EFFECT OF OVERRIDE SIZE ON FORECAST VALUE ADD

by

Jeffrey A. Baker
Bachelor of Science, Chemical Engineering, University of Illinois, 1987
Master of Business Administration, University of Michigan-Flint, 1995

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN SUPPLY CHAIN MANAGEMENT AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2018

(c) 2018 Jeffrey A. Baker. All rights reserved.
The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author

Jeffrey A. Baker
Department of Supply Chain Management
May 11th, 2018

Signature of Author

Dr. Sergio Caballero
Research Scientist
Thesis Advisor

Signature of Author

Dr. Nima Kazemi
Postdoctoral Associate
Thesis Advisor

Signature of Author

Prof. Yossi Sheffi
Director, MIT Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering
ABSTRACT

Business forecasting frequently combines quantitative time series techniques with qualitative expert opinion overrides to create a final consensus forecast. The objective of these overrides is to reduce forecast error, which enables safety stock reduction, customer service improvement, and manufacturing schedule stability. However, these overrides often fail to improve final forecast accuracy. Process mis-steps include small adjustments, adjustments to accurate statistical forecasts, and adjustments to match financial goals. At best, these overrides waste scarce forecasting resources; at worst, they seriously impact business performance.

This report offers a framework for identifying overrides that are likely to improve forecast accuracy. A new class of metrics, Dispersion-Scaled Overrides, was developed as an input to the framework. Other inputs included statistical forecast accuracy and auto-correlation. Classification techniques were used to identify whether an override was likely to create forecast value add. These techniques were found to be approximately 80% accurate in predicting success. This result suggests that using Dispersion-Scaled Overrides alongside common forecast accuracy metrics can reliably predict forecast value add. In turn, this can maximize the value that experts add to the business forecasting process.
Acknowledgements

My deep gratitude goes first to Dr. Sergio Caballero, whose guidance led to break-throughs which made this work possible. I’d also like to express my thanks to Dr. Nima Kazemi for his help and suggestions to improve upon this work.

To Dr. Chris Caplice, Dr. Eva Ponce, Dr. Alexis Bateman, and the entire SCx staff, thank you for your dedication to SCM MicroMasters program, which provided me the opportunity to attend MIT. You were the catalyst for this journey.

To my friends in the very first SCM blended cohort – we are the pioneers! I’ve never been part of such a diverse, yet cohesive team. My deepest thanks for your friendship and support on our journey. Apropos of our diversity, I express my thanks by saying Thank You. Gracias. Obrigado. Danke. ありがとう. धन्यवाद. วันยิ่ง บุญ. 多謝. شكرا. Terima kasih. Merci. Спасибо. Asante. Ευχαριστώ.

To my parents and step-parents, Drs. Bob and Margaret Baker, Dr. James Kirchner, and Kathe Kirchner, thank you for teaching me the value of education, and that the quest for knowledge is a life-long endeavor.

To Kim and Matt, we’ve always told you to make good choices. Education is one of them. Find your passion, follow it relentlessly, and enjoy the journey – at any age. I love you both.

To the love of my life, Chris. Your love and support throughout the years means the world to me, and without you I’d be nothing. You are my candle in the window. I dedicate this work to you.
Table of Contents

Abstract ........................................................................................................................................... 1
Acknowledgements .......................................................................................................................... 2
Table of Contents ............................................................................................................................ 3
List of Figures .................................................................................................................................. 5
List of Tables .................................................................................................................................. 6
1 Introduction ............................................................................................................................. 7
  1.1 Background - Demand Forecasting ..................................................................................... 7
  1.2 Research Problem ................................................................................................................. 8
  1.3 Current Knowledge Gap ................................................................................................. 9
  1.4 Methodology ....................................................................................................................... 9
  1.5 Hypothesis .......................................................................................................................... 10
  1.6 Summary ............................................................................................................................ 11
2 Literature Review .................................................................................................................. 12
  2.1 Forecasting and Supply Chain Management .................................................................. 13
  2.2 Do Overrides Improve Accuracy? ..................................................................................... 14
  2.3 Data Analysis and Metrics .............................................................................................. 17
  2.4 Literature Gap ................................................................................................................ 20
3 Methodology .......................................................................................................................... 23
  3.1 Time Series Data and Accuracy Metrics ........................................................................ 26
  3.2 Time Series Decomposition and Residual Analysis ...................................................... 28
  3.3 Autocorrelation ................................................................................................................ 31
  3.4 Dispersion-Scaled Overrides .......................................................................................... 31
  3.5 Opportunity Indicators ................................................................................................. 33
  3.6 Data Review ..................................................................................................................... 35
  3.7 Data Mining Analytics ..................................................................................................... 36
  3.8 Methodology Summary ................................................................................................. 38
4 Results and Discussion ........................................................................................................ 41
  4.1 Research Problem ............................................................................................................ 41
  4.2 Input Data Results .......................................................................................................... 42
  4.3 Opportunity Indicators ................................................................................................... 47
  4.4 Data Mining Analytics .................................................................................................. 49
List of Figures

Figure 1: Methodology Data Flow Overview ........................................................................... 25
Figure 2: Sample output from STL procedure ........................................................................ 29
Figure 3: Sample residuals from STL procedure ..................................................................... 30
Figure 4: Histogram of Residuals .............................................................................................. 30
Figure 5: Autocorrelation of residuals ..................................................................................... 31
Figure 6: Override Percent Histogram ..................................................................................... 42
Figure 7: Median FVA, by Quartile ......................................................................................... 45
Figure 8: Override Size/Direction vs. FVA ............................................................................. 46
Figure 9: Classification Tree .................................................................................................... 52
Figure 10: Variable Importance Plot ....................................................................................... 55
Figure 11: Standard Deviation Scaled Override vs. Forecast Value Add (%) ......................... 59
Figure 12: Framework for Maximizing Value Add .................................................................. 62
List of Tables

Table 1: Example of Scaled Override vs. Percent Override ....................................................... 21
Table 2: Overview of Data Elements .................................................................................................. 35
Table 3: Data Mining – Classification Techniques ........................................................................ 37
Table 4: Confusion Matrix .............................................................................................................. 37
Table 5: Mean and Median of Override Sizes .................................................................................. 43
Table 6: Accuracy Metrics for Statistical and Consensus Forecasts ............................................. 44
Table 7: Opportunity Indicators ......................................................................................................... 48
Table 8: Response and Predictor Variables ...................................................................................... 51
Table 9: Confusion Matrix for Classification Tree Method .......................................................... 53
Table 10: Confusion Matrix for Random Forest Method ............................................................. 55
Table 11: Confusion Matrix for Boosted Tree Method .................................................................... 56
Table 12: Variable Statistical Significance ...................................................................................... 57
Table 13: Confusion Matrix for Logistic Regression ........................................................................... 58
Table 14: Summary of Selected Metrics for Classification Techniques .............................................. 58
Table 15: Data file structure for Input_Data.csv .............................................................................. 69
Chapter 1 introduces the concept of demand forecasting and the challenges which it presents in the business environment. An overview of time series models and forecast accuracy measures is provided and serves as the basis for defining the research problem. The current knowledge gap in the literature is defined. Finally, a classification framework is proposed, and the research methodology designed to test it is described.

1.1 Background - Demand Forecasting

Demand forecasting is a critical element of supply chain management, and directly impacts a firm's ability to effectively and efficiently meet customer demand (Chopra & Meindl, 2010). Improving forecast accuracy is important because it drives optimum inventory levels, increased manufacturing efficiency, and improved attainment of customer service levels (M. A. Moon, Mentzer, & Smith, 2003).

Time series models are commonly used in business today to create future demand forecasts (Malehorn, 2005). These models utilize historical patterns in trend and seasonality to predict future values. The underlying assumption of time series forecasting is that the past patterns in the data will repeat themselves in the future. If there are new events or changes in the environment, integrating managerial adjustments, known as overrides, often reduce forecast error (Donihue, 1993).

However, there are several contrary studies which suggest overrides to the baseline statistical forecast may introduce bias, increase error, or are unnecessary (Lawrence, Goodwin, O’Connor, & Önkal, 2006; Eroglu & Croxton, 2010). Over-forecasting bias creates excessive inventory,
which increases holding costs and the working capital required to support it. Under-forecasting bias, while rare, creates the potential for stockouts. Because forecast error is directly related to the amount of safety stock required, increased forecast error will either lead to increased safety stock requirements, or reduced service levels at the same inventory level. Unnecessary overrides, which add little or no value, tie up scarce resources which could be better used elsewhere. Creating a classification framework to identify and minimize these forecast errors is the goal of this research.

1.2 Research Problem

Improving final forecast accuracy for products or aggregated groups of products is critical because it drives optimum inventory levels, increases manufacturing efficiency, and improves customer service levels. Despite the benefits of reduced forecast error, there is usually no guidance on which statistical forecasts are critical to review, or the magnitude of the override which is statistically different from the baseline forecast and therefore likely to improve accuracy. The business need for increased forecast accuracy, coupled with the lack of guidance on where to focus, has created common situations. First, overrides to the statistical forecast are made for products where it would have been better to accept the statistical forecast as the final forecast. Second, small overrides are entered, creating a final forecast which is indistinguishable from the statistical forecast due to underlying variation. Both situations represent wasted effort. So, the goal of forecast accuracy improvement must be balanced against the time and effort required.

Given the desire to improve forecast accuracy and large number of potential overrides, the research question then becomes this: Can a framework be created which identifies where efforts could improve forecast accuracy, or where they might add no value at all?
1.3 Current Knowledge Gap

While previous research has investigated the impact that overrides have on forecast accuracy, this research seeks to address two gaps in this field of study. First, there has been no research in scaling overrides based on dispersion measures. Dispersion-scaled overrides (DSO's) take into consideration the underlying variability of the baseline demand, creating a signal-to-noise style of metric. Second, there is no research which attempts to classify overrides as value added or non-value added based upon dispersion-scaled overrides and baseline statistical forecast accuracy.

1.4 Methodology

This research aims to develop a qualitative framework for classifying a proposed override as either value added or non-value added based on dispersion-scaled overrides, baseline statistical forecast accuracy, and autocorrelation.

To accomplish this, time series data from forecasting triples consisting of actual demand, statistical forecast, and final forecast are required. Based on the accuracy of the statistical forecast and the final forecast, an improvement metric can be calculated. This metric, which reflects the success of the judgmental override, is known as Forecast Value Added (FVA). FVA values can be compared to user-defined FVA thresholds (FVA_{crit}) to classify overrides as either value added or non-value added (NVA). FVA is the response variable of interest.

Predictor variables fall into two categories. The first is dispersion-scaled overrides. The actual product demand time series may be decomposed into trend, seasonal, and residual components. For the residual component, the dispersion measures of standard deviation, trimmed mean absolute deviation, and median absolute deviation are calculated. The judgmental override is divided by
these measures of dispersion, creating a signal-to-noise relationship between the override and the underlying variability in the time series.

The second category of predictor variables are opportunity indicators. One is autocorrelation in the residuals, which indicate if there is information not captured in the statistical forecast. The others are measures of baseline statistical forecast accuracy – Root Mean Square Percentage Error (RMSPE), Mean Absolute Percent Error (MAPE) and Median Absolute Percent Error (MdAPE).

To analyze the relationship between the Forecast Value Add response variable and the predictor variables, data mining classification techniques are used. Methods include classification tree, boosted tree, random forest, and logistic regression.

1.5 Hypothesis

It is hypothesized that the probability of creating a value-added override increases based on two or three key factors. The first factor is the accuracy of the baseline statistical model. If the statistical forecast is performing poorly, there are probably opportunities to improve upon it. Conversely, a well-performing model will be difficult to improve. The second factor is autocorrelation. If present, it represents information not reflected in a time series model. The third factor is the size of the override compared to the underlying variation. Overrides indistinguishable from the random noise of random are unlikely to add value.

Developing a classification framework which relates dispersion-scaled overrides, baseline forecast accuracy, and auto-correlation to forecast value add will enhance the efficiency and effectiveness of the business forecasting process. Efficiency will increase as small overrides are flagged as non-value added, and no time is spent entering them. Effectiveness will increase as entering an override
which is flagged as non-value add will prompt discussion and further rationalization. Furthermore, this framework will be self-correcting. Success in creating FVA overrides will make it more likely that future overrides are flagged as FVA. Conversely, poor judgement in entering overrides will make it more likely that future overrides will be flagged as non-value added (NVA).

1.6 Summary

Chapter 1 reviewed demand forecasting, provided an overview of time series models and forecast accuracy measures, and defined the research problem and current knowledge gap. The classification framework and the research methodology designed was described. Chapter 2 will provide an extensive review of the literature which is pertinent to this thesis.
2 Literature Review

Business forecasting is viewed as the starting point for supply chain management, especially as manufacturing strategies have transitioned from a push to pull. A pure push strategy is production-centric, and the manufacturer simply pushes product into the supply chain and attempts to sell it. The more common approach, however, is the customer-centric pull model. Customer demand is the primary driver, and the supply chain must synchronize accordingly.

In this environment, business forecasting is essential to understand customer demand. The better the final forecast represents true customer demand, the better the supply chain will be able to respond. This response must be both effective and efficient. Being effective means that the customer demand is satisfied; being efficient means that the response is done in a cost-effective manner which allows the supply chain to be profitable.

To create an accurate forecast of customer demand, businesses commonly combine quantitative techniques with qualitative expert opinion to create the final consensus forecast. A commonly used quantitative technique is time series modeling, which predicts future demand based on past demand. The name time series derives from the fact that in these models, time is the only driver. Therefore, future patterns of trend and seasonality will look past patterns of demand and seasonality. To the extent that this assumption holds true, forecasts will be accurate, and the supply chain may be managed to have both an efficient and effective response. However, due to competitive pressures and other business environment changes, this assumption often breaks down and the errors in the statistical forecast increase.
It is because of these breakdowns that judgmental overrides from business experts are often, although not always, desirable. However, overrides from these experts often fail to improve final forecast accuracy.

The Literature Review explores academic works in the areas of judgmental overrides to the forecast, error metrics, and forecast value added analysis. It concludes with a discussion of a gap in the literature, which this analysis seeks to fill.

2.1 Forecasting and Supply Chain Management

Forecasting plays a crucial role in supply chain management, as it impacts production scheduling, inventory management, and procurement (Chopra & Meindl, 2010).

To enable an efficient and effective response to customer demand, production scheduling and inventory management require accurate demand information. Accurate forecasts allow optimum synchronization between the manufacturing plan and inventory levels. This is the well-known economic order quantity, where the costs of manufacturing are balanced against the costs of holding inventory. When multiple products are run on the same line, sequence optimization can minimize the costs of sequence-dependent change overs. Conversely, poor demand information creates many issues. Errors in the forecast could cause unplanned schedule changes, creating waste and increased manufacturing costs as schedules are needlessly adjusted. Poor forecasting will result in either stock-out or excess inventory, or perhaps both. Either situation represents additional costs for a company, and is undesirable from a business standpoint (M. Moon & Alle, 2015). Businesses may avoid the risk of stock-out by carrying additional safety stock, but the amount of required for a given service level is directly impacted by the forecast error (Huang, Hsieh, & Farn, 2011).
Procurement, both short term purchases and long-term capital expenditures, need forecasting to accurately predict requirements. Research by Syntetos, Nikolopoulos, & Boylan (2010) demonstrated that forecast adjustments which drive error reduction have a significant impact on reducing the inventory level of raw materials. This effect is similar to that discussed above for finished goods. Increased inventory performance may then be used to either decrease inventory level at a consistent service level, or to increase service at the same inventory level.

Based upon the impacts discussed above, there is considerable interest in reducing forecast error. Given the limitations of statistical models, it is desirable to improve accuracy through judgmental overrides.

2.2 Do Overrides Improve Accuracy?

Judgmental overrides are commonly used in business forecasting, particularly when external factors are likely to create an impact (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). The typical business process is to create the baseline statistical forecast automatically in a software tool, and then have the forecast be reviewed by business personnel. While the review process is company specific, it often includes the personnel from the demand management, sales, marketing, and business leadership functions. The purpose for the review is that while a forecasting model may incorporate many aspects of demand – such as trend and seasonality – it does not incorporate all influences on demand. The are many influences where human input is desirable. These factors could include insights into competitor actions, customer expansion plans, changing consumer tastes, or a myriad of other influences which are difficult to incorporate into a forecasting model. Judgmental overrides are designed to improve final forecast accuracy by integrating expert opinion into the forecasting process.
The initial contribution to the literature was presented in a series of papers by Mathews and Diamantopolous (Mathews & Diamantopoulos, 1986, 1989, 1992), which documented the impact that judgmental overrides had on forecast error. Mathews & Diamantopoulos (1986) documented decreases in forecast error through judgmental overrides of the baseline forecast. More forecasts were improved than made worse, although it was noted that biased increased. In their 1989 work, they concluded that the effectiveness of forecast overrides was a function of the product, not the individual creating the adjustment. The 1992 study noted marginal improvements in error reduction and recommended a formal system of reviewing overrides.

Subsequent research by Fildes et al., (2009) investigated the impact of override size and direction upon forecast accuracy. Negative adjustments were found to be more likely to improve forecast accuracy. Positive adjustments were less likely to increase accuracy, and in many cases reduced accuracy. This behavior driven by two factors, the first being optimism bias. Forecasters and others who enter overrides typically want to show increases in demand. As a result, upward revisions are seldom challenged, and final forecast accuracy suffers. Negative adjustments, however, are usually undesirable. They are often challenged and revised to most accurately reflect the business situation, resulting in increased accuracy. The second driver is confusion between the forecast and the business plan. Forecasts are based on data and should be unbiased, while the business plan is aspirational and based on the desires of management. An incorrect, yet unfortunately frequent response is to enter an override to make the forecast match the business plan.

The size of the adjustment also played a critical role. The smallest quartile of adjustments did not reduce error, which represented a waste of the adjuster’s time. Large, negative adjustments usually decreased error, across all companies. Large, positive adjustments decreased error in
some companies, but in others they increased the error. Their recommendation was to avoid
small adjustments, which were likely due to adjusters seeing false patterns in the data, or the
adjuster trying to prove their role is needed in the process. Avoiding small adjustments would
free up time. Their work also highlight that wrong-sided adjustments are particularly damaging
for forecast accuracy.

Several number of papers were published by Franses and Legerstee. (P. H. Franses & Legerstee,
2009) concluded that 90% of forecasts are adjusted, and that there is a slight preference to
upward adjustments. Forecaster overrides were also related to their past override, i.e. they
develop consistent habits over time. Another point was that when evaluating the forecast,
adjusters fail to realize that past results, which predicts future demand, are already incorporated
to some extent into the current forecast. This leads to a double counting effect.

Franses & Legerstee (2010) concluded that experts put too much weight on their adjustments,
and that forecast accuracy could be improved by reducing the weights that forecaster adjustments
had on the final forecast. Overall, the adjusted final forecasts were no better than the statistical
baseline model.

Franses & Legerstee's (2011) study improved upon the 50% model + 50% manager adjustment
weighting scheme proposed by Blattberg and Hoch (Blattberg, 1990). Franses and Legerstee
suggested incorporating past performance of experts to modify the split ratio. Too novice or too
experienced tended to make larger adjustments, and both cases need to be considered in override
evaluation.

Expert opinion was found to decrease error where past model performance had been poor, but
not in other cases (P. H. Franses & Legerstee, 2013).
2.2.1 Situations Where Overrides Add Value, and Where They Don’t

A common issue with forecasting systems is that the ability to enter forecast overrides imply the need to override each forecast. Doing so often leads to increased errors. Goodwin, Fildes, Lawrence, & Stephens (2011) studied the impacts of placing limits on overrides. One group was allowed to enter overrides only during promotional periods, while another was prohibited from making small adjustments to the forecast, which would be deemed wasteful. The results showed that neither prohibiting small adjustments, nor limiting override to promotional periods improved forecast accuracy. The impact was to encourage over-forecasting when small adjustments were disallowed.

Trapero, Pedregal, Fildes, & Kourentzes (2013) showed improvement in accuracy when overrides occur during promotions, but that greater accuracy may be obtained through multi-variate analysis which includes promotional activity as an input variable instead of a qualitative adjustment. However, an ensemble approach of multi-variate modelling and expert opinion could further improve upon forecast accuracy.

Trapero, Kourentzes, & Fildes, (2015) further their work in advancements of multi-variate models as a replacement to expert opinion. Their dynamic regression models demonstrate improved accuracy, which frees up experts. However, they note no commercially available software to enable dynamic regression, leaving business forecasters the choice between multi-variate models and/or expert opinion.

2.3 Data Analysis and Metrics

In this section, common forecasting error metrics are described. Error metrics are needed to evaluate the impact of judgmental overrides on either increasing or decreasing the final forecast
accuracy. Forecast Value Added (FVA) analysis, which leverages error metrics, is then discussed as a technique to evaluate a judgmental override. Finally, seasonal-trend decomposition using Loess is described as the technique to understand residual variability in a time series – this is the error which cannot be explained by trend or seasonality.

2.3.1 Forecast Accuracy Metrics

Forecast error for time period t, \( e_t \), is the difference between the actual value (\( A_t \)) and the forecasted value (\( F_t \)).

\[
e_t = A_t - F_t
\]  

(1)

There are many scale-dependent metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MdAE). While they are useful for comparing errors in the same data set, they are not appropriate for comparison between data sets.

Percentage error \( p_t \) is calculated by dividing the error (\( e_t \)) by the actual value \( A_t \).

\[
p_t = \frac{e_t}{A_t}
\]

(2)

Scale-independent metrics based on percentage error are commonly used to compare performance between products and product groups which are dissimilar in size. Mean Absolute Percent Error (MAPE) is often recommended in textbooks (Bowerman, O’Connell, & Koehler, 2004). However, MAPE represents one of many scale-independent metrics which may be used to evaluate forecast performance. Other variants include Median Absolute Percent Error (MdAPE) and Root Mean Square Percentage Error (RMSPE), and are discussed by Hyndman & Koehler, 2006. RMSPE was also utilized by (P. H. Franses & Legerstee, 2010).
\[ RMSPE = \sqrt{\frac{\sum_{i=1}^{n} (Actual \ Demand_i - Statistical \ Forecast_i)^2}{n}} \]  

(3)

\[ MAPE = \frac{\sum_{i=1}^{n} |Actual \ Demand_i - Statistical \ Forecast_i|}{n} \]  

(4)

\[ MdAPE = Median \left[ \frac{|Actual \ Demand_i - Statistical \ Forecast_i|}{Actual \ Demand_i} \right] \]  

(5)

There are several downsides to using percentage measures however. In the cases of either intermittent demand, or highly skewed demand, the value for actual demand \( A_t \) could become zero; in those instances, \( p \) and these measures would be undefined. Values close to zero will also cause these measures to become over-inflated.

2.3.2 Forecast Value Add

The typical forecasting process includes several steps. The most common first step is the creation of a statistical forecast, which is based on historic demand observations. Once the statistical forecast is generated, it is then distributed to various functions, which create overrides, and ultimately the final consensus forecast.

Because any type of qualitative adjustment to the forecast requires effort on the part of the adjuster, it is always a business question if that effort has resulted in a positive impact to the forecast. If not, it has wasted both time and money (Gilliland, 2013).

While the FVA methodology appears straightforward, it is deceptively difficult to implement. Often no guidance is given to the functional personnel who are tasked with adjusting the forecast. In fact, it is often framed as a requirement for adjustments to be made in order to “correct” the statistical forecast, which is often viewed as “always wrong” (Singh, Raman, & Wilson, 2015).

2.3.3 Seasonal-Trend Decomposition using Loess (STL)

Seasonal-Trend Decomposition using Loess (STL) is a method for decomposing time series into seasonal, trend, and residual components. This method was developed by Cleveland et. al (1990), and has since been coded into R.

Hyndman and Athanasopoulos (2013) discuss the several advantages that STL has over traditional decomposition techniques. STL can handle all types of seasonality, and the seasonal component is allowed to change over time in case where many seasonal cycles are observed. It is robust to outliers, but while the impact on trend and seasonality is minimized, the residuals can be impacted.

2.4 Literature Gap

The literature describes the need to adjust statistical forecasts to maximize final forecast accuracy, and the fact that overrides commonly fail to do that. In addition, several papers describe a high-level relationship between the size of an override and forecast value add. However, the scaling simply ratios the override by the underlying statistical forecast. No
consideration is given to the underlying variability in the time series, or the performance of the statistical forecast.

This research advances the body of knowledge in two key areas. First, it develops a dispersion-scaled override (DSO), which serves as a signal-to-noise ratio for the override. This improves upon the previous methods of scaling by the original forecast.

Consider 2 cases listed in Table 1 below. In each case, the statistical forecast is 1,000, the override is 100, which in both case is a 10% override. However, in case 1, the standard deviation of the residuals is 250. This means that the override of 100 is only 0.4 standard deviation units away. In this case, it’s difficult to distinguish the signal from the noise.

In case 2, the standard deviation is 50, which means that the override is 2.0 standard deviations away. This very likely represents a clear signal and should have a higher probability of increasing the accuracy of the statistical forecast. It is important to note that the initial override must be based on managerial insight or some other expert opinion and must have rational supporting evidence for it.

Table 1: Example of Scaled Override vs. Percent Override

<table>
<thead>
<tr>
<th>Statistical Forecast</th>
<th>Override</th>
<th>Percent Override</th>
<th>Standard Deviation of Residuals</th>
<th>Scaled Override</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1 1,000</td>
<td>+100</td>
<td>10%</td>
<td>250</td>
<td>0.40</td>
</tr>
<tr>
<td>Case 2 1,000</td>
<td>+100</td>
<td>10%</td>
<td>50</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The second advancement is relating the dispersion-scaled override to various predictive indicators using data mining techniques. The combination of scaled overrides and data mining
techniques helps support the creation of an override evaluation framework which will be
generalizable to business forecasting.
3 Methodology

The methodology chapter describes the approach taken to understand the relationship between forecast value add and several input variables. It first describes using time series data to create the forecast value add response variable. It then describes how time series decomposition is used to create the first class of predictor variables, dispersion-scaled overrides. Standard error and autocorrelation analysis is then used to create the second class of predictor variables. From there, the use of classification techniques to create a framework for relating the response variable to the predictor variable is described. The goal of the framework is to increase the efficiency and effectiveness of the forecasting process.

To address this challenge, a framework was developed to classify proposed overrides as value added or non-values added based on two categories of predictor variables. The data mining techniques of classification tree, boosted tree, random forest, and logistic regression were chosen. These classification techniques have several advantages, including automatic variable selection, robustness to outliers, and no assumptions regarding linearity of relationships between predictor variables and the response variable (Shmueli, Bruce, Yahav, Patel, & Lichtendahl, 2018). In particular, classification trees are attractive because they are graphical in nature and easy to explain to non-technical business personnel.

This framework is appropriate for any company using time series forecasting techniques and managerial overrides. It requires time series forecasting “triples”: actual demand, statistical forecast, and final consensus forecast. These triples must be collected at the same level where the override is entered, as that is where the business rationale behind the override is developed and challenged through the consensus process. Typically, the consensus process occurs during sales
and operation planning, but this is not a requirement. If consistency is maintained, any level of the forecasting hierarchy may be used.

The first step in the methodology is to utilize the forecasting triples to create a predictor variable for each set of data points. The actual demand was used to determine the accuracy of statistical and final consensus forecasts. Then, those two accuracy values were compared to create an improvement metric known as Forecast Value Added (FVA). FVA reflects the success of the override in reducing forecast error. FVA was then compared to a user-defined FVA threshold, $FVA_{crit}$. Values less than this value are classified as non-value added, while those above the threshold are classified as value added. This is the response variable of interest.

The predictor variables fell into two categories. The first was dispersion-scaled overrides. The actual product demand time series was decomposed into trend, seasonal, and residual components. For the residual component, the dispersion measures of standard deviation, trimmed mean absolute deviation, and median absolute deviation were calculated. The judgmental override was divided by these measures of dispersion, creating a signal-to-noise relationship between the override and the underlying variability in the time series.

The second category of predictor variables were opportunity indicators. One was autocorrelation in the residuals, which would indicate if there was information not captured in the statistical forecast. The others were percentage-based measures of statistical forecast accuracy – Root Mean Square Percentage Error (RMSPE), Mean Absolute Percent Error (MAPE) and Median Absolute Percent Error (MdAPE).
Figure 1 below provides a useful tool to visualize the data flow and connectivity of the various data elements.

Figure 1: Methodology Data Flow Overview
3.1 Time Series Data and Accuracy Metrics

The starting point of the analysis is the collection of time series data on actual demand, statistical forecasts, and consensus forecast. These can be used to calculate forecast accuracy and forecast value add.

The three time series are collected from the company’s system of record, which could be anything from an enterprise resource planning (ERP) system to something as simple as Microsoft Excel. The input data file structure is provided in Appendix A: Data File Structure. Data is typically available within a few business days after month end close. A critical task is to ensure that the data is accurate, as this new data point is used in the generation of the statistical forecast. It was assumed that adequate data management techniques were in place to ensure accuracy of the collected data, as data cleansing is outside of the scope of this effort.

Knowing both the actual demand and the statistical and consensus forecasts, the accuracy of those forecasts may be determined. One commonly used accuracy metric is absolute percent error (APE). Absolute percent error determines the error which exists between a forecast and its actual value. The sign of the error is ignored, as the goal is to reduce the level of error.

\[
APE_t = \frac{|Actual_t - Forecast_t|}{Actual_t}
\]  

APE was calculated for both the Statistical and the Consensus forecasts. The difference between them can be used to calculate the Forecast Value Added (FVA) metric. FVA describes how forecast accuracy changes from one step of the forecasting process to the next (Singh, Raman, &
Wilson, 2015). In this analysis, FVA is measured as the change in absolute percent error (APE) that exists between the statistical forecast to the consensus forecast:

\[ FVA_t = (APE_{Cons,t} - APE_{Stat,t}) \]  

(7)

FVA was then compared to a user-defined FVA threshold, FVA\text{crit}. Values less than this value are classified as non-value added, while those above the threshold are classified as value added. This is the response variable of interest.

Knowing the statistical forecast and the consensus forecast also allows the override size to be calculated. The override is the adjustment to the statistical forecast, and is measured as the difference, in units, between the statistical forecast and the consensus forecast:

\[ Override_t = Consensus\ Forecast_t - Statistical\ Forecast_t \]  

(8)

Note that both positive and negative overrides occur and are tracked separately. The override may also be described as a percentage change difference based on the statistical forecast:

\[ Override\ Percent_t = \frac{Consensus\ Forecast_t - Statistical\ Forecast_t}{Statistical\ Forecast_t} \]  

(9)

The override percent metric is included to allow comparison to earlier work (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009).

The next step in the methodology is the use of time series decomposition to extract residual error, then calculate dispersion metrics on that error for further use.
3.2 Time Series Decomposition and Residual Analysis

The predictor variables in this research represent signal to noise ratios between an override and the underlying variability of the actual demand time series. It is therefore necessary to decompose the actual demand time series, extract the residuals, and determine their dispersion measures.

Given actual historical demand, a technique called Seasonal-Trend Decomposition using Loess (STL) may be employed to decompose the time series into those trend and seasonal components. An additional set of data, known as residuals, are also computed with this method. The residuals are the errors between the fitted values and the actual historical values of demand.

Dispersion measures were calculated for the residuals, and included root mean square error, trimmed mean absolute deviation, and median absolute deviation. The residuals were also tested for autocorrelation.

3.2.1 Seasonal-Trend Decomposition Using Loess

Actual demand (D) was decomposed into the time series components of residual error (e), trend (T), and Seasonality (S) using the Seasonal-Trend decomposition using Loess (STL) procedure (Cleveland, R. B., Cleveland, W. S., McRae, J.E., & Terpenning, I., 1990). STL is part of the “stats v3.4.3” package in R, and is often used to decompose historical demand data (Hyndman, R.J. & Athanasopoulos, G., 2012).

An example of the output from the STL procedure is shown in Figure 2 below:
The output of interest for this research were the residual errors, $e_t$. The residual error represents the difference between the original time series demand data ($D$) and the seasonal ($S$) and trend ($T$) components:

$$e_t = D_t - T_t - S_t$$

Once residuals were obtained, measures of their dispersion were calculated. This provided an estimate of how much variability there is in the underlying data pattern that cannot be explained through time series modelling. The values for standard deviation, mean absolute deviation, and median absolute deviation were determined. Their use as scaling factors for overrides is discussed in section 3.4 below.

Residuals were plotted over time to check that they were homoscedastic, i.e. they will not change over time. See Figure 3 below. Histograms of the residuals were plotted to visually check for a mean of zero and normal distribution. See Figure 4 below. Note that in many cases there were a few outliers.
Using these residuals, their measures of dispersion were calculated. The measures used were root mean square error, trimmed mean absolute deviation, and median absolute deviation.
3.3 Autocorrelation

Another critical check was for autocorrelation, which determines if there is a statistically significant relationship between time-lagged residuals. The presence of autocorrelation suggests there is information which has not been accounted for in the time series model and represents an opportunity for experts to provide a value-added override.

The Ljung-Box test (Ljung & Box, 1978) is implemented in the “stats v3.4.3” package in R, and was used to determine the probability values for autocorrelation. See Figure 5 below. The probability values were used directly in the analysis as predictor variables.

![ACF of residuals](image)

**Box-Ljung test**

- X-squared = 11.342,
- df = 10,
- p-value = 0.3315

Figure 5: Autocorrelation of residuals

3.4 Dispersion-Scaled Overrides

There are two classes of predictor variables in this research: Dispersion-Scaled Overrides and Opportunity Indicators. In this section, the former will be discussed, while the later will be discussed in section 3.5 below.

The override is the qualitative adjustment, in units, to the statistical forecast, as shown in (9) above. It is performed based on business insights, and the belief that the adjustment represents some
business situation that is not represented by modeling trend and seasonality. Examples include situations such as a new customer, an economic slowdown softening demand, or expansion into a new geography bringing in new demand. To compare the overrides among product groups of varying size, overrides were scaled. The scaling factors were the dispersion measures discussed in section 3.2.1 above; standard deviation, mean absolute deviation, and median absolute deviation. The scaled overrides are analogous to a signal-to-noise ratios and are discussed below.

The first approach is to scale the overrides by dividing through by the standard deviation of the residuals:

\[
\text{Dispersion Scaled Override}_t = \frac{\text{Override}_t}{\sigma_{\text{residuals}}}
\]  

(11)

While the standard deviation is a widely used measure of dispersion, it does have the drawback that it is sensitive to outlier values due to the squaring of the error terms.

A second approach is to scale the overrides by dividing through by the Mean Absolute Deviation (MAD) of the residuals. Because the mean is sensitive to outliers, the top five percent of extreme values were trimmed:

\[
\text{Dispersion Scaled Override}_t = \frac{\text{Override}_t}{\text{MAD}_{\text{residuals}}}
\]

(12)

A scaled override based on trimmed MAD would be less sensitive to outliers, since the underlying calculation is based on differences rather than the squared differences.

A third approach is to scale the overrides by dividing through by the Median Absolute Deviation (MdAD) of the residuals:
\[ Dispersion \text{ Scaled Override}_t = \frac{(\text{Override}_t)}{\text{MdAD}_{\text{residuals}}} \] (13)

A scaled override based on Median Absolute Error would be robust, and less impacted by extreme values.

The final approach to scaling is to simply divide the override by the statistical forecast; this:

\[ Forecast \text{ Scaled Override}_t = \frac{\text{Override}_t}{\text{Statistical Forecast}_t} \] (14)

This is a straightforward approach is the technique used in previous studies (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). Note that this is not a signal-to-noise ratio style of scaling.

3.5 Opportunity Indicators

Time series models only leverage past actual demand patterns in order to generate a statistical forecast. They assume that future patterns will take a similar form as they have in the past, and that they can be described using only level, trend, and seasonality.

There will also a residual error in every time series statistical forecast. If the error component is large, it is an indication that the time series assumptions have broken down, and that there are other forces at work. As error increases, so does the opportunity for improving upon that error.

For that reason, 3 common measures for the statistical forecast error (Hyndman & Koehler, 2006; Hyndman, R.J. & Athanasopoulos, G., 2012) are included in the analysis. These metrics are calculated for each product group over the entire time horizon and are an indication of how well the statistical model is performing over time. They are also a potential indicator that the model can be improved upon through expert opinion. In other words, a product group with a relatively
high error metrics represents a larger opportunity for improvement versus a more easily forecastable product group with a relatively lower error metrics.

Each of these metrics has been implemented in the functional time series analysis "ftsa v1.1" package in R.

\[
RMSPE = \sqrt{\frac{\sum_{i=1}^{n} (\text{Actual Demand}_i - \text{Statistical Forecast}_i)^2}{n}} \tag{15}
\]

\[
MAPE = \frac{\sum_{i=1}^{n} |\text{Actual Demand}_i - \text{Statistical Forecast}_i|}{n} \tag{16}
\]

\[
MdAPE = \text{Median} \left| \frac{\text{Actual Demand}_i - \text{Statistical Forecast}_i}{\text{Actual Demand}_i} \right| \tag{17}
\]

Finally, autocorrelation in the residuals is considered. Autocorrelation exists in time series error data when there is a statistically significant relationship between a value and its own previous historical values. If autocorrelation does exist, it represents an opportunity for improvement that the statistical model cannot capture. The Ljung-Box test for Autocorrelation is coded in R and was used in this analysis.
3.6 Data Review

This section presents an overview of the required data, the accuracy metrics employed, and the two classes of predictor variables used. Table 2 summarizes the data elements.

Table 2: Overview of Data Elements

<table>
<thead>
<tr>
<th>Category</th>
<th>Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Data and Accuracy Metrics</td>
<td>Actual Demand</td>
</tr>
<tr>
<td></td>
<td>Statistical Forecast</td>
</tr>
<tr>
<td></td>
<td>Consensus Forecast</td>
</tr>
<tr>
<td></td>
<td>Override Size</td>
</tr>
<tr>
<td></td>
<td>Absolute Percent Error, statistical and consensus forecasts</td>
</tr>
<tr>
<td>Response Variable</td>
<td>Forecast Value Add</td>
</tr>
<tr>
<td>Time Series Decomposition and Residual Analysis</td>
<td>Residuals</td>
</tr>
<tr>
<td></td>
<td>Root Mean Square Error of Residuals (RMSE_r)</td>
</tr>
<tr>
<td></td>
<td>Trimmed Mean Absolute Error of Residuals (trMAE)</td>
</tr>
<tr>
<td></td>
<td>Median Absolute Error of Residuals (MdAE_r)</td>
</tr>
<tr>
<td>Predictor Variables – Dispersion Scaled Overrides</td>
<td>Scaled Override, Root Mean Square Error</td>
</tr>
<tr>
<td></td>
<td>Scaled Override, trimmed Mean Absolute Deviation</td>
</tr>
<tr>
<td></td>
<td>Scaled Override, Median Absolute Deviation</td>
</tr>
<tr>
<td></td>
<td>Forecast Scaled Override</td>
</tr>
<tr>
<td>Predictor Variables – Opportunity Indicators</td>
<td>Root Mean Square % Error (RMSPE) of Statistical Forecast</td>
</tr>
<tr>
<td></td>
<td>Mean Absolute % Error (MAPE) of Statistical forecast</td>
</tr>
<tr>
<td></td>
<td>Median Absolute % Error (MdAPE) of Statistical forecast</td>
</tr>
<tr>
<td></td>
<td>Ljung-Box test for Autocorrelation</td>
</tr>
</tbody>
</table>
3.7 Data Mining Analytics

Several classification techniques were used to determine the relationship between the forecast value add response variable and the scaled override and opportunity indicator predictor variables. The R programming language was used to perform the analysis; code may be found in Appendix B.

The first step of the analysis was to pre-process the data. The data was randomly partitioned into a training set (60% of data) and testing set (40% of data) then both data sets were scaled based on the training data set. Data partitioning is necessary to avoid over-fitting, and to assess realistically how the model will perform on new data in the future. The 60 training /40 testing partition ratio is found in data mining textbooks (Shmueli et al., 2018).

The second step was to calculate a categorical response variable with only two outcomes. In this study, a threshold value $FV_A^{\text{crit}}$ of 5% was assumed. Any values $\leq 5\%$ were assigned a 0, indicating non value added (NVA). Values $> 5\%$ were assigned a 1, indicating that there was value added (FVA). In practice, $FV_A^{\text{crit}}$ would be an input parameter based on a company’s definition of value add.

Positive and negative overrides were analyzed separately, as it is hypothesized that larger scaled overrides add value, regardless of sign. Once pre-processing was completed, the data mining techniques shown in
Table 3 below were applied to the training data partition.

Table 3: Data Mining – Classification Techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Type</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td></td>
<td>Predict whether an override will create a Forecast Value Added result greater than the threshold value of 5%</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td>Boosted Tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The models created were then evaluated using the data from the testing partition, and a confusion matrix was generated for the predicted versus actual model classifications. See Table 4 below for the generalized confusion matrix format.

Table 4: Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>FVA (Predicted)</th>
<th>NVA (Predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVA (Actual)</td>
<td>t_p, True Positive</td>
<td>f_n, False Negative</td>
</tr>
<tr>
<td>NVA (Actual)</td>
<td>f_p, False Positive</td>
<td>t_n, True Negative</td>
</tr>
</tbody>
</table>

Given data on correctly classified true positives (t_p) and true negatives (t_n), and misclassified false positives (f_p) and false negatives (f_n), various measures of accuracy could be determined.
Accuracy is defined as the overall ability to correctly classify true positives and true negatives, and is calculated as per Equation (18):

\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (18)
\]

Sensitivity, also known as recall, measure how well the classification method can identify true positives. The calculation is shown in Equation (19):

\[
\text{Sensitivity} = \frac{tp}{tp + fn} \quad (19)
\]

Specificity is the ability to correctly classify false records. The calculation is shown in Equation (20):

\[
\text{Specificity} = \frac{tn}{tn + fp} \quad (20)
\]

Precision describes how often the classification is correct when it predicted to be correct. The calculation is shown in Equation (21):

\[
\text{Precision} = \frac{tp}{tp + fp} \quad (21)
\]

3.8 Methodology Summary

Chapter 3 began by describing the three input time series which serve as the foundation of the analysis—actual demand (D), statistical forecast (SF) and consensus forecast (CF). A fourth time series, override (OVR) was calculated as CF – SF. Comparing the statistical forecast and consensus forecast to the actual demand allowed their absolute percent errors (APE_{SF}, APE_{CF}) to
be calculated. Calculating the difference between $\text{APE}_{\text{CF}}$ and $\text{APE}_{\text{SF}}$ determined the forecast value add (FVA) of the override. If this value exceeded a threshold value $FVA_{\text{crit}}$, the override was categorized as value added. Otherwise, it was non-value added. This value add / non-value add categorization is the response variable of interest.

To develop a classification framework for the response variable FVA, three categories of predictor variables were created. The first was Dispersion-Scaled Overrides (DSO). The actual demand was decomposed into seasonal (S), trend (T), and residual error (e) components. Dispersion measures of the residual error component – root mean square error (RMSE), mean absolute deviation (MAD), and median absolute deviation (MdAD) – were calculated. Overrides were the divided by these dispersion measures to create DSOs.

The second category of predictor variables was statistical forecast accuracy. The Root Mean Square Percentage Error (RMSPE), Mean Absolute Percentage Error (MAPE) and Median Absolute Percentage Error (MdAPE) of the statistical forecast were calculated.

The final category was autocorrelation of the residuals, which was determined using the Ljung-Box test.

Classification Tree, Boosted Tree, Random Forest, and Logistic Regression were then used to classify overrides as value added or non-value added based on the predictor variables.

Chapter 4 will describe two key areas. The first is an analysis of the input times series, forecast accuracy, and the forecast value add of overrides. The second area is the outcome of the various classification techniques used to relate forecast value add to dispersion-scaled overrides, statistical forecast accuracy, and auto-correlation.
4 Results and Discussion

In this section the results of the analysis are presented. First, information on the input data set is presented, showing the sizes of the adjustments and the accuracy of the statistical and consensus forecasts. Forecast value add, being the response variable of interest, is also presented. The outputs are then discussed. Results from the various data mining approaches are reviewed, beginning with classification techniques before moving on to regression methods.

This chapter concludes by placing the research results from the previous chapter in context with the research problem, and with the broader literature. The results are synthesized, and the implications of those results are discussed. Finally, limitations of this study and recommendations for future work are provided.

4.1 Research Problem

Business forecasting frequently combines quantitative time series techniques with qualitative expert opinion overrides to create a final consensus forecast. The objective of these overrides is to reduce forecast error, which enables safety stock reduction, customer service improvement, and manufacturing schedule stability. However, these adjustments often fail to improve forecast accuracy. Process mis-steps include small adjustments, adjustments to accurate statistical forecasts, and adjustments to match financial goals. At best, these adjustments waste scare forecasting resources; at worst, they seriously impact business performance. The goal of this research was to create a framework for identifying overrides that are likely to improve forecast accuracy.
4.2 Input Data Results

The input to the analysis was monthly time series data at the product group level. Matched sets of forecasting triples – the statistical forecast, the final consensus forecast, and the actual demand – were collected. The difference, in units, between the monthly values of the statistical and final consensus forecasts allowed the override to be calculated, as per Equation (8) above. The override percent was then calculated by dividing through by the statistical forecast value, as per Equation (9) above. The override percent standardizes the override, allowing comparisons to be made between product lines which differed on an absolute scale. Figure 6 below shows the distribution of overrides from -100% to 200%. The data are right skewed, and 33 of the 703 overrides (4.7%) were greater than 200%.

![Override Percent Histogram](image)

Figure 6: Override Percent Histogram
Of the 703 triples collected, only 3 showed no override activity. The 700 overrides were almost equally distributed between positive (n=345) and negative (n=355) adjustments to the statistical forecast (see Table 5 below). Negative overrides were smaller and appear unskewed based on similarities of the median (-30%) and the mean (-31%). However, the positive overrides were larger, and show signs of several very large adjustments which skew the mean (85%) well above the median (43%).

Table 5: Mean and Median of Override Sizes

<table>
<thead>
<tr>
<th>Direction of Adjustment</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>345</td>
<td>85%</td>
<td>43%</td>
</tr>
<tr>
<td>Negative</td>
<td>355</td>
<td>-31%</td>
<td>-30%</td>
</tr>
</tbody>
</table>

The skewedness of the positive overrides suggests that an external factor is driving large adjustments and introducing bias into the consensus forecasting process. A likely candidate is the business pressure to have the forecast match the financial objectives of the organization.

Once the nature of the overrides has been understood, their effect upon the consensus forecast was investigated. To do so, accuracy metrics for both the statistical and the consensus forecasts, as compared to the actual demand, were calculated as per equations (16) and (17) above, and are summarized in Table 6 below. Mean Absolute Percent Error (MAPE) is a commonly used accuracy metric; Median Absolute Percent Error (MdAPE) is less frequently used but is more robust to extreme values.

For positive overrides (n=345), both MAPE and MdAPE increase between the statistical forecast and the final consensus forecast, 54.1% to 200.5%, and 14.5% to 55.5%, respectively. In other words, manual adjustment is causing an increase in the error. This is the opposite of what is
expected and is an indicator that the skewedness observed in the positive override adjustments has had a damaging effect.

For negative overrides \((n=355)\), both MAPE and MdAPE decrease between the statistical forecast and the final consensus forecast, 180.4% to 68.6%, and 36.3% to 24.4%, respectively. This result is as expected; manual intervention by experts should incorporate information unknown to the statistical model, and error should decrease.

Table 6: Accuracy Metrics for Statistical and Consensus Forecasts

<table>
<thead>
<tr>
<th>Direction of Adjustment</th>
<th>N</th>
<th>MAPE Statistical Forecast</th>
<th>Consensus Forecast</th>
<th>MdAPE Statistical Forecast</th>
<th>Consensus Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>345</td>
<td>54.1</td>
<td>200.5</td>
<td>14.45</td>
<td>55.5</td>
</tr>
<tr>
<td>Negative</td>
<td>355</td>
<td>180.4</td>
<td>68.6</td>
<td>36.3</td>
<td>24.4</td>
</tr>
</tbody>
</table>

The response variable forecast value add (FVA) was calculated as the pair-wise difference between the absolute present error (APE) of the final consensus forecast minus the statistical forecast, as per equations (6) and (7) above. The FVA median values for positive and negative adjustments were grouped separately, by quartile. The median values based on quartiles is shown in Figure 7: Median FVA, by Quartile, below.
Figure 7: Median FVA, by Quartile

The smallest 25% of both positive and negative overrides (n= 176) had negligible impact on FVA. These overrides would fall into the category of time wasters, as the effort put into creating them had no impact on consensus forecast accuracy but did require time.

Negative overrides did show value add, improving the accuracy of the consensus forecast. The second, third, and fourth quartiles showed median accuracy improvements of 18%, 23%, and 78%, respectively. The 78% improvement observed for the largest negative overrides suggest external events which were not predicted by the statistical model but were discerned by the experts entering the override. These results are what would be expected from an expert-opinion based override. The expert has additional information which the statistical forecast does not, and therefore is able to add value. The greater the information gap, the greater the potential for value add. For instance, if a customer goes out of business, sales personnel can immediately remove their demand from the
forecast. In a statistical model, it would take time for that new information to be reflected in the forecast.

Positive overrides resulted in median negative value adds of -23%, -61% and -191% for the second, third, and forth quartiles, respectively. This result is not expected and indicates that something besides additional business information is influencing overrides. The most likely cause of this behavior would be overriding the forecast to equal financial plans. While unexpected, similar results are described in the literature (Baecke, De Baets, & Vanderheyden, 2017; Fildes et al., 2009). Another view of the FVA data is to do so by quadrant, as shown in Figure 8 below. Median values were again used, in order to minimize the impact of extreme values.

Figure 8: Override Size/Direction vs. FVA
A similar trend appears here, as larger negative overrides tended to add the most value. The upper left quadrant shows that 217 overrides had an average adjustment of -34%, and increased accuracy by 49%. In the opposite, lower right quadrant, we see that larger positive overrides tended to reduce value considerably. In the quadrant, 292 overrides had an average adjustment of 49%, but reduced accuracy by 52%.

Based on the input data, there is considerable opportunity to create a framework to guide adjustments. This guidance would reduce the number of time-wasting small overrides, and flag other large overrides as potentially value destroying.

The next section reviews the data used to identify areas of opportunity to improve forecast value add. If there is opportunity, adjustments could be warranted. If there is little opportunity, the results of the statistical forecast should stand unadjusted.

4.3 Opportunity Indicators

A class of variables was developed to predict if there was an opportunity to increase forecast value add through adjustment. These predictor variables are based on autocorrelation and statistical forecast accuracy. The values, by product group, are summarized in Table 7 below.

If autocorrelation is present, it would not be modeled explicitly with time series techniques. This gap would therefore represent an area where an expert could add value. Autocorrelation probability values were calculated by performing a Ljung-Box test on the STL residual values. The values are shown in Table 7 below in the L-J PVAL column. Statistically significant autocorrelation (L-J PVAL values < 5%) only appeared in one group, BIL.
The next section reviews the classification techniques which were used to understand the relationship between the scaled overrides (discussed in Methodology section), the opportunity indicators discussed above, and the response variable – forecast value add.

4.4 Data Mining Analytics

The goal of the data mining analysis was to determine which combination of input variables best predicted the classification of the response variable, forecast value add (FVA). Two classes of predictor variables were used, as seen in
The other indicators are derived from the error in the statistical forecast. If large errors are present, the statistical model is not accounting for some external factors, which would represent another area where an expert could add value. Examples of these un-modeled factors could include promotions, specific customer demand changes, or overall demand changes based on economic conditions.

Product group Median Absolute Percent Error (MdAPE) values ranged from 9% to 110%. Mean Absolute Percent Error (MAPE) values ranged from 13% to 659%, and Root Mean Square Percentage Error (RMSPE) values ranged from 19% to 1,916%. While the definition of a “good” statistical forecast relies on numerous factors, the relative sizes of these errors are likely to be indicative of where improvement opportunities lie. Extreme values are also indicative of outliers which should be investigated.

Table 7: Opportunity Indicators

<table>
<thead>
<tr>
<th>Group</th>
<th>L-J PVAL</th>
<th>MdAPE</th>
<th>MAPE</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>999</td>
<td>33%</td>
<td>9%</td>
<td>13%</td>
<td>19%</td>
</tr>
<tr>
<td>FUT</td>
<td>46%</td>
<td>11%</td>
<td>16%</td>
<td>25%</td>
</tr>
<tr>
<td>HO</td>
<td>100%</td>
<td>11%</td>
<td>18%</td>
<td>27%</td>
</tr>
<tr>
<td>VOL</td>
<td>99%</td>
<td>14%</td>
<td>23%</td>
<td>34%</td>
</tr>
<tr>
<td>ES</td>
<td>93%</td>
<td>17%</td>
<td>24%</td>
<td>36%</td>
</tr>
<tr>
<td>TTI</td>
<td>33%</td>
<td>26%</td>
<td>40%</td>
<td>55%</td>
</tr>
<tr>
<td>GO</td>
<td>80%</td>
<td>16%</td>
<td>37%</td>
<td>74%</td>
</tr>
<tr>
<td>BIL</td>
<td>1%</td>
<td>23%</td>
<td>44%</td>
<td>101%</td>
</tr>
<tr>
<td>NEW</td>
<td>66%</td>
<td>17%</td>
<td>51%</td>
<td>105%</td>
</tr>
<tr>
<td>MOS</td>
<td>61%</td>
<td>27%</td>
<td>66%</td>
<td>116%</td>
</tr>
<tr>
<td>CUE</td>
<td>23%</td>
<td>14%</td>
<td>65%</td>
<td>160%</td>
</tr>
<tr>
<td>AG</td>
<td>19%</td>
<td>22%</td>
<td>80%</td>
<td>178%</td>
</tr>
<tr>
<td>OUT</td>
<td>50%</td>
<td>46%</td>
<td>112%</td>
<td>192%</td>
</tr>
<tr>
<td>IN</td>
<td>20%</td>
<td>29%</td>
<td>105%</td>
<td>198%</td>
</tr>
<tr>
<td>PIC</td>
<td>56%</td>
<td>68%</td>
<td>158%</td>
<td>261%</td>
</tr>
<tr>
<td>DAR</td>
<td>23%</td>
<td>37%</td>
<td>203%</td>
<td>516%</td>
</tr>
<tr>
<td>STI</td>
<td>51%</td>
<td>43%</td>
<td>291%</td>
<td>733%</td>
</tr>
<tr>
<td>KIL</td>
<td>98%</td>
<td>15%</td>
<td>232%</td>
<td>797%</td>
</tr>
<tr>
<td>RU</td>
<td>57%</td>
<td>110%</td>
<td>659%</td>
<td>1916%</td>
</tr>
</tbody>
</table>
Table 8 below. The first class is dispersion-scaled overrides (DSO). The three versions of dispersion-scaled overrides are signal-to-noise ratio metrics, where the override is divided by the measures of dispersion listed – root mean square error, trimmed mean absolute error, and median absolute error.

The second class, termed opportunity indicators, describe the accuracy of the statistical method. If the statistical accuracy is poor, it may be an indication of opportunity for improvement through override adjustments. Conversely, low statistical forecast error may be an indication that there is little room for improvement from an override.
### Table 8: Response and Predictor Variables

<table>
<thead>
<tr>
<th>Category</th>
<th>Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variable</td>
<td>Forecast Value Add. Set to “1” if FVA exceeds 5%, otherwise set to “0”</td>
</tr>
<tr>
<td>Predictor Variables:</td>
<td>Displacement-Scaled Override, Root Mean Square Error</td>
</tr>
<tr>
<td>Dispersion-Scaled Overrides</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Displacement-Scaled Override, Mean Absolute Error</td>
</tr>
<tr>
<td></td>
<td>Displacement-Scaled Override, Median Absolute Error</td>
</tr>
<tr>
<td></td>
<td>Forecast Scaled Override</td>
</tr>
<tr>
<td>Predictor Variables:</td>
<td>Root Mean Square % Error (RMSPE) of Statistical Opportunity Indicators</td>
</tr>
<tr>
<td>Opportunity Indicators</td>
<td>Mean Absolute % Error (MAPE) of Statistical forecast</td>
</tr>
<tr>
<td></td>
<td>Median Absolute % Error (MdAPE) of Statistical</td>
</tr>
<tr>
<td></td>
<td>Ljung-Box test for Autocorrelation</td>
</tr>
</tbody>
</table>

Two types of techniques were used to establish the response variable – predictor variable relationship. The first were classification techniques, which sought to predict a binary prediction if an override was either value added or non-value added. These techniques included classification tree, boosted tree, random forest, and logistic regression. The R programming language was used to perform the analysis; code may be found in Appendix B: R Code.

The first step of the analysis was to pre-process the data. The data was randomly partitioned into a training set (60% of data) and testing set (40% of data). Each model was then developed using the training partition. Once developed, the model’s predictive capability was tested using the separate validation partition. This allowed the predictive capability of the models to be determined and avoided over-fitting.

The next step was to create a categorical response variable based on comparing FVA values to a threshold value, $FVA_{crit}$. A user-defined threshold FVA value of 5% was assumed, but in practice, this would be an input parameter based on a company’s definition of value add. In this work, any
FVA values < 5% were assigned a 0, indicating non-value added, and those > 5% were assigned a 1, indicating value added.

4.4.1 Classification Tree

Classification trees benefit from their highly visible nature, which makes them easy to interpret and communicate to non-technical personnel. A classification tree was developed for the variables described in Section 4.3 above, using the 421 observations in the training data partition. The tree was pruned by adjusting the complexity parameter to correspond to the minimal error, resulting in 6 nodes. The results are seen in Figure 9: Classification Tree below.

![Classification Tree](image)

Figure 9: Classification Tree

Node 1 splits based on the override percent variable, using a value greater than or equal to -5.6%. If the adjustment size is greater than this value, the “Yes” response logic goes to Node 2 to the left;
otherwise the “No” response goes to Node 3. This indicates that the most important indicator in predicting forecast value add is the direction of the adjustment.

The next most important indicators were the dispersion-scaled overrides, either the mean-scaled (Node 2) or standard deviation-scaled (Node 3). The Node 2 (overrides $\geq -5.6\%$) branch seldom indicates forecast value add, and a small degree of overfitting may be occurring. Node 3 then splits based on a predictor variable, root mean square percentage error. This is the third most important factor.

The key discovery is that override direction, dispersion-scaled overrides and the predictor variable all play a key role in determining forecast value add. The other discovery is that autocorrelation does not appear to be a significant factor.

The 281 observations in the validation data partition were then used to create a confusion matrix for the classification tree method, and various key metrics were calculated as shown in Table 9 below. In this case, the accuracy is 80.07%. The ability to correctly predict cases where overrides added value (sensitivity) is 85.71%, while the ability to correctly predict cases where overrides do not add value (specificity) is 69.70%. Precision is 83.87%.

Table 9: Confusion Matrix for Classification Tree Method

<table>
<thead>
<tr>
<th></th>
<th>FVA (Predicted)</th>
<th>NVA (Predicted)</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVA (Actual)</td>
<td>156</td>
<td>30</td>
<td>83.87%</td>
</tr>
<tr>
<td>NVA (Actual)</td>
<td>26</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>85.71%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>69.70%</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>80.07%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

53
4.4.2 Random Forest

The random forest is a variant of a classification tree which leverages bootstrapping and aggregation to create multiple trees, and then combines the classifications using ensemble techniques. While this typically increases the accuracy of the model, there is a loss of interpretability. The random forest is a “black box” technique, and there is no single tree to display.

Random forests do provide a critical output known as the variable importance plot, which measures the impact of predictor variables. See Figure 10 below. We see that the Scaled Override based on standard deviation is the highest importance, consistent with the results from the classification tree. The other scaling factors, as well as the opportunity indicators, play a reduced role. The second most important variable is the size of the adjustment on a percentage basis. Autocorrelation appears to have very little impact. However, as seen in Table 7 above, autocorrelation is only present in one product group. In a generalized framework, it would be recommended to keep this predictor in the model.
The 281 observations in the validation data partition were then used to create a confusion matrix for the random forest method, and various key metrics were calculated, see Table 10 below. Compared to the classification method above, the random forest method increased accuracy to 81.85%, and increase sensitivity to 86.85%. The specificity improved to 71.4%, and precision increased to 86.39%. This is expected, as random forest combines the results from multiple trees into an ensemble model.

**Table 10: Confusion Matrix for Random Forest Method**

<table>
<thead>
<tr>
<th></th>
<th>FVA (Predicted)</th>
<th>NVA (Predicted)</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVA (Actual)</td>
<td>165</td>
<td>26</td>
<td>86.39%</td>
</tr>
<tr>
<td>NVA (Actual)</td>
<td>25</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td>86.85%</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td></td>
<td>71.4%</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>81.85%</td>
</tr>
</tbody>
</table>
4.4.3 Boosted Tree

The boosted tree is another variant of a classification tree which attempts to minimize misclassification. It is an iterative procedure which over-samples misclassified records. As with a random forest, this typically increases the accuracy of the model, but results in a loss of interpretability because there is no single tree to display.

The confusion matrix and key metrics for the boosted tree method were calculated, see Table 11 below. The accuracy of the boosted tree method was 80.78%, and the sensitivity increased up to 90.11%. This is expected, as the strength of boosted tree is to reduce the number of misclassified records. However, the gain in sensitivity came at the expense of specificity, which dropped to 63.64%, the lowest rate in the research. Precision was 86.39%.

Table 11: Confusion Matrix for Boosted Tree Method

<table>
<thead>
<tr>
<th></th>
<th>FVA (Predicted)</th>
<th>NVA (Predicted)</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVA (Actual)</td>
<td>164</td>
<td>36</td>
<td>86.39%</td>
</tr>
<tr>
<td>NVA (Actual)</td>
<td>18</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td>90.11%</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td></td>
<td>63.64%</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>80.78%</td>
</tr>
</tbody>
</table>
4.4.4 Logistic Regression

A key outcome of logistic regression is the determination of probability values for the variables under investigation. As seen in Table 12 below, the most statistically significant factor is the scaled override based on standard deviation of residuals. Three other factors are also statistically significant at over 95% confidence: The adjustment size, which is a percentage of the statistical forecast, the mean absolute percentage error (an opportunity indicator), and the root mean square of the percentage error (RMSPE). RMSPE was also seen to be a significant factor in the classification tree.

Table 12: Variable Statistical Significance

| Variable           | Pr(>|z|)     | Significance |
|--------------------|-------------|--------------|
| rmspe              | 0.0282      | 95% +        |
| mape               | 0.0476      | 95% +        |
| mdape              | 0.9423      |              |
| adj_size           | 0.0121      | 95% +        |
| scaled_ovr_stdev   | 0.00000000188 | 99.99% +      |
| scaled_ovr_mean    | 0.9196      |              |
| scaled_ovr_med     | 0.1384      |              |
| lj_pval            | 0.6975      |              |

Using the logistic regression approach, accuracy was 81.49%. Its true positive rate (sensitivity) is 97.80%, while its false positive rate is 45.12%.

The confusion matrix and key metrics for logistic regression were calculated, see Table 13 below. The accuracy of logistic regression was 81.49%, and the sensitivity increased up to 97.80%. As in the case for boosted tree, gains in sensitivity came at the expense of specificity, which dropped to 51.52%. Precision was 87.76.
Table 13: Confusion Matrix for Logistic Regression

<table>
<thead>
<tr>
<th></th>
<th>FVA (Predicted)</th>
<th>NVA (Predicted)</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVA (Actual)</td>
<td>178</td>
<td>48</td>
<td>78.76%</td>
</tr>
<tr>
<td>NVA (Actual)</td>
<td>4</td>
<td>51</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity  97.80%
Specificity  51.52%
Accuracy      81.49%

4.4.5 Summary of Classification Techniques

Four classification techniques were utilized: classification tree, random forest, boosted tree, and logistic regression. In Table 14 below, the summarized accuracy, sensitivity, specificity, and precision are presented:

Table 14: Summary of Selected Metrics for Classification Techniques.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Tree</td>
<td>80.1 %</td>
<td>85.7 %</td>
<td>69.7 %</td>
<td>83.9 %</td>
</tr>
<tr>
<td>Random Forest</td>
<td>81.9 %</td>
<td>86.8 %</td>
<td>71.4 %</td>
<td>86.4 %</td>
</tr>
<tr>
<td>Boosted Tree</td>
<td>80.8 %</td>
<td>90.1 %</td>
<td>63.6 %</td>
<td>82.0 %</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>81.5 %</td>
<td>97.8 %</td>
<td>45.1 %</td>
<td>78.8 %</td>
</tr>
</tbody>
</table>

4.4.6 Investigation of Standard Deviation Scale Override

Because the impact of the Standard Deviation Scale Override is pronounced, it was investigated directly, and the results as shown in Figure 11 below. Visual inspection suggests 3 overlapping regions. The first is the area with large negative override which increase FVA. Another is the area with large negative overrides, where there is a notable decrease in FVA. The third region occurs in the middle, where small negative and positive adjustments occur. These tend to have an overall net-zero impact.
Figure 11: Standard Deviation Scaled Override vs. Forecast Value Add (%)
4.5 Summary of Outcomes

The classification framework created was approximately 80% accurate in predicting whether an override would or would not create forecast value add above a user-defined threshold. Three key factors drive the probability of creating a value-added override. The first factor is the accuracy of the baseline statistical model. If the statistical forecast was performing poorly, there are probably opportunities to improve upon it. Conversely, a well-performing model was difficult to improve upon. The second factor was the size of the override compared to the underlying variation. Overrides indistinguishable from the random noise are unlikely to add value. The final factor was the direction of the override. Downward adjustments of the statistical forecast were more likely to add value, likely due to increased scrutiny. Upward adjustments, often driven by the desire to match financial goals, were more likely to degrade forecast accuracy. This directional bias is consistent with previous studies (Fildes et al., 2009).

4.6 Implications of Findings

Three key learnings arise from analyzing the input data and should be considered when entering an override. First, small overrides often create no forecast value add, and should be eliminated if possible. The second insight is that the size of the override has an influence on the potential forecast value add, with bigger overrides causing bigger changes in forecast value add. The final learning is a caveat to the second; large positive overrides have the potential to make the forecast worse due to optimism bias.

The results suggest that a classification framework which relates dispersion-scaled overrides, baseline forecast accuracy, and auto-correlation to forecast value add will enhance the efficiency
and effectiveness of the business forecasting process. This framework should be used to evaluate all potential overrides before they are applied to the baseline statistical model.

The efficiency of the consensus forecasting process will increase as small adjustments are flagged as non-value added, and no time is spent entering them. The net effect on overall forecast accuracy would be minimal, as any potential small gains from one override would be offset by small losses from another. However, this will free up time, allowing greater focus on areas where value is more likely to be added, which increases the effectiveness of the consensus forecasting process.

Effectiveness will increase, as entering an override which is flagged as non-value add will prompt further discussion, rationalization, and documentation. The discussion and rationalization of the underlying assumptions for an override is critical for two reasons. First, it focuses on what business factors are driving the gap between the baseline statistical model demand and the demand predicted by the experts. Second, it drives clarity on where the gap is coming from. For instance, if some of the increase is due to marketing, some is due to sales, and some is due to an economic upturn, clarity is required to ensure there is no double-counting of effects.

Furthermore, using this framework during the consensus process will drive business behavior that will be self-correcting over time. Success in creating FVA overrides, and conversely not creating NVA overrides, will make it more likely that future overrides will be correctly classified flagged as FVA. Conversely, poor judgement in entering overrides will make it difficult to classify future overrides as FVA, which would drive additional discussion during the consensus forecasting process.
A conceptual overview of the framework is presented in Figure 12 below. For cases where the statistical forecast error is low, or where the overrides are small, the override will be flagged as non value added. If an override is entered, it will need to be accompanied by sound business rationale. In cases where the override is sufficiently large and the statistical forecast suffers accuracy issues, override would be flagged as value added. Note that there still must be sound business rationale behind each and every override.

Figure 12: Framework for Maximizing Value Add

This chapter discussed the results of the research. It opened by re-stating the research problem, then reviewed the input data. The response variable and opportunity indicators were introduced as inputs to various classification techniques. Chapter 4 concluded with a graphical review of the framework concepts. In chapter 5, concluding comments are offered, including limitations and future research opportunities.
Conclusion

In this study, a classification framework for identifying value added overrides which improve forecast accuracy was created. This framework should be used during the consensus forecasting process to evaluate any override to the baseline statistical model. A new class of metrics, Dispersion-Scaled Overrides, was developed as an input to the framework. Other inputs included statistical forecast accuracy and auto-correlation. The classification framework was approximately 80% accurate in predicting success for the business studied in this research. Despite the presence of biased overrides, the framework demonstrated the capability to adjust.

The results suggest that using dispersion-scaled overrides alongside forecast accuracy metrics in a classification framework can reliably predict forecast value add. There are three areas of the business forecasting process which will be impacted. First, numerous small adjustments which typically do not add value will be avoided. Second, adjustments to already accurate statistical forecasts will be contraindicated. Finally, the forecast adjustment behavior will become self-correcting. Previous bad decisions will increase the likelihood of an override being flagged as non value add, which will increase the need to document and discuss underlying assumptions. Documentation and discussion will in turn increase forecast value-add.

There were limitations to this research. Additional data sets from companies in other industries would be beneficial to test the applicability under many business environments. While the time series models used as the baseline forecast are commonly used in business today, there are several other classes of forecasting models which were not investigated.
Further research in this area is warranted. The clearest opportunity is the extension to other forecasting techniques, such as multiple linear regression. This would require scaling the overrides with dispersion measures based on forecast fit metrics. Another possibility is to study promotion-driven businesses and incorporate promotional activity as a predictor variable. Additional data mining approaches need to be considered, including neural net models. The FVA response variable should be extended to a third category which represents forecast value destruction – those overrides which are likely to significantly decrease forecast accuracy.

This thesis created a framework to guide businesses in their efforts to maximize forecast accuracy while minimizing effort. This lays the groundwork for an efficient and effective business forecasting process. The impact will be seen in optimum inventory levels, increased manufacturing schedule stability, and improved customer service. Ultimately, business profitability will improve.
References

https://doi.org/10.1016/j.ijpe.2017.05.016


https://doi.org/10.1016/j.ijforecast.2009.02.005

https://doi.org/10.1016/j.ijforecast.2008.11.010

https://doi.org/10.1016/j.ijforecast.2008.11.009


## Appendix A: Data File Structure

Table 15: Data file structure for Input_Data.csv

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Group</td>
<td>Prod_Grp</td>
<td>chr</td>
<td>Product group or family. An aggregate level used in S&amp;OP meetings.</td>
</tr>
<tr>
<td>Year</td>
<td>Year</td>
<td>num</td>
<td>Year</td>
</tr>
<tr>
<td>Period</td>
<td>Period</td>
<td>num</td>
<td>Monthly / Weekly buckets. Ideally will see 3 complete cycles; 2 cycles + 1 period is minimum.</td>
</tr>
<tr>
<td>Demand</td>
<td>Act_Dem</td>
<td>num</td>
<td>Actual demand for the period</td>
</tr>
<tr>
<td>Stat Fcst</td>
<td>Stat_Fcst</td>
<td>num</td>
<td>Statistical time series forecast, based on actual demand history, at the same level of aggregation as override</td>
</tr>
<tr>
<td>Consensus Fcst</td>
<td>Cons_Fcst</td>
<td>num</td>
<td>Final forecast, based on judgmental overrides. Same level of aggregation as input - should not be disaggregated from a higher-level override or aggregated from lower-level overrides.</td>
</tr>
</tbody>
</table>
Appendix B: R Code

# setwd("/Users/Jeff/Documents/~MITx/~Thesis Research/Data")
setwd("/Users/Jeff/Documents/~MITx/~Thesis Research/Data/Test")

# Load required libraries
library(forecast)
library(readxl)
library(xlsx)
library(dbplyr)
library(TSrepr)
library(ftsa)

# Data File Structure
# Product Group         Prod_Grp
# Year                 Year
# Period               Period
# Demand               Act_Dem
# Stat Fcst            Stat_Fcst
# Consensus Fcst       Cons_Fcst

# Read the data from Excel into R (NEED TO PARAMETERIZE!!!)
fct.data <- read_excel("Input_Data_Master.xlsx", sheet = "Esca")

# Set FVA Limit (NEED TO PARAMETERIZE!!!)
# Test at 5%
fva.lim <- 0.05

# Create list of unique products
prods <- unique(fct.data[c("Prod_Grp")])
n <- nrow(prods)

# Loop through inputs and run calculations
for (i in 1:n){
dat <- subset(fct.data,fct.data$Prod_Grp==toString(prods[i,]))

#### Calculate Accuracy Metrics
dat$adj = dat$Cons_Fcst - dat$Stat_Fcst
dat$statpe = abs(dat$Act_Dem-dat$Stat_Fcst) / dat$Act_Dem
dat$conspe = abs(dat$Act_Dem-dat$Cons_Fcst) / dat$Act_Dem

#### FVA is the Response Variable
dat$fva = dat$statpe - dat$conspe

#### Predictor Variables - "Opportunity Indicator" Class
# based on Statistical Forecast Error
# Root Mean Square Percentage Error of the Statistical Forecast
dat$rmspe$ <- error(forecast=dat$Stat_Fcst, true=dat$Act_Dem, method="rmspe")

# MAPE and Weighted MAPE of the Statistical Forecast
dat$ae_adj$ = abs(dat$Act_Dem-dat$Cons_Fcst)
dat$wmape$ = 100*sum(dat$ae_adj)/sum(dat$Act_Dem)
dat$mape$ <- error(forecast=dat$Stat_Fcst, true=dat$Act_Dem, method="mape")

# Median APE of the Statistical Forecast
dat$mdape$ <- error(forecast=dat$Stat_Fcst, true=dat$Act_Dem, method="mdape")

tsData2$ <- ts(data=dat[, 4],
    start = c(2015,1),
    end = c(2017,11),
    frequency = 12)
stlplot$ <- stl(tsData2, t.window=12, s.window="periodic", robust=TRUE)
st.df$ <- as.data.frame(stlplot$time.series)

# Might need to append 2 rows, or ignore missing 2 values in STL output
while (nrow(st.df) < nrow(dat)) {
    st.df[nrow(st.df)+1,] <- NA
}
dat$seasonal$ = st.df$seasonal
dat$trend$ = st.df$trend
dat$remainder$ = st.df$remainder

# Metrics on remainder
dat$std_dev_r$ = sd(dat$remainder, na.rm = TRUE)
dat$mean_r$ = mean(dat$remainder, na.rm = TRUE, trim=0.05)
dat$med_r$ = median(dat$remainder, na.rm = TRUE)

# Scaled Overrides - another class of predictor variables
dat$adj_size$ = dat$adj / dat$Stat_Fcst
dat$scaled_ovr_stdev$ = dat$adj / dat$std_dev_r$
dat$scaled_ovr_mean$ = dat$adj / dat$mean_r$
dat$scaled_ovr_med$ = dat$adj / dat$med_r$

# ACF - the final predictor variable in the "Opportunity Indicator" class
acfdat$ <- Box.test(dat$remainder, lag=10, fitdf=0, type="Lj")
dat$ljpval$ = acfdat$p.value

# Last column is [ 0 | 1 ] Response Variable for FVA
p$ <- nrow(dat)
for (j in 1:p){
    if(dat[j,"fva"] < fva.lim) dat[j,"fva.resp"] = 0 else dat[j,"fva.resp"] = 1
### Testing Below ###

```r
if(sign(dat[j,"Act_Dem"] - dat[j,"Stat_Fcst"])< 0) dat[j,"cor_adj"] = 0 else dat[j,"cor_adj"] = 1
if(sign(dat[j,"Cons_Fcst"] - dat[j,"Stat_Fcst"])< 0) dat[j,"ovr_adj"] = 0 else dat[j,"ovr_adj"] = 1
```

#dir_accy.df <- confusionMatrix(dat$ovr_adj,dat$cor_adj)

### Testing Above ###

```r
}

write.csv(dat,paste(toString(prods[i,]),".csv"))
## Write out results
if(i==l) write.table(dat, file = "STLResults.csv", append=T, sep = ",", row.names=F)
else write.table(dat, file = "STLResults.csv", append=T, sep = ",", row.names=F, col.names=F)
```

### Load required libraries ###

```r
library(forecast)
library(dbplyr)
library(FNN)
library(ggplot2)
library(lattice)
library(class)
library(caret)
library(data.table)
library(e1071)
library(forecast)
library(ggplot2)
library(rpart)
library(rpart.plot)
```

### Data Loading and pre-processing ###

```r
# Read data
stldata.full.df <- read.csv("STL_Results.csv")

# Keep columns needed for data mining
don.drop.vars <- names(stldata.full.df) %in%
  "stat_pe", "cons_pe", "ae_adj",
  "seasonal", "trend", "remainder")
stldata.trimmed.df <- stldata.full.df[!don.drop.vars]

# fva.resp is our response, and needs to be a factor
stldata.df <- stldata.trimmed.df
fva.index <- which(names(stldata.df) == "fva.resp")
stldata.df[, fva.index] <- as.factor(stldata.trimmed.df[, fva.index])

# Set up data partitions - training (60%) and validation (40%). Sample random rows
train.index <- sample(rownames(stldata.df), dim(stldata.df)[1]*0.6)
valid.index <- sample(setdiff(rownames(stldata.df), train.index), dim(stldata.df)[1]*0.4)
train.df <- stldata.df[train.index, ]
valid.df <- stldata.df[valid.index, ]

# initialize normalized training, validation data, complete data frames to originals
train.norm.df <- train.df
valid.norm.df <- valid.df

# use preProcess() from the caret package to normalize columns.
norm.values <- preProcess(train.df[, ], method = c("center", "scale"))
train.norm.df[, -fva.index] <- predict(norm.values, train.df[, -fva.index])
valid.norm.df[, -fva.index] <- predict(norm.values, valid.df[, -fva.index])

# initialize a data frame with two columns: k, and accuracy.
accuracy.df <- data.frame(k = seq(1, 14, 1), accuracy = rep(0, 14))

##########################################################################
# compute knn for different k on validation.
for(i in 1:14) {
  knn.pred <- FNN::knn(train = train.norm.df[, -fva.index], valid.norm.df[, -fva.index],
                        cl = train.norm.df[, fva.index], k = i, prob = TRUE)
  accuracy.df[i, 2] <- confusionMatrix(knn.pred, valid.norm.df[, fva.index])$overall[1]
}
bnn <- which.max(accuracy.df$accuracy)

knn.pred <- FNN::knn(train = train.norm.df[, -fva.index], test = valid.norm.df[, -fva.index],
                      cl = train.norm.df[, fva.index], k = bnn, prob = TRUE)
accuracy.valid.df <- confusionMatrix(knn.pred, valid.norm.df[, fva.index])
### Regression Tree

Regression Tree

```r
ctrl=rpart.control(maxdepth = 30, cp = 0.001, minbucket = 1)
reg.tree <- rpart(fva ~ ., data = train.df[-c(3,6,7,8,14)], method = "anova", control = ctrl, xval = 20)

pruned.ct <- prune(reg.tree, 
  cp = reg.tree$cptable[which.min(reg.tree$ cptable[, "xerror"] ), "CP"])
length(pruned.ct$frame$var[pruned.ct$frame$var == "<leaf>" ])
prp(pruned.ct, type = 1, extra = 1, split.font = 1, varlen = -10)

# plot tree
rpart.plot(reg.tree)
prp(reg.tree, type = 1, under = TRUE, split.font = 1, varlen = 0, nn=TRUE, extra=1, 
  box.palette="auto", 
  fallen.leaves = FALSE, tweak = 1.3)
print(reg.tree)
printcp(reg.tree)
plotcp(reg.tree)
fancyRpartPlot(reg.tree)

# Determine model accuracy
accuracy(predict(reg.tree, train.df), train.df$fva)
```

### Classification Tree

Classification Tree

```r
ctrl=rpart.control(maxdepth = 5, cp = 0.01, minbucket = 1)
class.tree <- rpart(fva.resp ~ ., data = train.df[-c(1,3,6,7,8)], method = "class", control = ctrl, xval = 5)

# plot tree
rpart.plot(class.tree)
prp(class.tree, type = 1, under = TRUE, split.font = 1, varlen = 0, nn=TRUE, extra=1, 
  box.palette="auto", 
  fallen.leaves = FALSE, tweak = 1.3)
print(class.tree)
printcp(reg.tree)
```

# Predict and Generate confusion matrix

```r
class.tree.pred.train <- predict(class.tree,valid.df[-c(1,3,6,7,8)],type="class")
confusionMatrix(class.tree.pred.train, valid.df$fva.resp)
```
```r
cms.ct <- confusionMatrix(class.tree.pred.train, valid.df$fva.resp)

################### Random Forest
library(randomForest)
rf <- randomForest(fva.resp ~ ., data = train.df[-c(1,3,6,7,8)], ntree = 500,
mtry = 4, nodesize = 5, importance = TRUE)

## variable importance plot
varImpPlot(rf, type = 1)

## confusion matrix
rf.pred <- predict(rf, valid.df[-c(1,3,6,7,8)])
confusionMatrix(rf.pred, valid.df$fva.resp)

########################### Boosted Tree
library(adabag)
library(rpart)
library(caret)
boost <- boosting(fva.resp ~ ., data = train.df[-c(1,3,6,7,8)])
bt.pred <- predict(boost, valid.df[-c(1,3,6,7,8)])
confusionMatrix(bt.pred$class, valid.df$fva.resp)

########################### Logistic Regression
library(gains)
train.lr.df <- stldata.dft[train.index,c(2,4,5,9:14)]
valid.lr.df <- stldata.dft[valid.index,c(2,4,5,9:14)]

logit.reg <- glm(fva.resp ~ ., data = train.lr.df, family = "binomial")
options(scipen=999)
summary(logit.reg)

# use predict() with type = "response" to compute predicted probabilities.
logit.reg.pred <- predict(logit.reg, valid.lr.df, type = "response")

# first 10 actual and predicted records
data.frame(actual = valid.lr.df$fva.resp[1:10], predicted = logit.reg.pred[1:10])
```

75
gain <- gains(valid.lr.df$fva.resp, logit.reg.pred, groups=10)

# plot lift chart
plot(c(0,gain$cume.pct.of.total*sum(valid.lr.df$fva.resp))~c(0,gain$cume.obs),
       xlab="# cases", ylab="Cumulative", main="", type="l")
lines(c(0,sum(valid.lr.df$fva.resp))~c(0, dim(valid.df)[1]), lty=2)

# compute deciles and plot decile-wise chart
heights <- gain$mean.resp/mean(valid.lr.df$fva.resp)
midpoints <- barplot(heights, names.arg = gain$depth, ylim = c(0,4),
                           xlab = "Percentile", ylab = "Mean Response", main = "Decile-wise lift chart")

# add labels to columns
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

# confusion matris=x
confusionMatrix(ifelse(logit.reg.pred > 0.5, 1, 0), valid.df$fva.resp)

Data Analysis

# Histogram
library(ggplot2)
ggplot(data=stldata.full.df, aes(stldata.full.df$adjsize)) +
  geom_histogram(aes(y = ..density..), breaks=seq(-1,2, by=0.1),
     col="red", fill="green",
     alpha = .2 ) +
  geom_density() +
  labs(title= "") +
  labs(x="Adjustment Size (%)", y="Percent") +
  xlim(c(-1,2)) +
  theme_bw()

mean(subset(stldata.full.df, adj_size > 0)$adj_size)
mean(subset(stldata.full.df, adj_size < 0)$adj_size)
median(subset(stldata.full.df, adj_size > 0)$adj_size)
median(subset(stldata.full.df, adj_size < 0)$adj_size)

ovr.pos <- subset(stldata.full.df, adj_size > 0)
ovr.neg <- subset(stldata.full.df, adj_size < 0)
ovr.non <- subset(stldata.full.df, adj_size == 0)
index.pos <- rownames(ovr.pos)
index.neg <- rownames(ovr.neg)
index.non <- rownames(ovr.non)
accuracy(stldata.full.df$Stat_Fcst, stldata.full.df$Act_Dem)
accuracy(ovr.pos$Stat_Fcst, ovr.pos$Act_Dem)

76
accuracy(ovr.neg$Stat_Fcst, ovr.neg$Act_Dem)
accuracy(ovr.non$Stat_Fcst, ovr.non$Act_Dem)

library(TSrepr)
# mase(real, forecast, naive)

library(ftsa)
error(forecast=ovr.pos$Stat_Fcst, true=ovr.pos$Act_Dem, method="mape")
error(forecast=ovr.pos$Stat_Fcst, true=ovr.pos$Act_Dem, method="mdape")
error(forecast=ovr.neg$Stat_Fcst, true=ovr.neg$Act_Dem, method="mape")
error(forecast=ovr.neg$Stat_Fcst, true=ovr.neg$Act_Dem, method="mdape")
error(forecast=ovr.pos$Cons_Fcst, true=ovr.pos$Act_Dem, method="mape")
error(forecast=ovr.pos$Cons_Fcst, true=ovr.pos$Act_Dem, method="mdape")
error(forecast=ovr.neg$Cons_Fcst, true=ovr.neg$Act_Dem, method="mape")
error(forecast=ovr.neg$Cons_Fcst, true=ovr.neg$Act_Dem, method="mdape")