Rossmann Sales Forecasting

Peter Broadstone, Keith Castelino, Eirik Fosnaes, Huy Le

June 06, 2019

Abstract

This paper details the methods and results obtained in forecasting sales for Rossmann, a leading German drug store. The forecasting approach involves the use of time series regression compared with benchmark methods using data provided on sales, customers, promotions and holidays for each Rossmann location.

Introduction

Forecasting sales is extremely important in the business world because of the effect that is has on so many parts of the business. If a company plans for more sales than they get, they are left with excess inventory, which can be very costly. If a company expects less business than they get, they may lose out on potential sales and in turn lose customers down the road. An accurate sales forecast allows a store to meet demands while also avoiding excess costs of keeping too much inventory in stock.

Sales forecasting is especially crucial when it comes to retail because of its fast-paced environment and the importance of having items in stock when customers need them. For our project, we are developing a sales forecast for Dirk Rossmann GmbH. Rossmann is Germany's second-largest drug store chain with over 3,000 stores in Europe.

We explored many different methods in our attempts to produce an accurate forecast for Rossmann. This included using best subset regression to select the best combination of variables and testing a time series regression model using the best model. We used MASE to evaluate the accuracy of our forecast.

We also looked at the data in different ways when developing our forecasts. We first produced a forecast with all the stores grouped together. We then looked at individual stores to test our forecast. We picked one store that was not open on Sundays (Store 1) and one store that was open every Sunday (Store 85).

Literature Review

With the forecasting of retail sales being so important to a business' success, there have been many studies in the past that have attempted to effectively forecast retail sales. Many have forecasted at the product level, like in the article[1] "Retail Demand Management: Forecasting, Assortment Planning and Pricing" (Vaidyanathan, 2011). Many have forecasted at the company level or even at the national level, for example[2] "Forecasting Aggregate Retail Sales: The Case of South Africa" (Aye, Balcilar, Gupta and Majumdar, 2015).

The article[3] "Forecasting time series with complex seasonal patterns using exponential smoothing" (Livera and Hyndman, 2009) introduces and discusses an exponential smoothing model that improves upon previous exponential smoothing models. The new modeling framework incorporates Box-Cox transformations, Fourier series, and ARMA error correction. The model introduced in the article has proven to be useful in multiple studies where other forecasting methods were not effective

We felt that it would be important to get some background on different methods for forecasting retail sales and how they compare. An article that discusses methods of forecasting that we found interesting is "Forecasting with Statistical Models and a Case Study of Retail Sales" (Bechter and Rutner, 1978). The article discusses methods of forecasting[4] such as economic models, time series models, and ARIMA models and compares the forecasting accuracy of ARIMA with that of two economic models designed for forecasting

retail sales. The results of comparison showed that ARIMA did not forecast retail sales any better than the naïve model and the economic models were superior

Description of Data

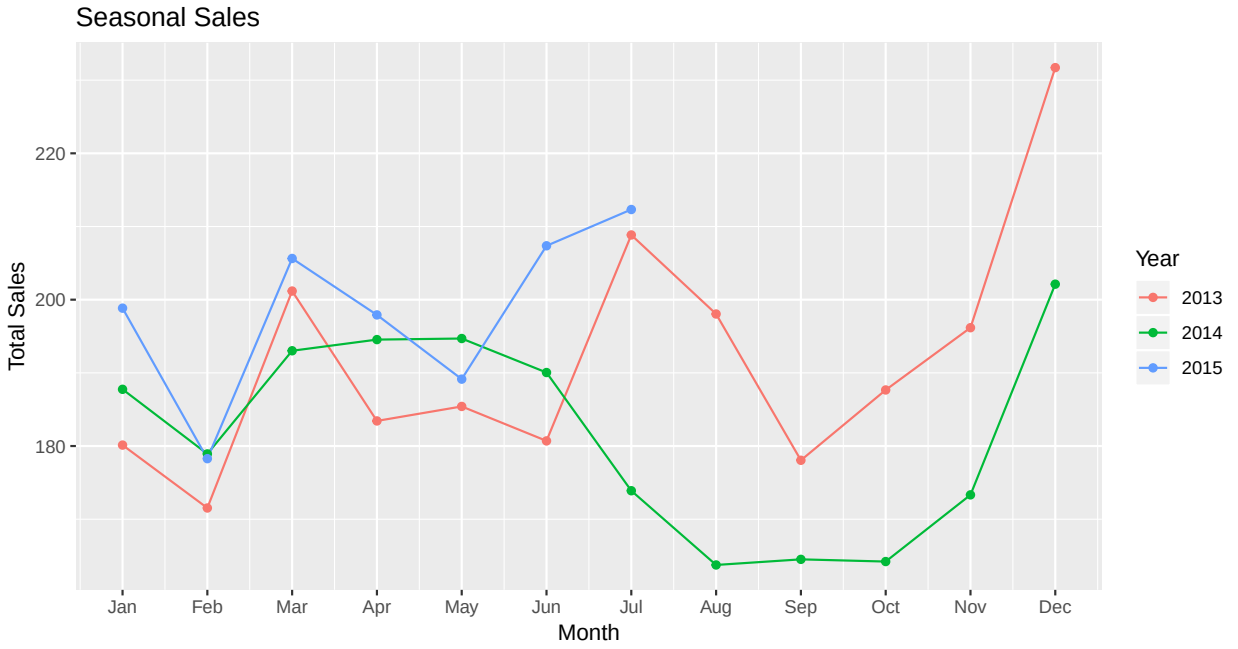
The dataset we use for this project is from the Kaggle competition Rossmann Store Sales - Forecast sales using store, and promotion data. The dataset was provided with data on 1115 stores located across Germany. The data included daily sales records of 942 days for each store from 1st Jan 2013 to 31st July 2015, each record has nine variables, the description of the data set is shown in the table below:

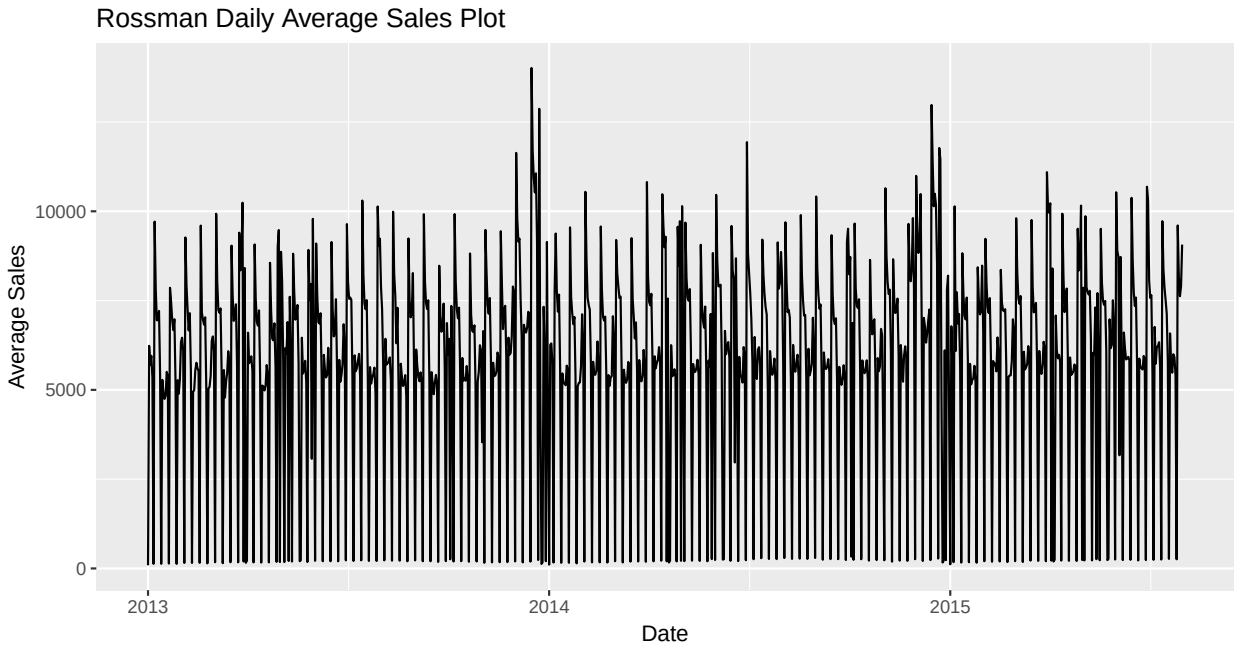
- Store: Each store in the dataset has a unique ID
- DayOfWeek: Varies from 1 to 7 corresponding to a week going from Monday to Sunday
- Date: Sales date
- Sales: The turnover of a store on given day
- Customers: The number of customers who visited the store on given day
- Open: Indicates whether a store is open (1) or closed (0); some stores are closed on Sundays
- Promo: Indicates whether a store is running a promo (1) or not (0) on any given day
- StateHoliday: Indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends; holidays are either public holidays (a), Easter (b), Christmas (c) or not a holiday at all (0)
- SchoolHoliday: Indicates if a store was affected by the closure of public schools

There are no missing values in the dataset. However, there are 180 stores missing 184 days of data in the middle of the series from 1st July 2014 to 31st Dec 2014.

Sales on a given day are a maximum of EUR 41,551 as there are days when stores were closed in the dataset corresponding to the number of customers visited on that day. A total of 17% of sales records are on closed days. Over the period, Rossmann store only 38% applied promotion. It looks like school holiday does not has a significant effect on Rossmann Sales since there is only 18% of the sales record were affected by the closure of public schools.

Exploratory analysis indicated that there are strong seasonal patterns in the dataset - weekly seasonality as well as annual seasonality. The sales performance for 2013, 2014 and 2015 have the same patterns. There is no apparent trend in the data over this time period.





Below are descriptive statistics of sales and customers by each day of the week. The first table represents the mean, median, standard deviation, minimum and maximum for each store by day of the week with two different variables. The second table represents the same metrics, but the numbers are derived from the total sum of customers and sales by day for all stores combined.

The first table shows that the most customers enter the store on Monday's. This is backed up by having the highest mean and median. This makes sense, because most stores are closed on Sunday's and people are most likely buying things on Monday's that they might have bought on Sunday's but can't because stores are closed. Monday is also the day with the highest sale and shows that the customers entering the store spend equivalent to what they do on other days.

Sunday is for natural reasons the worst performing day, given that most stores are closed on this day. The standard deviation is quite large compared to the mean because some stores are allowed to have open on

$$\widehat{y_t} = \beta_0 + \sum_{i=1}^{m} \beta_i X_{ti}$$

Figure 1: Time Series Linear Regression

Sunday's and therefore have normal sales data on this day.

Interestingly, Saturday is the second worst performing day. This is strange because it is natural to believe that customers are aware that stores will be closed the following day and would want to buy necessary items the day before this.

Descriptive Statistics by Day Of Week - Quantity						
Day of Week	Variable	Mean	Median	St.Dev	Min	Max
Friday	Customers	743.2	682	410.1	0	5,494
Monday	Customers	813.1	748	449.6	0	5,387
Saturday	Customers	657.1	571	387.8	0	4,762
Sunday	Customers	35.8	0	284.6	0	5,145
Thursday	Customers	697.5	646	416.1	0	5,297
Tuesday	Customers	761.0	680	396.3	0	7,388
Wednesday	Customers	721.6	651	385.9	0	5,106
Friday	Sales	6,723.3	6,434	3,101.0	0	38,722
Monday	Sales	7,809.0	7,310	4,016.5	0	41,551
Saturday	Sales	5,847.6	5,410	2,874.0	0	31,683
Sunday	Sales	204.2	0	1,613.2	0	37,376
Thursday	Sales	6,247.6	6,020	3,209.8	0	38,367
Tuesday	Sales	7,005.2	6,463	3,142.0	0	34,692
Wednesday	Sales	6,555.9	6,133	2,944.4	0	33,151

[a] Number of records in Table: 1017209

Econometric Model

We used a multilinear regression model to predict the total daily sales of Rossmann stores. The model was trained on the data which was aggregated from all of Rossmann stores excluding 180 stores which don't have complete data. This training set is the sales data from January 1 2013 to June 19 2015. Then we used model to predict the total sales of Rossmann stores up to 6 weeks in advance (from June 20 2015 to July 31 2015). After that we compared the result from multilinear regression model to the result from Average, Naïve, Seasonal Naïve and Drift Method.

The econometric method that we have finalized for this paper is based on time series regression. We will be using the best subset regression method to arrive at an ideal model, by optimizing for AIC. The mathematical formulation for a time series regression is as follows:

The model we select for this paper will used to forecast sales for all of Rossmann's stores as a whole. We will also address the use case for individual stores to show what happens when we run a model on only one store at a time.

In order to ensure that the model we select is at least better than the most simplest methods, we are comparing our model against the following simple forecasting methods:

- Average Method, which uses the mean of the training period as a forecast

- Naive Method, which uses the last observation as forecast
- Seasonal Naive Method, which uses the last observation from the same seasonality
- Drift Method, which assumes that the trend from the past will continue in the same vein. It derives a forecast using the following mathematical formula:

First, We used linear regression with all variables customer, promo, open, StateHoliday, SchoolHoliday, and Day_of_week to obtain a baseline for predicted model. We did not use trend and seasonal dummy variable since base on the time series plot there is no trend in the data and seasonal dummy variables are colinear with Day_of_weeks. The result of first model is show in the figure below:

Summary of Time Series Linear Regression				
term	estimate	std.error	statistic	p.value
(Intercept)	-2.330912e+05	2.820928e+04	-8.262927	0.0000000
customers	1.186935e+01	1.422232e-01	83.455831	0.0000000
day_of_week	1.868771e+04	7.275641e+03	2.568531	0.0104124
open	-2.160737e+06	1.060461e+05	-20.375443	0.0000000
promo	4.113381e+05	3.073906e+04	13.381610	0.0000000
state_holiday	-1.129942e+05	2.806845e+04	-4.025667	0.0000628

[a] Number of records in Table: 730

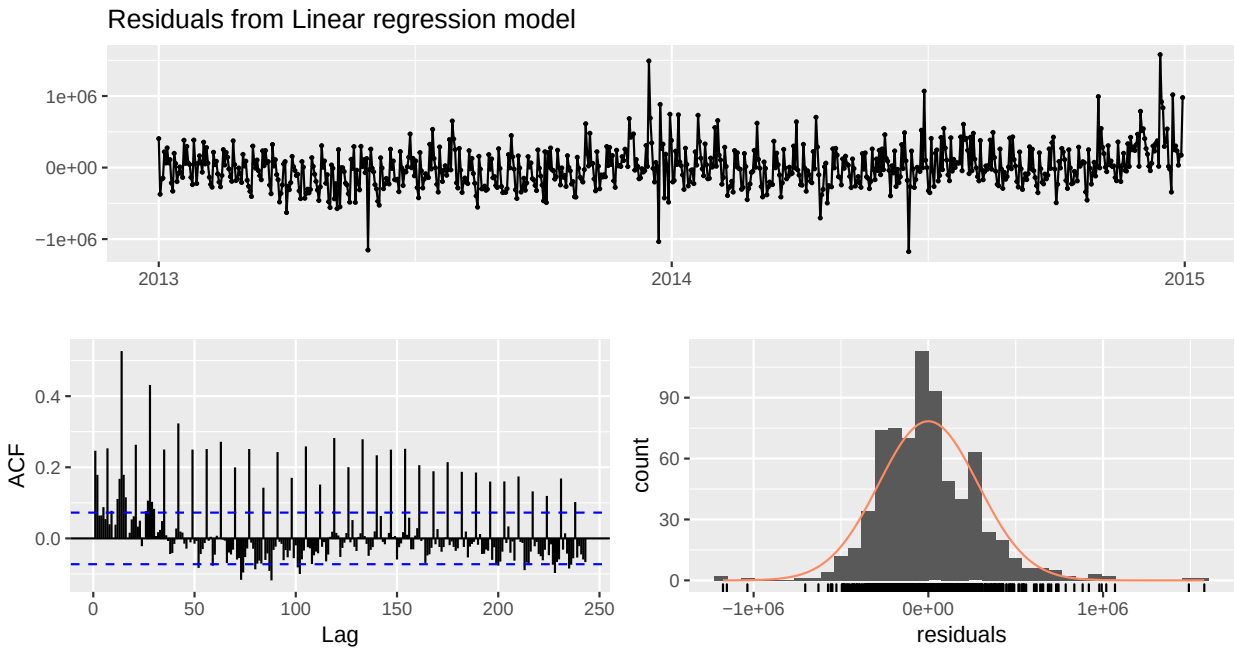
We then used the olsrr library and applied the best subset regression for variable selection. We used AIC and R-Squared measures as the metric for choosing predictors. We want to find the model with the lowest value in AIC and R-Squared, we seek the model with the highest value. The error metric we used the measure the model accuracy was the root mean square error – RMSE.

Results

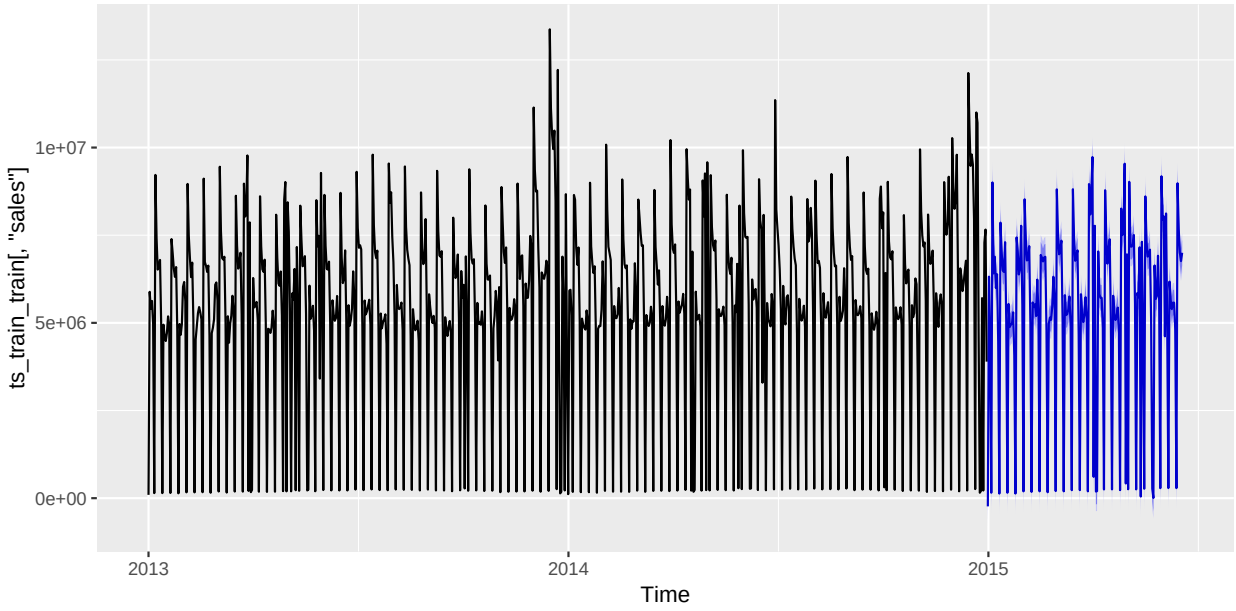
By optimizing for the lowest value of AIC in the best subset regression, we have arrived at a model (8) that takes into account the following variables:

- customers
- sales day Monday
- sales day Thursday
- sales day Friday
- sales day Saturday
- whether the majority of stores are open or not
- whether promos are run are not
- whether the majority of stores were affected by a state holiday

The above model tells us when everything is kept constant, the beta estimates for every variables are significant at a 5% level. The Adj-R2 of the model is 0.9915 that mean our model can explain 99% of total sales of Rossmann Stores. The sales are heavily influenced by the day of the week, due to which the model predicts that stores lose more than EUR 2,000,000 in revenue when they are open, everything else remaining constant. Sales are also shown to increase by EUR 546,000 when promos are run. In addition, Monday and Saturday will have higher sales than other day of weeks since the Sales Increase are 88,900 and 400,000 on Monday and Saturday respectively. On the other hand, revenue falls by more than EUR 87,900 when there is a state holiday. In comparing the regression model we have used with the benchmark methods, we can confirm that this model performs far better than any of the benchmark methods that is measured against, based on a MASE of 0.12 and RMSE 406562.5, which is the lowest of all models used.




```
##
##  Breusch-Godfrey test for serial correlation of order up to 146
##
## data:  Residuals from Linear regression model
## LM test = 397.18, df = 146, p-value < 2.2e-16
```



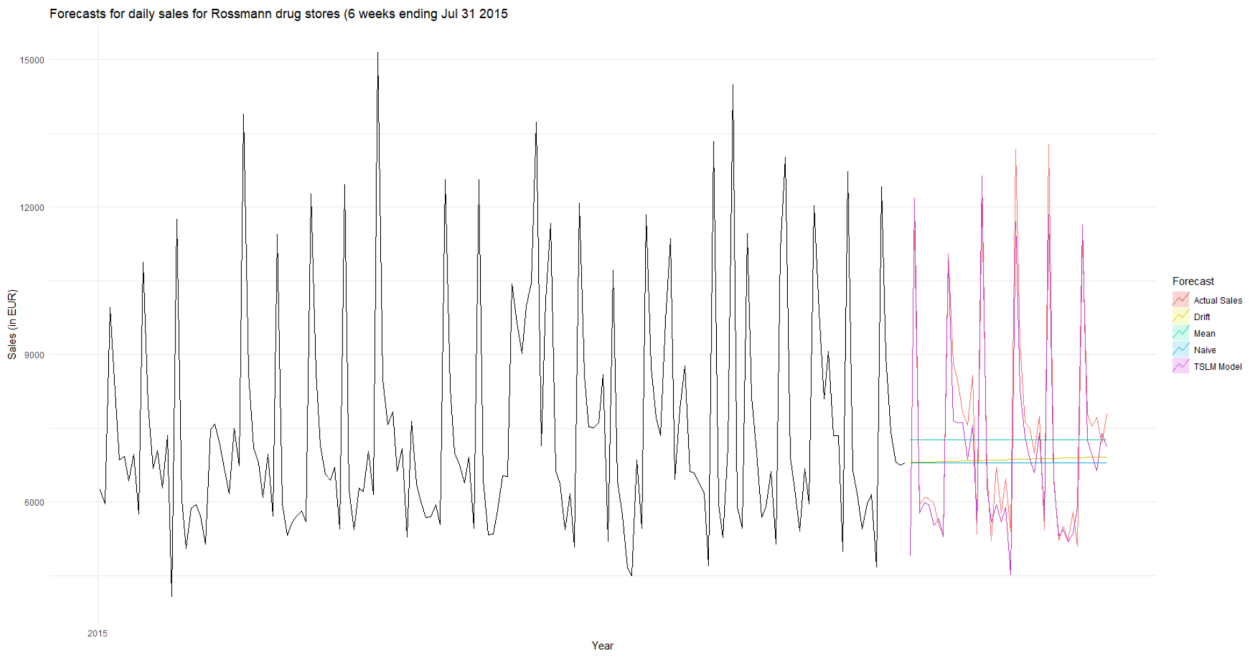


Figure 2: Forecast

	Mean method	Naive Method	Drift Method	Regression
ME	1.269525e+05	1.648934e+06	1.200751e+06	1.601708e+05
RMSE	2.729343e+06	3.186248e+06	2.984746e+06	3.408150e+05
MAE	2.012705e+06	2.911095e+06	2.633872e+06	2.274773e+05
MPE	-3.835633e+02	-2.476669e+02	-2.870120e+02	2.576582e+00
MAPE	4.108391e+02	3.123252e+02	3.405614e+02	7.159274e+00
MASE	7.606448e-01	1.100166e+00	9.953972e-01	8.596860e-02
ACF1	-5.967450e-02	-5.967450e-02	-5.480540e-02	2.024930e-01
Theil's U	4.092851e-01	5.550436e-01	5.138016e-01	6.715150e-02

As for what happens when we run this model on only one store, we have two separate scenarios - one for a store not open on Sundays and another for a store open on Sundays. Using Store 1 as an example of the former, we find that our best model consists of the variables customers, day of the week and promo. Using Store 85 as an example of the latter, we find that only customers and promo are part of our best model. In each case, the regression model performs better than the benchmarks when optimizing for the lowest MASE.

The forecast we have arrived at using the test set is very close to the actual sales values for the test time period of 6 weeks ending July 31 2015, as illustrated below.

Conclusion

In conclusion we ended up using alinear regression model, which gave us the lowest MASE and RMSE. The graph clearly shows that this was the highest performing method, because it had the lowest training error.

This model also achieved the best balance between overfitting and underfitting. It makes sense that this model was the highest performer because it takes trend, sesonality and variation into account. The mean and naive models are not doing this and because of this they will draw a straight line going forward. Given

that our data fluctuates so much, these methods did not provide us with a great forecast. As seen from the daily sales forecast graph the TSLM Model is following actual sales very well and appear to follow the fluctuation in sales.

There were several variables that weren't included in our dataset that we felt would have helped us formulate a more effective forecasting model. For example, we are interested in a variable that gauges how well the economy is doing throughout the data. Although many of the purchases made at Rossmann may be necessities, we still feel that if the economy does better, sales will also increase. We also feel that some of the variables used could be expanded upon. Promotion proved to be useful in our final model, but it would have been even more valuable to know the type of promotion. While it is not always reasonable to have all the information you want, we feel that we could have produced a more effective forecasting model with more information.

In terms of how this paper could be used for other situations, this paper could be used to predict retail drugstore sales in different countries as well. Most countries in Europe have similar purchasing patterns and it would be interesting to see how effective our model would be in different countries.

Sources

[1] Retail Demand Management: Forecasting, Assortment Planning and Pricing

[2] Forecasting Aggregate Retail Sales: The Case of South Africa

[3] Forecasting time series with complex seasonal patterns using exponential smoothing

[4] Forecasting With Statistical Models and a Case Study of Retail Sales