Forecasting with Machine Learning

by

Tamara Lessley

A Capstone Project Submitted to the Faculty of

Utica University

August 2022

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in
Data Science

Abstract

All industries and domains require clean data to make accurate data driven decisions and predictions. The data set utilized in this project was sourced from a dairy manufacturer and was sales data. The data spans 2019 through June of 2022, and encompasses one product category, one brand, and one package size. There are other dimensions present in the dataset, regarding a 'sold to' hierarchy (e.g., Corporate Entity, Banner, Regional, and Partner), and a small variety of material ids. Interspersed in the data are major data events that altered patterns in the data. These factors contributed to the strategy of how to treat and test the data.

Three explicit time series machine learning models were utilized in this project - namely seasonal autoregressive integrated moving average (SARIMA), error trend seasonal (ETS), and exponential smoothing functionality from Tableau. The data was cleansed minimally and maximumly and sent through each of the three models to see how the yielded forecast results would compare for purposes of accuracy, efficiency, and alignment with an industry metric. The data was cleansed using a combination of tactics. Some of those where the following: transposition, filling in missing values, and normalizing the major data events using a scalable tracking system with mitigation for each circumstance.

All the models performed more accurately and effectively using the maximumly cleansed data set. All three models forecasted values such that any of these models could be used in a production environment.

Keywords: Data Science, Dr. Michael McCarthy, Alteryx, python, continuous improvement

Acknowledgments

This project would not have been possible without the love and support of my family, friends, cohorts, and coworkers. They all encouraged me, gave feedback, and gave me grace when it was needed. This project is dedicated to my loving husband, Philip Lessley, who has always encouraged me to do my best and learn more. This is also dedicated to my dad and mom - Duane and Patty Johanson - without them, I would not have had the amazing education and upbringing possible. It was because of my dad's encouragement that I continued down this path. He passed on August 26, 2020, and would have loved to have seen this come to fruition.

Table of Contents

List of Illustrative Materials

Introduction

Background

The recently implemented enterprise resource planning (ERP) system at a food manufacturing company is limited in its ability to scrub and wrangle historical data during the forecasting process of sales data.

There were significant data events that occurred in the historical data set (past three years) – namely COVID-19 increases to sales, the ERP system was implemented in August 2021, causing an interruption in customer ordering patterns, and a cyber-event in March 2022, that halted product sales for ten days, and then changed sales volume for the two months occurring afterwards.

Problems to Solve

There were three problems to solve in this project. The first was to identify the specific scenarios in which scrubbing and wrangling of the historic data was required, and to what degree it should be used. The second problem was the use of a high-level wrangling method of weighted averages, that allows for the tracking of all major data events, and the best mitigation for each situation. The third problem was to create a comprehensive wrangling process that allows for the tracking of all major data events, the treatment, and remediation for each situation.

Research Question

Wrangling data in forecasting and across all types of data processes is typically applied to duplicate records, missing items, irrelevant data points, noisy data, anomalies, and data points outside the norm (e.g., beyond 2 standard deviations). Wrangling is synonymous with cleansing; the data is massaged in some manner. It is something that happens regularly to data sets, otherwise the disruptive elements of dirty data infect processes downstream, and will result in

unreliable and inaccurate outcomes, poor reporting, or loss of confidence from downstream users. Extract, transform, and load (ETL) procedures are developed by data engineers (typically). These ETL procedures focus on normalizing, cleansing, and wrangling data for those downstream to utilize. This project aimed to find the most efficient and accurate combinations of wrangling practices to normalize the data, that will result in the best return on investment because the company's current system is not optimized for ETL procedures resulting in time-consuming data wrangling. A great return on investment can be seen in the efficiencies gained, the improved accuracies of the forecasts, and the efficacies of the metrics. The scenarios utilized here were compared against the standard metric for the given food manufacturer:

> 'last three weeks actuals' divided by the 'forecast for the last three weeks' multiplied by one hundred.

This research also seeks to identify scenarios where data wrangling is not a good use of resources, and therefore, a poor return on investment. This project needs to be mindful of the seasonality of sales cycles and products. Involving the demand elasticity of these products as well, may aid this project and its success.

Literature Review

Theme 1: Data Cleansing/Wrangling/Transformation/Normalization

Data streams through many systems as it travels from origination to destination. Depending on how the users of the data need the data to behave, the wrangling and transformation process will potentially look different from one data set to another. Data cleansing is such an important part of the data modeling process, otherwise the downstream modeling and reporting will be displaying and using incomplete or nonstandard data. Cleansing

the data is non-negotiable and must be performed to avoid the common issue of "garbage in, garbage out."

Typically, data scientists use 60% of their time preprocessing data (e.g. wrangling, cleansing, normalizing) before modeling it. Before cleaning the data, ensure familiarity to confirm which elements are relevant and which are not. There are eight basic cleansing steps that should be applied to most data sets for natural language processing (NLP).

The first step is to remove duplicate entries (Roldos, 2021). The second step is to remove any known irrelevant data. Items such as personally identifiable information (PII), URLs, HTML tags, canned text, and excessive blank space. The third step is to standard capitalization before modeling and understand any data ambiguity (e.g., a person named Bill, vs an invoice bill). The fourth step is to convert data types and ensure proper data type application. Often numbers and dates can be stored as strings, it is important to convert these items to their proper type before modeling. The fifth step is to clean any superfluous formatting. If data comes from various data sources, the data can vary in look and format. The sixth step is to fix any typos, weird punctuation or ascii characters, and ensure currencies (or any other measurements) are uniform. Step seven, if applicable, especially if using NLP (natural language processing), ensure all language elements are in the same uniform language – one language chosen. The eighth step is to deal with missing elements. According to Roldos, there are two options, one is to remove the missing value items, or option two is to input the missing data in the empty cells (Roldos, 2021).

The data may then move from the cleansing stage to a wrangling or transformational stage, where the data is formatted, updated, added to, aggregated, or pivoted to show different angles of the same data. The data is normalized to make elements appear cohesive, especially when blending data from various sources.

Theme 2: Continuous Improvement

Striving to eliminate waste in any industry is key to driving down expenses and increasing net income. One way to eliminate waste is to implement continuous improvement. There are many different methods that can be useful such as Six Sigma, Kaizen, Lean Development, and Agile Development. There are valuable components of each. Consider Kaizen, the premise is to rectify issues as they are found. This keeps teams on track while focusing on the leanest practices (User, 2022). Everyone contributes to the continuous nature of being aware and focusing on decreasing waste. Being vigilant, responsive, and corrective to issues as they arise, contributes to better customer satisfaction as well, because the feedback of the customer is given a priority position.

Process improvement can create a domino effect, improved processes translate to more team efficiencies, less mistakes, less rework, more accuracy, and therefore a higher quality of output (User, 2022). If continuous improvement were implemented in the forecasting food manufacturing world, imagine the improvements that could be made to the operations, the supply chain, and all the data points up and downstream. Not to mention the decreased waste of any perishable items.

Edwards Deming applied the scientific method to business processes. Deming created a cycle of improvement applicable to manufacturing but can be applied to many more domains and industries. The cycle begins with the planning phase. The key stakeholders must first understand the definition of quality, what changes can bring improvement, are there any predicted outcomes or can anything be predicted (Henshall, 2020).

The second phase of the cycle is to perform small tests and perform iterative changes to test the variables. Every plan, prediction, test, variable, and increment should be documented.

Understand the output and how things are improved or not, with the tests. The third phase is to study the outcomes. Check to see whether they match the predictions. Compare the results of the tests against the actuals. Check that all the angles are compared and find the places in which the iterative testing proved worthwhile.

The fourth and final stage of the cycle is to implement the recommended changes that proved to be successful. Track the performance of these changes over time and verify that all is improved as expected. And wrap it up with documentation of the entire cycle and the implementation steps. All the steps combined are the Plan, Do, Study, and Act (PDSA) cycle (Henshall, 2020).

Considering the PDSC cycle alongside the process of operationalization of forecasting, there are some continuous efforts that can be applied to the data and operations to enforce lean practices. This project aims to look at many angles in which forecasting can be wrangled to arrive at a more optimum process - reducing waste, time, and resources.

Theme 3: Cognitive Bias

Cognitive bias is the effect of people's feelings and experiences affecting their judgment (Fallmann, 2021). Humans have naturally occurring cognitive biases. Those biases are inadvertently applied within data processing functions, like machine learning.

There is the study of cognitive science, which is the compilation of psychology, neuroscience, and cognitive neuroscience. Consider that artificial intelligence (AI) functions like a basic version of the human brain. The machine learning (ML) branch of AI deals with data predictions (Fallmann, 2021).

Humans tend to rely on information that supports their current beliefs, this is the cognitive bias of rationalization. Suppositions are tricky for humans and programs created by

humans to deal with. When ML processes encounter data that is beyond their knowledge base, it creates memory leakage. The process may end up using old, or stale data to resolve the anomalous data.

Humans are fallible, and therefore, AI is as well. Cognitive biases are innate in human existence. The more that developers know about cognitive bias, the better they can create their processes to account for those aspects (Fallmann, 2021).

Many people conflate the term robot with AI. However, considering that AI is ubiquitous, it appears in cell phones with assumed text, assumed email addresses, and it is present in many vehicles as well with lane assist and predictive stopping.

AlphaGo beat Lee Sedol (a genius) in chess. And IBM's Watson program competed and won on Jeopardy (Kim, 2021). Humans are in awe of these feats. Humans tend to compare success based on the difficulty according to human standards. We fail to realize that robots, and machines are created with limited purpose, and defined boundaries. Humans tend to anthropomorphize machines and their accomplishments, putting them on a pedestal. We behave qualitatively, whereas AI behaves quantitatively (Kim, 2021).

The AI operates using the algorithms installed. Algorithms are simply recipes for processes. We are surrounded by algorithms constantly. Whether doing laundry or making food. We must decide how big of a load to put in the laundry, how much soap, temperature of the water, the duration of the cycle. Humans perform many algorithms effortlessly. Therefore, when humans program AI, it takes several algorithms to be successful at a task or project (Kim, 2021).

Often there are misalignments of understanding between humans and the processes executed by AI. We need to adjust our expectations relative to what the machine is designed to do. Since our understanding needs to be in alignment with expectations, it is pivotal that

cognitive bias and anthropomorphization be considered. Only then can humans set the appropriate direction for technology, and policies around technology (Kim, 2021).

Theme 4: Time Series Models

The long short-term memory (LSTM) network model was discussed. This model is great at retaining useful historical data and deciding which is relevant or not, which makes LSTM good with time-series predictions. LSTM has three gates in the modeling process. The first gate is called 'forget gate'. This gate looks at the previous hidden state and the new input data. Everything is assigned a value from zero to one. Anything near zero is irrelevant, and anything near one is relevant. Those items that are relevant are weighted heavier and held onto. Those items that are irrelevant are weighted lighter and will be less influential.

The second gate takes the same inputs as the 'forget gate,' however it treats them differently, in the fact that it applies a tanh function, so the outputs are between negative one (1) and positive one (1). This creates the 'new memory network.' This component dictates how much to update the cell state. The 'new memory network' is fed through the 'input gate,' and the 'input gate' acts as a filter to decide which items are relevant or not. Assigning values from zero to one and weighting those items appropriately and then added into the cell state (aka long-term memory).

The last gate of this process is called 'output gate'. At this stage, the long-term memory items are known and complete. Next is to decide the new hidden state. The 'output gate' acts as a filter to show those relevant items. Again, items that are close to zero are irrelevant, and those closer to one are more relevant. After these 3 gates, this gives the new hidden state. To obtain the predictions, the three gated steps need to be performed multiple times, then a linear layer is applied to show the future sales (Dolphin, 2021).

Choosing the appropriate key performance indicators (KPI) from the given forecast model is not always straightforward. There are common approaches for assessing forecast accuracy (i.e., finding errors) with various time series models. The first method discussed is mean absolute percentage error (MAPE). This is the sum of the individual absolute errors divided by the demand (actual sales each period). Otherwise known as the average of the percent of errors. This is a poor indicator of errors. MAPE divides the errors individually by the demand; therefore, high errors during low demand will greatly impact MAPE.

The next method discussed is mean absolute error (MAE) which is the absolute error. This relies on finding the middle point of the data, also known as the median. Low MAE scores are ideal. The RMSE (root mean squared error) is the square root of the average squared error. One drawback of RMSE is that it emphasizes larger errors a little more than MAE – since MAE is protected against outliers. If MAE results in a high bias (i.e., the average error), then use RMSE. If there are low demand items on a weekly basis, then aggregate them to a higher level and see if forecasting can be performed at that higher level (e.g., month, quarter) (Vandeput, 2021).

Another time series model worth investigating is error trend seasonal, which is also referred to as (ETS). There are many forecasting models within the exponential smoothing family. Other members of the ETS family include seasonal trend decomposition (STLM), and theta decomposition forecasting (THETAF). Each of these models accepts three parameters — one for error type, trend type, and one for season type. Each of those type categories can be any of the following options: additive, multiplicative, or none. The permutations of these three parameters with the allowable values gives the entire exponential smoothing family. The three models listed here are the most popular. The ETS model, as the name indicates, gives model

error rate, the trending aspects of the data, along with the seasonality component built in (Ellis, 2016).

Theme 5: Machine Learning (ML) for Forecasting

Common demand forecast tactics and reasons why companies use other methods. The first method is traditional statistical forecasting. It is widely used because so many people are comfortable with Microsoft Excel, and it interacts easily with most ERP (enterprise resource planning) systems. Excel can easily perform time-series forecasts. This is great for stable markets and typically needs two years' worth of data to produce reliable results. The traditional statistical models in Excel do not work well for new products, short-term products, or volatile markets. Therefore, they are very limited in their scope of application.

Another method for demand forecasting is machine learning (ML). This is great for volatility and learning from multiple data sources — historical, financial, social media, weather, — unstructured or structured data sources. ML can operate on new products, short cycle products, and weather/seasonal issues. ML also works great in fast-paced and changing environments. The third method is to use predictive analytics – which is akin to mixing the statistical models and the ML models. These do involve complex machine learning algorithms, and any results need to be verified and interpreted by humans. This method requires significant computing power and resources (AltexSoft Inc, 2020).

Demand planning or forecasting is the act of predicting the future of products and/or services. The goal is to increase accuracy of forecasting by using a combination of ML techniques. The data set used in this research article is about air passengers over time. The data set is first checked for stationarity; if the data set were stationary, it would not have trends, seasonality, or cyclical patterns. The data set should be verified that is not stationary. Therefore,

the author, proceeds to create a model for the demand forecasting using a Monte Carlo Simulation (MCS) (Thete, 2022). The seasonal autoregressive integrated moving average (SARIMA) is developed for this data set. The parameters are run through looking for the best permutations and combinations therein. The model is then trained using those parameters. The model output is plotted to see the prediction versus the actual data. Then the errors are plotted. They have a bias because they center around one, not zero. To reduce the errors, a rolling forecast is applied. This will recreate the SARIMA after each observation is received. This will seed the training data. Then append new observations and iterate through. The new updated model is plotted versus the original data. The updated set of errors is plotted. The Laplace distribution is closest to the set of errors that one is chosen for moving forward with. Thete creates a function to do the MCS with the Laplacian distribution — therefore the function is a rolling forecast Monte Carlo simulation (2022). The author then runs the data through this model with 1,000 MC simulations. The predictions are plotted, along with the min and max range. The data is trained on this model. The accuracy of this model turns out to be that 100% of the demand will fall within the min/max range established here (Thete, 2022).

External variables and internal variables that are important to consider in demand planning (i.e., forecasting) in manufacturing. External variables such as international relations, government policies, and internal variables like the technical level of the company. The technology level of the company can be derived by viewing the patent applications, the sales volume, and the research and development. Dou explained that while random forest, and the gray model can be useful for demand forecasting, the best method was deep learning, and auto-regressive prediction (2021). Those models were compared to the efficacy of Long Short-Term Memory (LSTM) network (Dou, 2021).

A case for using external data in a Walmart scenario is presented (Riveroll, 2020). In general, companies need to look at external data sources when creating and testing ML models. External data can add so much context and help with accuracy (lower the errors). If all entities along the supply chain were to cooperate, this would ease the use of external sources and the large benefit that can come from such a blend. The first step in this process is to define the target, which in this case is to forecast the sales for Walmart stores and improve accuracy as external components are added to the model. The external variables commonly added include dollar index, oil price, and news about Walmart (e.g., promotions). The second step is to make the model. Run the model on the basic Walmart data to see what the baseline predictions and errors look like. For the purposes of this exercise, the author focuses on one store to apply external data components (Riveroll, 2020).

To gather and blend those external items, OpenBlender.io is used by Riveroll (2020). The first component blended are those items about Walmart promotions. The blended components were run through the model and compared to the original output. The accuracy improved by 12% with the addition of this one external component. The next components added was other external ideas (e.g., dollar index, oil). After blending those data with the components already used, the output of that was a 24% improvement to the overall model for the test-case store. The improved model (with all the added context) runs. The results show that, overall, those externally added components added vast improvements for the majority of the Walmart stores' sales forecasts. Therefore, it is with minimal effort that external data can be added to the model and the accuracy tends to improve (Riveroll, 2020).

Comparing various predictive analytics to help the manufacturing industry to make better decisions, which in turn will help with inventory, purchasing, planning, product availability, and

therefore profits. The literature presented in this article showed that for perishable goods, such as fruits and vegetables, a time series analysis, such as SARIMA (seasonal autoregressive integrated moving average), accounts for trends and seasons. Data cleansing is important to ensure that anomalies are accounted for and treated in a meaningful way. For example, for missing data, they used simple arithmetic average (SAA) to fill in the missing values. And looking for any noisy elements, they focused on anything that was beyond two standard deviations away from the norm. If it was found to not be linked to an event, promotion, or holiday, then it was normalized in a similar fashion as the missing values process. It was concluded that accurate demand forecasting is best if using a combination of domain data, plus external data (e.g., weather, promotions downstream of the manufacturer) to help with understanding the grander context in which to build the model (Falatouri, 2022).

Theme 6: Forecasting in the Food Industry

There is a history of the slow adoption of technology in the food industry. The resolution is to implement machine learning and artificial intelligence using current technology. The earliest tech was the barcode system in 1974. Shortly thereafter, in 1981, there was the electronic data interchange (EDI) that allowed information to flow digitally, instead of using paper. In 2017, blockchain emerged and acted as a digital ledger for all products. It recorded the data across multiple transactions – real-time information. The data was transparent and reliable. The counterpoint for block chain, is that the cost of the process and the fact that not many companies implement it.

The most recent tech innovation to reach the food industry is that of the cloud-based demand forecasting in 2019. This utilizes machine learning and artificial intelligence to predict demand for various products along the food supply chain. Machine learning allows for enhanced

communications and allows for large volumes of data to be utilized in the process. Artificial intelligence can collate data from weather sources, social media, events holidays, buying patterns. To get more adoption from the entities in the block chain, there needs to be a sharing of information for all the partners that are connected through supply chains. They need to work together to gather data and utilize demand forecasting (Traasdahl, 2020).

Overfitting issues arise with time series analysis. Meal prep kits are perishable and have time-sensitive needs for transportation and demand forecasting. The goal is to not overfit and use a few processes to determine which method works best for this data set. For the meal kit companies, the goal is to forecast for ten weeks, and to find the best method that will show the most decay while simultaneously showing low drop-out rates. The data is first normalized, missing values filled in, and category fields are encoded. Subsequently, hyperparameters are chosen by running a learning rate finder. To reduce training time, the highest population size is chosen. The seven resulting models are plotted to look for the desired results. Training and validation data are then plotted to compare. The method continues until the training and validation are fitted. The encoded category data needs to be rewritten to be consumed by the Random Forest Generator, LGBM (light gradient boosting machine) regressor, and the XGB (extreme gradient boosting) regressor processes. Comparing the RMSE (root mean square error) from all these methods, the best result actually came from the Random Forest (Iyer, 2020).

There are dilemmas placed on the food service industry and the supporting supply chains during the COVID-19 pandemic. Over half of food suppliers invested in technology during the pandemic, knowing that predictive analytics is vital to make more accurate forecasts and stay competitive (Scioscia, 2021). Statistical algorithms are used to advance the purchasing power of suppliers (and all entities on the supply chain). It allows those companies to look at internal data

(e.g., sales patterns, trends) plus external data (e.g., weather patterns, agricultural impacts) to detect hidden patterns or trends. It is important to be able to detect which products are in demand, and what the farmer or supplier lead times might look like. This allows for advanced knowledge, which will lead to better purchases to keep up with the demands of the various items needed.

There are various methods for forecasting depending on the length of time necessary – long, medium, or short-term forecasts. Focusing on ice as the product to investigate, the paths of medium, or short-term forecasts were chosen to investigate. Actual demands are compared to: moving average, simple exponential smoothing, double exponential smoothing, triple exponential smoothing, and root mean square error (RMSE). Running the data through all methods and then comparing the results, the RMSE was the most accurate (Guinoubi, 2022).

The current situation for the food industry is complex —inflation, increased competition, and increased need for resource and transportation costs. Food manufacturers have such small profit margins, it is hard for them to combat these growing issues. There are many variables to this equation — costs, volume, downstream effects, promotional periods, and potential loss of providers, customers, and/or suppliers. There are negative side effects of not solving these issues — namely running out of inventory. Lack of stock accounted for nearly $6 billion dollars of losses each year (Hennel, 2006). The fix for these issues comes from improved demand forecasting accuracy. If food manufacturers can lean on statistical models, with the knowledge of supply chains, and the agile mindset along with technology, they will have the edge in the industry (Hennel, 2006).

Analysis of all possible artificial intelligence (AI) methods that can be applied to supply chain data for the purposes of demand forecasting. Mediavilla (2022) explored various

publications and articles from the last 5 years, collating the evidence of what was being utilized, what was working and what created the most success in the manufacturing arena. They classified all the methods found, across the machine learning world. The most prevalent and successful methods were the following: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Artificial Neural Network (ANN). Even though many companies were successful, they focused on the retailer portion of the supply chain. The forecasting needs to be performed as a collaborative effort that encompasses the entire breadth of the supply chain. All parts are interrelated and need to operate together, otherwise the data can get distorted and the bullwhip effect can rear its head (Mediavilla, 2022).

Reducing waste in the food industry is a large concern. 30-80% of food is wasted globally (retail accounting for greater than 10% of that waste). Exploring different machine learning methods to help reduce the waste in the supply chain, these methods were compared: long short-term memory network, feedforward neural networks, support vector regression, random forests, and Holt-Winter statistical model. Those models were compared to the observed sales data. The machine learning models ultimately provided accurate forecasts. The best method in this study was that of the long short-term memory. It had the best results with respect to its root mean squared error (RMSE) (Migueis, 2022).

Future market demands are parsed into three segments for an equilibrium model: farm markets, wholesalers to wholesalers, who then sell to retailers. There was supply-side modeling performed. Note there is a dairy price support program, where notably the price of cheese set must be greater than or equal to that of government cheese. There are many variables in running these models, such as the following: demographic, spending patterns, advertising, income of customers, age, seasonality, time trends, race, and the energy cost index. The root mean square

error was used. The conclusions centered around how advertising costs played the largest role in the variables when it came to the forecast results (Schmit, 2010).

Important aspects of the manufacturing world, such as perishable items, need to be analyzed in a different manner; the lifecycles are variant. It is best to use algorithms such as ARIMA (Autoregressive Integrated Moving Average) and Holt-Winters (HW) models to forecast perishable items. The HW forecast showed the best results in this specific case, even though a variety of other methods were utilized in comparison: clustering, K-nearest neighbor, artificial neural networks, regression analysis, support vector machines, support vector regression, and mixed approaches as well (Seyedan, 2020). The bullwhip effect is real and needs to be addressed for any manufacturer. The bullwhip effect occurs in supply chains, when small changes happen at the retail end of the supply chain that cause progressively larger disturbances throughout the systems. Small changes in customer demand can result in exaggerated deltas in the providers upstream (e.g., manufacturer and suppliers) (Seyedan, 2020).

Typically, dairy product consumption was inelastic (e.g., relatively stable in pricing) because supply and demand were historically not volatile. However, more recently, the government is no longer the largest customer of the dairy industry, the supply and demand has started to shift. With the shift in demand, the prices start to rise higher and higher. That results in companies selling less and promoting those higher priced dairy products less (Merlo, 2015).

Data Activities and Review

Data

There were nine data sets available for use in this project. Data set number four was chosen and encompassed one package size, one brand, and one product category. This data set summarized the historical sales occurring at a food manufacturing company for one brand,

product category, and package size. The data set included customer dimension information such as sold to partner, with the dimensions for the sold to partner, being the sold to corporate entity, sold to banner, and sold to regional. The time element in our data set spanned from January 2019, through June 2022. All research was performed with this in mind. Therefore, many items surrounding time series models stood out.

There were four unique *MaterialID*s, along with a variety of corporate entities, banners, regionals, and partners to which this product was sold. In data set four, there are 189 columns. There are eight dimension fields and 181 columns with headers for weekly dates and the contents are the units sold per week (see Table 1 for the transposed version). There are column headers for all units sold in the given week, starting with the date column of Saturday, January 5, 2019, and the maximum date of the set was Saturday, June 25, 2022. The manufacturing company groups sales by week, from Sunday to Saturday.

Table 1. Updated Data Dictionary of Data Set Four

Field Name	Number of Distinct Values	Data Type	*Value Example*	Type Of Variable
PackageSizeName	1	string	*P_15*	Independent (X)
MaterialID	4	string	*M_129*	Independent (X)
Brand	1	string	*B_3*	Independent (X)
ProdCategory	1	string	*C_2*	Independent (X)
SoldToCorporateEntity	74	string	*SC_1147*	Independent (X)
SoldToBanner	84	string	*SB_1162*	Independent (X)
SoldToRegional	106	string	*SR_1216*	Independent (X)
SoldToPartner	254	string	*SP_2129*	Independent (X)
Date	181	date	*1/1/2022*	Independent (X)
UnitsSold	continuous	float	*5*	Dependent (Y)

In the entire data set, across the four *MaterialID*'s, the *SoldtoCorpEntities*, there are 871 rows of data that show the permutations of those combinations of fields. All dimensional elements of the data that are potentially sensitive (e.g., *PackageSizeName*, *MaterialID*, *ProdCategory*, *SoldtoCorporateEntity*, *SoldToBanner*, *SoldToRegional*, and *SoldToPartner*) were completely obfuscated. The data string values were converted to generic alpha-numeric placeholders in the Value Example column of Table 1.

The data set has only one product, that is C_2. The variations will come in the *MaterialID*'s, and the 'sold to' components. It was noted that among the corporate entities, not all of them purchase each period, nor are they present in the entire data set timeframe. Therefore, some 'sold to' entities do not have *UnitsSold* for each week, month. A simple aggregation pivot by year of the data with a heatmap was constructed to visualize the corporate entities that had the highest volumes of units sold in the data set (see Table 2). The sorted data identifies the corporation with the highest volumes at the top. The larger the volume, the deeper red in color the cell appears. The smaller the volume, the deeper the blue hue.

Table 2. Base Data - Sum of Units Sold by Corporation – Top 20

Sum of *UnitsSold*	Year				
SoldToCorporateEntity	**2019**	**2020**	**2021**	**2022**	**Grand Total**
SC_1147	1,803,441	2,615,585	2,771,047	1,278,044	8,468,117
SC_609	561,743	785,317	913,595	364,854	2,625,510
SC_861	553,614	732,824	781,583	332,468	2,400,489
SC_21	493,900	677,906	716,643	338,745	2,227,194
SC_24	368,425	513,078	592,346	231,145	1,704,994
SC_1177	243,043	288,802	281,665	121,930	935,440
SC_101	71,280	283,400	349,791	155,601	860,072
SC_1146	191,864	238,816	245,603	116,018	792,300
SC_1116	154,842	177,457	218,945	111,707	662,952
SC_160	108,047	136,694	149,043	61,938	455,722
SC_705	77,098	120,550	144,837	65,561	408,046
SC_1072	81,146	124,666	137,118	49,075	392,005
SC_1152	118,303	118,891	109,749	41,170	388,113
SC_688	75,766	109,589	121,859	55,530	362,744
SC_1030	92,116	108,447	109,878	41,942	352,383
SC_1009	74,505	106,746	111,656	54,176	347,083
SC_948	60,590	84,578	81,918	31,581	258,667
SC_124	61,609	76,091	74,465	41,122	253,286
SC_427	48,193	63,450	69,821	29,223	210,687
SC_991	41,116	52,999	52,444	23,681	170,240

For example, it was found that corporations “SC_1147”, and “SC_609” had the highest volumes across the entirety of time. Curiously, SC_1147 outpaces the nearest entity by more than three times the volume. This one customer contributes 30% of the sales of the given product (B_3).

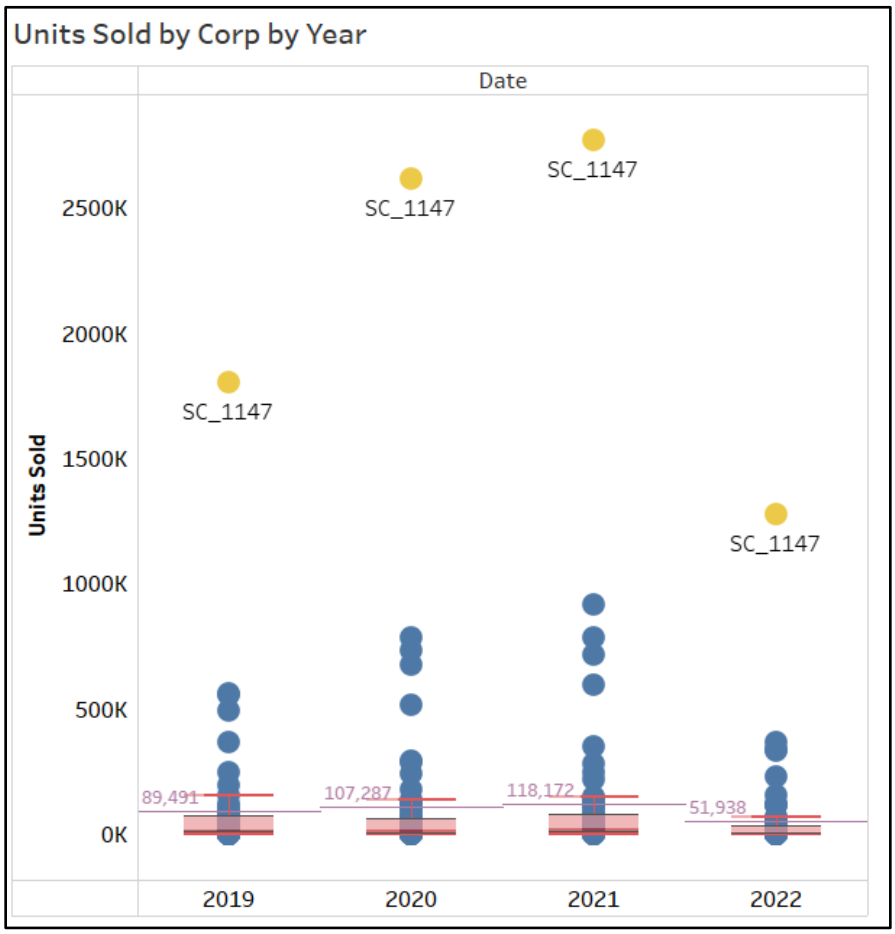
Exploratory Data Analysis (EDA)

All 181 date headered columns needed to be transposed into one singular column for ‘date’. Using Alteryx (version 2022.1), the dimension columns were held steady, and all the dates were brought into one column using the transpose function. Therefore, the data set was condensed into a more vertical shape, with only 10 fields, as shown in Table 1. This new data set contained 157,651 rows.

The exploration of the data showed that this one product category was most popular with one corporate entity. Running this updated data set through the EDA process, the following statistics were discovered (Figure 1). *UnitsSold* field is the only numeric field. The mean sum of *UnitsSold* from 2019 - 2022 were 89,491, 107,287, 118,172, and 51,938, respectively (see purple lines and values, per year in Figure 2). The median values for *UnitsSold* from 2019 - 2022 were 15,276, 15,093, 20,336, and 8,045 respectively (see the middle black horizontal line of the box and whiskers in Figure 1).

This dataset has many *UnitsSold* to a variety of corporate entities between the aggregate values of zero to 1 million. However, corporate entity SC_1147 far surpasses all other corporate entity activity. If we focus on the most voluminous of entities here and can forecast for the largest, then the same procedures can be performed for the remainder of the entities.

Figure 1. Data Set Four - Box and Whiskers Plot



Because of the aforementioned lack of sales for some corporate entities across time, there were many empty *UnitsSold* cells per line item in any given week, within the data set. Additionally, for the years of 2020 - 2022, there were two duplicate week headers for the beginning weeks of January. Those seemingly duplicate weeks were to be treated as week 52 and week 53 (as happens in the manufacturing industry). Those two were aggregated and treated together for this project purposes.

Intermixed in this data were impactful events that caused large deltas in the data. Table 2 shows the listing of those events, and their timings. There are many possible ways to combat these events when dealing with the data and attempting to normalize situations as these. A few of

the options were to take an average, a weighted average, a moving average, or any other number of options in that vein. For this project, the mitigation method chosen was a simple average looking at the previous year's data in the same period. For example, the data event of ERP Implementation impacted sales from 8/1/21 through 8/31/21. Therefore, the previous year's average, that is the average per corporate entity and *MaterialID*, from 8/1/20 through 8/31/20 was applied to August of 2021, to help counteract the effect that the ERP Implementation had on the data in 2021.

For purposes of this project a simple approach was taken, of using a SQL query to join the event table (Table 3) with the transposed data set (Table 1), to normalize the data events for future years' forecasting evaluations. It is recommended that with any large database, those events be tracked in a cogent manner, such that they can be referred to in future times. This allows for lessons learned, as well as this normalization with querying. This could be productionalized with the use of function objects standing for each mitigation method, and a stored procedure to ingest all data and run it through a stored procedure to manufacture the normalized data for comparison, and input into the machine learning models, further downstream. This historical tracking mechanism of historic data events can be applied to all data sets that were involved in the disruptive event, not just the data set in this project.

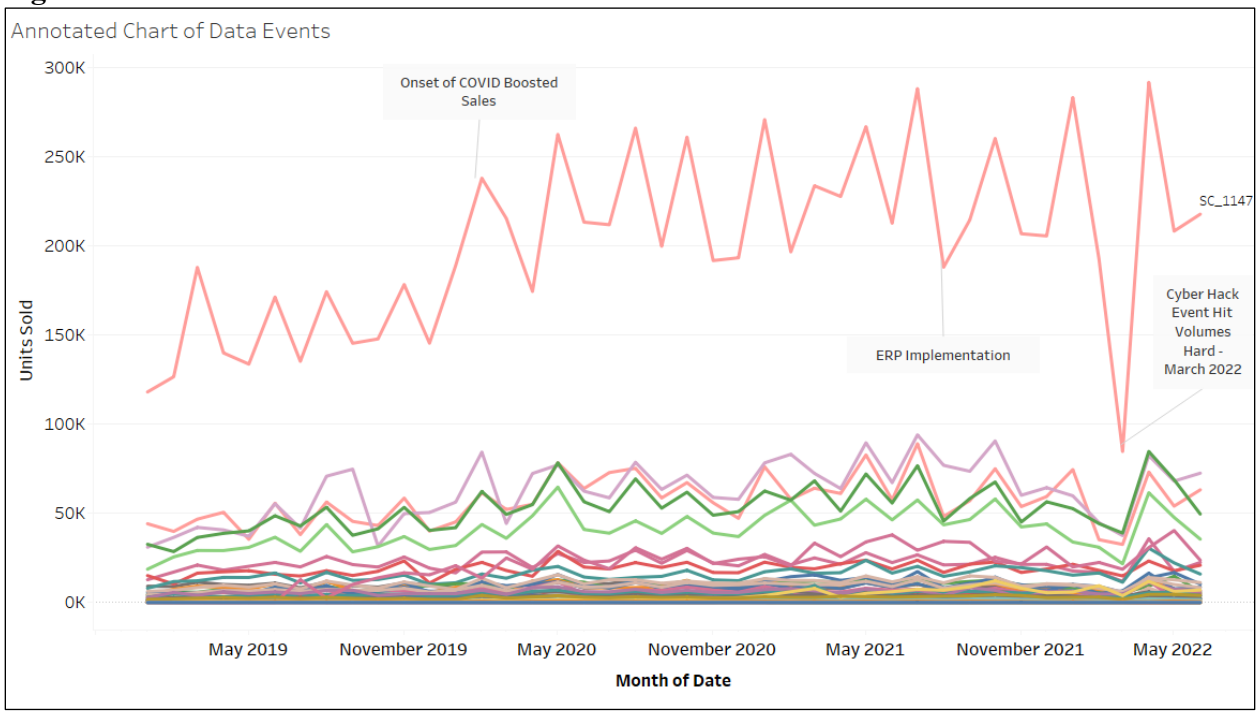
Table 3. Data Event History Tracking Table

Data Event Name	Event Start Date	Event End Date	Event Notes	Mitigation Method	Mitigation Notes
COVID-19	03/01/20	06/30/20	Increase in sales in the first quarter of the pandemic	Previous Year Average	Take average from same period (Start - End Dates) the previous year & replace this period
ERP Implementation	08/01/21	08/31/21	Decrease in activity due to onboarding new system and processes		
Cyber hack	03/01/22	03/10/22	Systems down for 10 days - no activity		
Sales activity post Cyber Hack - M1	03/11/22	04/30/22	Increase in April		
Sales activity post Cyber Hack - M2	05/01/22	05/31/22	Decrease in May		

The events were plotted to show the impact over time (Figure 2). These events can be seen clearly, and they are most likely one-time impacts. It is not desired that the one-time events dictate the forecasts for future years (Figure 2)

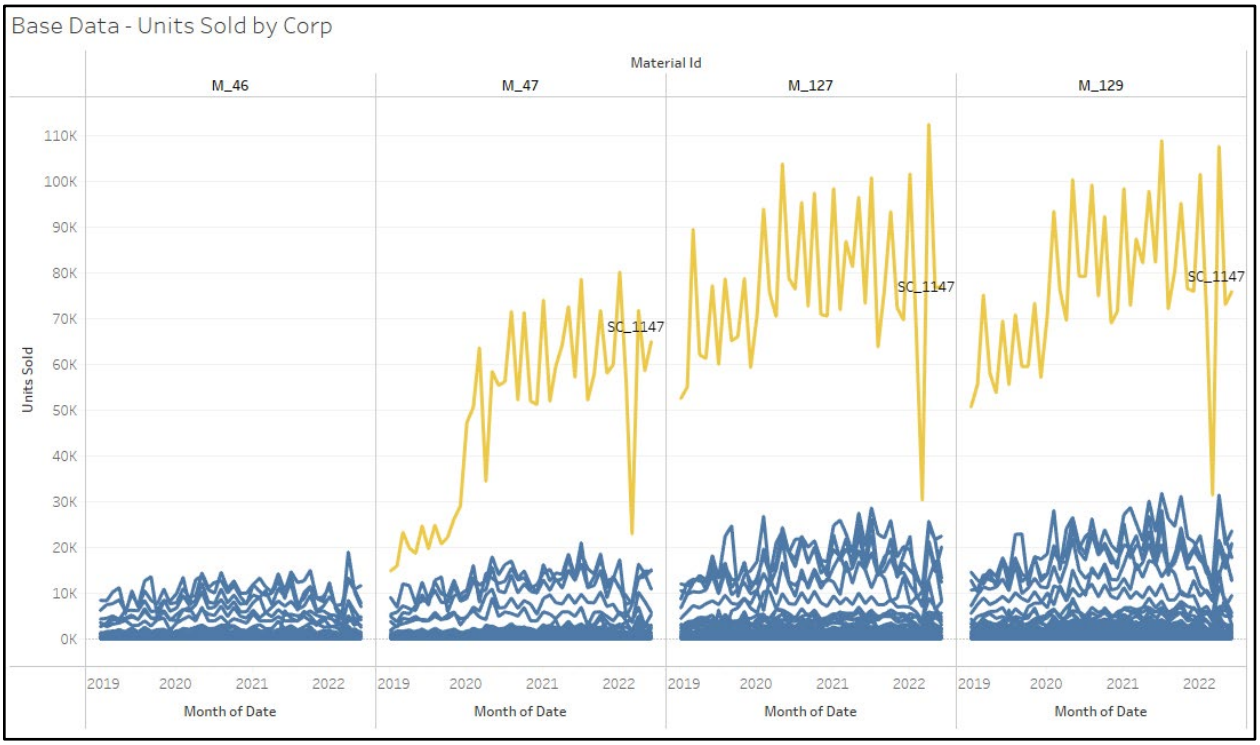
From Figure 3, SC_1147, is the largest volume contributor for *UnitsSold*. For this reason, this company acted as the test case for all experimentations and modeling as the project continued. This approach is conservative, if the results will accommodate the most copious company in the set, then it can easily represent the remainder of the companies.

Figure 2. Annotated Chart of Data Events



Not only will that one company be used as the test case, but to narrow the focus of the investigation, the *MaterialID* focused on will be M_129. As seen in Figure 3 below, the graph from Figure 3 has been sliced by *MaterialID*, to see its interplay with the data. It has a fair portion of the volume of the transactions and was a conservative place to start.

Figure 3. Chart of Units Sold by Corporation & MaterialID



Models and Methods

Considering that the data set has many independent variables, with one dependent variable, a few models were discussed in use for this project. The goal is to decrease the error aggregation and become more accurate with the resulting model functionality. This helps with overall confidence in the model, as the accuracy improves.

This project needed a model that would show the dynamic nature of the data, SARIMA was evaluated as a contender, as well as the ETS model. Both models were selected due to their credibility noted in the article by Sevedan (2020).

Minimally cleansed and normalized data will be that data that has the minimum effort put in, for the data to be put through the given model. Whereas, the data with much more normalizations applied (i.e., maximumly cleansed data), and then executed through the various

models and judged for efficacy. These approaches will be tested and iterated through, just as discussed in Demings' PDSA cycle of continuous improvement.

With any of these models, there would need to be a way to compare the model behavior with minimally cleansed data, versus having the model use more maximumly cleansed and normalized data. The goal was to run the two versions of the updated data set through the SARIMA and ETS models and compare the results from all (Table 4). This simple approach allowed the viewing of the results at both ends of the spectrum.

Table 4. Approach Matrix

	SARIMA	ETS
Minimally Cleansed/Normalized Data	* *Compare timing, accuracy, and feasibility* * *Result set A*	* *Compare timing, accuracy, and feasibility* * *Result set B*
More Maximumly Cleansed/Normalized Data	* *Compare timing, accuracy, and feasibility* * *Result set C*	* *Compare timing, accuracy, and feasibility* * *Result set D*

Wrangling Performed

Knowing that SC_1147 is the largest corporate entity, the data will be filtered to view results on this one corporate entity's activity. Also, that *MaterialID* of M_129 is a large driver of transactions, these two filters will be applied to the data, to narrow the view of the research and allow hyperfocus on those results. The results for the larger data set can be extrapolated from this one large, most voluminous company.

The wrangling of the 'minimally cleansed data' means the following:

1. The date columns have been transposed into one singular column (as mentioned earlier)

2. The fields for both Units Sold and Date were renamed
3. The date field was converted to a date type field
4. All NULL Units Sold cells were filled in with zeros
    a. Akin to what Roldos (2021) recommends in step eight
5. The duplicate weeks in January were aggregated together
6. Filtered to corporate entity SC_1147 *MaterialID* for M_129
7. Removed the last week from the data set because it was incomplete
8. Segregated the next to last 3 weeks, to allow for the testing and comparison of the actuals to the forecasts, to compute the industry metric (noted above from the food industry company).

The normalization of the 'maximumly cleansed data' includes the following steps:

1. Steps 1 through 5 in the minimally cleansed data routine
2. Standardized the historical data events using the database tracking table (Table 3) with prescribed mitigation methods per event described previously
    a. [this step not included in minimally cleansed data routine]
3. Filtered to corporate entity SC_1147 and *MaterialID* for M_129
4. Removed the last week from the data set because it was incomplete
5. Segregation of the next to last 3 weeks, to allow for the testing and comparison of the actuals to the forecasts, to compute the industry metric (noted above from food industry company)
6. Used a moving weighted average of the units sold field value.
    a. Organizing the data by date ascending, each row was evaluated - meaning that the previous row value for UnitSold was added to the following row

value for UnitSold, then that sum was divided by two, to find the simple rolling average.

b. Once the rolling average was accomplished, then those values were weighted such that those older values were less weighty than the newer values.

c. [this step not included in minimally cleansed data routine]

The first two approaches used the SARIMA model with the minimally cleansed (result set A) and more maximumly cleansed data (result set C). The SARIMA model was programmed to look at the data at a frequency of weekly. The seasonal lag term of 1, and the lag of the moving average term was also set to 1 week. This model was set to allow for drift.

The next approach used the ETS mode (result sets B and D). This was programmed for a weekly frequency like the SASRIMA. The error type, trend type, trend dampening, and the seasonal type options were set to auto, meaning that Alteryx would run the additive and multiplicative versions and find the best combination of those components to utilize across the various permutations and Akaike Information Criterion (AIC) values produced. The AIC is a method for evaluating how well the model fits the data. It consists of the number of independent variables, and the maximum estimate of the model.

Findings

The findings from these approaches with the minimally cleansed and maximumly cleansed data sets were reasonable in a couple ways. Firstly, the results from the minimally cleansed data through the SARIMA model (result set A) showed a RMSE of 3,434.408, and a MAE of 2,256.469. This model ran in 23 seconds (Table 5).

Table 5. Summary of SARIMA Results

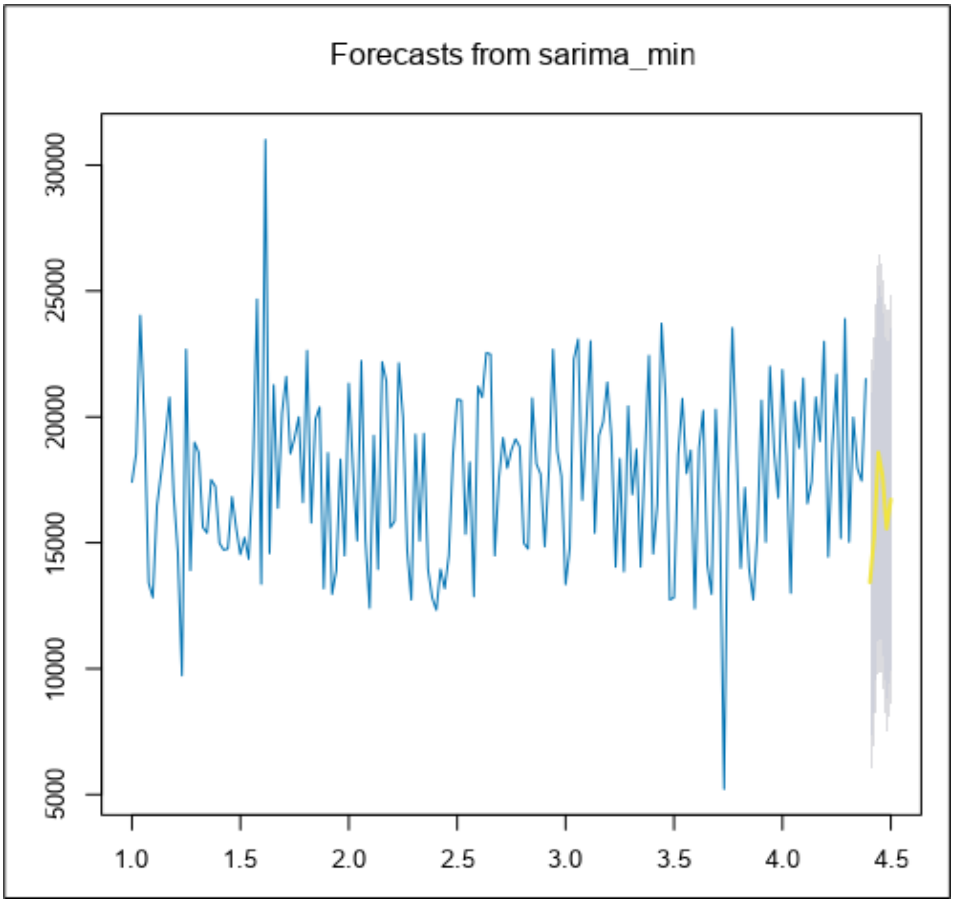
Approach		SARIMA			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specific Metric**
A	**MIN Cleansed Data**	3,434.4078	2,256.4692	110.738%	23 seconds to run
C	**MAX Cleansed Data**	1,469.4755	1,037.6113	92.571%	19.2 seconds to run
% increase in accuracy		*57%*	*54%*		*20% increased efficiency*

Applying the "industry metric" (sum of 3 actual weeks / sum of 3 forecast weeks * 100), the industry metric for forecast average was 110.74%. Consider that the goal of the industry is to meet the needs of their customers and at the same time not have too much inventory on hand that might perish. Ideally, the goal is to order enough to supply to the customers downstream, to be able to fulfill all orders. The industry metric is one way to gauge if the forecast method is reasonable considering the needs of the customers. The industry metric works best when the value approaches 100% - meaning that the forecasted amounts are on par with what is actually sold to the customers downstream. If the industry metric is less than 100, that implies that the forecast amounts will be larger than the actuals and will allow for the fulfillment of all the corporate entity orders, plus extra inventory. Whereas, if the industry metric is over 100%, that means that the forecasted amounts will be less than the actuals needed, and the food manufacturing company would not be able to fulfill their orders. Table 6 shows the calculation of the industry metric on result set A. See Figure 4 to see the minimally cleansed data forecasts plotted in the yellow and gray area. The yellow line is the forecasted amounts, whereas the grey plotted area is the 95% confidence interval around the forecast.

Table 6. Industry Metric Calculation on Result Set A

Year	Week	Actual	Forecast
2022	23	14,985	15,263
2022	24	20,520	18,594
2022	25	21,555	17,670
Grand Total		57,060	51,527
	Industry Metric		**110.738%**

Figure 4. SARIMA Forecast Plotted - Minimally Cleansed Data



Putting the maximumly cleansed data through this SARIMA model resulted in a more accurate (smaller) RMSE of 1,469.475, with a MAE of 1,037.611. This model ran faster with this data set than the minimally cleansed data - this executed in 19.2 seconds. This is result set C. However, applying the industry metric, that value was more conservative, that value was 92.571.

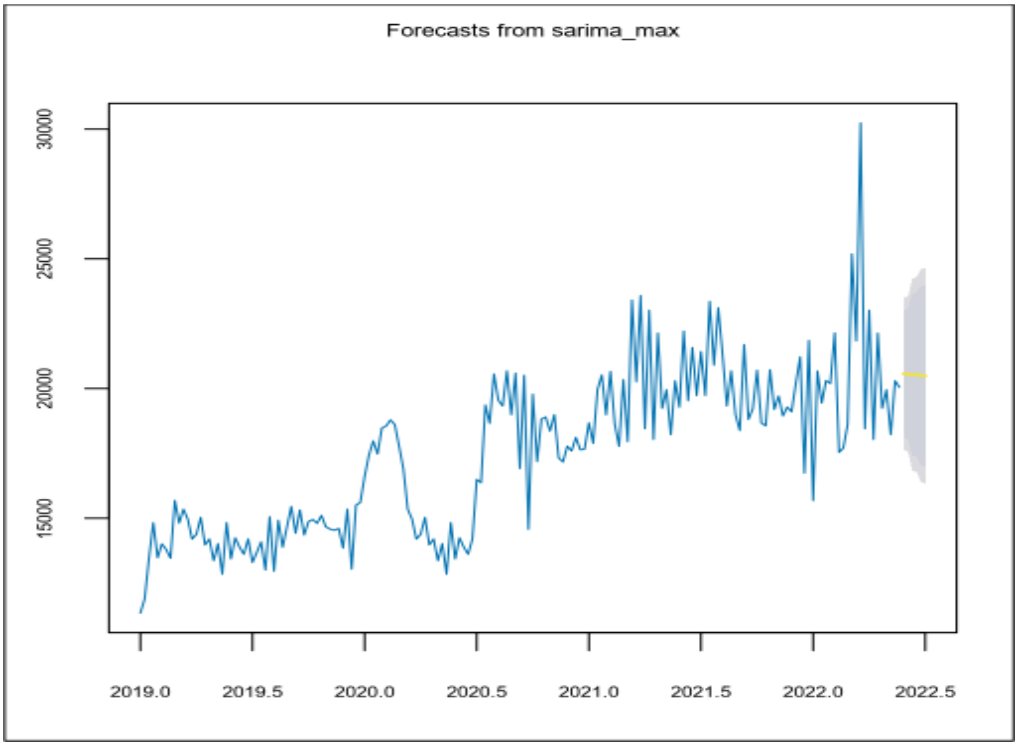
Which implies that if the food manufacturer had used the forecasted figures from this model, there would have been a surplus of product of this *MaterialID*. See Table 7 for the summary of this calculation.

Table 7. Industry Metric Calculation on Result Set C

Year	Week	Actual	Forecast
2022	23	14,985	20,569
2022	24	20,520	20,522
2022	25	21,555	20,548
Grand Total		57,060	61,639
	Industry Metric		**92.571%**

See Figure 5 to see the maximumly cleansed data forecasts plotted in the yellow and gray area. The yellow line is the forecasted amounts, whereas the grey plotted area is the 95% confidence interval around the forecast. Also, see Table 5 for the collective details regarding these two SARIMA approaches.

Figure 5. SARIMA Forecast Plotted - Maximumly Cleansed Data

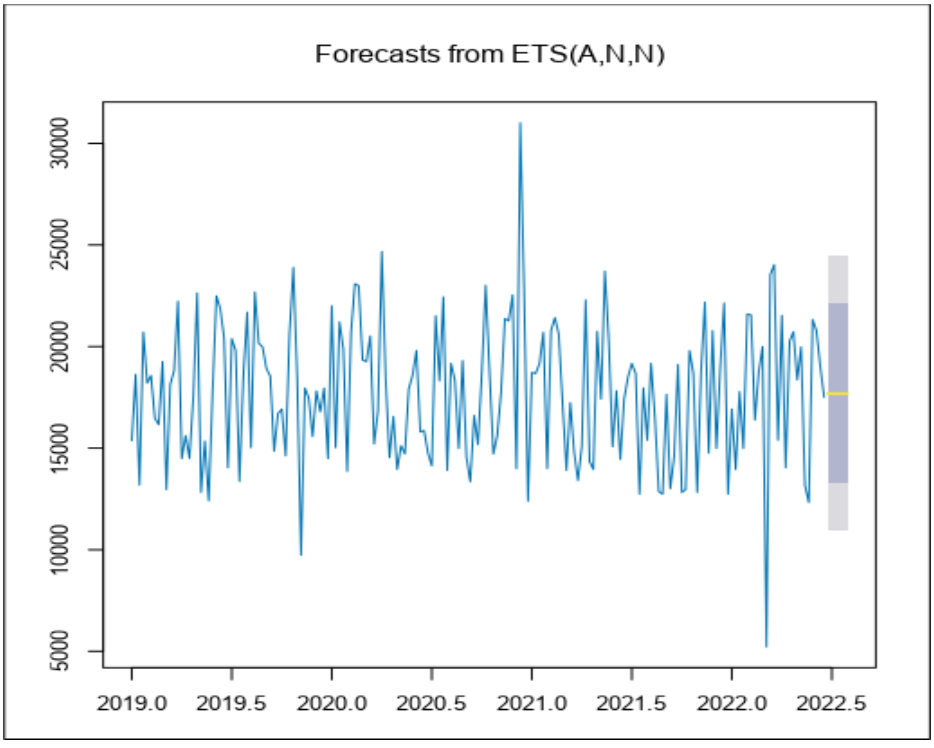


The next two approaches were executed using an ETS model. Putting the minimally cleansed data through the model, the RMSE resulted in 3,441.789, and the MAE was 2,807.048. The model finished within 16.9 seconds. This was the result set B. The industry metric was 107.655%, see Table 8 for the calculation of the industry metric for result set B. See Figure 6 to see the forecast values plotted for this first ETS approach, in gray and yellow. The yellow line is the forecasted amounts, whereas the grey plotted area is the 95% confidence interval around the forecast.

Table 8. Industry Metric Calculation on Result Set B

Year	Week	Actual	Forecast
2022	23	14,985	17,668
2022	24	20,520	17,668
2022	25	21,555	17,668
Grand Total		57,060	53,003
	Industry Metric		**107.655%**

Figure 6. ETS Forecast Plotted - Minimally Cleansed Data



When the maximumly cleansed data was run through this model, the RMSE was more accurate than the ETS model ran with the minimally cleansed data, with a result of 1,783.407 and a MAE of 1,249.334. This data (result set D) ran through in 15.4 seconds. The industry metric however, in this case was conservative at 95.618 (Table 9).

Table 9. Industry Metric Calculation on Result Set D

Year	Week	Actual	Forecast
2022	23	14,985	19,892
2022	24	20,520	19,892
2022	25	21,555	19,892
Grand Total		57,060	59,675
	Industry Metric		**95.618%**

Like the SARIMA model ran with maximumly cleansed data, the industry metric was more conservative. See Figure 7 for the forecast results of the second ETS approach. Also, see Table 10 for the summary of the ETS model runs.

Figure 7. ETS Forecast Plotted - Maximumly Cleansed Data

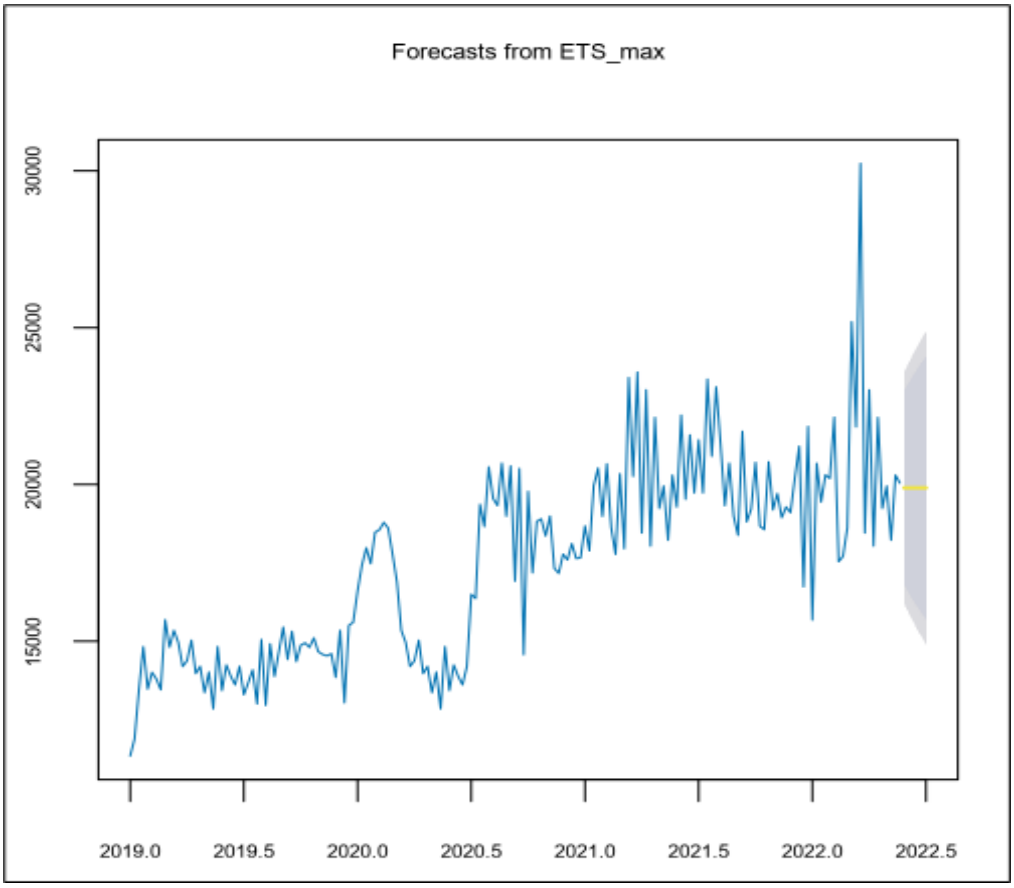


Table 10. Summary of SARIMA and ETS Results

Approach		SARIMA			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specific Metric**
A	**MIN Cleansed Data**	3,434.4078	2,256.4692	110.738%	23 seconds to run
C	**MAX Cleansed Data**	1,469.4755	1,037.6113	92.571%	19.2 seconds to run
	% increase in accuracy	*57%*	*54%*		*20% increased efficiency*
Approach		ETS			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specifi Metric**
B	**MIN Cleansed Data**	3,441.7892	2,807.0483	107.655%	16.9 seconds to run
D	**MAX Cleansed Data**	1,783.4071	1,249.3343	95.618%	15.4 seconds to run
	% increase in accuracy	*48%*	*55%*		*10% increased efficiency*

Is there a way to obtain more efficiencies, with more accuracy, within the industry metric standards, that is also a simple application of forecasting principles? The last attempt at making a better model was simply use the forecasting abilities built into Tableau. The minimally cleansed data was run through Tableau using a triple exponential smoothing model. The RMSE of this was 2.984, and the MAE was 1.989. The data source (if connected to a database or the like) will refresh in less than 5 seconds. If the data set grows to more than one million, it will take upwards of 1 minute to refresh. The industry metric was 98.531% (see Table 11), which means that the

ordering would be sufficient and ensure that there is plenty of product for the corporate entities. See Figure 8, and Tables 12 and 13 for the summary of these results.

Table 11. Industry Metric Calculation - Tableau Exponential Smoothing - Minimally Cleansed Data

Year	Week	Actual	Forecast
2022	23	14,985	20,027
2022	24	20,520	19,381
2022	25	21,555	18,503
Grand Total		57,060	57,911
	Industry Metric		**98.531%**

Figure 8. Tableau Exponential Smoothing – Forecast Plot - Minimally Cleansed Data

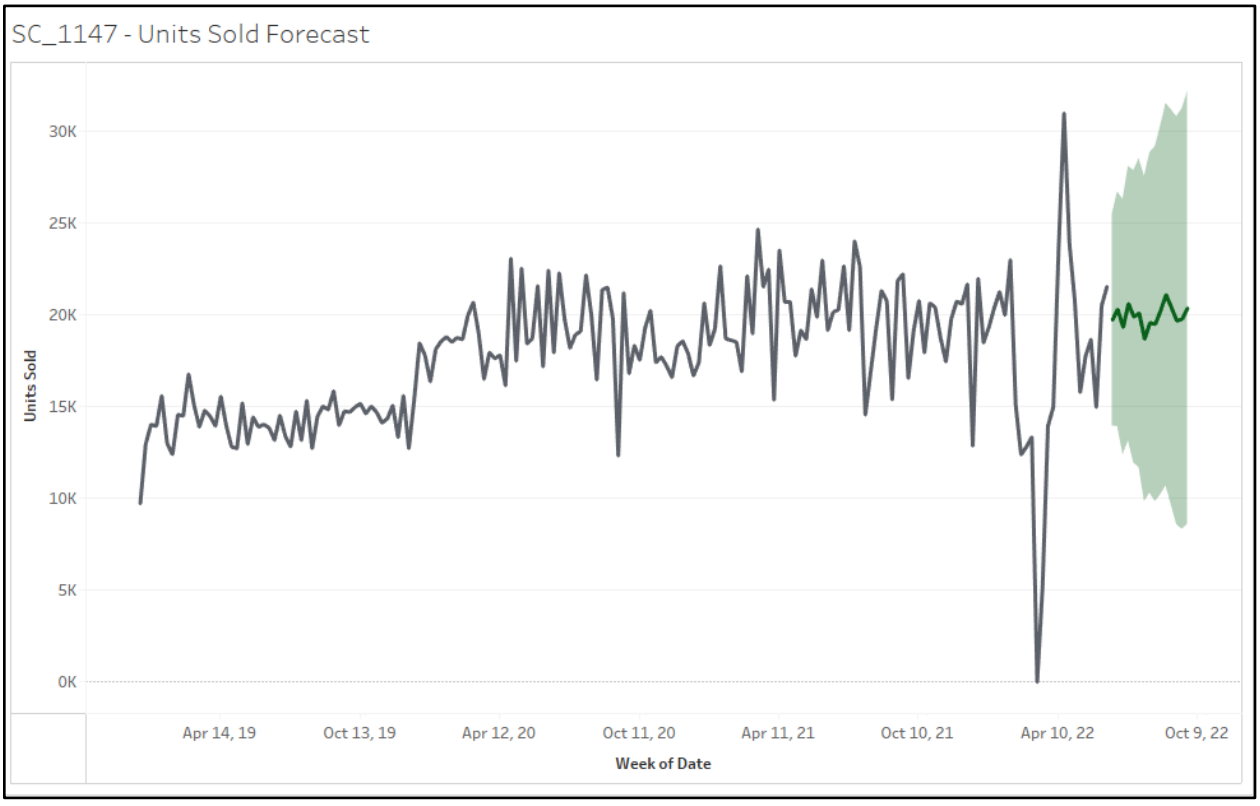


Table 12. Tableau Exponential Smoothing - Forecast Descriptive Statistics - Minimally Cleansed Data

All forecasts were computed using exponential smoothing.

Sum of Units Sold

Column	Model			Quality Metrics					Smoothing Coefficients		
SoldToCorporate Entity	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
SC_1147	Additive	Additive	Additive	2,984	1,989	0.85	11.1%	2,884	0.448	0.000	0.041

Table 13. ableau Exponential Smoothing - Forecast Model Summary - Minimally Cleansed Data

Options Used to Create Forecasts

Time series: Week of Date
Measures: Sum of Units Sold
Forecast forward: 18 weeks (May 29, 2022 – September 25, 2022)
Forecast based on: December 30, 2018 – May 22, 2022
Ignore last: 4 weeks (May 29, 2022 – June 19, 2022)
Seasonal pattern: 13 week cycle

Sum of Units Sold

Column	Initial	Change From Initial	Seasonal Effect				Contribution		
SoldToCorporate Entity	May 29, 2022	May 29, 2022 – September 25, 2022	High		Low		Trend	Season	Quality
SC_1147	20,027 ± 29.2%	-7.4%	August 28, 2022	6.1%	July 31, 2022	-5.6%	0.7%	99.3%	Poor

Using the maximumly cleansed in Tableau, the RMSE grew to 3.047, the MAE decreased to 1.819, and the industry metric was even more conservative at 91.982% (see Table 14). The increase of the RMSE is slight and in the opposite direction of the previous dataset ran through this model. However, the MAE decrease is a positive outcome and improves the accuracy of the results. The run time of this is the same as the previous Tableau forecasting model. See Figure 9, and Tables 15 and 16 for the summary of these results. Also, see Table 17 for the full summary of the stats of these two data sets run through the Tableau exponential smoothing model.

Table 14. Industry Metric Calculation - Exponential Smoothing - Maximumly Cleansed Data

Year	Week	Actual	Forecast
2022	23	14,985	21,418
2022	24	20,520	21,175
2022	25	21,555	19,441
Grand Total		57,060	62,034
	Industry Metric		**91.982%**

Figure 9. Tableau Exponential Smoothing - Forecast Plot - Maximumly Cleansed Data

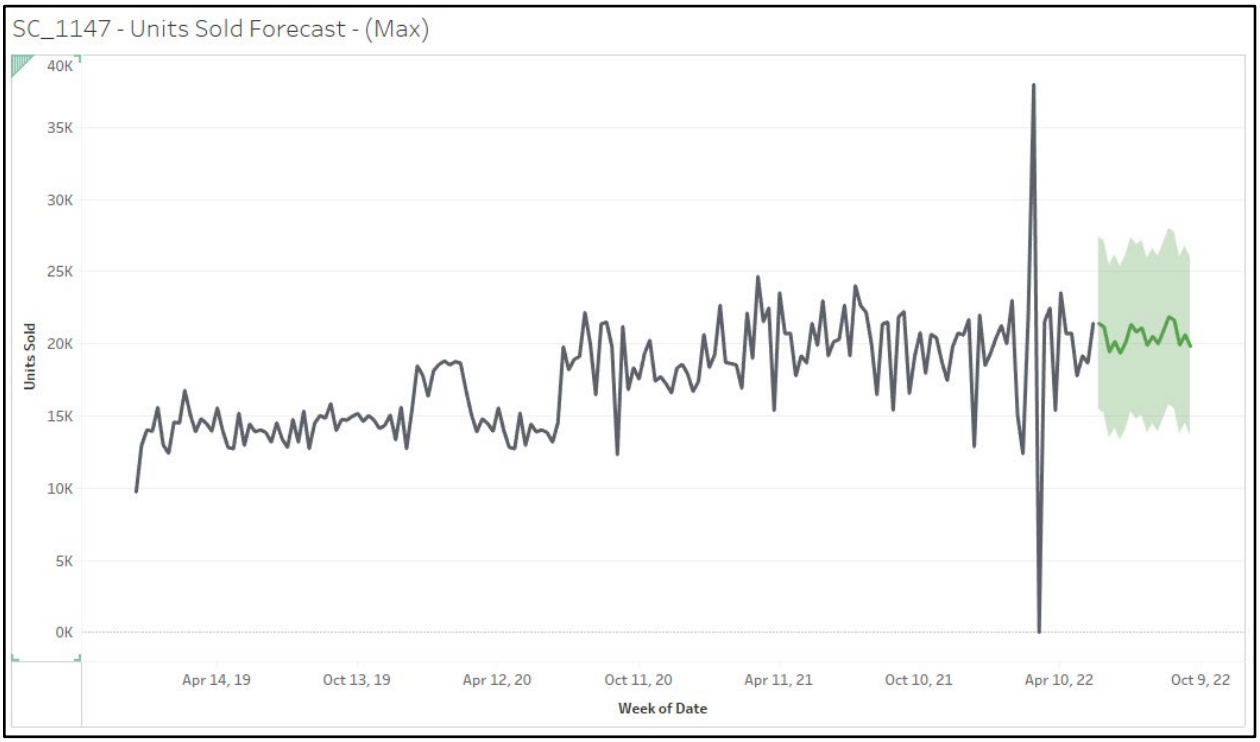


Table 15. Tableau Exponential Smoothing - Forecast Descriptive Statistics - Maximumly Cleansed Data

All forecasts were computed using exponential smoothing.

Sum of Units Sold

Column	Model			Quality Metrics					Smoothing Coefficients		
SoldToCorporate Entity	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
SC_1147	Additive	Additive	Additive	3,047	1,891	0.78	10.2%	2,892	0.057	0.000	0.030

Table 16. Tableau Exponential Smoothing - Forecast Model Summary - Maximumly Cleansed Data

Options Used to Create Forecasts

Time series: Week of Date
Measures: Sum of Units Sold
Forecast forward: 18 weeks (May 29, 2022 – September 25, 2022)
Forecast based on: December 30, 2018 – May 22, 2022
Ignore last: 4 weeks (May 29, 2022 – June 19, 2022)
Seasonal pattern: 13 week cycle

Sum of Units Sold

Column	Initial	Change From Initial	Seasonal Effect				Contribution		
SoldToCorporate Entity	May 29, 2022	May 29, 2022 – September 25, 2022	High		Low		Trend	Season	Quality
SC_1147	21,418 ± 27.9%	-7.5%	August 28, 2022	5.8%	September 25, 2022	-4.9%	5.6%	94.4%	Ok

Table 17. Summary of Tableau Results

		Exponential Smoothing Tableau			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specific Metric**
Tableau Min	**MIN Cleansed Data**	2.98400	1.98900	98.531%	< 5 secconds to refresh
Tableau Max	**MAX Cleansed Data**	3.04700	1.81910	91.982%	< 5 secconds to refresh
	change in accuracy	*-2%*	*9%*		

If these models were pushed to production, they would have dynamic connections to data sources that would automatically update on the required cadences. The ETL steps and subsequent modeling components would be developed and run automatically daily with a scheduling set of jobs, depending on the software applications available. These models assume the backend systems and capabilities exist to stand up these models. It is also assumed that there is adequate reporting and ordering software to push the model data downstream to populate those systems that handle the ordering. The timing found in each case, was the length of time to put this one small data set through each model, as if they were productionalized. Considering that this data set is a small portion of the data available from the food industry company, one can simply take the percentage gains on each scenario and apply them company-wide, if desired.

Results Discussion

The best results for accuracy, efficiency, and efficacy stemmed from the Tableau exponential smoothing model. The results from both the SARIMA and ETS (see Table 9), had promising directionality when it came to RMSE and MAE figures for the maximumly cleansed data. Both measurements had greater accuracy when utilizing the maximumly cleansed data set. The SARIMA MAE had a decrease in errors equivalent to 54%. The SARIMA RMSE decreased in aggregate errors by 57%. The ETS MAE had a decrease of errors by 55%. And the ETS RMSE had a decrease of errors equal to 48%. A decrease in aggregate errors translates into more accuracy for the model. This helps with overall confidence in the model, as the accuracy improves.

Consider the efficiency gains of the maximumly cleansed data versus the minimally cleansed data. For the SARIMA model, the maximumly cleansed data ran 20% faster. The ETS model handled the maximumly cleansed data 10% faster than the minimally cleansed data.

It is seen that the maximumly cleansed data ran more quickly and had smaller error aggregations than the minimally cleansed data. Also, the industry metric for both maximumly cleansed data scenarios was more conservative than that of the minimally cleansed data results. All good signs for both the SARIMA and ETS models.

The Tableau Exponential Smoothing results were more efficient, accurate and conservative than the SARIMA and ETS model outputs. Looking at Table 18, it is seen by the red highlighted cells that the lowest RMSE, the lowest MAE, the most conservative industry metric, and the quickest timing can all be found within the two Tableau models and their resulting metrics.

Table 18. Summary of Results from All Model Approaches

Approach		SARIMA			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specific Metric**
A	**MIN Cleansed Data**	3,434.40785	2,256.46920	110.738%	23 seconds to run
C	**MAX Cleansed Data**	1,469.47555	1,037.61134	92.571%	19.2 seconds to run
	% increase in accuracy	*57%*	*54%*		*20% increased efficiency*
Approach		ETS			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specifi Metric**
B	**MIN Cleansed Data**	3,441.78929	2,807.04831	107.655%	16.9 seconds to run
D	**MAX Cleansed Data**	1,783.40712	1,249.33436	95.618%	15.4 seconds to run
	% increase in accuracy	*48%*	*55%*		*10% increased efficiency*
		Exponential Smoothing Tableau			
Result Set	**Result Set Description**	**RMSE**	**MAE**	**Industry Metric**	**Specific Metric**
Tableau Min	**MIN Cleansed Data**	2.98400	1.98900	98.531%	< 5 seconds to refresh
Tableau Max	**MAX Cleansed Data**	3.04700	1.81910	91.982%	< 5 seconds to refresh
	change in accuracy	*-2%*	*9%*		

There is a wide variety of options available for the analyst that will be reviewing this data daily. In this summary (Table 18), there are six options of paths to develop toward more efficient or more accurate results. The company must look at the pros and cons of each and decide

accordingly as to what path suits their needs best. Table 19 displays the steps that would be taken daily if these models were operationalized and used by the analysts at the company.

Each of these models, if productionalized, most of the work would be automated to allow for the most efficient use of the analyst's time. The automation would take care of sourcing the appropriate data each day, computing the results of each model, sending the data to the ERP, and publishing any necessary reports for the analysis of the end users. The analyst would receive automated notifications or reports when those daily processes were complete and ready for their review. The analyst would then review the results of the forecasted units per each dimension in the desired data set and ensure that the ERP was equipped with those results to ensure that all entities downstream in the supply chain had the required quantities.

Table 19. Summary of Models, Steps, & Time Allotted

Model	Steps	Automated System Time/Load / Day	Analyst Time / Day	Assumptions
SARIMA + Minimally Cleansed Data	1) The system will automatically run this process at the top of the day (typically shortly after midnight) 2) The system will publish the results to the ERP &/or send reports to the end users 3) Analyst will analyze results and apply or adjust the order fulfillment system accordingly	23 seconds x the permutations (871 noted at the onset of the project) = 5.56 hours	The time it takes to analyze data and perform any updates to ERP.	1) Model is published to a production location 2) Applications are all synthesized together for automated communication and push/pull of data 3) Daily schedulers allow for all systems to pass data downstream across systems 4) Data sources are dynamically sourced to allow for the daily operations to occur 5) Automation & subscription of results is a built-in feature of the application(s) 6) This is the most conservative time estimate, efficiencies could be gained in the development process to bring this down substantially.
SARIMA + Maximally Cleansed Data		19.2 seconds x the permutations (871 noted at the onset of the project) = 4.64 hours		
ETS + Minimally Cleansed Data		16.9 seconds x the permutations (871 noted at the onset of the project) = 4.09 hours		
ETS + Maximally Cleansed Data		15.4 seconds x the permutations (871 noted at the onset of the project) = 3.73 hours		
Exponential Smoothing + Minimally Cleansed Data	1) The Tableau Server will have an extract refresh set up to dynamically pick up the newest data for the day and pull it in. 2) The system will publish the results and send reports to the end users 3) Analyst will analyze results and apply or adjust the order fulfillment system accordingly.	5 seconds x the permutations (871 noted at the onset of the project) = 1.21 hours		
Exponential Smoothing + Maximally Cleansed Data				

Conclusion

Three models were utilized - SARIMA, ETS, and a triple exponential smoothing model in Tableau. The approaches compared a minimally cleansed data set versus a maximumly cleansed data set and their behavior through the three models. The results were compared, for one test case - one corporate entity and one *MaterialID*.

The operations process of this project focused on normalizing the given data, and then comparing it to even further normalized data. This process assumed that this normalization process would lead to a beneficial outcome. This is confirmation bias. Humans have beliefs and tend to search for evidence that supports their suppositions.

In addition, this project selected one *SoldToEntity* to focus on, in addition to the already narrow focus of having one product category represented in the data set. Theorizing that if the models suited that one corporate entity, then it would be equally effective to all other entities. This is selection bias. It is probable that if these forecasting models were applied to all corporate entity data, there could be a wide array of results, and not necessarily as pointed as the results for this one corporate entity.

Given these ever-present biases in this project, it is imperative that all results be judged in the light of the fact that the utilized data elements are very narrowly focused and do not necessarily represent the results of the entire population.

It should also be noted that this dataset was fully obfuscated, no distinguishing labels or customer names were made known, and the name of the company that provided the data will not be published. Therefore, there are no ethical issues regarding the creation of this project and the subsequent documentation dissemination.

Whenever handling data, one should be considerate of all ramifications and downstream effects of any data normalization processes. There is always responsibility to safeguard the data and its integrity. There is also the need to consider the phrase ‘do no harm’. Physicians live by that code, as should all data stewards. The end goal should make data stronger and the processes surrounding the data stronger and more reliable. Then all consumers of that data can have more confidence and have more effective processes as a result. More effective processes can result in efficiency gains, less waste, and greater accuracy.

The result sets were compared for accuracy, efficiency, and efficacy regarding the industry metric. The highlights of this project start with the minimally cleansed data sets faring well when compared to the industry metric. The maximumly cleansed data sets were even closer to the industry metrics - which would result in conservative buffers of product when fulfilling orders for the customers downstream.

The most efficient runs were the results of using the maximumly cleansed data sets. Ten to twenty percent gains on efficiency at this level of investigation, could potentially lead to overall gains at a company ordering level. Small efficiencies, as noted by the Edwards Deming method, can lead to long lasting gains downstream.

The one model that not only fared well for accuracy, efficiency, and efficacy, was the Tableau exponential smoothing model. This model realized good accuracy, very conservative industry metrics, and the run/refresh time was minimal if placed into production. This was the simplest of all the models, with the lowest level of development time necessary. The exponential smoothing model is the recommended model to utilize for the food industry company. Sometimes the best solution is the simplest one.

References

AltexSoft Inc. (2020). *Demand forecasting methods: Using machine learning and predictive analytics to see the future of Sales.* Medium. Retrieved July 16, 2022, from https://medium.com/datadriveninvestor/demand-forecasting-methods-using-machine-learning-and-predictive-analytics-to-see-the-future-of-137b2342f6c4

Dolphin, R. (2021). *LSTM networks: A detailed explanation*. Medium. Retrieved July 13, 2022, Towards Data Science. from https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

Dou, Z., Sun, Y., Zhang, Y., Wang, T., Wu, C., & Fan, S. (2021). Regional Manufacturing Industry Demand Forecasting: A Deep Learning Approach. *Applied Sciences*, 11(13), 6199. https://doi.org/10.3390/app11136199

Ellis, P. (2016). *Error, trend, seasonality - ETS and its forecast Model Friends*. free range statistics. Retrieved August 6, 2022, from http://freerangestats.info/blog/2016/11/27/ets-friends

Falatouri, T., Darbanian, F., Brandtner, P., & Udokwu, C. (2022). Predictive analytics for demand forecasting – A comparison of sarima and LSTM in retail SCM. *Procedia Computer Science*, *200*, 993–1003. https://doi.org/10.1016/j.procs.2022.01.298

Fallmann, D. (2021). *Council post: Human cognitive bias and its role in AI*. Forbes. Retrieved August 6, 2022, from https://www.forbes.com/sites/forbestechcouncil/2021/06/14/human-cognitive-bias-and-its-role-in-ai/?sh=326787eb27b9

Guinoubi, S., Hani, Y., & Elmhamedi, A. (2021). Demand forecast; a case study in the Agri-Food Sector: Cold. *IFAC-PapersOnLine*, *54*(1), 993–998. https://doi.org/10.1016/j.ifacol.2021.08.191

Hennel, M. (2006). Food manufacturers turn to demand forecasting to gain a competitive edge. *Food Manufacturing, 19*(3), 20-20,22. Retrieved from http://ezproxy.utica.edu/login

Iyer, N. (2020). *Forecasting food demand*. Medium. Retrieved July 11, 2022, from https://towardsdatascience.com/forecasting-food-demand-applying-neural-networks-to-the-meal-kit-industry-6f1e3b2207aa

Kim, S. (2021). Ai and robots are a minefield of cognitive biases. *IEEE Spectrum*. Retrieved August 6, 2022, from https://spectrum.ieee.org/humans-cognitive-biases-facing-ai

Mediavilla, M. A., Dietrich, F., & Palm, D. (2022). Review and analysis of Artificial Intelligence Methods for demand forecasting in supply chain management. *Procedia CIRP*, *107*, 1126–1131. https://doi.org/10.1016/j.procir.2022.05.119

Merlo, C. (2015). *Why dairy demand has become more elastic - dairy herd*. Why Dairy Demand Has Become More Elastic. Retrieved July 30, 2022, from https://www.dairyherd.com/news-news/new-products/why-dairy-demand-has-become-more-elastic

Miguéis, V. L., Pereira, A., Pereira, J., & Figueira, G. (2022). Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and machine learning. *Journal of Cleaner Production*, *359*, 131852. https://doi.org/10.1016/j.jclepro.2022.131852

Riveroll, F. (2020). *Immensely improving every 'Walmart Sales' demand forecasting model*. Medium. Retrieved July 16, 2022, from https://medium.com/analytics-vidhya/immensely-improving-every-walmart-sales-demand-forecasting-model-3e9449f892ea

Roldos, I. (2021). *8 effective data cleaning techniques for better data*. MonkeyLearn Blog. Retrieved July 14, 2022, from https://monkeylearn.com/blog/data-cleaning-techniques/

Henshall, A. (2020). *How to use the Deming cycle for continuous quality improvement: Process street: Checklist, Workflow and SOP software*. Process Street. Retrieved August 1, 2022, from https://www.process.st/deming-cycle/

Schmit, T. M., & Kaiser, H. M. (2006). Forecasting fluid milk and cheese demands for the next decade. *Journal of Dairy Science*, *89*(12), 4924–4936. https://doi.org/10.3168/jds.s0022-0302(06)72543-7

Scioscia, J. (2021). *How foodservice suppliers can use predictive analytics to meet demand*. QSR magazine. Retrieved July 13, 2022, from https://www.qsrmagazine.com/outside-insights/how-foodservice-suppliers-can-use-predictive-analytics-meet-demand

Thete, J. (2022). *A stochastic model for demand forecasting in python*. Medium. Retrieved July 16, 2022, from https://medium.com/mlearning-ai/a-stochastic-model-for-demand-forecating-in-python-a1b568b80b94

Traasdahl, A. (2020). *How a History of Slow Technology Adoption Across Food Supply Chains Nearly Broke Us.* Food Safety Tech. https://foodsafetytech.com/column/how-a-history-of-slow-technology-adoption-across-food-supply-chains-nearly-broke-us/

Seyedan, M., Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *J Big Data* 7, 53. https://doi.org/10.1186/s40537-020-00329-2

User, G. (2022). *How to implement the kaizen method with your team*. Stormboard. Retrieved August 1, 2022, from https://stormboard.com/blog/implement-kaizen-method-team

Vandeput, N. (2021). *Forecast KPI: RMSE, Mae, Mape & Bias*. Medium. Retrieved July 15, 2022, from https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Appendices

GitHub Repository: https://github.com/tammylessley/tammylessley/tree/Utica-University