, as shown in the tables above, the performance of the seasonal ARIMA model does converge towards the baseline model as the forecast horizon increases, suggesting the the model may "memorize" the data for short sample periods. Furthermore, it should also be noted that the interpretability of the seasonal ARIMA model is poor. This may be a problem for the client whom may become confused with the coefficients presented with the model. Nonetheless, based on validation performance this model is selected for forecasting in the next section.

Furthermore, the other model selected is the multiplicative Holt-Winters model. The multiplicative Holt-Winters model was selected in preference of the ARIMA (0,1,1) model, the second best performing model other than the baseline, was because it is a substantively different model than the seasonal ARIMA model. This is to allow greater variability for the client to suit their business needs. However, this multiplicative Holt-Winters model actually under-performs when compared to the baseline random walk model. This suggests that the random walk model may be appropriate for this problem. It does in fact suggest the the other models have over-fit to the training data and therefore, perform poorly when forecasting out-of-sample data. Nonetheless, the relative performance of the multiplicative Holt-Winters model suggests that weighting recent observations more heavily may be appropriate for forecasting the sales data. It also has the advantage of being relatively simple to understand which is beneficial to the client. However, this model may not fully capture the seasonal variations within the data. As a result, this may explain why the multiplicative Holt-Winters model has a larger MSE and MAPE

when compared to the seasonal ARIMA model. Therefore, based on the results obtained in the tables above, only the seasonal ARIMA model is selected for forecasting.

6 Forecast

Based on the validation results the seasonal ARIMA(2,1,0)(0,1,1) model was selected for an analysis of the forecasts. The client requirements outline the need for the presentation of forecasts for six weeks of daily sales following the last period in the dataset.

Prior to doing this it is important to restate the assumption made in Section 2. More specifically, sales totaling $0.00 for a date were removed from the dataset. This is generally due to the store being closed due to the day of the week (i.e. Sunday). Therefore, the forecasts presented in this next section ignore the days at which Rossmann Store 1 is closed due to it being Sunday. In reality, it can be assumed that if the given day is Sunday, then the forecasted sales for that day should be zero.

Nonetheless, a fan chart was constructed for the seasonal ARIMA model. This fan chart was used consecutive prediction intervals of 99%, 90% and 75%. The shading on the fan chart is darkest for the prediction interval for 75%. The result is shown in Figure 13.

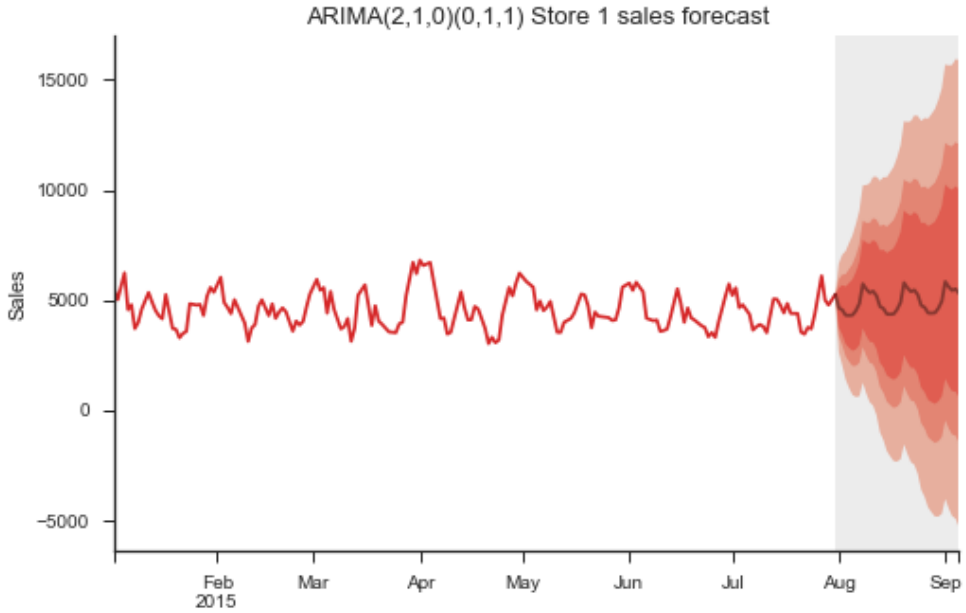

Figure 13: Seasonal ARIMA (2,1,0)(0,1,1) Fan Chart

The fan chart presented in Figure 13 shows the range of values a particular forecast can take for a given confidence interval. As a can be observed, for a given confidence interval, the interval gets wider with the increasing horizon. This is due to the increased uncertainty about future values. To demonstrate this, a fan chart was also constructed for the horizon of 26 weeks as shown in Figure 14 in the Appendix. Comparing these two figures it can be observed the increased horizon results in increased uncertainty with the intervals increasing in width for an increasing forecast horizon.

However, it should also be noted that these fan charts can be misleading. Taking Figure 13 it can be seen that for a particular threshold horizon each of the intervals contain values which are negative. This is not strictly correct as it doesn't make sense

to have negative sales (excluding refunds of purchases). Therefore, this will need to be properly communicated to the client when advising them on the proposed 6 week sale forecast.

Furthermore, it is also beneficial to consider the point forecasts based on the analysis conducted. The optimal point forecast under the squared error loss is given by the conditional expectation given the training data. For example, using the model proposed, it is possible to forecast the sales for every day of the week, six weeks in advance. These forecasts could be used for many purposes, including staff rostering and stock level planning. Therefore, the point forecasts for this period are contained within Table 3

Table 3: Store 1 Point Forecast Week 5-6

Date (2015)	**31/08**	**01/08**	**02/09**	**03/09**	**04/09**	**05/09**	**06/09**
Day	Mon.	Tues.	Wed.	Thur.	Fri.	Sat.	Sun.
Sale Forecast ($)	4957.34	5841.47	5605.37	5441.50	5512.46	5299.62	0.00

As can be seen in Table 3 the point forecasts using the seasonal ARIMA(2,1,0)(0,1,1) model were found for the period five to six weeks after the last recorded sale for Store 1. Firstly, a value of $0.00 was inserted for Sunday, knowing the store is closed. Furthermore, using the point forecasts based on the model, it can be seen that the quietest day is Monday with the busiest being Tuesday. The total sales for Store 1 approximately decreases throughout the week.

This is important information for the managers of Rossman Store 1. This will allow for more accurate planning to allow them to make more informed decisions about their stock levels and the number of staff they have on duty for these particular days.

7 Conclusion

Based on the client's requirements of forecasting the sales of Rossmann Store 1 the Seasonal ARIMA (2,1,0)(0,1,1) model is put forward. This model was found to have the lowest mean squared error and mean absolute percentage error when tested on unseen data. This was in comparison to several other candidate models and in relation to a random walk baseline model.

Using this selected model will allow the client to make more informed forecasts in relation to the sales of Store 1. For instance, the client can observe a fan chart at different confidence levels for their desired horizon or alternatively view the point forecasts for a selected week. As such, the managers of this store are able to make decision regarding the staff levels and amount of stock required with greater confidence. Overall, the seasonal ARIMA model may improve business results through increased visibility of their forecasted sales for this particular store.

8 References

Nau, R. (2014). *Introduction to ARIMA models.*

A Fan Charts

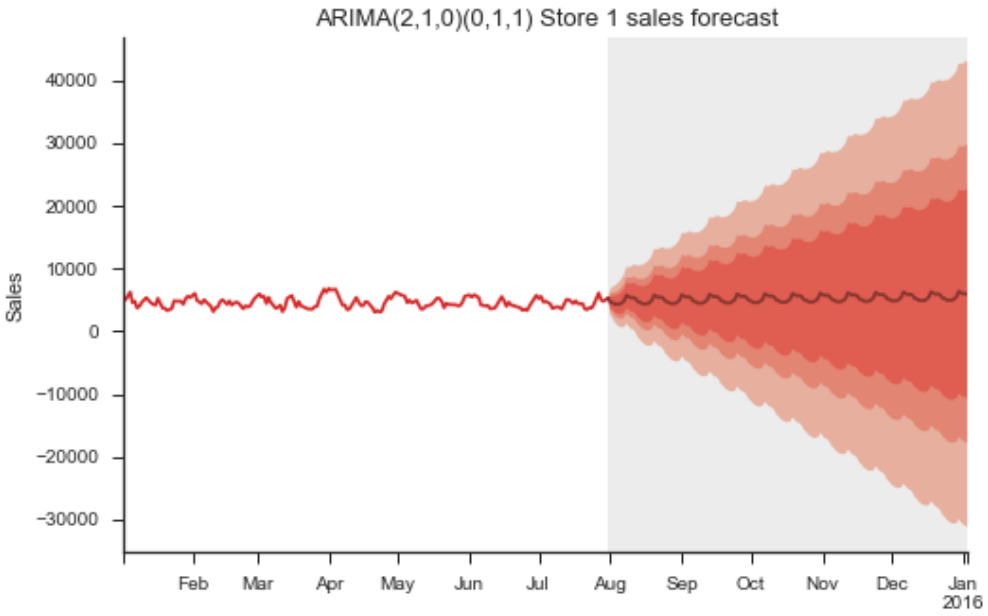

Figure 14: Seasonal ARIMA (2,1,0)(0,1,1) Fan Chart

B ARIMA Identification

The identification of the appropriate ARIMA specification requires stepwise investigation of the nature of the time series in question, specifically what transformations make the data stationary.

Initially, a data frame was created to store the original data, first difference, seasonally difference and fire and seasonally differenced series for further investigation.

The first difference reveals that there is somewhat still a slow decay occurring in the graph. In particular as can be seen in Figure 16 there is a spike at 12, 24 and 36, indicating seasonality of m=12.

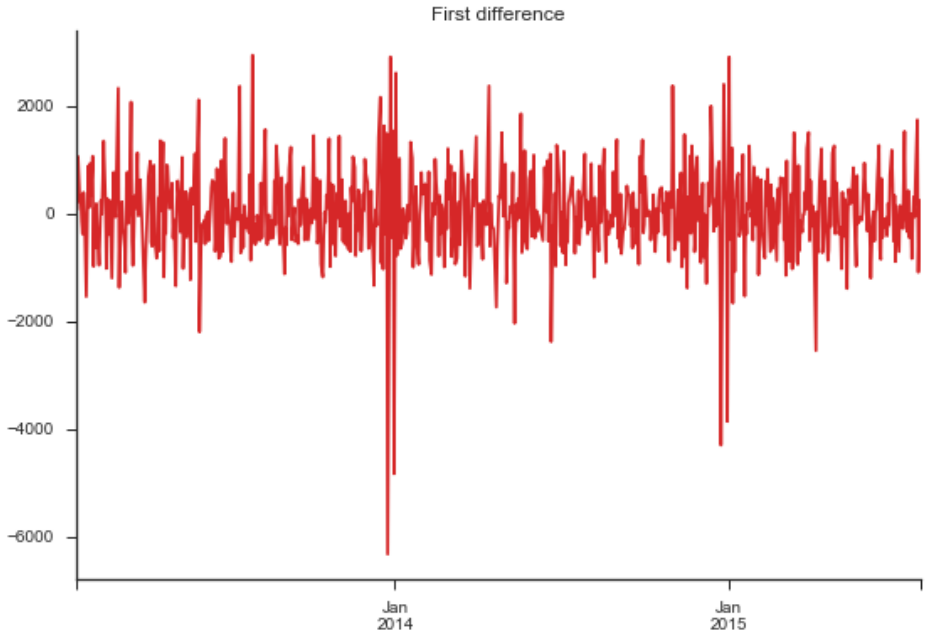

Figure 15: First Difference

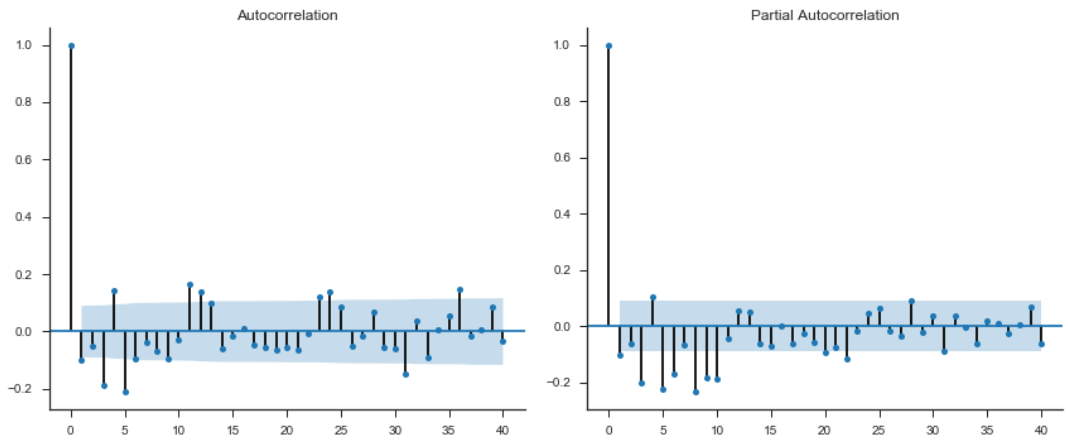

Figure 16: First Difference ACF and PCF

The next stage involves trialling a seasonal difference, to see if the series becomes stationary. Figure 18 shows that the series is clearly still not stationary, with a lag before

the plot reaches zero. This indicates that both differences are needed to make the series stationary.

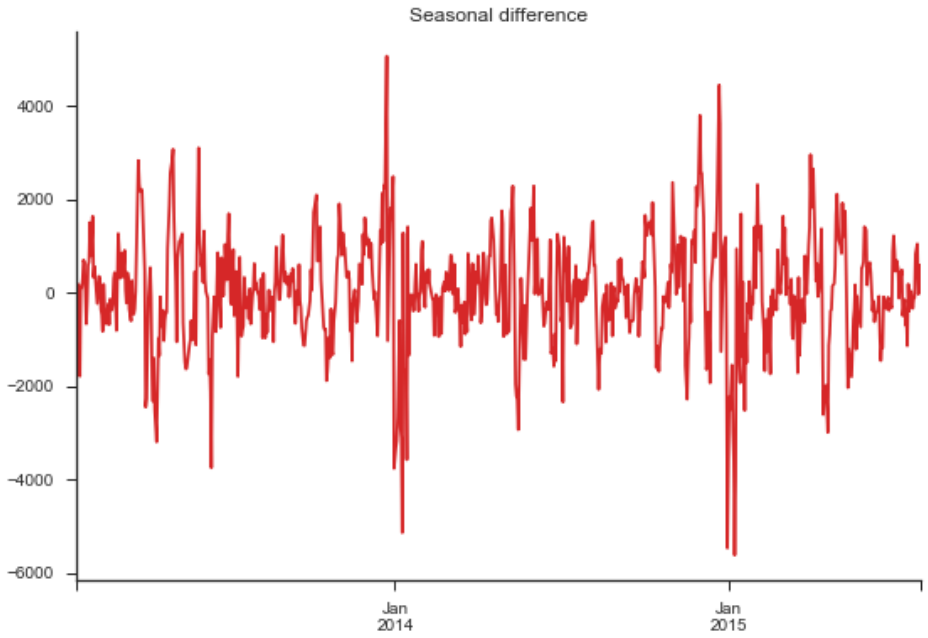

Figure 17: First Seasonal Difference

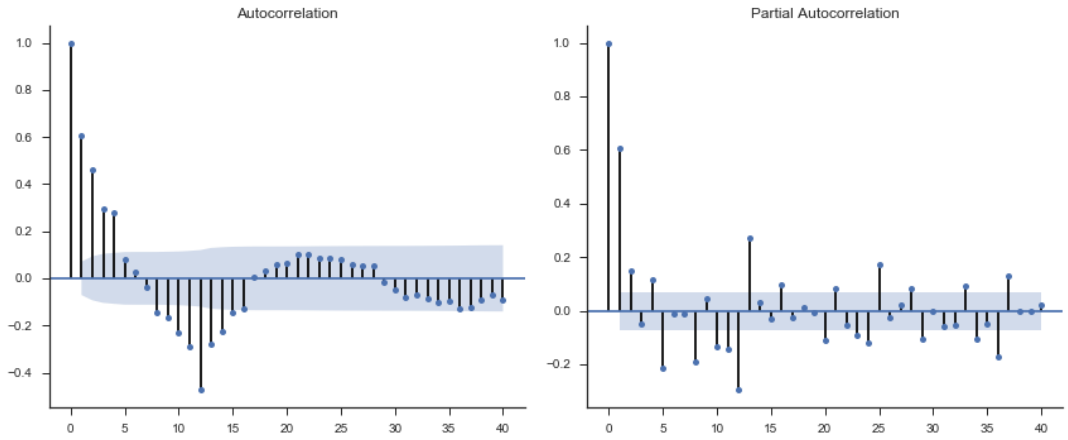

Figure 18: First Seasonal Difference ACF and PCF

The first and seasonally differenced series plots suggest that an AR(2) model should be used given that the autocorrelation in Figure 20 decreases gradually and the partial autocorrelations have a lag of 2. However notably there is no seasonal AR or MA pattern to be noticed from these plots.

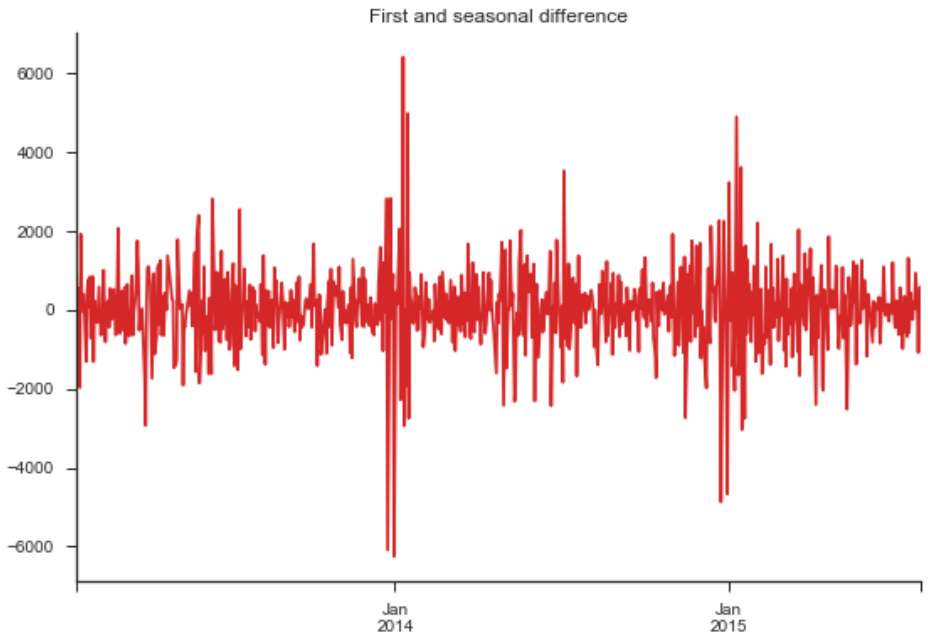

Figure 19: First and First Seasonal Difference

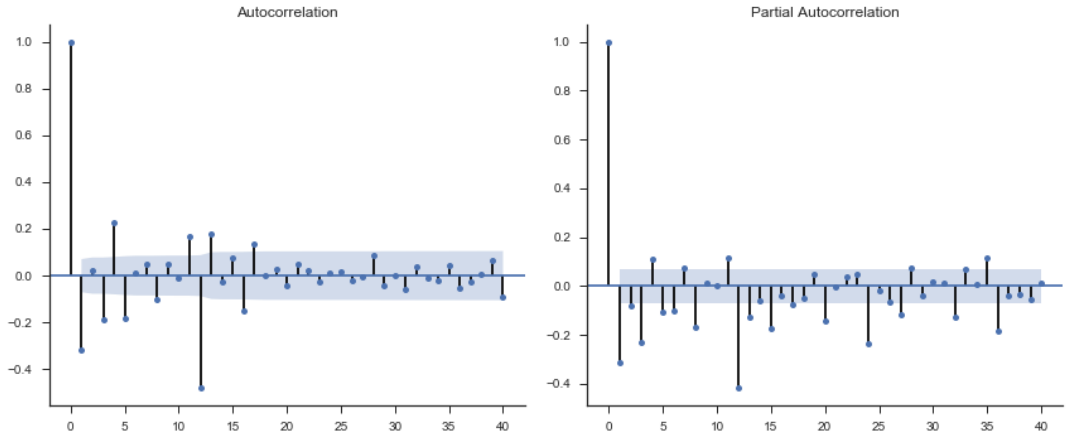

Figure 20: First and First Seasonal Difference ACF and PCF

Figure 21 shows the residuals for the remaining autocorrelations after the AR(2) model has been fit. The plots here seem to suggest a seasonal MA(1) model as there is significant lag in autocorrelations at lag 12, 25 and 36, and lag in partial autocorrelations at lag 12 and 24. This again confirms the seasonality of m=12.

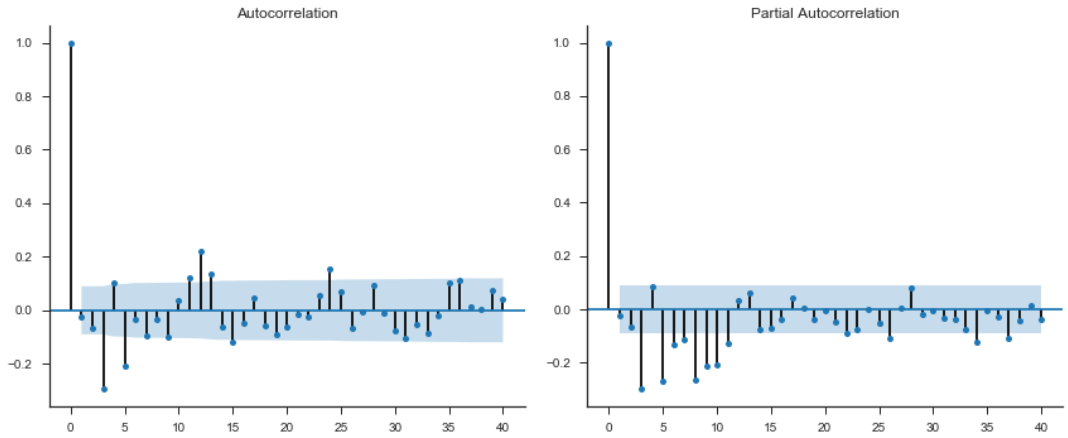

Figure 21: First and First Seasonal Difference, AR(2) ACF and PCF

C Additional Validation Results

Table 4: 2 Week Validation Results

Model	**MSE**	**MAPE**
Random Walk (Baseline)	696.267	10.490
SES	901.836	14.160
Additive Holt-Winters	729.000	14.929
Multiplicative Holt-Winters	672.710	13.193
Log Additive Holt-Winters	701.100	14.153
ARIMA (0,1,1)	689.788	10.453
Seasonal ARIMA (2,1,0)(0,1,1)	496.339	8.907

Table 5: 26 Week Validation Results

Model	**MSE**	**MAPE**
Random Walk (Baseline)	661.255	11.378
SES	922.091	15.653
Additive Holt-Winters	857.43	15.858
Multiplicative Holt-Winters	818.82	14.905
Log Additive Holt-Winters	867.32	15.997
ARIMA (0,1,1)	672.588	11.423
Seasonal ARIMA (2,1,0)(0,1,1)	639.677	11.309