Forecasting future sales at La Favorita supermarket

Fouad Yared
Hunter College, Stat 790: Case Seminar, Fall 2020

Introduction

Businesses have a strong incentive to develop accurate forecasting models. As noted in Harvard Business Review, "Consistently accurate sales forecasts are *gold*. They deliver the revenue predictability that is essential for companies to accelerate their growth and success. Unfortunately, consistently accurate sales forecasts are rare."

This paper will focus on forecasting the number of products sold at La Favorita, a large supermarket chain in Ecuador. (The data set is publicly available on Kaggle. [1] The train data set, used in this paper, begins January 1, 2013 and ends August 15, 2017.)

Data preparation

To prepare the data, it is important to know the questions of interest to aggregate sales data in an appropriate way. Are we interested in daily, weekly, or monthly forecasts? Are we interested in overall product sales, product sales by category, or product sales of individual items? Are we interested in the sales of a store, a region of stores, or all stores?

- This paper will focus on forecasting daily sales by category at "Store 1" of La Favorita.
- The exploratory analysis will look at daily, monthly, quarterly, and yearly aggregations of sales data, both overall and for each product category.

The main questions of interest are:

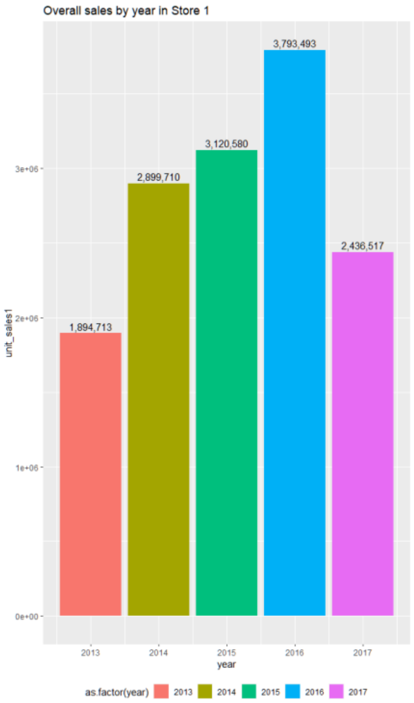
- What time-related patterns do we see in the data?
- Do overall or product category sales tend to follow a similar pattern across time?
- What are the best models to predict overall or product category sales?

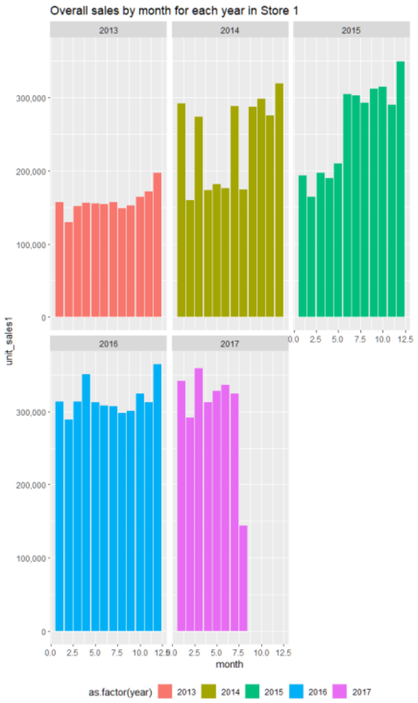
Exploratory analysis

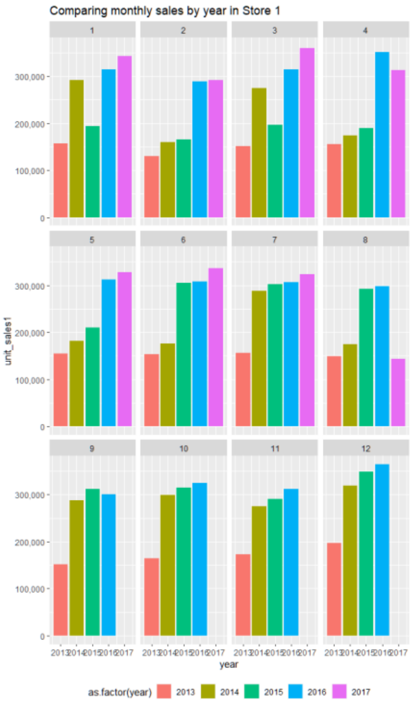
Although sales data will be modeled by day, we will first explore overall patterns in the data. The left chart below shows that the number of products sold each year increased from 2013 to 2016. The largest increases occurred between 2013-2014 and 2015-2016. (Data for 2017 end on August 15, 2017; the other years have data for the entire year.)

The middle chart below shows monthly sales across years. Sales in 2013 gradually rise, sales in 2014 fluctuate, and sales in 2015 have a dramatic rise that remains until we have incomplete data in August 2017. The large increase in yearly sales between 2013-2014 is attributed to the large increase of sales in fall and winter months (January, September-December 2014). The large increase in yearly sales between 2015-2016 is attributed to the large increase in late winter and spring months (January-May 2016).

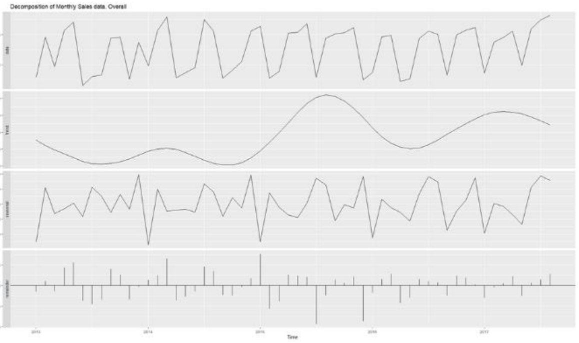
When comparing the same months from 2013 to 2017, we see that monthly sales tend to rise over time (e.g., sales in February increased each year, although to different degrees).

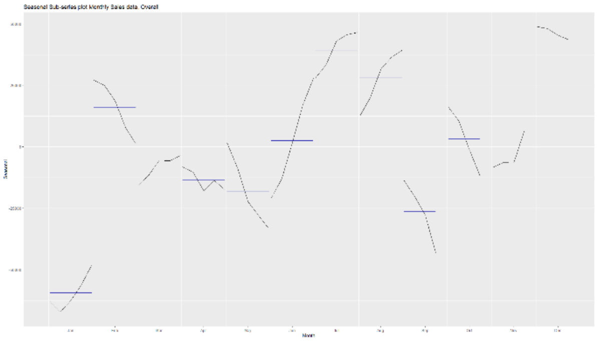


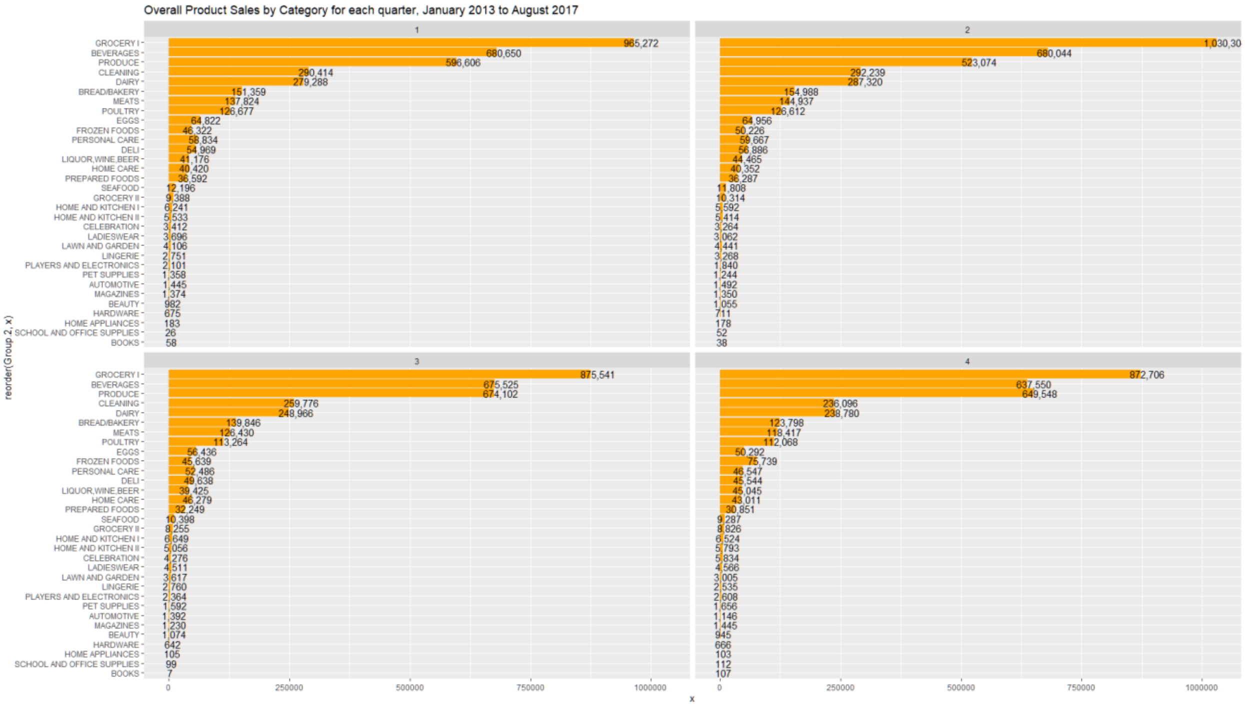




The plot on the left of the image below looks at the X11 decomposition of the overall trend, seasonal patterns, and remaining factors of the monthly sales data. The seasonal pattern shows fluctuations throughout the year and the largest decrease occurs between December and January. The remainder plot shows that unusual observations occurred in January, July, and December 2015. The seasonal sub-series plot, shown in the image below on the right, shows the rise and fall of monthly sales. Sales tend to rise from January to March, fall from April to May, rise dramatically from June to August, fall from September to October, then rise in November to December (where the most sales in a single month occur).

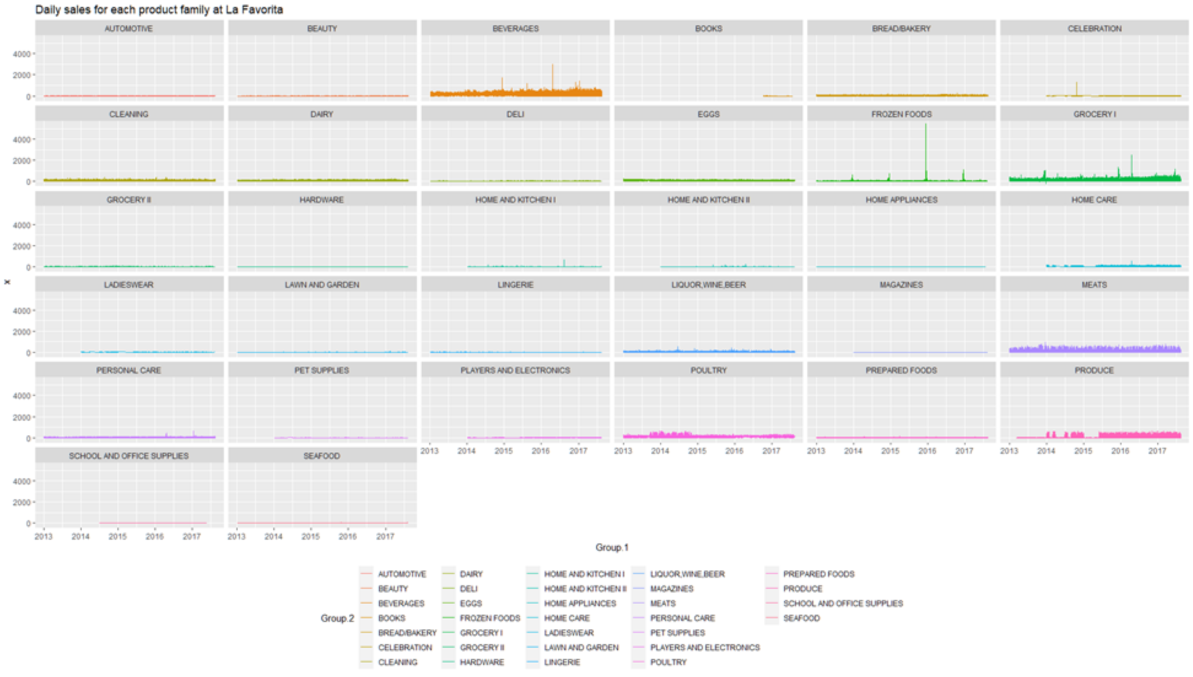




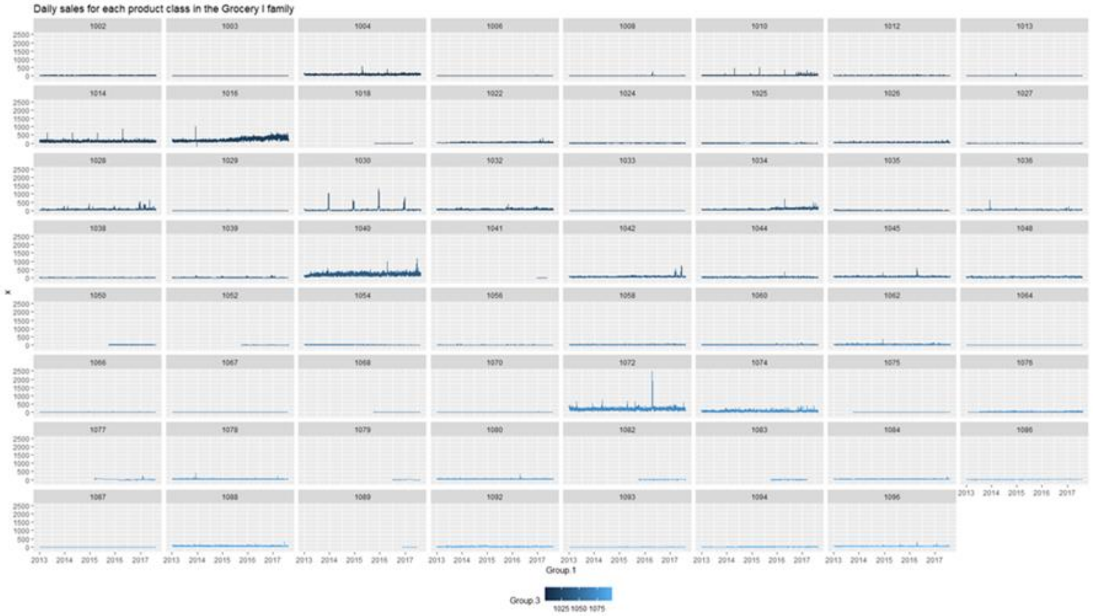


Quarterly sales data for each product category are above. Sales in specific product categories tend to be consistent across different quarters. Categories selling the most products in Q1 - Grocery I, Beverages, Produce, Cleaning, and Dairy – retain their positions each quarter.

The plot below shows the daily sales data for each product category. Product sales tend to be consistent across time. There are large spikes for the Beverage and Grocery I categories which may be attributed to holidays. It is unclear why there was a very large number of Frozen Food sales in early January 2016.



Daily sales of product classes within the Grocery I product family are plotted below. (The names of the classes were excluded from the data set. Instead, numbers were provided.) Some classes - specifically 1016, 1040, and 1072 - tend to sell many units while most other classes do not sell much at all. Although sales of product classes (or individual items) will not be modeled, product classes are analyzed to better understand how, when aggregated, they form product categories.

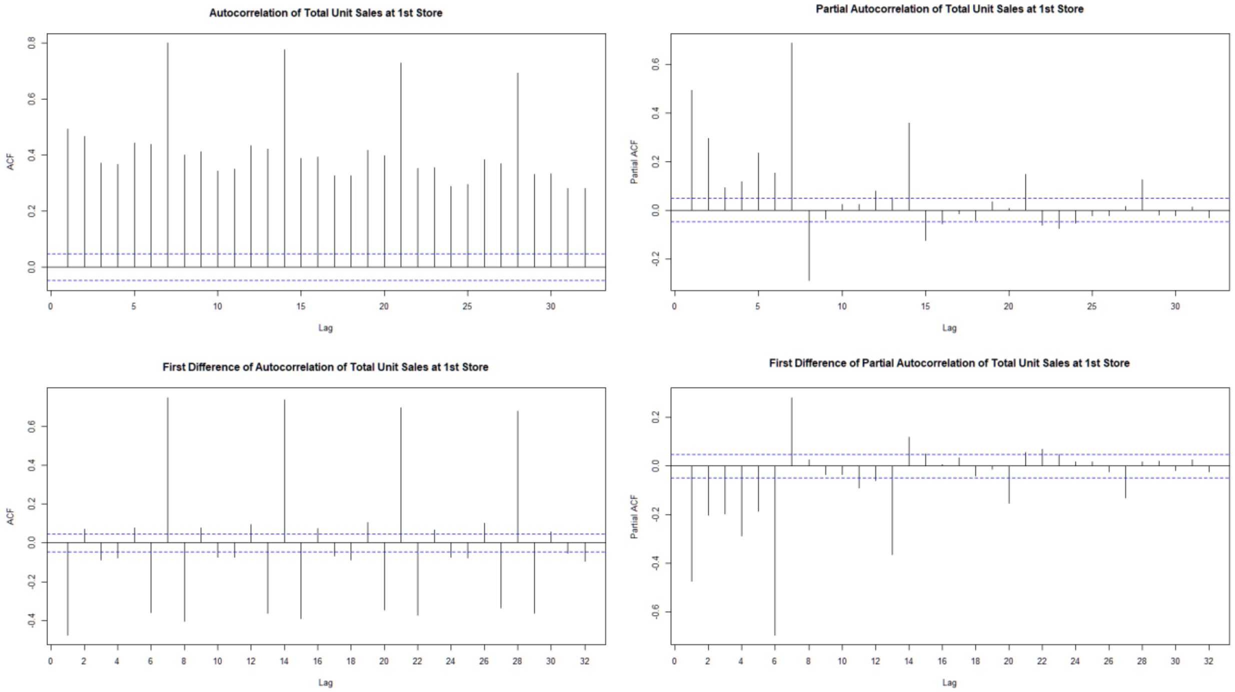


Autocorrelation plots were used to evaluate the relationship of sales data for a particular day against its lagged values (sales data for one to thirty-one days prior). The autocorrelation graph in the top-left panel in the image below shows that present day sales data tend to be related mostly to seven days prior sales data and other intervals of seven days (14, 21, 28 days).

Partial Autocorrelation adjusts the autocorrelation plot by removing the correlation from shorter lags. The partial autocorrelation can be found in the top-right of the below image. We see that within the first week, values tend to decline until seven day lag. The partial autocorrelation plot reduces some of the large lags found in the autocorrelation plot.

To evaluate whether the daily sales data has stationarity (a constant mean, variance, and autocorrelation structure over time), we use the KPSS Unit Root Test, which objectively determines whether differencing is required. The test-statistic value is 122.94 and the 2.5% critical value is 0.574. Since the test-statistic value is greater than the critical value, we determine that differencing is required. Only one difference is required as the root test produces a test-statistic less than the critical value when testing the differenced sales data.

Differencing is used to make the data stationary by computing differences between adjacent observations. The autocorrelation and partial autocorrelation plots are found on the bottom-left and bottom-right panels in the below graphic. The differenced autocorrelation plot shows a seven-day recurring pattern around days 1, 6, and 7. The differenced partial autocorrelation plot becomes less negative then oscillates above and below zero.

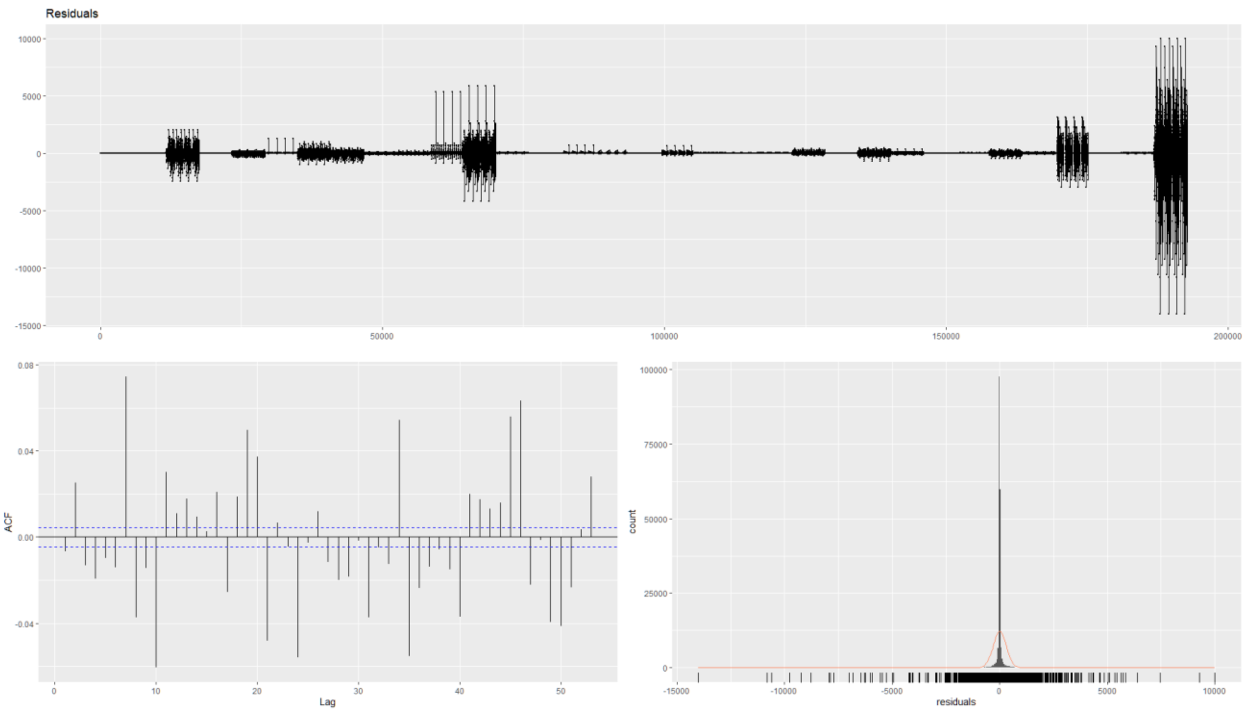


Methods: Model building

Hierarchical time series models were used to forecast the number of products sold both overall and by category. Models were trained on data from January 1, 2013 to December 31, 2016 and forecasts were evaluated from January 1, 2017 to July 31, 2017.

Many different models were used to forecast the first seven months of data in 2017. Several Arima models were fit to the data using both fixed and flexible parameters. (In an Arima model, p is the number of autoregressive terms, d is the number of nonseasonal differences for stationarity, and q is the number of items used for the moving average. Seasonal adjustments, which include values of P, D, and Q, were not evaluated. Auto.Arima models find values for p, d, and q that minimize AIC; a model may select a different set of parameters for each category.) Other models include ETS (which looks at seasonality, trends, and error), a random walk model, a croston model (which performs simple exponential smoothing), a mean model, and a time series linear model (which uses trend and seasonal characteristics to fit a linear model).

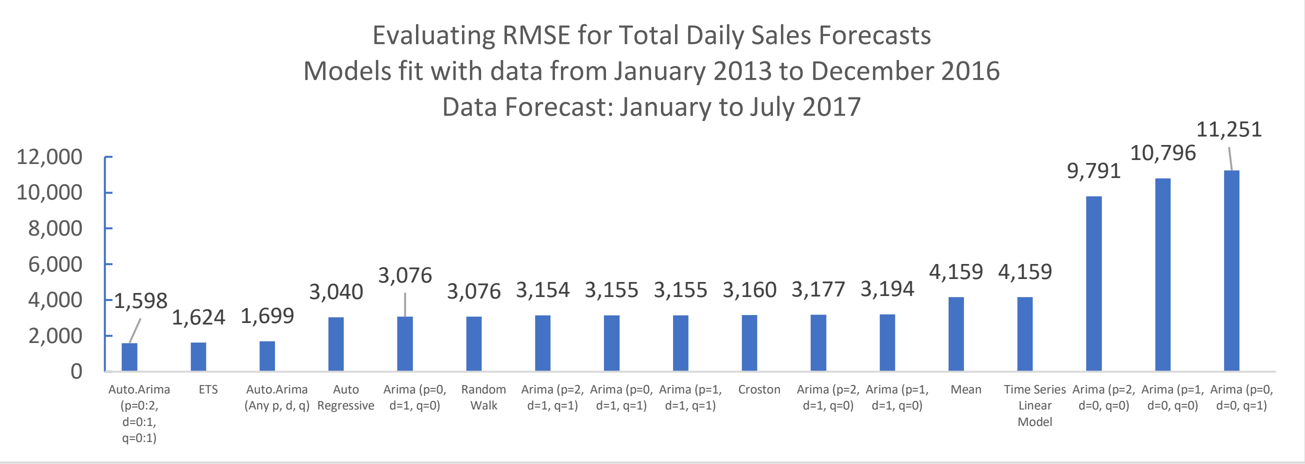
Residuals for the ETS model are shown below. These include residuals plotted against time, the autocorrelation of residuals, and a histogram of residuals. The residual plots of the ETS, constrained Auto.Arima, and unconstrained Auto.Arima functions appear very similar. The residual plot against time shows that residuals tend to be very small or large which may be attributed to certain holidays or outliers for individual days. The autocorrelation plot of residuals shows a similar autocorrelation plot to the one described earlier after taking the first difference of the data. The histogram of residuals reveals there are some very large residuals.



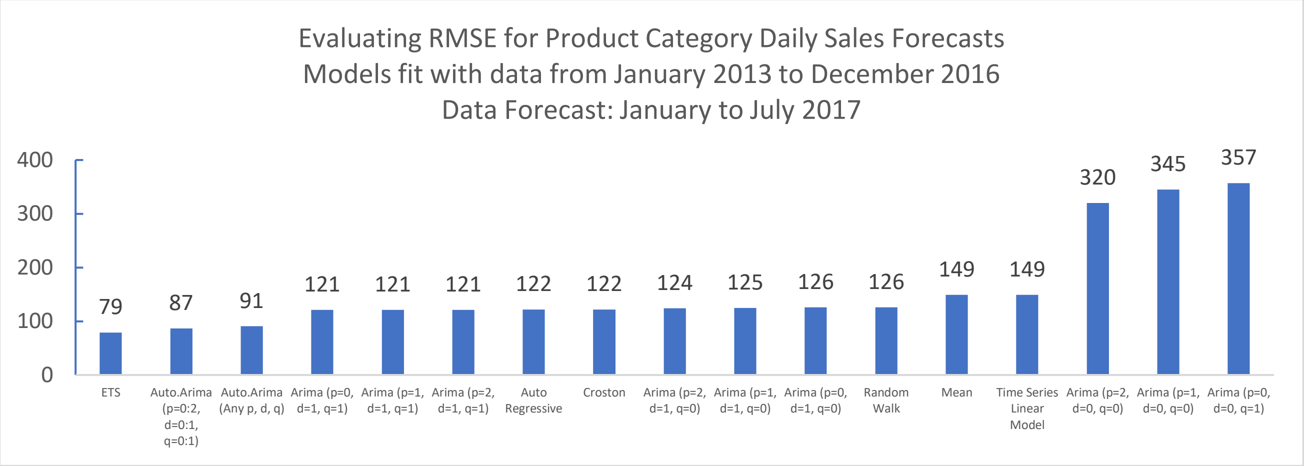
Results: Model evaluation

The root mean squared error (RMSE) was used to evaluate the total daily sales forecasts of different hierarchical models. The three models that performed far better than all other models were an Auto.Arima model that had constraints, the ETS model, and an Auto.Arima model without constraints. The constraints of the Auto.Arima model were p between 0 and 2, d between 0 and 1, and q between 0 and 1. The Auto.Arima function performed better when its values were constrained. Notably, the Arima functions that had a difference of 1 and a moving average value of 1 performed much better than those that did not. While most variables did better than the Mean prediction, only four did better than a random walk (which is the same as an Arima p=0, d=1, q=0).

It is worth noting that two ARIMA models failed to run and are excluded from this evaluation. These include ARIMA models with pdq values of 1,0,1 and 2,0,1.



A similar pattern emerges when evaluating the RMSE of daily sales forecasts of product categories. ETS performs best, followed by the constrained Auto.Arima, and the unconstrained Auto.Arima models.



Discussion

Time-related patterns found in Store 1 of the La Favorita data show that the number of products sold tends to increase each year (except for 2017 which does not have complete data). Years that saw the biggest increases in the number of sales were 2013-2014 and 2015-2016. Compared to the prior year, the largest increases in 2014 were in January and from September-December and the largest increases in 2016 were in January-May. Further research would need to be done to find out why there was a large increase in sales in these months. It may be attributed to seasonal product promotions or other factors.

After removing seasonal effects of monthly sales data – which tend to be highest in December and June, lowest in January, and consistent across other months – we see that monthly sales tended to increase most in the first half of 2015 then level off in the second half of 2015 until March 2016.

In looking at quarterly sales by product category, we see that the number of products sold in each category is consistent across different quarters. The categories that sell the most each quarter are, ranked from first largest to fifth largest: Grocery I, Beverages, Produce, Cleaning, and Dairy.

When evaluating product category sales by day, we see that figures are generally consistent over time with a few relatively large exceptions in the Beverage, Grocery I, and Frozen Foods categories. Further analysis can look into whether the large spikes are attributed to holidays or something unusual (regarding the large number of Frozen Food products sold in early January 2016).

Daily sales data were found to have high autocorrelations with seven-day lags. Taking partial autocorrelations removed the correlations of shorter lags, which led to the seven-daay recurring pattering of lags 1, 6, and 7. A unit root test was used to confirm the need for differencing since the daily sales data was found to be non-stationary. A one-level difference between adjacent observations was used to make the data stationary in order to aid model development.

Hierarchical time series models were fit on data aggregated by product category. The date range of the models fit was from January 1, 2013 to December 31, 2016. Residuals tended to be related to different time intervals, so they will need to be further evaluated. Models were forecast for the next seven months (January 1, 2017 to July 31, 2017) and were evaluated using RMSE. Models that performed best when evaluating RMSE for Total Daily Sales and Product Category Sales were the constrained Auto.Arima model (p=0:2, d=0:1, q=0:1), the ETS model, and the unconstrained Auto.Arima model (any p, d, q). Arima models with a differencing component (d=1) performed much better than those without it (confirming the need to take a difference of the data for stationarity).

Sales forecasts can be further analyzed using the three models listed above along with other covariates which may affect product sales, including whether a particular item was on sale and the oil price. Comparing the accuracy of predictions at different time increments could be helpful to forecast finer levels of product sales (e.g., sales of product classes or individual items within product categories) since daily sales for many product classes or individual items are zero and are hard to produce estimates for. Data for more La Favorita stores can also be evaluated to see how consistent the sales trends are across different stores.

References

Data Set

1. Corporación Favorita (2017). Corporación Favorita Grocery Sales Forecasting [Data set]. Kaggle. https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data

Other References (Listed Alphabetically)

2. Atrebas. (2019). *A data.table and dplyr tour.* https://atrebas.github.io/post/2019-03-03-datatable-dplyr/#summary Accessed on 12/14/20.
3. Cryer, J. D. & Chan, K-S. (2008). *Time Series Analysis: With Applications in R.* Springer Science. https://link.springer.com/book/10.1007/978-0-387-75959-3
4. Cowpertwait, P. S. P. & Metcalfe, A. V. (2009). *Introductory Time Series with R.* Springer Science. https://www.springer.com/gp/book/9780387886978
5. Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice, 2nd edition*, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 12/14/20.
6. Hyndman, R.J., & Athanasopoulos, G. (2019) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 12/14/20.
7. *NIST/SEMATECH e-Handbook of Statistical Methods*. Stationarity. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm Accessed on 12/14/20.
8. Shipley, L. (2019, August 28). *How to Make Your Sales Forecasts More Accurate.* Harvard Business Review. https://hbr.org/2019/08/how-to-make-your-sales-forecasts-more-accurate
9. Wang, E. (2020). *Introduction to tsibble.* Rstudio. https://cran.rstudio.com/web/packages/tsibble/vignettes/intro-tsibble.html