Use Case: Demand Forecasting

White Papers

OBJECTIVE

The objective of this project was to forecast the future quarterly revenues of one of the leading CPG companies in the United States. The forecast horizon for this project was Q2, Q3, and Q4 of 2020. The underlying value of revenue forecasting stems from the ability to budget expenses more deliberately, plan for growth, improve production scheduling, understand customer behavior, and improve cash flow and credit management.
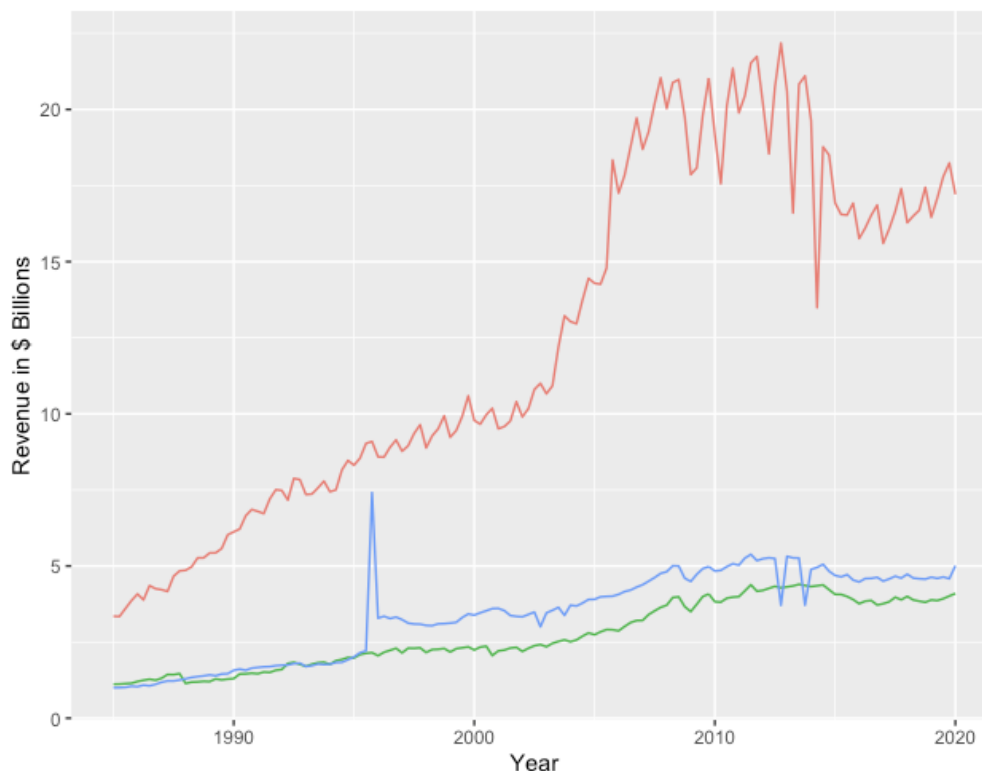
DATA

To carry out our objective, we extracted the company's quarterly revenue data from Q1 of 1985 to Q1 of 2020, resulting in a total of 141 evenly spaced and sequential observations. In addition, we also extracted the quarterly revenue data of the company's two main competitors for the same time period. The reason for this was to use the data of the other companies as inputs in our time series regression models.

PREPROCESSING

Since our dataset was extracted in a way that ensured it had no missing values, we did not have to overcome the obstacle of handling missing values in our data. It is crucial to have a dataset without missing values, especially for forecasting, as time series algorithms tend to function poorly with datasets that contain missing values.

However, we had to perform a few preprocessing steps to make the dataset usable for our time series algorithms. For instance, we sorted the records such that the first row contained the oldest observation and the last row contained the newest observation, as this step is necessary for fitting and forecasting.

TIME SERIES ELEMENTS

The blue line in the time series plot above represent our target company's revenue over time. The green and red lines correspond with the revenues of the company's two competitors. The sales data for all three companies have a general upward trend. Furthermore, it seems as though their quarterly sales are not seasonal. These conclusions are consistent with the following correlogram of ACFs of our target company. Note that the other two correlograms were almost identical in shape and magnitude.
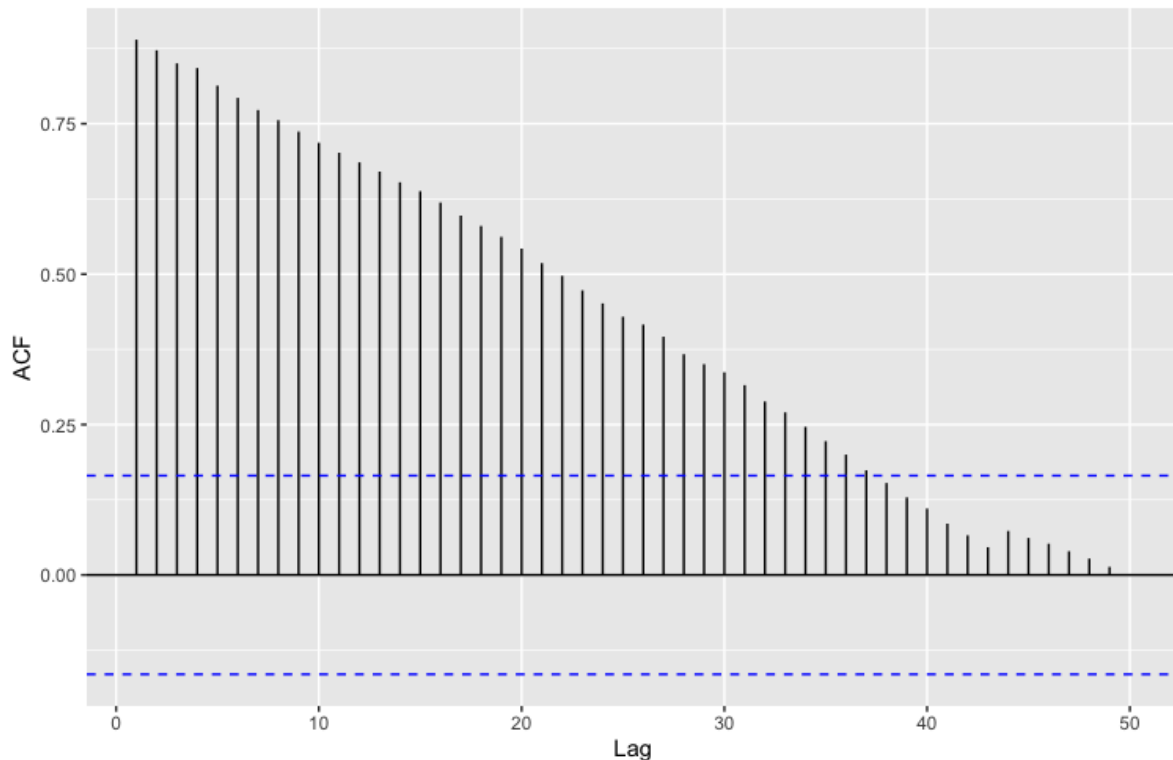
We can infer that the time series has a trend element, since the ACFs are close to 1 at the start, and they gradually decrease towards 0. In addition, since there are no spikes at multiples of 4 time lags (quarterly), we can conclude that there is no seasonality in this time series.

MODEL BUILDING

After splitting the dataset into a training set with the first 80% of the observations and a testing set with last 20% of the observations, we built a total of 38 forecasting models on the training data. We used the following algorithms to build the forecasting models:

- Naïve methods: naïve model, seasonal naïve model, and naïve model with drift
- Smoothing methods: simple average model, exponential smoothing model with optimal parameters and initial forecast values that minimize SSR, exponential smoothing models with tuned hyperparameter values of alpha, Holt-Winter's multiplicative model, Holt-Winter's additive model, and Holt method
- ARIMA Models: non-seasonal ARIMA(p,d,q) models fit with exhaustive search and rough search
- Linear Regression models: linear regression models with all possible combinations of competitor sales as inputs
- Linear Regression models with lagged predictors: linear regression models with all possible combinations of predictors lagged 1 quarter and 2 quarters for the two competitors

- **Non-linear Regression models:** after performing five non-linear transformations on each predictor (1/predictor, log(predictor), sqrt(predictor), predictor^2, and log-log exponential), we created non-linear regression models using each transformation on its own. We also combined the best transformation for each predictor.

MODEL EVALUATION

After building the models on the training set, we used the models to forecast the observations in the testing set. Subsequently, we evaluated the forecast values using five accuracy measures: Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percent Error (MPE), and Mean Absolute Percent Error (MAPE). The results are as follows:

S.No	Methods	ME	RMSE	MAE	MPE %	MAPE %
	FORECASTING ACCURACIES					
1	Seasonal Naïve	-0.17	0.63	0.56	-3.76	12.05
2	Naïve	0.99	1.04	0.99	20.80	20.80
3	Drift	0.63	0.75	0.63	12.92	13.10
4	Simple Average	1.64	1.66	1.64	34.53	34.53
5	Exponential Smoothing (Optimal)	0.01	0.30	0.20	-0.24	4.35
6	Exponential Smoothing (alpha = 0.1)	-0.13	0.33	0.26	-3.20	5.73
7	Exponential Smoothing (alpha = 0.3)	-0.06	0.30	0.23	-1.75	5.00
8	Exponential Smoothing (alpha = 0.5)	0.23	0.37	0.28	4.40	5.83
9	Exponential Smoothing (alpha = 0.7)	0.53	0.61	0.56	10.95	11.80
10	Exponential Smoothing (alpha = 0.9)	0.84	0.89	0.85	17.52	17.80
11	Holt Winters Additive	-0.77	0.91	0.83	-16.90	18.05
12	Holt Winters Multiplicative	-0.94	1.06	0.96	-20.53	21.01
13	Holt Method	-0.75	0.89	0.81	-16.50	17.63
14	ARIMA (0,1,1) - Exhaustive	-0.76	0.90	0.82	-16.64	17.77
15	ARIMA (0,1,1) - Rough Search	-0.76	0.90	0.82	-16.64	17.77
16	Reg: Comp1 + Comp2	-0.51	0.59	0.51	-11.19	11.19
17	Reg: Comp2	-0.56	0.64	0.56	-12.23	12.23
18	Reg: Comp1	0.31	0.54	0.42	6.13	9.05
19	Reg: Trend	-1.18	1.29	1.18	-25.80	25.80
20	Reg: Comp1Lag1 + Comp2Lag1	-0.48	0.59	0.48	-10.70	10.70
21	Reg: Comp1Lag2 + Comp2Lag2	-0.48	0.57	0.48	-10.46	10.46
22	Reg: Comp1Lag1 + Comp2Lag2	-0.47	0.59	0.47	-10.50	10.50
23	Reg: Comp1Lag2 + Comp2Lag1	-0.44	0.53	0.44	-9.65	9.65
24	Reg: Comp2Lag1	-0.61	0.70	0.61	-13.40	13.40
25	Reg: Comp1Lag1	0.24	0.52	0.38	4.75	8.28
26	Reg: Comp2Lag2	-0.68	0.76	0.68	-14.90	14.90
27	Reg: Comp1Lag2	0.16	0.38	0.30	3.29	6.48
28	Reg: 1/Comp1	0.51	0.60	0.56	10.54	11.82
29	Reg: Log(Comp1)	0.27	0.45	0.36	5.40	7.67
30	Reg: Sqrt(Comp1)	0.26	0.48	0.36	5.08	7.76
31	Reg: Comp1^2	0.50	0.73	0.62	10.33	13.35
32	Reg: Log-Log Comp1	0.13	0.50	0.32	2.44	7.03
33	Reg: 1/Comp2	0.06	0.27	0.15	0.81	3.18
34	Reg: Log(Comp2)	-0.32	0.41	0.32	-7.11	7.19
35	Reg: Sqrt(Comp2)	-0.46	0.54	0.46	-10.15	10.15
36	Reg: Comp2^2	-0.63	0.73	0.63	-13.62	13.62
37	Reg: Log-Log Comp2	-0.95	1.01	0.95	-20.55	20.55
38	Reg: Sqrt(Comp1) + 1/Comp2	0.10	0.33	0.20	1.79	4.37

The best model was the non-linear regression model that used 1/Competitor#2 as the predictor variable. This model had a MAPE of 3.18%, implying an accuracy of 96.82%. The second-best model was the exponential smoothing model with optimal parameters, which yielded a MAPE of 4.35% (95.65% accuracy).

FUTURE FORECASTS

We analyzed our top two models and found that it is not practical to use a regression model that does not have lagged predictors to make future predictions. This is because such a regression model would require us to forecast future revenues of our competitors and then use the forecasts as inputs to predict our company's future revenue. In other words, a forecast of our company's Q2 sales using the best non-linear regression model would require our competitor's Q2 sales as an input. As the additional layer of forecasting adds extra uncertainty and variability, we decided that it is better to use the second-best model: exponential smoothing with optimal parameters. This choice is justified by the negligible difference between the two models' accuracies.

Accordingly, we rebuilt the algorithm on the entire dataset, thereby using all the observations (not just training or testing set), and computed the forecasts for the next three quarters. The graph below shows the original time series in black, the fitted time series in red, the forecasts for the next three quarters, and the confidence intervals for the forecasts (80% and 95%).
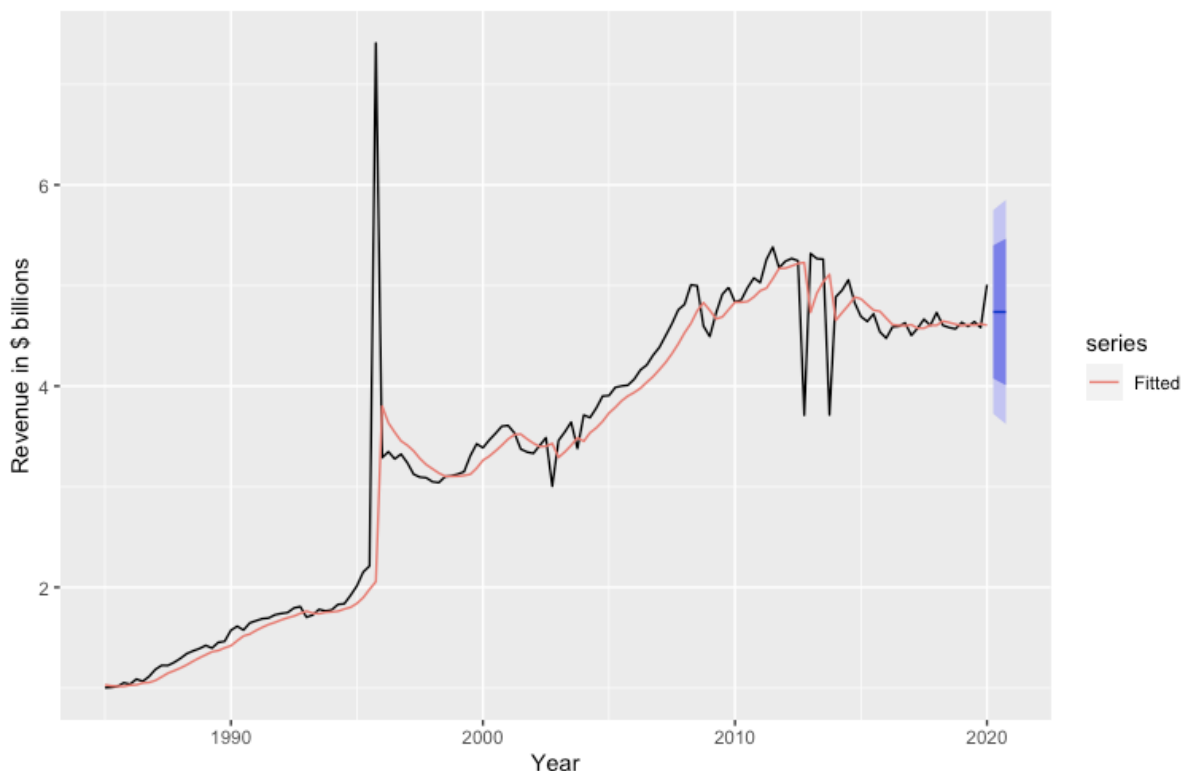
As seen in the figure above, the exponential smoothing model did an excellent job at fitting the historical data. It is important to note that this model gives the same forecast for the entire forecast horizon because of the way the algorithm is built. Therefore, for accurate results, we would have to update the model each quarter to give the next quarter's sale forecasts.

Finally, as this model had a 95.65% accuracy while forecasting the sales from Q1 of 2013 to Q1 of 2020 (testing set), we can expect this model to perform similarly in the future.