A Framework for Analyzing Forecast Accuracy Metrics

by

Yalu Wu
B.S. Management Science, Massachusetts Institute of Technology, 2009
B.S. Economics, Massachusetts Institute of Technology, 2009

Submitted to the MIT Sloan School of Management and the Civil Engineering Department in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Civil Engineering

In conjunction with the Leaders for Global Operations Program at the Massachusetts Institute of Technology

June 2015

Signature of Author ______ Signature redacted
MIT Sloan School of Management, Department of Civil Engineering
May 8, 2015

Certified by ______ Signature redacted
Stephen C. Graves, Thesis Supervisor
Professor of Management Science, MIT Sloan School of Management

Certified by ______ Signature redacted
David Simchi-Levi, Thesis Supervisor
Professor of Civil Engineering, Department of Civil Engineering

Accepted by ______ Signature redacted
Heidi Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Accepted by ______ Signature redacted
Maura Herson, Director of MBA Program
MIT Sloan School of Management

This page intentionally left blank.

A Framework for Analyzing Forecast Accuracy Metrics

by

Yalu Wu

Submitted to the MIT Sloan School of Management and the Civil Engineering Department on May 8, 2015 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Civil Engineering.

Abstract

Demand Planning forecasts at Nike, Inc. are used by many groups: Supply Planning/Materials Planning, Sourcing, Categories/Merchandising, Finance, S&OP, and Sales. These groups take forecasts as an input to make key decisions. Forecasts, by nature, will be inaccurate. There are two big unknowns to answer as Nike considers how to improve forecast accuracy: 1) how accurate can or should forecasts become (target setting) and 2) what are the causes and impacts of inaccuracy.

However, the first step to addressing these questions is to understand and measure forecast accuracy metrics in a consistent way across Nike's various Demand Planning groups. This project investigates the following through the design of a Tableau dashboard

- which metrics should be reviewed (accuracy, bias, volatility, etc.)
- how they should be computed (what to compare)
- at what level of aggregation for which groups
- at what level of detail for which groups (category, classification, etc.)
- over how many seasons
- with which filters

In addition to aligning on forecast accuracy metrics, the project also focuses on the dashboard design (determining the most appropriate structure/views, how information is laid out or presented, and the use of labels and color) and on setting the long-term vision for viewing and using forecast accuracy metrics through researching and outlining the process for root cause analysis and target setting.

Thesis Supervisor: Stephen Graves
Title: Professor, Management Science

Thesis Supervisor: David Simchi-Levi
Title: Professor, Civil Engineering

This page intentionally left blank.

Acknowledgements

I would like to first thank my coworkers at Nike for their support with this project. In particular, special thanks to Luciano Luz, my project supervisor, and Laura Baker, my project sponsor, who were always available to answer questions and provide guidance for the project, and without whose help this project would not have been possible.

I would also like to thank my advisors at MIT, Professor Stephen Graves and Professor David Simchi-Levi, for their time, advice, and assistance through the project.

In addition, thank you to my fellow LGOs, John Kang, David Jacobs, and Ryan Jacobs, for their invaluable feedback and friendship during our time at Nike and through the LGO program. Along a similar vein, I would like to thank and acknowledge Alessandra Perez Mak and Jonathan Dobberstein.

Finally, I would like to thank my family and in particular Michael for their support in everything I do.

This page intentionally left blank.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Definitions

Category	Type of athletic activity, e.g. Running, Training, Basketball
CGP	Category Game Plan, a 2-year planning process under the futures business model
FOB	Freight on Board
Geography	Nike designated areas in which the company does business (6 separate areas)
GFP	Global Footwear Planning, planning organization within Nike for footwear
LGO	Leaders for Global Operations, a MIT dual-degree program focused on operations
Product Engine	Type of product by production type: Footwear, Apparel, or Equipment
SKU	Stock Keeping Unit, a distinct product
Tableau	Data Visualization Software
Tooling	Equipment used to form the outsole of a sandal/shoe

1 Introduction

The research for this project was conducted during an LGO internship at Nike. The project focused on determining how forecast accuracy metrics need to be viewed and analyzed across the organization. This chapter introduces the purpose of the project.

1.1 Problem Statement and Motivation

Nike operates in many markets, causing demand planning to be decentralized – long-range forecasts are developed by product lines to better align with product development, while short-range forecasts are developed by geography to better align with retail accounts and selling. Planners use forecast accuracy metrics, which compare forecasted demand to actual retailer bookings, to determine the performance of their forecasts.

These metrics, however, are not standardized across groups, leading to difficulty in comparing and improving forecasts. In addition, within each group, the level of product aggregation at which these metrics should be measured, to best inform decision-making, is oftentimes unclear. Consequently, a cross-functional team was formed to align on these metrics and the processes surrounding them. Essentially, the team aims to establish when to measure which metrics, at what level(s) of aggregation, and how these metrics should be viewed/compared.

To facilitate the conversation around how forecast accuracy data should be viewed and analyzed, a Tableau dashboard was developed. The dashboard served as a pilot and helped set the vision for how forecast accuracy metrics are used within the organization in the future.

1.2 Project Goals

The primary goal of the project was to reach a consensus on how forecast accuracy metrics should be viewed and analyzed across the organization. This was facilitated through the design of a forecast accuracy dashboard, which initiated conversations with and feedback from the various demand planning groups. The ultimate goal was for key leaders to accept the vision and incorporate this framework of measuring forecast accuracy into their regular Planning Directors' review process.

The project also aimed to set the long-term vision for improving forecast accuracy, specifically by setting an approach for understanding the impacts of forecast inaccuracy and the target setting of forecasts.

Vision for Forecast Accuracy Dashboard

A first step to improving forecasts is to understand and measure forecast accuracy metrics in a standard, consistent manner over time. In doing so, initiatives carried out to improve forecast accuracy can be evaluated compared to a baseline. In addition, aligning on forecast accuracy metrics across different groups at Nike allows for the opportunity to compare and to transfer learnings.

The Tableau dashboard will enable greater visibility of forecast accuracy metrics across the organization. Because the dashboard is a global tool that caters to various stakeholders (senior leadership, the product engines, and the geographies), it provides a common platform for individuals to initiate discussions. An example of such a discussion would be how forecasts done by the geographies in the short-range could align to the forecasts by product classification done by the product engines in the long-range in a way that ensures forecast accuracy does not drop at the handoff point.

Long Range Vision for Improving Forecast Accuracy at Nike

An additional project goal was to set the long range vision for forecast accuracy metrics through the process of aligning metrics through the dashboard design. The dashboard enables the demand planning function within Nike to look at forecast accuracy metrics in a consistent manner and set the stage for root cause analysis, inaccuracy impact cost analysis, and target setting. Although the dashboard is in essence only a measurement tool, these next steps that build upon the dashboard then allow the demand planning function to better understand what level of accuracy are acceptable or "good enough" at certain gates/geographies/categories. The root cause and impact analysis allow demand planning to better understand the tradeoff in cost and resources between implementing mechanisms that could improve accuracy (root cause piece) and the financial impact of forecast inaccuracy. Section 7 Root Cause, Forecast Inaccuracy, and Target Setting provides a more detailed analysis of the long range vision.

1.3 Approach

An initial three-day in-person workshop set the stage for alignment on forecast accuracy metrics and formed a team with representatives from different groups across the organization. At the conclusion of the workshop, the team set out four key long-term priorities: a new tool/scorecard with slice and dice capabilities (i.e. the ability to drill down to view the data at various levels of aggregation), capabilities to capture root cause, visualization, and target setting. In addition, the team suggested a set of initial alignments, which are explained in Section 4 Methodology.

To understand how different groups need to view or use forecast accuracy metrics, the current state and process was analyzed. Best practices that were localized within groups were adopted or recommended for the ideal future state. The Tableau dashboard was then developed and iterated upon through two rounds of feedback and user testing with all the groups, to provide the final alignment for forecast accuracy metrics.

1.4 Thesis Overview

In this paper, Section 2 will provide a background of Nike as a company and how forecast accuracy fits into Nike's planning process. Section 3 contains a brief Literature Review that looks at forecasting in the retail industry, metrics and dashboards, and change management of processes in a large organization. Section 4 goes over the methodology for this project, while Section 5 discusses the dashboard and Section 6 discusses the final dashboard results. Finally, Section 7 delves into the long term vision aspects of forecast accuracy (regarding root cause, inaccuracy, and target setting). Sections 8 and 9 discuss the implementation of the dashboard into the forecasting review process and recommendations for future research.

2 Forecast Accuracy in Nike Planning's Process

2.1 Background of the Company

Nike, Inc. was founded in 1964 and today has annual revenues of $26 billion. As a sportswear company, Nike markets a wide variety of athletic footwear, apparel, and equipment under its own brand and its subsidiaries, which include Converse, Hurley, Jordan Brand. In 2013, Nike joined the Dow Jones industrial average.

Nike is best known for its innovative products and clever marketing campaigns. The company's focus has been to continue to develop great products and build its brand. Throughout the decades, innovations such as Nike Air, Zoom, Free, Flywire, and Flyknit have contributed to the company's brand recognition and overall success.

Most of the manufacturing of Nike products is done in Asia. In the late 1990s, Nike received significant negative publicity surrounding the labor practices of its subcontracted factories. Since the early 2000s, Nike has made significant efforts to counter this image and has in recent years become a leader on labor and sustainability-related issues (Nisen 2013). On labor, for example, Nike does not engage in seasonal hiring or firing of workers in the factories, thus guaranteeing a lower turnover rate and a higher quality product.

2.2 Supply Chain Structure at Nike

Nike is primarily focused on product design, development, and marketing. The company does not manufacture or sell its products, with some notable exceptions[1]. Many factories that produce Nike products are located in Asia due to cheaper manufacturing costs, which result in long lead times.

Nike works with over 900 factories in 50 countries, engaging approximately 1 million workers. The supply chain is complex because of long product creation timelines, long lead times, short product lifecycles, large product assortment, and a global customer base.

2.3 Futures Business Model

Nike operates with 4 primary business models: futures, replenishment, quick turn, and custom. The seasonal futures model comprises most of Nike's business by volume; under this model, retailers place orders several months ahead for a particular season and receive new products (styles) every season. The replenishment model includes business lines such as Always

[1] Manufacturing: Nike operates a select few manufacturing sites. Selling: Nike does sell directly through online (Nike.com) and physical stores, such as Niketown or Nike Factory Store

Available, which focuses on long lifecycle products (on average spanning several seasons) and provides retailers with weekly replenishment. The Quick Turn model focuses primarily on quick turnaround products such as NFL jerseys and the Custom model on services such as NikeID, which allows customers to design and personalize their own merchandise.

This project specifically focused on demand planning and forecast accuracy within Nike's core business, the seasonal futures business model. The planning process under this business model is approximately 2 years, as can be seen in the Category Game Plan (CGP) illustrated in Figure 1 below.

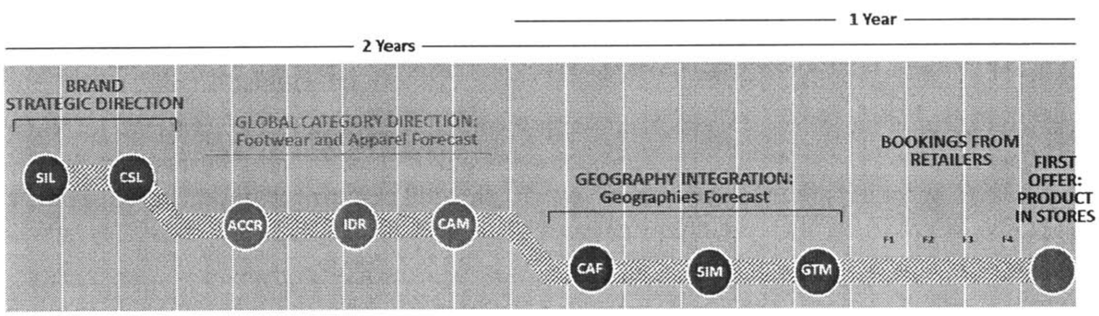

Figure 1: Category Game Plan

The CGP is an outline of planning activities or milestones that occur during every season. The circles in the figure above represent "gates", or milestones within the process for product, merchandising, demand and supply planning, and sales teams. Many groups execute certain activities at particular points in the timeline of each season's CGP. For confidentiality reasons, the names of the gates are not disclosed in this thesis.

Under this futures business model, Nike is internally organized by four distinct levels:

1) Global functions (Finance, Supply Chain, Planning, etc.),

2) 7 Categories responsible for product development (Running, Basketball, Football, Sportswear, Athletic Training, Women's Training, and Action Sports),
3) 3 Product Engines responsible for production (Footwear, Apparel, Equipment),
4) 6 Geographies responsible for execution (North America, Western Europe, Central & Eastern Europe, Japan, Greater China, and Emerging Markets)

2.4 Forecast Accuracy Metrics

How Forecasts Are Used

Demand Planning creates forecasts throughout the timeline of the CGP. These forecasts are used all over the organization by informing (serving as an input to) key planning decisions made by other groups such as Supply Planning/Materials Planning, Sourcing, Categories/Merchandising, Finance, S&OP, and Sales. On the long-term supply planning side, forecasts affect capacity and tooling decisions, sourcing decisions, and raw material purchases. On the short-term selling side, forecasts affect shipping/transit decisions (the need to air freight), inventory decisions, and whether there were lost sales. Figure 2 and Figure 3 show where forecasts are submitted relative to key apparel and footwear supply planning decisions. In the figures, the diamonds represent a forecast, and the position of the diamonds on the CGP represents the timing at which forecasts are submitted (immediately after the gates in the CGP), with the colors indicating which groups are responsible for the forecast at that stage. The gray boxes indicate planning decisions based off of that forecast. The F1, F2, and F3 milestones in the graphics indicate the timing at which retailers place their orders (for example, F1 is the "first futures" deadline by which retailers must place their orders with Nike).

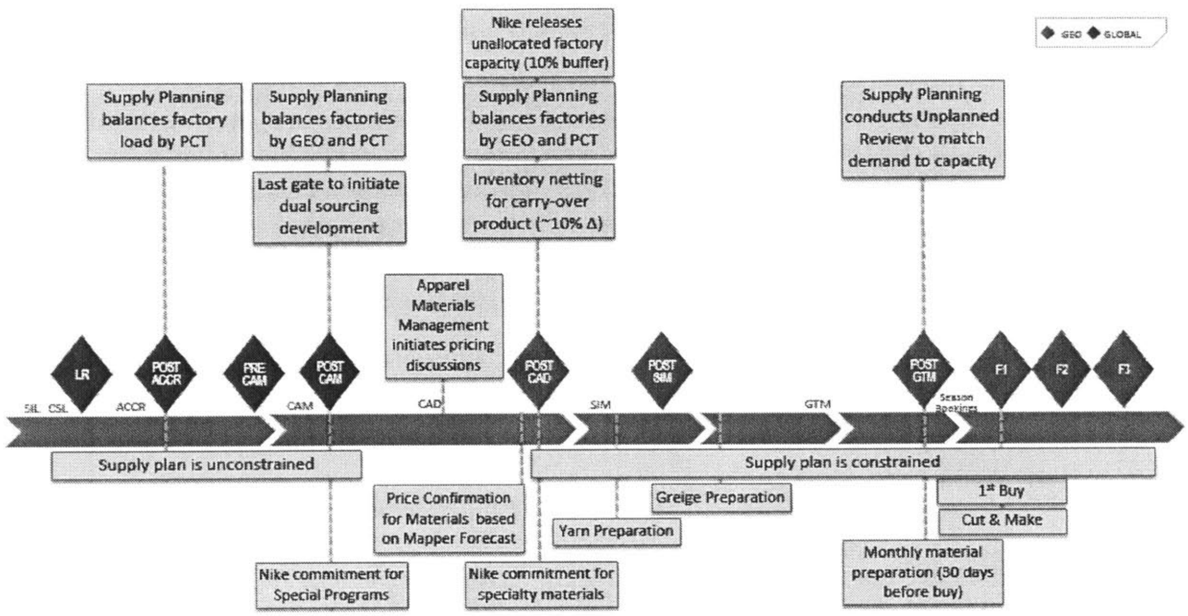

Figure 2: Key Apparel Planning Decisions

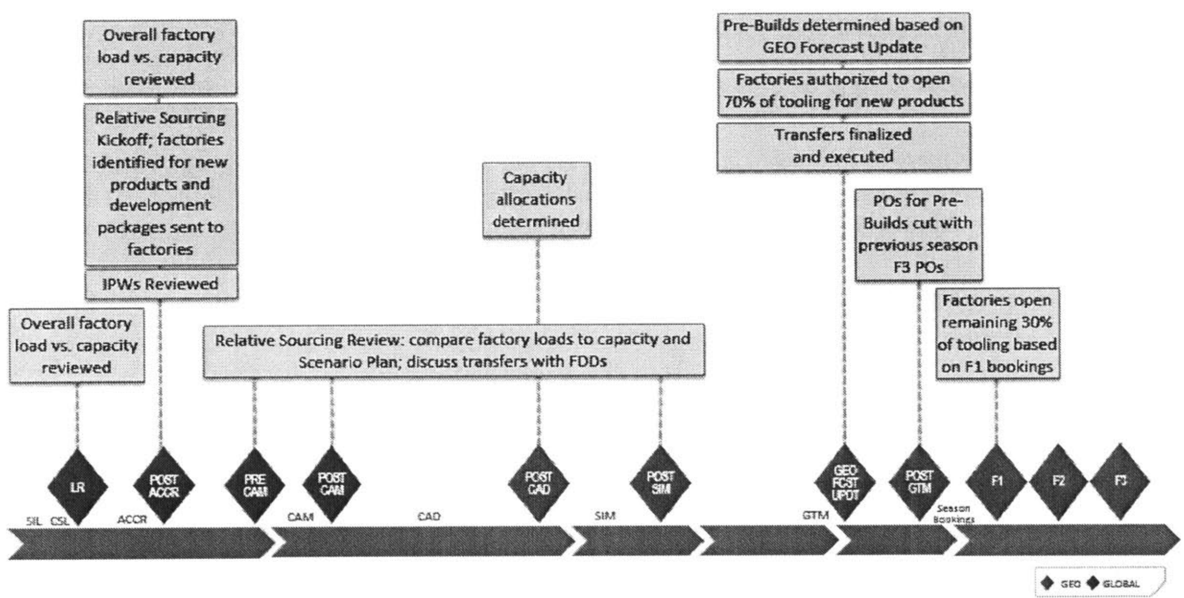

Figure 3: Key Footwear Planning Decisions

How Forecasts Are Created

For every season, long range forecasts (corresponding to ACCR through CAM) are developed by Product Engines that work with product development to form a top-down approach to forecasting. Forecast numbers for specific products are dictated primarily from a product creation perspective. Short range forecasts (corresponding to CAF through GTM), on the other hand, are developed by Geographies that work with retail accounts in those geographies to form a bottom-up approach. Enough retail accounts must pledge to buy a product for it to remain part of the plan.

In the long-range, the Footwear product engine plans at the global level because Footwear has a global source base whereas Apparel plans at the geography level in the long-range because Apparel sources by geography. Planners forecast according to the CGP in Figure 1 and "submit" their forecast at each gate. They submit projected volumes for each product or groups of product that will be sold for that entire season, and forecast at more aggregate levels if the product line is yet to be finalized. Planners are constantly revising forecasts between gate submissions and on average revise their forecasts twice a week.

Forecasting is inherently a difficult task for Nike, because the company is essentially introducing new products each season (every three months). Although products can be similar season to season (i.e. there is a lineage of products), in general "the footwear and apparel industry is challenged with forecasting in a world of long lead times and short product lifecycles. Since the majority of footwear and apparel production is outsourced and produced in Asia, it is common for order lead-times to be three to six months. When this is coupled with an industry that depends on current fashion trends, forecasting can be quite difficult." (Axline & Lebl 2007)

Studies indicate forecasts are more accurate when they combine the perspectives of many different groups. Thus, geography demand planners take inputs from Merchandising and Sales[2], and also look to similar products and styles from previous seasons to inform their forecasts.

[2] Merchandising can dictate which products will be heavily promoted, while Sales is closest to the retail accounts. Nike has considerable influence in leading the customer (setting the hot item) rather than the other way around.

Using these various inputs (or constraints), they then combine this information into a single number for a given product for that season.

How Forecast Accuracy Metrics Are Calculated

Retailers place orders, or bookings, several months prior to the first offer date (when products are in stores). As can be seen from the CGP, retailers can start placing orders shortly after GTM. The F1, F2, F3, and F4 milestones refer to snapshots of futures/retailer orders (e.g. F1 is called the first futures submit) and indicate the orders/demand received at that time. In computing forecast accuracy metrics, these orders represent the actual demand. Thus, the "actuals" represent *sell-in* demand, what retailers buy from Nike, versus *sell-through* demand, what the consumers buy from stores.

Forecast accuracy metrics at Nike refer to the following: accuracy, bias, and volatility. While the project initially started by focusing on accuracy alone, it soon became clear that this was not sufficient to meet group's needs, as detailed in 6.1 Final Dashboard Assumptions. There are many ways to calculate accuracy, bias, and volatility, as mentioned in Section 3 Literature Review, but the final aligned definitions are provided below.

Forecast Accuracy

Accuracy is computed by comparing a forecast at a particular point of the CGP with actual orders at a particular point of the CGP. For example, comparing a PostGTM forecast with a F1 actual yields the forecast accuracy at PostGTM vs. F1. Forecast accuracy is calculated as 1 – forecast error, as shown in Figure 4 below. If the forecast error (the part of the equation within the parenthetical) exceeds 100%, then the accuracy is 0%. For example, for a product with 75 units forecast and 25 units actual, the forecast error would be 200% and the accuracy would be 0%.

$$1 - \left(\frac{\Sigma \,|(\text{Forecast Qty} - \text{Actual Demand Qty})|}{\Sigma \text{ Actual Demand Qty}} \right) * 100$$

Figure 4: Forecast Accuracy Calculation

The summation in the equation above applies depending on the level of forecast aggregation being measured. For example, at the most basic level, one can calculate forecast accuracy for each product (at the style-color level, so no summation in the equation above). However, many times that is too low of a level, so forecast accuracy is calculated at a higher level (for example, all cotton tee products). In that case, the summation is over all relevant products and the forecast error calculated here is essentially a weighted MAPE. This was deemed most appropriate for the organization so that products with greater volumes have a greater weight in the forecast accuracy calculation and to account for zero volume products (line drops). An example of this is given in the following Levels of Aggregation section.

Forecast Bias

Bias is computed in a similar way.

$$\left(\frac{\Sigma \,(\text{Forecast Qty} - \text{Actual Demand Qty})}{\Sigma \text{ Actual Demand Qty}} \right) * 100$$

Figure 5: Forecast Bias Calculation

This metric is simply the non-absolute error and indicates whether the forecast has under-forecasted or over-forecasted. It does not look over multiple seasons, and is thus different from a more traditional definition of bias such as tracking signal, which "measures the degree to which a forecast is consistently low or high" (Webster 2009, p. 68).

Both accuracy and bias adhered to the definitions that were set firm-wide two years ago in a previous initiative to standardize metrics across the company (called *InfoAttack*) and are widely used by many Nike groups.

Forecast Volatility

Groups measure volatility in many different ways, as discussed in 6.1 Final Dashboard Assumptions. However, the final aligned definition is simply: the change in forecasted demand across all significant gates in the CGP. The preferred method of visualization, though, is to view a comparison of the absolute forecasted demand across these gates (versus percentage change or differences in forecasted demand across gates).

Levels of Aggregation

The accuracy and bias calculations will differ based on the level of calculation, also known as the level of aggregation. Figure 6 and Figure 7 illustrate how the forecast accuracy number differs if the calculation is done at the style-color level versus at the style level. In the following example, the numbers represent imaginary demand of a specific product for an entire season.

In below example, we calculate the error at style color level
Therefore we calculate the absolute error for each individual color
Then add all the absolute errors together; 151 in below example
Then we calculate the error % against the Actual qty (224) to get to the Error % (67%)
Accuracy = 1 minus the Error % = 1 - 67% = 33% error at style color

	Style A				
	Color 1	Color 2	Color 3	Color 4	TOTAL
Forecast	75	0	25	75	175
Actual	25	50	75	74	224
Abs Error	50	50	50	1	151
Error (%)	200%	100%	67%	1%	67%
Accuracy (%)	0%	0%	33%	99%	33%

Figure 6: Forecast Accuracy at Style-Color Level

In below example, we calculate the error at style level
Therefore we calculate the absolute error for each style
224 - 175 = 49 in below example
Then we calculate the error % against the Actual qty (224) to get to the Error % (22%)
Accuracy = 1 minus the Error % = 1 - 22% = 78% error at style

	Style A				
	Color 1	Color 2	Color 3	Color 4	TOTAL
Forecast	75	0	25	75	175
Actual	25	50	75	74	224
Abs Error	50	50	50	1	49
Error (%)	200%	100%	67%	1%	22%
Accuracy (%)	0%	0%	33%	99%	78%

Figure 7: Forecast Accuracy at Style Level

As we can see, the higher the level we measure forecast accuracy at, the more accurate the forecast will appear. This is because errors at the lower level, when aggregated, will compensate or make up for each other. In addition, adding more dimensions or fields to the data (such as geography, or sub regions) will also segment and further disaggregate the data, which will result in lower forecast accuracy figures than the same data at a more aggregate level.

3 Literature Review

Determining how to measure forecast accuracy is a process that must be customized to each organization and can depend heavily on the industry. Figure 8 shows a general recommended approach to aligning on tracking forecast accuracy.

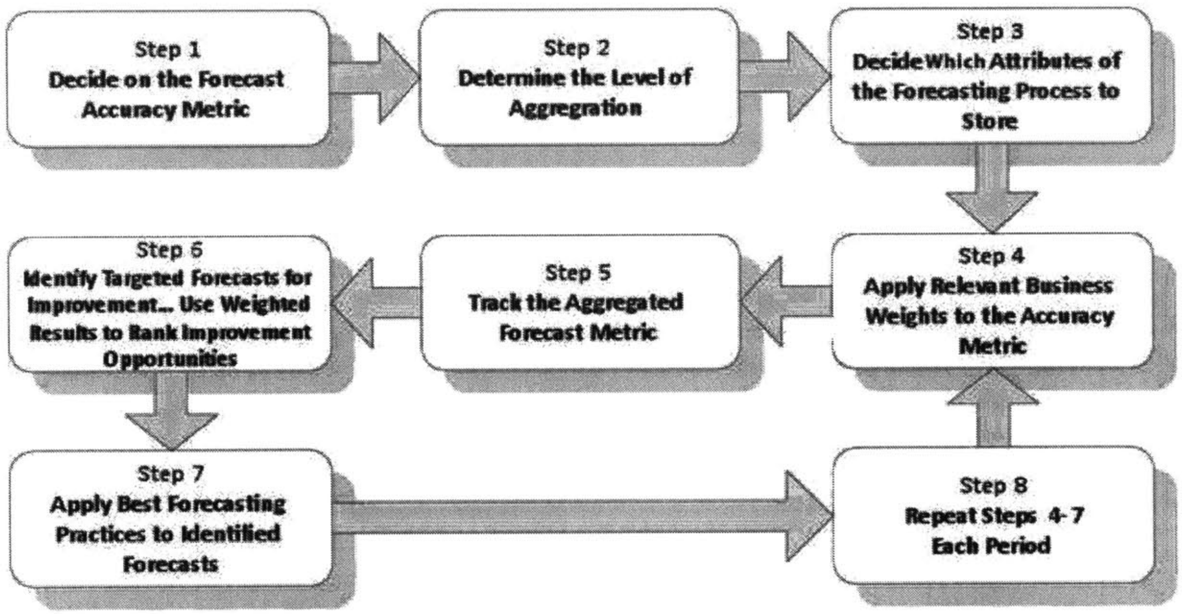

Figure 8: Tracking Forecast Accuracy

In this section, we review existing literature on forecasting practices (in particular in the retail industry), forecast accuracy metrics, target setting for forecasts, the use of dashboards for metrics in an organization, and change management in process changes. While there are common approaches and recommendations for each of these areas, especially in forecasting practices and forecast accuracy metrics, there is rarely an overall best practice or a long-term vision/process for standardizing the measurement of forecast accuracy across a company and ultimately improving accuracy. That is, while there are sound guidelines in each of these areas, they are oftentimes not well tied together in an actionable framework, which is what this project aims to do.

Thus, the following sections present existing best practices in each of these areas.

3.1 Current Retail Industry Practices for Forecasting and Demand Planning

As mentioned previously, forecasting in the retail industry is a hard task due to short product lifecycle and relatively long lead times. In addition, the ability to compare firms' forecasting processes is limited since practices across industries are so varied and industries differ on how much information they disclose. Stephan Kolassa (2008) describes this problem as "endemic in the retail market and makes benchmarking very difficult."

However, Armstrong's Principles of Forecasting provide some basic forecasting frameworks that are used in various industries and companies, including at Nike. For example, some recommendations are to tailor the level of data aggregation (or segmentation) to the decisions and to decompose the problem into parts (use a bottom-up approach by forecasting each component and then combining).

In addition, forecasting in the retail industry not only utilizes data such as past sales (of a specific product, where available) and current economic outlook, but must also consider the marketing promotions and objectives of the firm as a whole.

3.2 Forecast Accuracy Metrics

One question concerning forecasting is "how to measure?" There are many ways of computing forecast accuracy. Some commonly used metrics are: mean deviation (MD), mean absolute deviation (MAD), mean squared error (MSE), root mean squared error (RMSE), mean percent error (MPE), and mean absolute percent error (MAPE), as shown in Figure 9 below.

$$MD = \frac{\sum_{t=1}^{n} e_t}{n} \qquad MAD = \frac{\sum_{t=1}^{n} |e_t|}{n}$$

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n} \qquad RMSE = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n}}$$

$$MPE = \frac{\sum_{t=1}^{n} \frac{e_t}{D_t}}{n} \qquad MAPE = \frac{\sum_{t=1}^{n} \frac{|e_t|}{D_t}}{n}$$

Figure 9: Common Forecast Accuracy Metrics

While each metric has limitations, MAPE is a popular choice amongst firms, as the Foresight Fall 2008 issue points out: "the published surveys employ the MAPE – or a close variation thereof – as the 'standard' metric for forecast accuracy. In fact, there is little consensus on the 'best' metric for sales forecast accuracy. While the MAPE is certainly the most common measure used in sales forecasting, it does have serious shortcomings: asymmetry, for one, and error inflation if sales are low." As noted in Section 2.4 Forecast Accuracy Metrics, Nike employs a weighted MAPE to account for the error inflation for styles with low (or zero) sales.

A related question is what to measure – Accuracy? Bias? Unit Change? The same Foresight issue points out that "overforecasts and underforecasts of the same degree may have very different cost implications, depending on the industry and the product. Excess inventory may cost more than lost sales (as with short-life products like fresh produce, or high-tech items that quickly become obsolete), or it can be the other way around (e.g. for canned goods or raw materials). The MAPE and its variants, which treat an overforecast of 10% the same as an underforecast of 10%, may not adequately address the real business problem. KPIs that explicitly address over- and underforecasts may be more meaningful to forecast users."

Another common question concerning forecasting is: what level of aggregation to measure at? As mentioned in the previous section, a common recommendation is to aggregate at the level at which key decisions are being made. For example, "The first decision we need to make is the level of detail in which to aggregate our forecasts. We chose to aggregate items at the style level – essentially aggregating all sizes of an otherwise identical product – and predict demand for each style; the main reason why we do this is because pricing is set by style." (Johnson, Hong, Simchi-Levi 2013).

3.3 Target Setting

(Note: Target setting here, and throughout this paper, refers to targets or goals for forecast accuracy, as opposed to a more commonly used interpretation which refers to the sales target for the product.)

Setting forecast accuracy targets is a tricky task. By definition, the more accurate the forecasts the better – however, setting a target of 100% across all segments is not reasonable. How should one determine the level at which to set accuracy targets, for a given segment of the market (product line/geography)?

One idea is to develop several methods for forecasting (several different models, some based on historical data where possible) and use the highest accuracy number as the target. However, Bunn and Taylor (2001) have argued that using a relative measure provides no indication regarding how much improvement there could be beyond the best model. Instead, they argue that the goal should be to consider the forecast error; separating that error into irreducible and reducible (model) error, one can then minimize reducible error.

Mentzer and Moon (2004), however, argue for looking at the complete picture/the complete supply chain instead of just increasing accuracy or reducing error and note that "we should estimate the potential impacts on returns on shareholder value *prior to* making investments in sales forecasting and demand planning improvements. We make a serious mistake in sales forecasting and demand management when we go to upper management with a proposal to spend money to improve forecast accuracy, without indicating what 'impact' it will have on revenues, supply chain costs, and, subsequently, on shareholders."

3.4 Use of Dashboards in Business Intelligence

While improved forecasting practices are critical to any organization, it is equally important to have alignment and transparency across groups. Dashboards have become an easy way to review key metrics across an organization, at all levels and especially for senior management. Dagan (2007) asserts that "it is often not until your organization has achieved this level of visualization that the real value of your business intelligence infrastructure is realized."

Dashboards are "useful tools because they can leverage visual perceptions to communicate dense amounts of data clearly and concisely." (Smith 2013). Smith also recommends a set of steps to constructing and maintaining dashboards, as shown in the following figure.

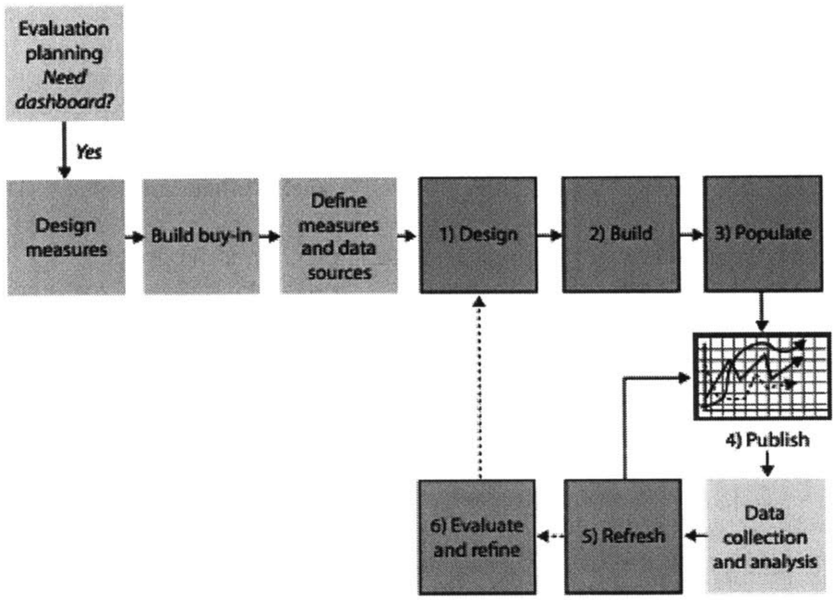

Figure 10: Common Forecast Accuracy Metrics

3.5 Change Management and Alignment Across Groups

Launching a new (or improved) dashboard application across the organization requires the buy-in of key stakeholders. As Chiang (2009) notes, "when you look at the data first, you risk choosing an incorrect delivery platform for your organization, one that may not address the needs of your end users… a more successful approach is to start by collaborating with stakeholders and asking: 'what reports or information do you want the end user to see?'"

In addition, in any large organization with multiple demand planning groups, it may be difficult to align the different groups to one common forecasting technique or process. Chase (2013) shares that he "was successful in creating change only when [he] relied on data and analytics to validate assumptions and provide a mechanism to develop strategies… members of the demand management team have always had to demonstrate that analytics outperforms judgment, and it has always been difficult to change long-held opinions."

4 Methodology

Through the dashboard design process, key initial assumptions about which metrics to use, when to measure, and what levels of aggregation to use were examined and tested. The following sections describe the primary steps which include the team and feedback process, initial assumptions and data, and mapping the current state and envisioned future state.

4.1 Overview

This project was launched with the formation of a cross functional team (comprised of individuals from various demand planning groups) focused on forecast accuracy. Through a 3-day workshop, the team compared high level processes and roles across groups, discussed targets for the forecasts, the timing of retail order bookings (the "actuals" in the forecast accuracy metric calculation), etc. The team identified four top priorities that would benefit their groups coming out of the workshop: 1) develop a new Global Scorecard (the Tableau dashboard) with slice and dice capabilities (the ability to drill down into the data), 2) root cause capabilities, 3) visualization and 4) target setting. The terms Global Scorecard and Tableau dashboard are essentially synonymous in that both refer to Global Planning's reporting/presentation of Forecast Accuracy metrics, henceforth Global Scorecard will refer to the original tool and Tableau dashboard to the new tool.

The team formed from the 3-day workshop consisted of demand planning managers and analysts from across the organization. All the primary groups (Global Planning, Product Engines, Geographies, and a few others) were represented. This team continued to meet after the workshop every two weeks to continue the discussions and action-items surrounding forecast accuracy alignment.

The Tableau dashboard design was able to take advantage of this structure to align on key initial assumptions for forecast accuracy metrics and to work with the various groups for feedback. Ultimately, the groups solicited for feedback (part of the process outlined in Section 4.2 Dashboard Design and Feedback Process) included the Footwear and Apparel Product Engines and all the Geographies (North America, Europe[3], Greater China, Japan, and Emerging Markets). Feedback was also obtained from senior demand planning leaders and the Global Planning group on a more ad-hoc basis.

4.2 Dashboard Design and Feedback Process

The Tableau dashboard development followed a series of steps, as illustrated in Figure 11 below.

Figure 11: Dashboard Design Methodology

The analysis of the current state ranged 5-6 weeks (excluding an initial 3-day workshop) and the data gathering and initial design 4 weeks. The feedback portion of the dashboard design spanned a total of 11-12 weeks and was conducted in three parts: a first round of feedback presenting the dashboard to key team members in various groups, a second round of feedback of a similar nature, and user testing of the dashboard. The first and second rounds of feedback were separated by approximately 4 weeks (during which changes were incorporated to the dashboard design) and the user testing followed 3 weeks after the second round of feedback meetings. The

[3] Nike is broken down into 6 geographies, but in all the data systems for reporting Forecast Accuracy Metrics Europe is not broken down between Western Europe and Central and Eastern Europe.

finalization of the dashboard design was a 2 week process that followed a 2 week period of user testing.

Feedback was collected from key team members and oftentimes additional individuals in those Geographies or Product Engines. In each round of feedback, each group had the opportunity to see the dashboard views that they would most likely use and to provide feedback. Types of feedback ranged from the types of charts that would be useful to have to the level of aggregation that users felt they needed for their analysis.

4.3 Initial Dashboard Assumptions

To support the development of the forecast accuracy dashboard in Tableau, the team aligned on a set of initial assumptions based on their discussions:

1) Incorporate the forecast accuracy metric
2) Measure forecasts at PostACCR, PostCAM, PostCAF, and PostGTM gates
3) Measure actuals at F1 and F4
4) Measure forecast accuracy at project code level for Geographies (short-term), at project classification type (PCT) level for Apparel and at tooling level for Footwear (long-term).

Incorporate the Forecast Accuracy Metric

The forecast accuracy metric was deemed the most important to all groups and it was initially recommended that the dashboard development start with this metric (before bias, volatility, or unit change).

Measure Forecasts at Specific Gates

Because the CGP structure contains three gates in the long-range and three gates in the short-range at which Product Engines and Geographies forecast (respectively), the team determined that forecast accuracy should be measured at PostACCR and PostCAM (the earliest and latest points at which the Product Engines forecast) and at PostCAF and PostGTM (the earliest and latest points at which the Geographies forecast).

Measure Actuals at F1 and F4

As we can recall from the CGP, actual bookings are represented as F1 (first futures booking), F2, F3, F4, and EOS (End of Season). The determination of which actuals to measure at (to use for forecast accuracy) was a more complex decision. Currently at Nike, this is not aligned across different Geographies. Table 1 below illustrates the pros and cons of measuring at F1 and F4 as compared to measuring solely at F3. Ultimately, there were more advantages to measuring at F1 and F4 as compared to F3.

	PROS	CONS
F1 & F4	• Timely • Planner's memory is fresh • Better ability to react. The directional feedback after F1 could impact seasonal submit • F1 is the last chance to impact Sourcing decision • F1 can be used as input to next year's Global Post-ACCR deadline • F4 timing aligns with next season's F1 (with some variations by PE) • F4 can be used as an input to next' year's Global Post-CAM	• Business is not closed at F1 • FW: partially booked • AP: mostly booked • F1 does not inform well for some categories • Fleece in FA, Baseball (EQ) in HO • F2 and F3 for FW are still at forecast • It could require multiple deep dives in the same season

	• F4 aligns with the process of other tools • Forecast Alignment, Flow Dashboard	
F3	• F3 is final booking deadline - Truer measurement • Futures bookings received for all PEs • One deep dive per season which is easier to organize	• Less timely • Two months later than F1 - DP and Categories have moved on by then • Planner's memory is not fresh • Actions cannot impact current season • Retrospective only hindsight

Table 1: Pros and Cons of Measuring Forecast Accuracy at F1&F4 vs. F3

Aggregation Level for Forecast Accuracy Calculation

Determining which levels of aggregation to measure forecast accuracy was also a complex decision. A proposal was made to measure forecast accuracy at the level key decisions are made, as this would most closely tie in with the impact of having inaccurate forecasts. The following figure shows some of these key decisions for Apparel and Footwear.

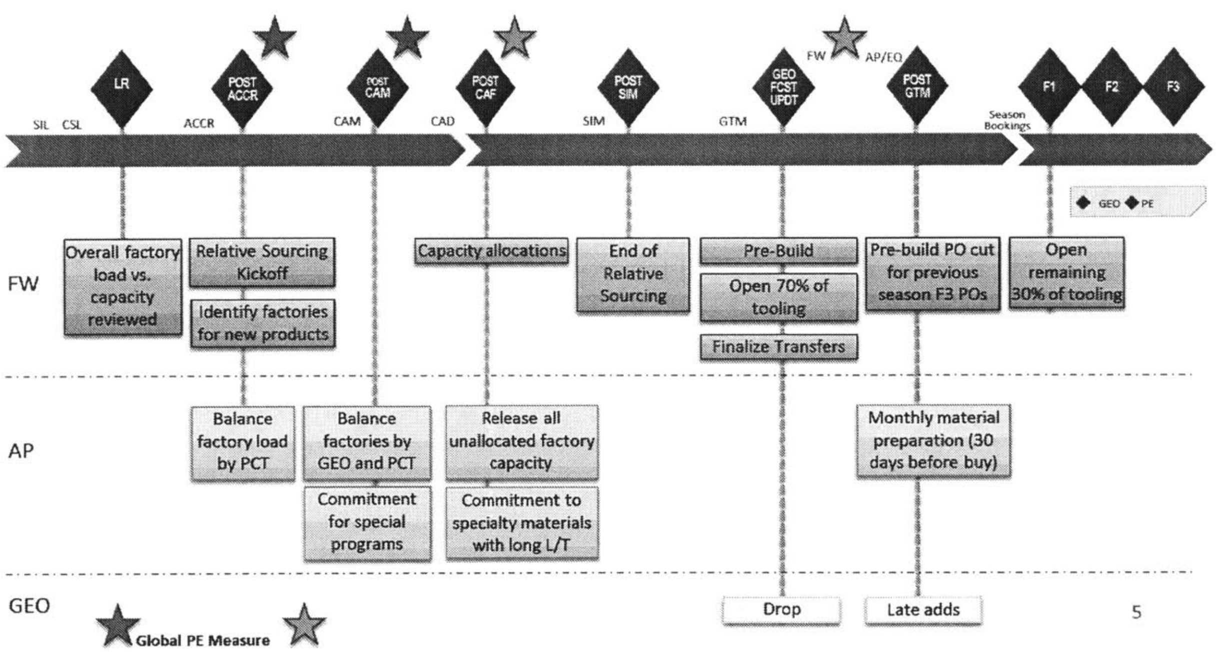

Figure 12: Key Planning Decisions

Figure 12 shows that many key planning decisions are made at the Product Classification Type level (PCT, level that distinguishes between different product types, for example Cotton Tees) for Apparel and at the Tooling level (type of machine needed to manufacture for Footwear) for Footwear. Thus, the group aligned to measure forecast accuracy metrics at the PCT level for Apparel and at the Tooling level for Footwear.

The Geographies aligned to measure forecast accuracy metrics at a more disaggregate Style Code level or an equivalent level (measuring accuracy at the level of each style, within each product classification type). However, it had been found that style codes were blank (marked as unknown) for the records submitted prior to Style Code assignment (when preliminary codes are still used in each of the Product Engines). Subject Matter Experts (SMEs) in each of the Geographies and Product Engines had agreed that Project Code, a data field that in the short-range corresponds to the preliminary codes used in each of the Product Engines, would be an appropriate proxy for style code at PostGTM.

4.4 Data Mapping and Sources

The data for the Tableau dashboard comes from a Cognos package that in turn pulls from a Teradata database. Most groups use Cognos for ad-hoc reports, but the Footwear Product Engine uses a different system entirely called DataMart; thus one goal of the dashboard is to ensure the same source of data.

4.5 Current State Analysis

Forecast Accuracy Reporting

A current state analysis was conducted to understand how the different Product Engines and Geographies look at forecast accuracy metrics at Nike and what sorts of tools and reports they utilize. The tools that are available to the groups are a global scorecard in Cognos, customized group dashboards in Tableau, and Cognos ad-hoc reports.

The original Global Scorecard in Cognos is a static report with colors representing ranges of forecast accuracy. The Scorecard has visibility across Geographies and Categories and is able to compare across Product Engines for each Geography-Category. However, none of the groups currently use the Global Scorecard for various reasons. The Product Engines do not utilize it because the Scorecard only shows forecast accuracy metrics in the short-range at PostGTM. The Geographies also do not utilize the Global Scorecard because they need to be able to drill down into the forecast accuracy numbers at many levels of aggregation (not just Category) to be able to understand what drives accuracy at a particular level.

Due to the limitations of the Global Scorecard and the need to view forecast accuracy metrics at a more detailed levels of aggregation, a few groups such as the Footwear Product Engine and the North America Geography built their own customized dashboards. These dashboards provide metrics at levels of aggregation specific to their groups and measure forecast accuracy at points of the CGP that make the most sense for them; thus, these customized views provide greater visibility into the forecast accuracy metric for that particular group (which is better for root cause analysis) but does not provide comparability or visibility across groups.

Ad-hoc reports are also available in Cognos, which allow groups to pull information at the right level they need to analyze. For groups that do not have custom dashboards, the Global Scorecard serves as a high-level overview of forecast accuracy metrics, while the ad-hoc reports help with deep diving and root cause analysis.

Forecast Accuracy Reach

While demand forecasts are used by many parts of the organization, such as supply planning and sourcing, demand planning is the only group that reviews forecast *accuracy* metrics. Demand planners analyze forecast accuracy metrics to understand how to plan better and to create more accurate forecasts in the future. Sourcing and supply planning models and frameworks take forecasts as givens/inputs – that is, they do not treat a product differently if accuracy for that product is 40% versus 80%, but focus solely on the demand volume. Thus, the overall system is currently more reactive than predictive, correcting for changes in demand with factory adjustments instead of potentially planning for them.

In reviewing forecast accuracy metrics, demand planners generally look at low forecast accuracy numbers (either of a particular product or at a particular point of time in the planning process) and attempt to trace the cause through root cause analysis. This could lead planners to change how they forecast in the next season, for example placing less of a reliance on the sales forecast or putting more emphasis on the accounts' indications of orders. Some inaccuracies depend on other groups (for example, a line drop or add late in the process warrants a discussion with the Merchandising team) and the ability to effect any changes/improvements could depend on demand planning's ability to influence other groups' processes.

4.6 Elements of the Future State

The Tableau dashboard, as a global tool, is meant to replace and serve as an upgrade or enhancement of the Global Scorecard. However, the dashboard is not meant to replace the customized dashboards of specific groups or the use of ad-hoc reports, though there may be overlap in some of the views reported. The dashboard provides senior leadership, demand planning managers, and demand planners with a tool to view/measure high-level forecast accuracy metrics that are aligned across groups.

By better understanding forecast accuracy metrics, the demand planning function can better understand the root causes that lead to forecast inaccuracy and how forecast inaccuracy impacts the supply chain. Forecasts will never be 100% accurate, but in order to set targets, it is important to understand the tradeoff between spending the resources to address a root cause of forecast inaccuracy and the cost of that inaccuracy upstream or downstream (upstream refers to costs that relate to improving a forecast, while downstream refers to costs incurred resulting from an inaccurate forecast). In other words, if the forecast accuracy of a particular Category in a particular Geography is 60%, is that good enough? If we want to get to 70% accuracy, how do we get there (which root cause(s) do we need to address) and what is the benefit of that 10%?

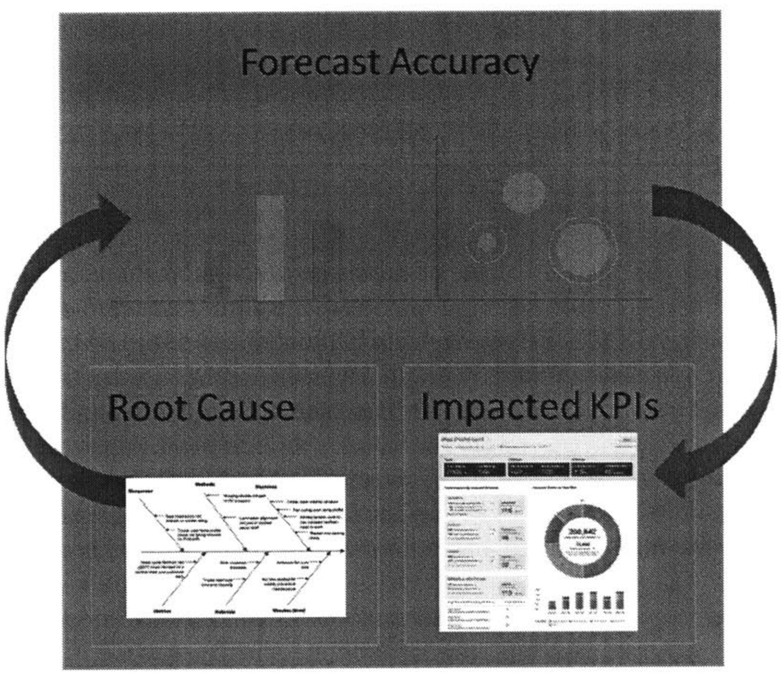

Figure 13: Future State of Forecast Accuracy Framework

Figure 13 describes the ideal relationship between root cause analysis, the forecast accuracy metrics, and other impacted key performance indicators (KPIs). The forecast accuracy dashboard cannot sit alone but instead must be supported by robust root cause analysis and clear quantifiable impact analysis. The tool can then achieve its goal of enabling the demand planning function with first, root cause analysis to understand how to improve forecast accuracy (sometimes through changes in other groups/functions that fall outside of demand planning) and second, by justifying those changes or influencing those groups by quantifying the benefit to the organization.

Improved forecast accuracy will translate to cost savings, as more accurate forecasts will lead to a reduction in costs associated with expediting or rebalancing (whether that be materials, production, or shipping). Although resources need to be added to the system (better measurements, tracking, and changes in processes), additional transparency and a more holistic system-wide view will allow the planning organization to balance the additional input costs of improving forecasts with the potential savings on the sourcing and supply planning side, with the ultimate goal of determining how to minimize overall costs in the system. As such, this transparency will provide for a more informed target setting method.

As mentioned previously, the current use of demand forecasts is reactive rather than predictive. With improved forecasts, this may still be the case. However, because certain products inherently have more or less stable demand than others, eventually the system could reach a state where products with higher or lower accuracy figures follow different sourcing processes.

5 Dashboard Design and Visualization

This section explores the design and visualization components of the forecast accuracy dashboard, specifically delving into how the data and Nike's organizational structure in demand planning shapes the way in which views are designed and how the dashboard is meant to be used across various groups.

5.1 Medium (Tableau)

The original Global Scorecard was designed in Cognos, where the ad-hoc forecast accuracy reports and demand planning training materials also reside. Thus, one option was to develop a dashboard in Cognos to keep all forecast accuracy reports accessible in one location. Another option was to use an external tool to develop the dashboard. While there are several options, the most obvious was to develop a Tableau dashboard, as Tableau is a tool widely used and supported across Nike.

The option to develop the dashboard in Cognos was ultimately rejected in favor of Tableau due to the ease of dashboard development in Tableau as compared to Cognos. In addition, Tableau has greater visualization capabilities in terms of creating views to filter and to slice and dice the data.

5.2 Design (Layout, Views)

The dashboard caters to a wide audience consisting of senior leadership, demand planning managers, and analysts who are in the Product Engines and Geographies. While some views are relevant across groups, in general each group only needs to use a specific set of views that contain information pertinent to their group. For example, the Apparel Product Engine needs to see all their metrics by the Apparel classification type, PCT, but the Footwear classification type would be irrelevant. Thus, in determining dashboard layout, the views can be separated into three sets: Apparel, Footwear, and Geographies that view metrics by PCT, Primary Platform Group, and Project Code respectively.

Figure 14 shows the Cover Page/Table of Contents for the Forecast Accuracy Dashboard. There are four sets of views: Executive (for senior leadership), Apparel, Footwear, and Geographies. The separation of these views also provides each group of key stakeholders with a customized list of views to focus on.

Figure 14: Forecast Accuracy Dashboard Cover Page/Table of Contents

The list of recommended views is relatively consistent across three of the sets of views (Apparel, Footwear, and Geographies). All three groups need to view the forecast accuracy metrics across gates, across seasons, at a certain gate, and through a classification lens. The Executive view, on the other hand, is very high level and provides an overall summary of the forecast accuracy metrics. The views under this set oftentimes compare across Geographies, whereas the views under the Geography set only allow a user to filter for specific Geographies.

Across Gates

Although planners mostly analyze forecast accuracy metrics at one particular gate (and oftentimes two), senior leadership and some demand planning managers emphasized the need to look at forecast accuracy metrics across multiple gates. This is especially true when looking at forecast accuracy metrics in the long-range through the lens of the Apparel and Footwear Product Engines, where you can compare metrics at PostACCR, PostCAM, PostCAF, and PostGTM. This comparison, as shown in Figure 15, allows planners to ensure that forecasts closer to PostGTM are more accurate. In addition, the comparison enables planners to take notice and do root cause analysis when forecasts do not consistently improve. One common drop in forecast accuracy often occurs at PostCAF, when forecasting is handed over from the Product Engines to the Geographies.

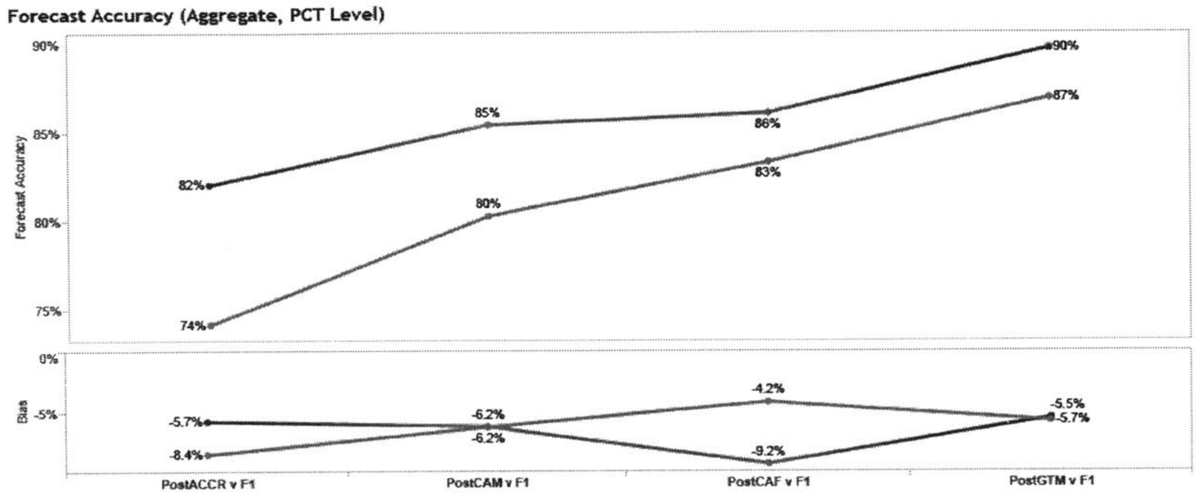

Figure 15: Forecast Accuracy Across CGP Gates

Figure 15 shows forecast accuracy and bias metrics across four comparisons (four different forecast gates, all compared to F1 actual bookings). In this Apparel-only case, forecast accuracy is calculated as described in Section 2.4 Forecast Accuracy Metrics, at the PCT level with each "observation" being a different product classification type. For example, the 82% accuracy at PostACCR v F1 indicates that there was only 18% forecast error when comparing the PostACCR forecast with the F1 bookings at the PCT level. The -5.7% bias figure indicates that the forecast at PostACCR is under forecasting (forecasted demand was less than actual demand) by 5.7% when compared to the F1 actuals.

In the above figure, the colored line[4] represents the metrics for the current season while the gray line represents the metrics from the same season in the previous year. Demand planners emphasized that they most commonly benchmark how their forecasts perform using a year-over-year comparison.

In this "across gates" view, planners cited the need to see forecast accuracy and demand in the same view. In addition, because these metrics can be analyzed at many different levels,

[4] Colors in the line correspond to colors in the CGP; these colors are held consistent throughout the dashboard

for example further disaggregated by Category or Classification, a standard view needed to be developed across all levels of aggregation. This consistent view is presented in Figure 16 below.

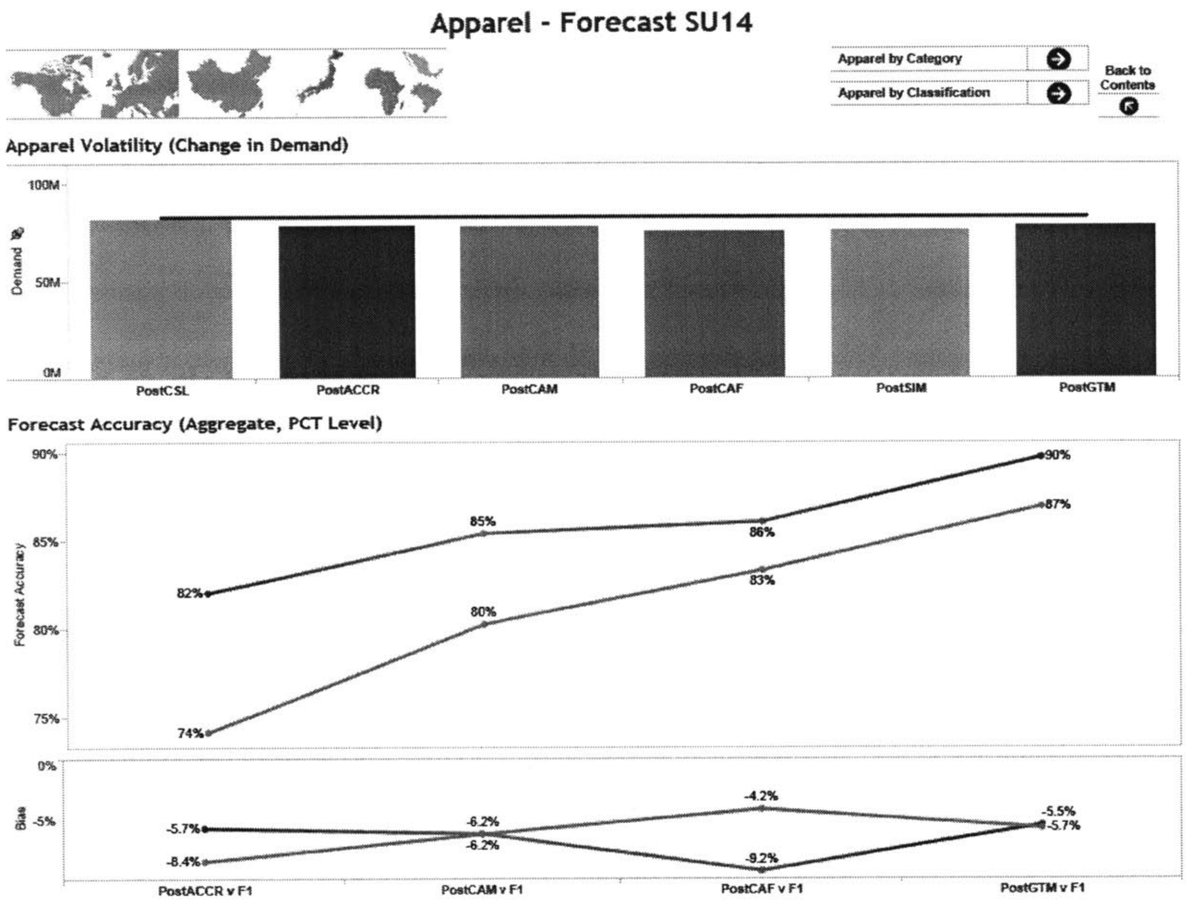

Figure 16: Standard View for Forecast Volatility (Change in Demand), Accuracy, and Bias Across Gates

The top graph shows demand across various gates in the CGP and the bottom graph accuracy and bias. While accuracy and bias are measured at four specific gates, the Planning Directors felt that volatility needed to be represented as the change in forecasted demand across significant gates in the CGP; thus, the demand graph shows the volume at PostCSL in the long-range and PostSIM in the short-range in addition to the four other gates. In the demand graph, the gray horizontal line across the top shows F1 demand across all the forecasts, highlighting the difference between the actual and the forecast at that gate.

In each of these views, the level of product aggregation at which accuracy is measured is indicated both by the heading and by the filters available. For example, the view shown in Figure 16 indicates that accuracy is measured at the PCT level for Apparel and also contains Geography filters on the bar in the upper left (showing North America, Europe, Greater China, Japan, and Africa, in that order). These Geography filters allow a user to view these same charts/statistics, but only for a particular geography – for example, one could view forecast accuracy for Apparel Japan at the PCT-Geography level of aggregation. Thus, this indicates that the accuracy numbers presented here at the aggregate level (with no Geography filter activated) are in fact at the PCT-Geography level rather than at the PCT level. Because of the Geography filter, these figures at the aggregate level are actually the weighted average accuracy at the PCT-Geography level, weighted by Geography volume (if North America has a greater volume of product, the accuracy of North America's products will have more bearing on the aggregate accuracy figure). In other words, without the Geography filters, the accuracy numbers would be different in that they would be at a more aggregate level, the PCT level (global instead of by geography), and likely higher.

Because the calculation of forecast accuracy means a different view at each level of aggregation, each view is accompanied with a set of buttons in the upper right that allow demand planners to toggle to different levels of aggregation. In the Category view, for instance, demand planners can view volatility, accuracy, and bias broken down in each Category, and can filter to only see Categories they are interested in. This is extremely useful as demand planners are often responsible for several Categories. The Category view then has additional levels of aggregation to toggle through – a demand planner can conduct a deep dive to the Core Focus level, which is a level under Category.

Across Seasons

The structure of the standard view persists through the dashboard, even in the view Across Seasons as shown in Figure 17. This view also allows the user to filter by Geography, and allows the user to select which seasons and PCTs (here shown as Orig Src Type Desc) to include.

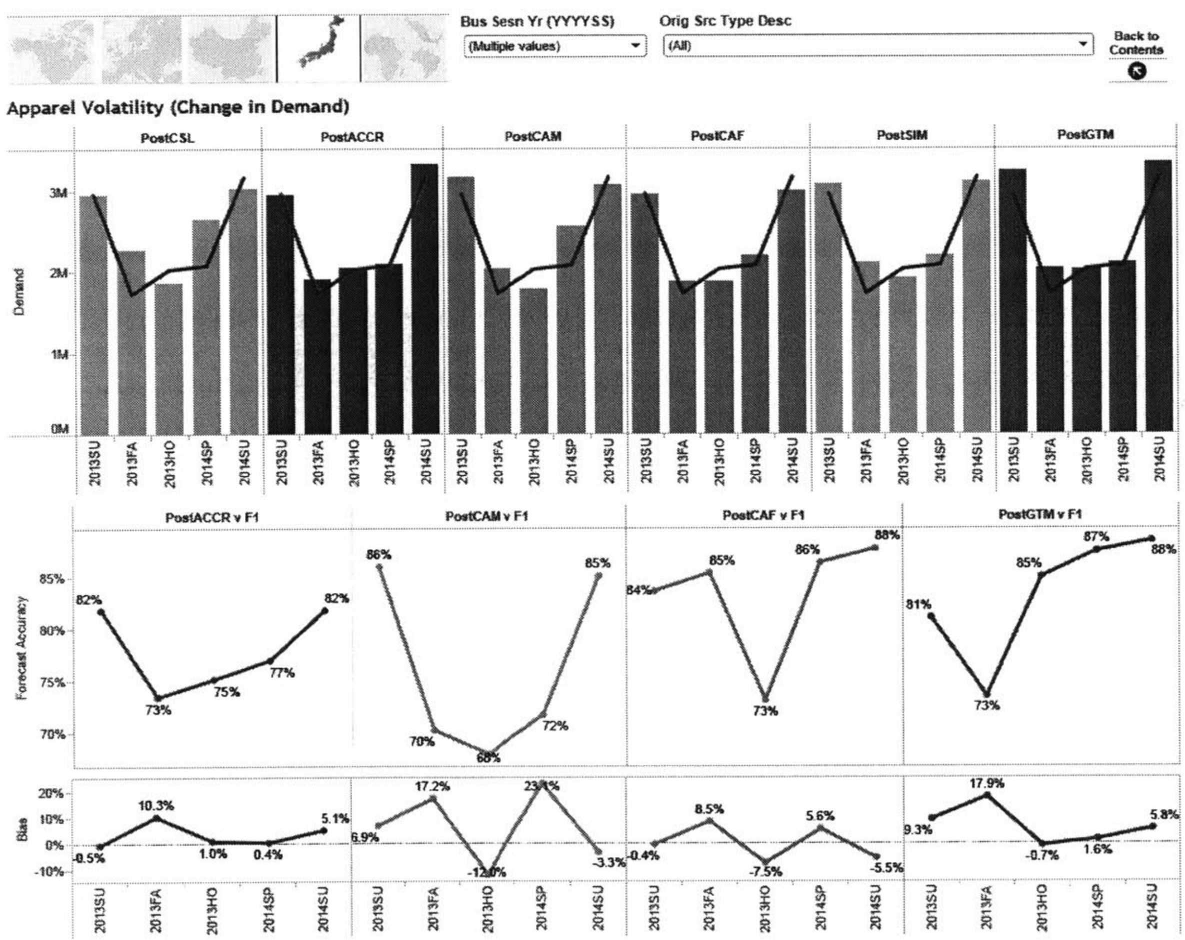

Figure 17: Standard View for Across Seasons – Japan filter example

Many demand planners and demand planning managers found the Across Seasons view extremely useful as it allowed them to see seasonal trends in demand and accuracy. They cited the need to have visibility across at least 5 consecutive seasons so that the year-over-year view would also be available.

In addition, this view allows one to have visibility across gates for multiple seasons – one can see that in Figure 17 above, while forecast accuracy increases across gates for the Summer 2014 season, it does not do so consistently for the Holiday 2013 season (dropping from 75% at PostACCR to 68% at PostCAM and then steadily creeping up). This type of visual analysis quickly provides the demand planner with insight into an area to dig in deeper to understand the causes.

In the long-range, the Across Seasons graph would also allow planners to measure the impact of an initiative to improve forecast accuracy. For example, if in a particular Category within a particular Geography there was an initiative to partner with Merchandising to reduce the number of product line adds and drops, all else being equal the effect should be visible as an increase in forecast accuracy starting in the season in which the change occurred.

Detailed View

The following graph shown in Figure 18 is an example of another view that is prevalent in the dashboard, across the sets of views (Leadership, Apparel, Footwear, and Geography views).

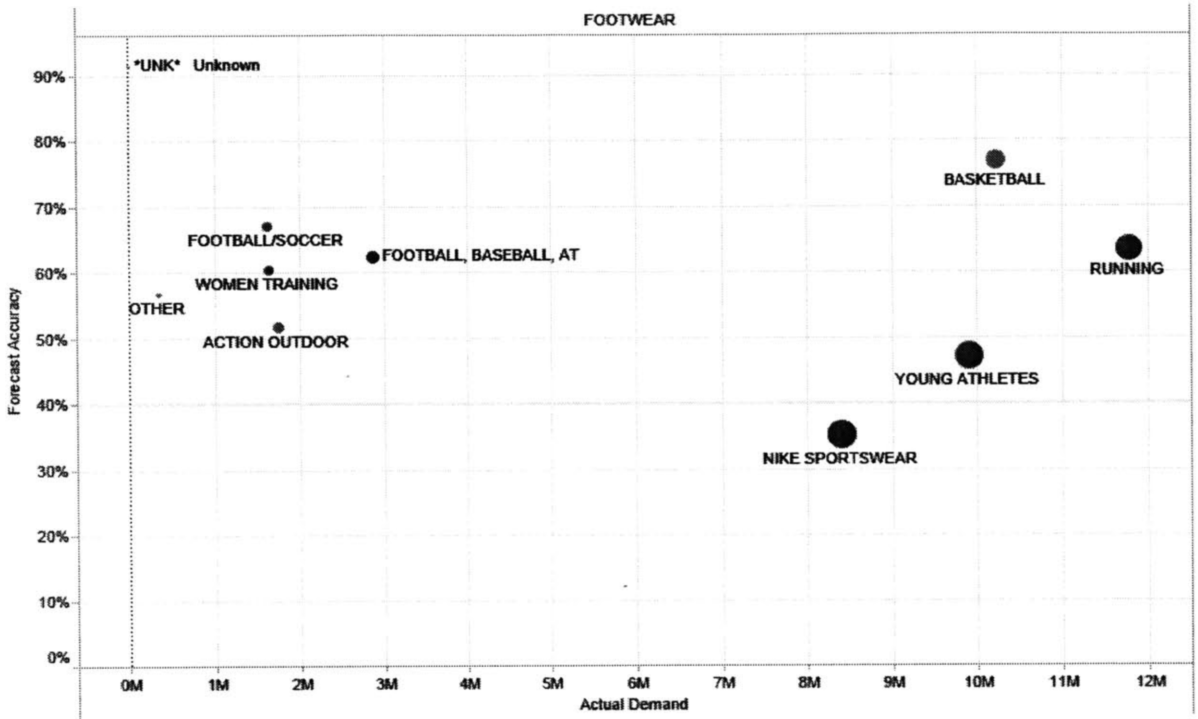

Figure 18: Detailed Accuracy vs. Demand Bubble Chart

This graph shows the forecast accuracy and actual demand for each category of product. The size of the bubble represents the absolute error for that particular Category. For example, we can see that the Nike Sportswear Category of product (the brown circle) has an approximately 34% forecast accuracy, an approximate 8.4M in actual demand, and a relatively large error (would need to mouse over for the exact error number). The level of aggregation at which forecast accuracy is calculated would be indicated by the title (not shown here) and the filters – here, forecast accuracy is calculated at the Geography-Category-Product Engine level, meaning that these forecast accuracy numbers reflect the weighted average accuracy across Geographies, broken out by Category and by Product Engine (Figure 16 shows for Footwear Product Engine only).

In interpreting this graph, one would expect to see the higher volume Categories with higher accuracy, as higher volume products are in general more stable and easier to forecast. Thus, one hopes to see all the bubbles towards the upper left corner of the graph. The size of the bubbles, the absolute error, is of course correlated with the position on the x-axis, the actual demand.

The graph allows demand planning managers and senior leadership to quickly see forecast accuracy metrics across Categories. This graph was not a newly developed view of the dashboard, but was taken from existing reports that the demand planners currently utilized in their review processes. In fact, while many of the views are newly developed, many demand planners wanted to see in the dashboard certain graphs or charts that they were already familiar with and that were already integrated with their review processes. Including these views in the dashboard increases the adoption rate and helps ensure that the dashboard will be utilized.

5.3 Dashboard Usage

As mentioned in the previous section, each group (Senior Leadership, Apparel, Footwear, and Geographies) has a list of custom views. The intent is for demand planners to begin with those sets of views to get a high-level understanding of the forecast accuracy metrics for that season, and then conduct further analysis as needed (either utilizing the dashboard or through another tool, such as running ad hoc reports). One benefit of having all the forecast accuracy metrics in one dashboard is the opportunity for demand planners in one area to explore the views in another area (e.g. for demand planners in a particular Product Engine to look at the metrics by Geography), and to have cross-functional discussions on improving forecast accuracy.

Because the dashboard is meant to replace the Global Scorecard, it only provides a high-level, global overview of forecast accuracy metrics and does not support customized views across Geographies or Product Engines. Thus, it does not replace specialized dashboards or is able to provide an in-depth analysis of forecast accuracy metrics. Instead, it provides a unified view and a common platform for discussion.

5.4 Visualization

Overall Design

The forecast accuracy dashboard is unique from many other Tableau dashboards at Nike for a couple of reasons: 1) the calculation of forecast accuracy requires the ability to see accuracy at many different levels of aggregation, requiring a different view for each and thus implying the need for a large and extensive dashboard and 2) the need for the stakeholders to slice and dice the data in many different ways means that the dashboard is, for many views, a tool to customize your own view (based on which Category, Geography, and specific filters you would like to include) rather than a dashboard on which the user can gain additional information by clicking on relatively static items.

Thus, making the dashboard visually appealing is a challenge. To the first point above, while many dashboards in Tableau only contain 4-5 views (with several different charts each), this dashboard must support 3-4 different levels of aggregation for each view, for each group (to reiterate here, examples of views are "Across Gates" or "Across Seasons" and groups are "Executive," "Apparel", "Footwear," and "Geographies"). Thus, this dashboard supports around 40 different pages to navigate to, justifying the need for the Cover Page to include a Table of Contents and leading to an extensive navigation system. To the second point above, although good design and usability are often linked, in many cases simple tasks such as aligning graphs to be more intuitive and user friendly to or managing real estate on specific views are difficult when the same view must be used by Europe, which currently in our dataset only has one SubRegion versus Emerging Markets, which has nine.

Visualization of the Standard View

Returning to the standard view shown in Figure 16, demand is shown as a bar chart, while accuracy and bias are shown as line charts. Although at Nike Demand is oftentimes visualized also as a line graph in various reports, it was deemed most appropriate to show forecasted demand using bars to accommodate charts where it may be important to show product breakdown (through a stacked bar chart, for example) and to provide contrast with the F1 actual target, which is represented as a horizontal gray line across all the relevant gates. In charts where demand is visualized as a line graph, the x-axis usually includes both the forecasted and actual demand gates – thus, the user must compare across the graph left-to-right rather than below-above.

Accuracy and bias, on the other hand, are presented as line charts. Unlike demand, which is a total or an accumulation of the demand of many products, the forecast accuracy and bias figures reflect one number, a percentage difference. Thus, the charts are represented as points connected by lines – the points show the accuracy or bias number at that particular gate and the line helps the demand planner visually see whether the trend is increasing or decreasing.

Many demand planners expressed the need to be able to compare accuracy and bias figures to last year's figures on the same chart (a year-over-year comparison for the same season). Because targets are currently not well defined, especially at a detailed level (for example Geography-Product Engine-Category), demand planners often look to last year's accuracy and bias figures to assess whether their numbers have improved. The last year's figures are presented in gray as to not dominate the chart.

A couple of demand planners also requested the year-over-year view shown in circles in Figure 19 (the lines not in circles represent the standard view). In this view, the line connects the accuracy number from the same season last year to this year's at the same gate. However, most demand planners felt that the standard view's method of connecting the accuracy number across gates for the same year was more intuitive, so this requested change was not adopted for the dashboard.

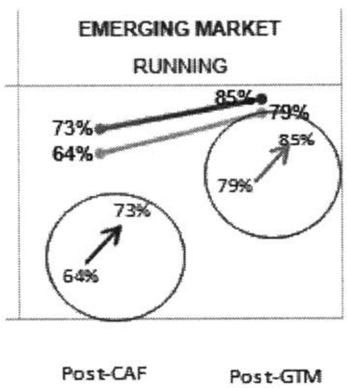

Figure 19: Alternate Year-Over-Year View

Color

The use of color was a difficult choice. Because red and green are traditionally associated with targets, which are not incorporated into the dashboard, demand planners expressed a preference to omit those colors to avoid confusion. Ultimately colors were selected to match those in the CGP – long-range metrics are presented in various tones of orange while short-range metrics are presented in various tones of blue.

Labels

Originally, in favor of a clean dashboard design, measure value numbers (e.g. numbers on the points for forecast accuracy) were omitted. However, some planners emphasized that they would be likely not only to use the dashboard on a computer (where mouseover functionality is enabled) but also to print out specific views to bring to meetings. Thus, they requested that all accuracy and bias charts include clear labels so that planners do not have to estimate/interpret the numbers along the line from the y-axis.

Examples of Omitted Chart Designs

Throughout the development phases of the dashboard, many charts were explored and ultimately rejected. In particular, during the first phase of dashboard feedback, a wide variety of charts (pareto charts, box and whisker plots, weighted bar graphs) were explored and evaluated by demand planners.

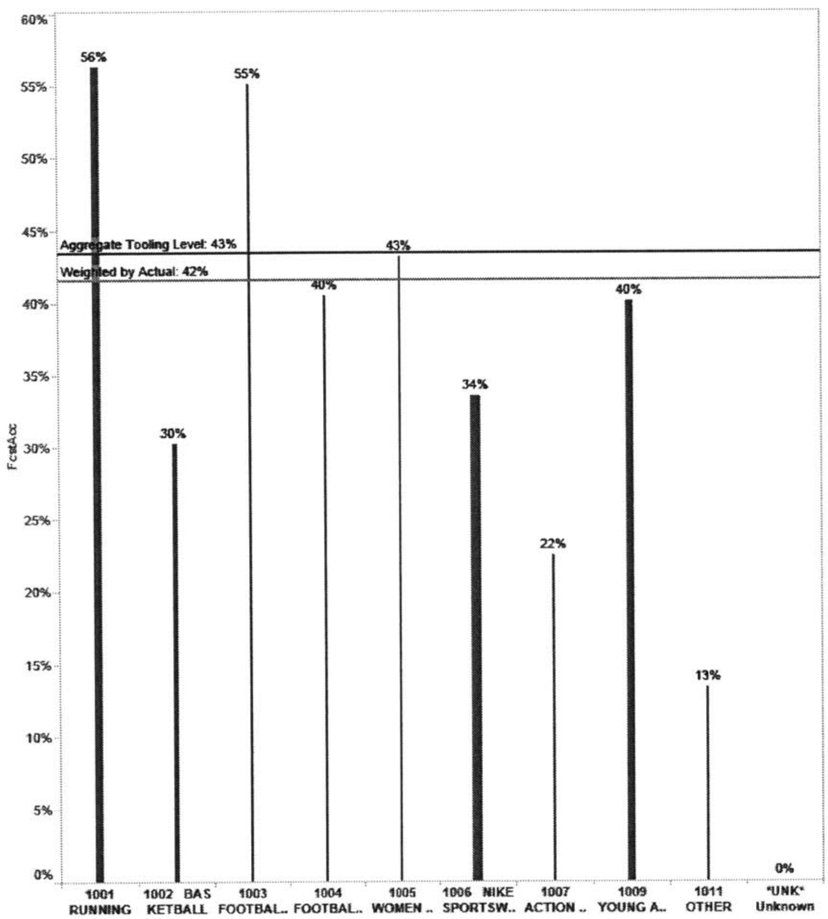

Figure 20: Weighted Bar Chart Example

The weighted bar chart in Figure 20 shows an example of a potential view to display forecast accuracy across categories. In this chart, the thickness of the bars indicates the forecasted demand while the y-axis shows accuracy. The gray reference line shows a weighted average (42%) forecast accuracy across all categories, while the red reference line shows the forecast accuracy measured at a more aggregate (non-category) level.

Although Figure 20 provides a significant amount of information in one chart, demand planners surveyed expressed the opinion that the chart is not immediately clear to the first-time user (it attempts to present too much information at once). In addition, while the comparison of accuracy at a different level of aggregation is useful to see via the red reference line, it is somewhat confusing to planners who are used to seeing charts where all metrics are at the same level of aggregation.

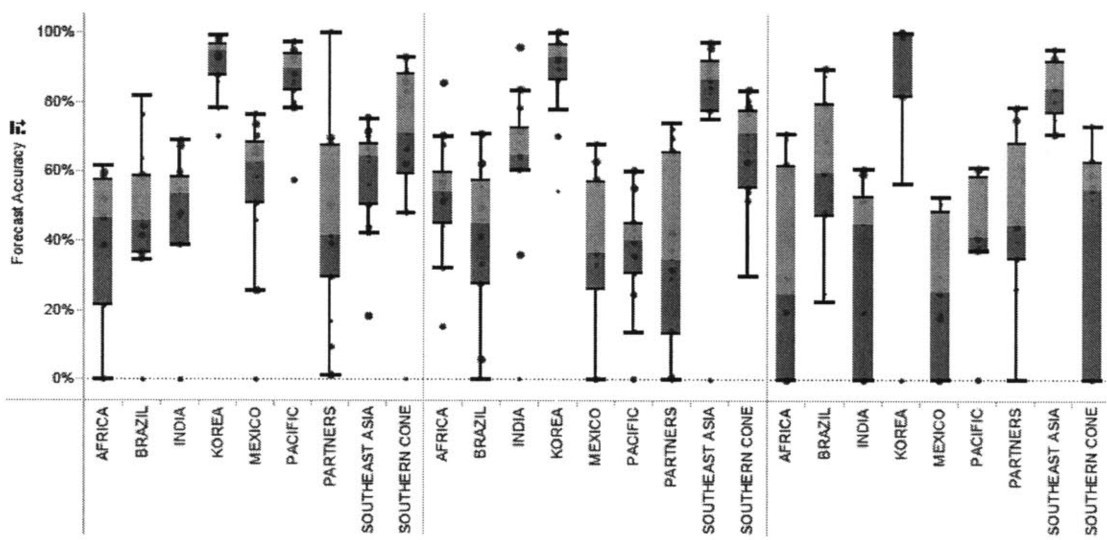

Figure 21: Box and Whisker Plot Example
Various Panels represent different Product Engines (Apparel, Footwear, Equipment), not shown here

Another chart that was explored was the Box and Whisker plot shown in Figure 21. This chart received the most negative reviews, as planners were either unsure how to interpret the box and whiskers or felt that those quartiles were irrelevant. Many planners felt that it was not important to understand which Categories fell in the upper quartile for a particular SubRegion; it was more important to understand which Category had what level of accuracy.

6 Dashboard Results

6.1 Final Dashboard Assumptions

Section 4.3 Initial Dashboard Assumptions presented the following set of initial assumptions for the dashboard design:

1) Incorporate the forecast accuracy metric
2) Measure forecasts at PostACCR, PostCAM, PostCAF, and PostGTM gates
3) Measure actuals at F1 and F4
4) Measure forecast accuracy at Project Code level for Geographies (short-term), at Project Classification Type (PCT) level for Apparel and at Tooling level for Footwear (long-term)

The following list presents the set of final assumptions after incorporating the data into the dashboard and receiving feedback from various groups on how the dashboard would be used:

1) Include forecast accuracy, bias, and show volatility through change in demand
2) Measure forecast accuracy and bias at PostACCR, PostCAM, PostCAF, and PostGTM but show change in demand (volatility) at PostCSL, PostACCR, PostCAM, PostCAF, PostSIM, and PostGTM (all major gates)
3) Measure actuals at F1 and F4

4) Measure forecast accuracy at Project Code level for Geographies (short-term), at Project Classification Type (PCT) level for Apparel and at Primary Platform Group level for Footwear (long-term)

The following sections explore the changes between the initial and final assumptions.

Metrics Included

Initially, the dashboard only incorporated the forecast accuracy metric. However, demand planning analysts and managers stressed that forecast accuracy is not meaningful without the related demand/volume information – higher volume products need to have and should have higher accuracy, whereas lower volume products are harder to forecast and are not as important. In addition, demand planners noted the importance of understanding whether the forecasts were above or below the actual demand (the bias metric), as this has a key impact on supply and material planning decisions. Thus, demand and bias were incorporated in the dashboard.

Demand planners, especially senior planning directors, also emphasized the need to view forecast change across gates in the CGP (the volatility metric). Upon further discussions, however, it became clear that various groups used different definitions of volatility. Because of this discrepancy, the team underwent a term alignment process to choose between three ways of viewing volatility: 1) viewing demand over various gates in the CGP and noting the change as volatility (maintaining the y-axis as demand), 2) calculating volatility as the absolute percent variance of the forecast at one gate as compared to the forecast at the previous gate, or 3) calculating volatility as the ratio of standard deviation to the mean. Of these three methods, the first one was adopted as it was the most basic and clearest way of expressing volatility while also allowing the user to easily visualize demand.

Thus, the dashboard includes three key metrics: forecast accuracy, bias, and volatility as shown through change in demand. The need to display these three metrics results in the standard view shown in Figure 16.

Gates to Measure at – Forecasted Demand

The team initially aligned to measure forecasts at four gates: PostACCR, PostCAM, PostCAF, and PostGTM. This alignment was primarily set to establish a standard across groups for measuring forecast accuracy and bias, and the current dashboard still shows these two metrics at these four gates. However, the need to view volatility meant that all significant gates should be included for the demand chart; thus, the PostCSL and PostSIM gates were added for this chart only.

Gates to Measure at – Actual Demand

The team also initially aligned to measure actual demand at F1 and F4. This decision was outlined in 4.3 Initial Dashboard Assumptions and did not change through the rounds of feedback sessions.

The dashboard initially excluded Nike Golf, as the Golf business is run separately and demand planning in Golf needs to view forecast accuracy by comparing the actuals at the End-of-Season. While some discussion occurred around incorporating Golf into the Forecast Accuracy Dashboard, ultimately it was concluded that for now Golf would be excluded in favor of keeping the actuals only at F1 and F4.

Level of Aggregation

For the Geography and Apparel views, the initial assumptions to measure forecast accuracy at project code level and at PCT level for Apparel were kept. Project Code aligned closely to style, which is the level at which demand planners in the Geographies plan (style and style-color). In Apparel, each article of apparel corresponds to one distinct PCT which allows planners to make broad judgments and conclusions about the classification type.

However, for Footwear, it readily became apparent that Tooling was not a good initial choice for the level of aggregation. While Apparel has approximately 55 different PCTs, Footwear has approximately 900 different Tooling codes, meaning that planners could not easily make conclusions about various types of shoes. In addition, the Tooling codes had relatively obscure descriptions (e.g. 1411 or OS394120-1), making it hard to interpret. To complicate matters, as mentioned in Section 4.4 Data Mapping and Sources, the Footwear Product Engine primarily uses data from DataMart for internal reporting. Thus, many fields used by Footwear for other purposes could not be used for forecast accuracy metrics because they do not currently exist in Teradata.

After much discussion with the Footwear Product Engine, the team came to a decision to align on measuring forecast accuracy metrics at the Primary Platform Group level (of which there are 15 different types). Although this field only represents 50% of all Footwear products, at this time it is the best workaround.

Although the team aligned to measure forecast accuracy metrics at these levels for the Geographies, the Apparel Product Engine, and the Footwear Product Engine, it is important to understand that the level of aggregation at which forecast accuracy or bias is being measured at depends also on the fields and filters present. For example, while at the Geography level accuracy and bias are always measured at the Project Code level, the numbers will differ between the Category view and the Core Focus view because the Category view is more aggregate (Core Focus is a subset of Category). If we look at accuracy figures for the Basketball Category and compare those numbers to the Basketball Core Focus and the Jordan Core Focus, we will see that at the Category level the figures are higher than at the Core Focus level – this is because if the Basketball Core Focus had under-forecasted and the Jordan Core Focus had over-forecasted, these differences would have been canceled out at the more aggregate Category level. Thus, even though we are measuring at the Project Code level, we are essentially looking at one set of metrics at the Project Code-Category level and one set of metrics at the Project Code-Core Focus level.

On top of this, if we choose to add filters such as by SubRegion (a subset of Geography), by Product Engine, or by AA Filter, we may be looking at accuracy not at the Project Code-Category level anymore, but at the even more disaggregate level of Project Code-Category-Sub Region-Product Engine-AA. Because forecast accuracy and bias are essentially row-level calculations (the error is computed for each row of the data), additional fields that break up the data into a more disaggregate form pushes the user to view these metrics at a more disaggregate level.

In the dashboard, the level at which forecast accuracy and bias metrics are calculated is indicated by the filters and fields available. Thus, if the user is able to filter by Category-Sub Region-Product Engine-AA, then the metrics are calculated at that level. Views in which the level of aggregation may be ambiguous were clearly labeled.

7 Root Cause, Forecast Inaccuracy, and Target Setting

As mentioned in Section 4.6 Elements of the Future State, the dashboard serves as a tool to measure forecast accuracy metrics and visualize areas for root cause analysis. This section will discuss in greater detail the steps that follow: root cause analysis, forecast inaccuracy impact determination, and target setting practices.

7.1 Root Cause Current State, Process, and Future State

Currently the Global Product Engines, Footwear and Apparel, do not employ any root cause analysis to their forecast accuracy metrics. The Geographies have differing processes and varying root cause buckets/categorization. At present, the Japan and Emerging Markets geographies do not have a defined root cause process, but Greater China, Europe, and North America do.

Although these three Geographies have different processes, the underlying process is relatively similar. Demand planners in each group look at the top 5 or 10 styles with the highest absolute error in each Category-Product Engine. They then deep dive into these styles to determine the root cause and classify each style's error to a root cause bucket. The largest buckets are often the same between Geographies: late style adds/drops, lack of information from Sales, or lack of information due to a new style. However, these three Geographies also have many root cause buckets that are specific to their group.

The future state of root cause analysis is one where Geographies are using the same processes and root cause buckets. To the extent that it makes sense, the Product Engines should also aim to be consistent. To achieve this, the three Geographies that currently have root cause processes in place should align in the following manner: first agree upon common root cause buckets to use and second align processes to support these buckets. Following this, the standard practice can be rolled out to the Geographies that currently do not do any root cause analysis and any relevant learnings can be applied to the Product Engines as well.

One additional caveat is that root cause analysis is likely to be different when comparing forecast accuracy at different gates. For example, a root cause analysis on the forecast accuracy of PostGTM vs. F1 will need different root cause buckets than one on the forecast accuracy of PostCAF vs. F1, because the causes of forecast inaccuracy will be different at those two gates. Thus, the Geographies will need to create two sets of root cause buckets (currently root cause analysis is only done comparing PostGTM forecast accuracy metrics).

7.2 Forecast Inaccuracy Costs Long-Term Vision and Recommendation

Determining the impact of forecast inaccuracy is inherently a difficult task. Although forecasts serve as an input to many functions such as supply planning and sourcing, assigning a cost is not easy because of the varied impacts of forecast inaccuracy.

Forecasts can be inaccurate in the following detrimental ways: forecasts are consistently higher than actuals (consistently here means across gates in a given season), lower than actuals, or change in between gates after certain key decisions (for example, capacity planning) have already been made. The following lists the impacts of increased demand, decreased demand, and changes in demand (demand swings).

<u>Increased Demand</u>

1) Increased tooling in the factories, resulting in reduced category margins

2) Late delivery, resulting in possibly air freighting products to the geographies to deliver on time
3) Lost sales, if Nike does not allow customers to place an order because it cannot be produced in time

<u>Decreased Demand</u>

1) Decreased production/factory utilization, resulting in idle workers which could increase in increased FOB (freight on board) cost in future seasons

<u>Changes in Demand</u>

1) Transfers in production quantity and locations, which can result in additional tooling and late deliveries
2) Excess materials as transferring materials from country to country is difficult

However, these inaccuracies do not always have the same impact; oftentimes they are absorbed. For example, consider an increase in demand (the original forecast was too low). There are a few possible outcomes: demand is absorbed by the original factories (no impact but factories are above the ideal utilization), demand must be transferred to other factories (factory rebalancing, which can result in no additional cost or could result in transfer costs and possible delays), demand cannot be fulfilled because of material unavailability (resulting in lost sales). In these three scenarios, only the transfer costs and delays are relatively quantifiable. Lost sales are difficult to quantify because in many cases Nike negotiates with the customer to take another product. This is possible because Nike has strong market power and this practice is known as demand shaping.

However, it is difficult to tease out the effect of forecast inaccuracy on factory rebalancing and delays because these events happen for other reasons as well. Those reasons could be Nike related or could be macro events (political unrest, strikes). There are currently no metrics within sourcing or supply planning that track or tag events that happen solely due to forecast inaccuracy, thus making it difficult to separate out those costs.

Despite these challenges, it is critical for Nike to understand the cost of forecast inaccuracy. Without this crucial piece of information, demand planning groups cannot set targets for what an acceptable forecast accuracy should be for a particular Geography-Category – whether the target should be 80% for North America Cotton Tees, for example, or 70%.

Thus, a recommendation would be for the organization to investigate how to track the impacts of inaccuracy. This should start with information mapping, first understanding the current processes in a detailed fashion and identifying specific areas where forecast inaccuracy impacts costs but those costs are not being captured as such. For example, this could entail looking in detail at the process that occurs when a factory is overloaded because of an increase in demand and seeing what metrics are captured or how costs are flagged when tooling changes. After understanding the processes and where the gaps are in capturing costs that relate to forecast accuracy, one can then begin to make recommendations on new metrics to add (and possibly process changes to support these new metrics).

In addition, once the metrics are in place, analysis can be done on historical data to answer questions such as "what percentage of time does an increase in demand result in tooling adds? In air freighting?" Nike may determine, for a particular type of product, that 50% of the time increases in demand are absorbed, but 20% of time tooling adds are necessary, 20% of the time air freighting is necessary, and 10% of the time the customer does not receive the product they ordered. Then, for this particular type of product, one can determine an approximate expected cost for the demand increase.

Furthermore, the cost of inaccuracy is not necessarily linear to the forecast inaccuracy. That is to say, if a decrease in forecast inaccuracy from 85% to 80% results in an expected extra cost of $0.25 per shoe, a decrease in forecast inaccuracy from 80% to 75% may result in a larger expected extra cost, say $0.50 per shoe (the numbers used here are entirely made up for this example). Thus, with the right data and metrics, analysis can be done to determine how forecast inaccuracy over various thresholds has related to cost historically, and the predictive power of this type of analysis (whether the historical analysis is indicative of what may happen in the future, whether it captures trends that are inherent to the system). This type of analysis will be incredibly informative for target setting.

This first approach assumes measurement of forecast inaccuracy costs by the resulting reactive actions taken to correct for forecast inaccuracy. Alternatively, a different approach is to plan separately for products with differing forecast accuracy levels. If, based on historical forecast accuracy data, we can determine that the demand of some products is relatively stable (high accuracy) and the demand of others is not (low accuracy), then we could potentially plan sourcing and inventory through varying processes with different associated costs. The cost of forecast inaccuracy could then be determined by comparing how plans would change for differing assumptions of forecast accuracy, for example how much more or less inventory or capacity is needed for similar products that have different historical accuracy figures. The proactive planning approach (based on historical forecast accuracy of similar products) is likely to be more cost-efficient than the reactive planning approach (responding to forecast inaccuracy).

7.3 Target Setting Analysis and Feedback

Forecast accuracy targets should be set based on the level of forecast aggregation (how detailed or aggregated the forecasts in question are) as well as the costs associated with forecasts being off (i.e. the importance of those forecasts). However, currently the second is not well measured or known.

Currently, targets are set depending on the level of aggregation and benchmarking to previous years' forecast accuracy numbers (as a measure of what can be achieved or attained). While some have suggested targets with yearly improvements of accuracy, such as 1% or 3%, it is unclear how this would be sustainable (realizable) or how it would level off at a certain level (and at which level?). Thus, key opportunities in this area are to better understand (a) the costs of forecast inaccuracy (and separately, the cost relating to forecasts being off or inaccurate and the cost of volatile forecasts) and (b) using the costs of forecast inaccuracy to better inform target setting, how these targets would compare to targets set with previous years' forecast accuracy numbers.

There is indeed a question of whether forecasts can be improved upon at all. In some years (for a given Geography-Category), forecasts have certainly been more accurate than others. However, it is not clear that anything was done differently in those years from the forecasting side – the actuals may simply have matched the predicted orders better. Thus, the implementation of the Tableau dashboard, which will allow better/more visible tracking over time, will allow managers to ascertain whether forecast accuracy improves from a baseline given certain initiatives that Nike is considering using to improve the forecasting practice (more statistical forecasting for stable, wholesale product lines, for instance). For some products, how

Nike forecasts now may simply be the most accurate that product can be, even with improved methods. Thus, while there is little data or evidence to show that Nike can improve upon their current forecasts, the general thought is that there is room for improvement and this can only be proven (or not) with improved forecast accuracy tracking/visibility along with varying forecasting practices.

8 Implementation Strategy

At the end of this project, the dashboard was presented to the planning directors and a number of key leaders within Nike's planning organization. Certain views from the dashboard were selected for review in the quarterly planning director's review process, meaning that the updated, aligned views from the dashboard tool would serve as the metrics against which forecasts would be judged.

This project (initial dashboard design and implementation into the review process) was deemed Phase I of the dashboard development and a Phase II was started which will involve usage of the dashboard (automated updating of future seasons along with quarterly reviews, as well as continuous improvements to the dashboard) and process alignment of root cause analysis across different groups, which will ultimately be incorporated into the dashboard.

9 Conclusions and Recommendations

Improving demand planning forecast accuracy is a long process in any company, in any industry. Because it is not the most urgent nor the most interesting of tasks, it can oftentimes be overlooked. However, measurable improvements to forecasts can have enormous financial impacts throughout the supply chain and to the bottom line.

This project focused on the measurement of forecast accuracy metrics, specifically on the alignment of metrics across different demand planning groups and the use of a dashboard to view these metrics consistently across groups and with senior leadership. The project established a framework regarding the long-term vision for improving forecast accuracy, and how measurement and alignment are critical, initial steps, to this process.

There are several next steps to this project: refine the dashboard (now live and in usage) through continuous feedback, align on root cause analysis and incorporate into the dashboard, quantify costs of forecast inaccuracy, and use the root cause analysis and inaccuracy cost analysis to determine how to set targets for forecast accuracy. While these steps should initially be covered in this order, this is not an inherently linear process and some iterations and refinements of certain steps should be expected.

Within these next steps, there are a couple of distinct possible follow-on projects. The most important of these would be a project to quantify the costs of forecast inaccuracy, on the sourcing side. This would involve working closely with manufacturing and sourcing and coming up with new metrics that would track specifically the costs associated with a revision in forecasts or with forecasts that are extremely off from the actual bookings. Another similar project would be to quantify the costs of forecast inaccuracy on the demand side (understanding how the retailers' product needs may not be met or quantifying the compromises they must make due to inaccurate forecasts).

Another interesting topic would be to quantify the costs of a potential proactive planning supply chain (designed in such a way to proactively plan for products with varying forecast accuracy) as compared to current costs and to costs of an even more flexible or reactive supply chain.

These projects are suggested over projects to improve forecast accuracy directly. Often, there can be an excessive focus on forecasting within an organization, as forecasts are the starting point and drive other major assumptions. Instead, the idea is to shift the primary focus from improving the forecast to instead minimizing the cost implications related to inaccurate forecasts and to better understand the costs, risks, and tradeoffs that forecasts are part of.

Finally, in the market that Nike finds itself in – one with long lead times, continuously new products, and a large number of SKUs – improving forecast accuracy will only bring the organization so far. This initiative must also be combined with other strategic supply chain initiatives, such as reducing lead time/time in the supply chain for products, introducing postponement to a greater number of products (already done for quick turn jerseys, for example), reducing the number of SKUs (especially those with low volume and low margins), etc. Many of these initiatives are not so straightforward either in the area where supply chain strategy meets business strategy. For example, while it may seem compelling to reduce SKU complexity by eliminating low volume SKUs, product variety may be a strong attracter of customers (as is often the case in the fashion/retail industry).

Works Cited

Armstrong, J. S. (2001). Standards and Practices for Forecasting.

Axline, J. E., & Lebl, B. J. (2007). Leveraging Downstream Data in the Footwear / Apparel Industry by, 1–65.

Bunn, D. W., & Taylor, J. W. (2001). Setting accuracy targets for short-term judgemental sales forecasting. *International Journal of Forecasting*, *17*(2), 159–169. doi:10.1016/S0169-2070(00)00090-X

Chase, C. W., Jr. (2013). Demand-Driven Forecasting.

Chiang, A. (n.d.). Creating Dashboards : The Players and Collaboration You Need for a Successful Project, 14(1), 59-64.

Dagan, B. (2007). Dashboards and Scorecards Aid in Performance Management and Monitoring, (September), 23-28. Doi:10.1002/gas

Hoover, J. (2009). "How To Track Forecast Accuracy To Guide Forecast Process Improvement". *Foresight*, *summer*(14), 17–24.

Johnson, K., Hong, B., Lee, A., & Simchi-Levi, D. (2013). Analytics for an Online Retailer : Demand Forecasting and Price Optimization, (2012), 1–34.

Kolassa, S. (2008). Can we obtain valid benchmarks from published surveys of forecast accuracy? Foresight, (11).

Lapide, B. Y. L. (2013). Demand-shaping, (November), 4–6.

Mentzer, J.T., Moon, M. A. (2004). Managing the Sales Forecasting Process, 1–42.

Nederpel, L., Operations, I., & Nike, S. (2012). Improving the close-out supply forecast accuracy at Nike Inc ., (February), 1–71.

Nisen, M. 2013. "How Nike Solved Its Sweatshop Problem." Business Insider. May 9. http://www.businessinsider.com/how-nike-solved-its-sweatshop-problem-2013-5

Poirier, C. C. (n.d.). Forecasting, Demand Management, and Capacity Planning, 1–18.

Promise, T. (2013). The Promise and Pitfalls of Big Data, (August), 4–6.

Smith, V.S. (2013). Data Dashboard as Evaluation and Research Communication Tool, (140), 21-45. Doi:10.1002/ev