Comparison and Financial Assessment of Demand Forecasting Methodologies for Seasonal CPGs

by

Burak Gundogdu, BS, MBA

and

Jeffrey Maloney, BS

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2019

Signature of Author: ____________________________________
Department of Supply Chain Management
May 10, 2019

Signature of Author: ____________________________________
Department of Supply Chain Management
May 10, 2019

Certified by: ____________________________________
Tugba Efendigil
Research Scientist
Capstone Advisor

Accepted by: ____________________________________
Dr. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Comparison and Financial Assessment of Demand Forecasting Methodologies for Seasonal CPGs

by
Burak Gundogdu
and
Jeffrey Maloney

Submitted to the Program in Supply Chain Management
on May 10, 2019 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Forecast accuracy is an ongoing challenge for made-to-stock companies. For highly seasonal fast-moving consumer packaged goods (CPGs) companies like King's Hawaiian, an improved forecast accuracy can have significant financial benefits. Traditional time series forecasting methods are quick to build and simple to run, but with the proliferation of available data and decreasing cost of computational power, time series' position as the most cost-effective demand forecasting method is now in question. Machine learning demand forecasting is increasingly offered as an improved alternative to traditional statistical techniques, but can this advanced analytical approach deliver more value than the cost to implement and maintain? To answer this question, we created a three-dimensional evaluation (cube search) across five unique models with varying pairs of hyper-parameters and eight different data sets with different features to identify the most accurate model. The selected model was then compared to the current statistical approach used at King's Hawaiian to determine not just the impact on forecast accuracy but the change in required safety stock. Our approach identified a machine learning model, trained on data that included features beyond the traditional data set, that resulted in a nearly 4% improvement in the annual forecast accuracy over the current statistical approach. The decrease in the value of the safety stock as a result of the lower forecast variation offsets the incremental costs of data and personnel required to run the more advanced model. The research demonstrates that a machine learning model can outperform traditional approaches for highly seasonal CPGs with sufficient cost savings to justify the implementation. Our research helps frame the financial implications associated with adopting advanced analytic techniques like machine learning. The benefits of this research extend beyond King's Hawaiian to companies with similar characteristics that are facing this decision.

Capstone Advisor: Tugba Efendigil
Title: Research Scientist

ACKNOWLEDGMENTS

We would like to thank the community of the Supply Chain Management program at MIT for giving us a once in a lifetime experience to work with and learn from such inspiring people. We are especially grateful to our advisor, Dr. Tugba Efendigil, for her leadership and insight throughout this project. This project would not have been possible without the support and participation of the sponsoring company, King's Hawaiian. A special thanks to the leadership at King's Hawaiian; Mark, John, Dan, Tony and Joe, whose vision and commitment to employee development and knowledge have made this project a reality. We would also like to thank the King's Hawaiian o'hana who have contributed to this project and graciously lent support when called upon: Nicolas, Bill, DG, Jordan, and Amelia.

Burak & Jeff

I am honored to have worked on this project with a partner as talented as Burak and thank him for his patience and the knowledge he has gifted me throughout this process. I am truly grateful to my friends and family who have supported me on this adventure. I'm sure Burak feels the same.

Jeff

I would like to thank my imaginary cat "Schroding-cat" for always being there, or not.

Burak

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation

For businesses that operate from a demand forecast, achieving a low forecast error can allow them to carry less safety stock and confers a significant financial advantage. Conversely, a high forecast error requires greater investment in safety stock inventory to cover demand variation. High levels of safety stock can become debilitating for small firms that do not possess adequate financial strength and may pose a barrier to entry into markets for new firms. A high forecast error can lead to missed sales and be detrimental to customer relationships and ultimately damage the brand.

High forecast errors also lead to supply chain inefficiencies. Companies can suffer from increasing logistics costs due to relocating product throughout the network and shipping product from a sub-optimal region for a higher cost. Forecast volatility causes a bullwhip effect upstream to manufacturing, which can result in operational inefficiencies through frequent adjustments to production plans. The higher volatility complicates upstream planning with vendors, resulting in inefficient order sizes and increased costs for expedited orders.

Improving forecast accuracy is a challenge that many companies face. It is especially important to consumer packaged goods (CPG) companies that are built-to-stock, as they must invest in finished goods to cover forecasted demand while minimizing their financial exposure. Build-to-stock companies that have strong seasonal demand must begin planning and building inventory earlier, and with less information, than similar non-seasonal CPG companies.

Advancements in forecasting software and analytics have provided modern businesses with many options. The promise of advanced forecasting methods is a more accurate forecast that will yield financial savings to the firm by reducing forecast error and uncertainty. However, advanced forecasting techniques require an investment in not just the software, but the increased amounts of data required to run the model. Additionally, the company must account for the cost of the increasingly sophisticated personnel responsible for initializing and maintaining an advanced demand forecasting solution. A company must select a forecast process that matches their demand pattern while being cognizant of the ongoing costs for forecast software and all supporting processes. The development of a framework that can narrow the selection process based on potential financial impact can help companies invest capital in a solution that is sized for their business model.

Many companies currently face just such a challenge. As a highly seasonal build-to-stock business, King's Hawaiian must rely on a forecast to build inventory sufficient to cover their peak demand and carry safety stock to ensure high customer service levels. Forecast accuracy becomes even more important when additional constraints such as capacity, obsolescence, and seasonality are considered. Forecast-driven companies like King's Hawaiian need to understand whether they can expect to see an appreciable decrease in forecast error from the adoption of a machine learning forecast. They also need to understand whether the potential savings, resulting from reduced safety stock due to a lower forecast error, will offset the increased costs of data and technical expertise.

1.2 Problem Statement

Many methodologies for demand forecasting are available to companies. Traditional methods include, but are not limited to, expert opinions, market research, and statistical methods (which are typically built in Excel) to better understand trends in the marketplace.

One alternative to traditional methods is applying machine learning to demand forecasting. Even though this is hardly a new idea, advanced machine learning techniques have only recently started to take their place in supply chain professionals' toolbox. That change can mainly be attributed to the recent rise of artificial intelligence/machine learning applications' popularity in supply chain management due to the potential profitability (Chui et al., 2018) and the increased ease of access and implementation of complex machine learning models with new techniques and software (e.g. more than fifty thousand Tensorflow repositories on Github as of January 2019).

The main question we will answer in this capstone project is whether the improvement (if any) in demand forecast accuracy from a machine learning process versus traditional statistical methods is significant enough to justify the increased costs.

While machine learning methods have been heavily researched, there is a lack of academic studies on the financial link between the benefit and costs of applying machine learning to demand forecasting. Therefore, we will demonstrate an approach that can help companies make cost-effective decisions when it comes to implementing a machine learning forecast, in addition to determining which machine learning methods provide the best results and what data to include in the models for companies similar to King's Hawaiian.

1.3 Summary of Methodology

The approach for this capstone is to identify, through current academic literature, existing machine learning methods used for demand forecasting. We compared these techniques to the current statistical method being used at King's Hawaiian.

Since machine learning models can leverage large amounts of information, data in addition to King's Hawaiian's historical shipments was incorporated. In addition to the shipment data out of the ERP system, third-party consumption data that are being purchased were also used. To augment the machine learning, publicly available data sets for socio-economic factors (e.g. income, education level) and severe weather events such as hurricanes or winter storms were also captured. The relative values of these additional data attributes were classified by machine learning to determine the relevancy of incorporating the information in the demand forecast.

Five different machine learning demand forecasting methodologies -- support vector regression, artificial neural network, random forest, gradient boosting, and k-nearest neighbors regression -- were run over multiple iterations with various attributes to generate many demand forecasts. The accuracies of these models were compared to the current statistical model and evaluated using standard forecast error metrics.

When a solution was found that generated a lower forecast error than the current statistical methodology, the change in projected safety stock as a result of the new forecast error was calculated. The resulting financial savings were compared against the incremental costs for the additional data, technical and personnel-related costs required to generate a forecast through more technologically advanced techniques. This comparison allowed for the evaluation of the net savings that can be expected from safety stock reduction, for a similar company, as a result of adopting a machine learning demand forecast.

1.4 Company Background

King's Hawaiian is a national Hawaiian Foods company with a primary focus on dinner rolls. A national producer of consumer goods, King's Hawaiian is a family owned and operated business that was started as a bakery and restaurant in Hilo, Hawaii in the 1950s. King's Hawaiian entered the mass-produced CPG stage in the 1970s with the opening of their first production facility in Torrance, California. Today King's Hawaiian operates on both coasts with three production facilities and a national sales footprint (King's Hawaiian Marketing Department, 2019).

King's Hawaiian is a highly seasonal consumer product, most popular during the Thanksgiving, Christmas, and Easter holiday seasons. They contract a network of frozen warehouses and carriers to store and deliver their product to national retailers and redistributors. As a build-to-stock company, King's Hawaiian relies heavily on its Sales and Operations Planning process to correctly align its production capabilities with the forecasted demand (King's Hawaiian Supply Chain Department, 2019).

To deliver on its brand promise of being irresistible, King's Hawaiian ensures a 99.5% service level by holding safety stock to cover the volatility in the demand forecast. As the company continues to grow, both in sales and unique offerings for customers, the utilization of their facilities increases. The above-mentioned factors increase the pressure for an accurate forecast to deliver product to its customers while minimizing the amount of capital invested as safety stock for the seasonal demand peaks.

2. LITERATURE REVIEW

2.1 Overview

Machine learning is a very broad topic and a very dynamic area of research where new techniques are continuously being developed. For that reason, the general academic approach in the studies we have examined is comparing a few select machine learning techniques to traditional approaches. We also came across some broader comparative studies, but they are limited in number and by no means include all possible machine learning approaches.

The general consensus is that machine learning based methods are promising in improving demand forecasts even if they do not necessarily perform better than traditional methods under every scenario. The challenge for demand forecasters is finding out which machine learning methods can perform better than the legacy approaches for the underlying demand of their specific industry and product, and how to make sure that the improvement in forecasts is significant enough. Even though some papers mention that applying machine learning is costlier than staying with traditional methods, e.g. by sharing computation times, we did not come across a study that implements a detailed cost-benefit analysis of using machine learning to improve supply chain demand forecasts.

We also noticed that there is no consistency in terms of terminology. A good example is the classification of linear regression as both a machine learning based technique and also a traditional method due to its wide usage both in academia and industry. In this section, we followed the academic authors' classification of methods.

2.2 Machine Learning in Time Series Forecasting

Artificial intelligence in general and machine learning specifically are highlighted as having significant potential to improve costs over traditional analytical techniques (Chui et al., 2018). The field of supply chain management and the consumer packaged goods industry are identified as among the best candidates for adoption of these techniques, projecting some of the largest cost savings as a result. The cases Chui et al. reviewed for demand forecasting that implemented artificial intelligence to identify underlying causal drivers show a 10-20 percent improvement over existing methodologies. While acknowledging challenges around these more advanced methods, they conclude that there is a clear argument for the value of augmenting existing analytical capabilities with machine learning and artificial intelligence.

Ahmed, Atiya, Gayar, & El-Shishiny (2010) applied multiple machine learning models to M3 time series competition data to compare the major models using approximately a thousand time series and argue that multilayer perceptron and the Gaussian process regression perform better than Bayesian neural networks, radial basis functions, generalized regression neural networks (kernel regression), k-nearest neighbors regression, CART regression trees and support vector regression.

Even though machine learning is considered a very promising field in improving time series forecasting, some studies argue that they will not consistently improve over the forecasts using traditional techniques. Makridakis, Spiliotis, and Assimakopoulos (2018) find that post-sample accuracy of traditional methods is better than that of machine learning methods for all forecasting horizons examined using 1045 M3 competition time series.

2.3 Studies in Traditional and Machine Learning Based Demand Forecasting

Many authors have applied machine learning to demand forecasting and compared forecast accuracy against the accuracy of traditional forecasting methods. Hribar, Potocnik, Silc, & Papa (2018) show that recurrent neural networks are more accurate than empirical models in forecasting natural gas demand, and they suggest implementing a linear regression model instead when computational resources are limited. Machine learning based demand forecasting using recurrent neural networks and support vector machines can show better performance than traditional techniques (naïve forecasting, trend, moving average, and linear regression), but it should be noted that improvement in forecast accuracy with those more complex models is not always more statistically significant than using linear regression (Carbonneau, Laframboise, & Vahidov, 2008). Saloux and Candanedo (2018) also argue that advanced machine learning approaches – decision trees, support vector machines, artificial neural networks – improve demand forecasts compared to traditional linear regression models.

Machine learning based demand forecasting can also be used to improve forecasts when demand is intermittent. Neural networks can perform better than Croston, moving average and single exponential smoothing when demand is irregular, as in spare parts supply chain (Amirkolaii, Baboli, Shahzad, & Tonadre, 2017). Vargas and Cortes (2017) claim that artificial neural networks perform well in the sample period, but they are less steady than ARIMA in post-sample period forecasting of irregular spare parts demand.

For a demand forecaster who agrees that machine learning based demand forecasting can perform better than traditional approaches, the next challenge is to pick the right machine learning technique. Therefore, part of the literature surveyed focuses on comparing machine learning methods with each other. Gaur, Goel, and Jain (2015) implement a simulation using Walmart shipment data set and a confusion matrix as a performance metric to compare nearest neighbors and Bayesian networks for demand forecasting in supply chain management and argue that the latter performs better. Guanghui (2012) implements support vector regression and radial basis function neural networks to forecast demand of weekly sales for a large paper enterprise and shows that support vector regression performs better. In a study comparing three machine learning approaches to forecast demand of bike-sharing service, linear regression performs better than neural networks and random forests, although using a longer time period than the one-month used in the study (i.e. more data) is suggested to understand real performance of more complex approaches. Combining multiple techniques to build ensemble or hybrid models is also common when implementing machine learning in general. Efendigil, Onut, and Kahraman (2009) use an adaptive neuro fuzzy inference system to utilize both neural networks and fuzzy modeling to get more accurate demand forecasts. Johansson et al. (2017) use ensembles of machine learning algorithms for operational demand forecasting and argue that extreme learning machines provide the best accuracy.

2.4 Studies Focusing on FMCG Demand Forecasting with Machine Learning

Forecasting demand for fast moving consumer goods (FMCG) is challenging due to reasons like shelf-life and seasonality. For the same reasons, this is an area where applying machine learning techniques can result in better forecasts as machine learning models can capture complex relationships. The studies we have analyzed regarding applying machine learning to demand forecasting of FMCG products show promising results. While a large selection of machine learning algorithms has been implemented, neural networks and support vector machines are among the common ones. Table 2.1 summarizes papers that generally focus on demand forecasting of different FMCG products and summarizes the methods and features used along with the conclusions.

Table 2.1: *Literature Overview. Review and summary of similar demand forecasting studies*

Title	Forecasted FMCG product / industry	Methods	Features	Conclusion
A Comparative Study of Machine Learning Frameworks for Demand Forecasting (Mupparaju, Soni, Gujela, Lanham, 2018)	Grocery store items	Moving average, gradient boosting, factorization machines, three versions of deep neural networks	Sales, date, store and item numbers, promotion, store location, categorization, clustering, family item and class, perishability	Neural network model with a sequence-to-sequence approach to forecast the sequence of 16-day sales using the sequence of previous 50-day sales as input performs best
A comparison of sales forecasting methods for a feed company: A case study (Demir and Akkas, 2018)	Five products of a feed company	Moving average, exponential smoothing, Holt's linear method, Winters's method, artificial neural networks, support vector regression	Product sales	Support vector regression produces the best results
A Comparison of Various Forecasting Methods for Autocorrelated Time Series (Kandananond, 2012)	Demand of six different products from a consumer product company (e.g. dishwasher liquid, detergent)	Artificial neural network (multilayer perceptrons (MLP) and radial basis function (RBF)), support vector machine (SVM), and autoregressive integrated moving average (ARIMA)	Historical demand at t-1, t-2,..., t-10	Support vector machines perform better than ARIMA and artificial neural networks
Artificial Neural Networks/or Demand Forecasting: Application Using Moroccan Supermarket Data (Slimani, El Farissi, Achchab, 2015)	Supermarket sales	Focuses on finding the optimal Multi Layer Perceptron (MLP) structure for demand forecasting	Daily order quantities	MLP with hidden layers 4, 4, 4 performs best
Forecasting Seasonal Footwear Demand Using Machine Learning (Liu and Fricke, 2018)	Footwear	Regression trees, random forests, k-nearest neighbors, linear regression and neural networks	Store count, month, lifecycle month, gender, AUR, year, basic material, MSRP, color, lifecycle, cut description, product class description	Ensemble methods (median and average) and random forests gave the best predictive performance
Oral-Care Goods Sales Forecasting Using Artificial Neural Network Model (Vhatkar and Dias, 2016)	Oral-care	Back-Propagation Neural Network Model	Past sales	Neural network models can be used for predicting the sales of FMCG products
Predictive Demand Models in the Food and Agriculture Sectors: An Analysis of the Current Models and Results of a Novel Approach using Machine Learning Techniques with Retail Scanner Data (Pezente, 2018)	Sugar commodity demand (utilizes products with sugar component)	Linear regression, ARIMA, artificial neural networks	Consumption, price, volume, GDP, population	Neural netrowk models were significantly more accurate
Support Vector Regression for Newspaper/Magazine Sales Forecasting (Yu, Qi, Zhao, 2013)	Newspaper/magazines	Traditional regression, support vector regression	Store type, occupation, education, age, gender, and income	Support vector regression performs better
Utilizing Artificial Neural Networks to Predict Demand for Weather-Sensitive Products at Retail Stores (Taghizadeh, 2017)	Weather-sensitive retail products	The multilayer perceptron, time delay neural networks, recurrent neural networks, bagging, linear regression	Sales of potentially weather-sensitive products, weather	The multilayer perceptron (MLP) with the back propagation learning algorithm performs best
Sales forecasting using extreme learning machine with applications in fashion retailing (Sun, Choi, Au, Yu, 2008)	Fashion products	Batch steepest descent backpropagation algorithm with an adaptive learning rate (GDA), the gradient descent momentum and adaptive learning ratio backpropagation (GDX), extreme learning machine (ELM) and extreme learning machine extension (ELME)	Sales data of one kind of fashion clothes, color, size, and price	ELM and its extension perform better

Research into the application of machine learning in the field of demand forecasting is still in its early stages and has shown mixed results. There is no one-size-fits-all approach when it comes to replacing traditional demand forecasting methods. The selection of the machine learning model, and the hyperparameters that guide it, play a significant role in the final forecast accuracy.

3. DATA AND METHODOLOGY

This section discusses the collection, preparation, and analysis of the various data sources required for the machine learning demand forecast. The goal was to selectively and iteratively evaluate relevant regional data to increase the predictive performance of a machine learning forecast, which required processing and aligning the data from internal and external data sources. Sources, features, and level of resolution of the different data sources were identified.

Also detailed is the preparation of the different data sources including aggregation, normalization, standardization, and feature selection using the random forest machine learning model, for input into the various machine learning forecast models. The different machine learning models (support vector regression, artificial neural network, random forest, gradient boosting, and k-nearest neighbors regression), hyper-parameters and feature sets evaluated are discussed, as well as the methodology for measuring their performance against the current Holt-Winters statistical forecast. The overall process is mapped out in Figure 3.1.

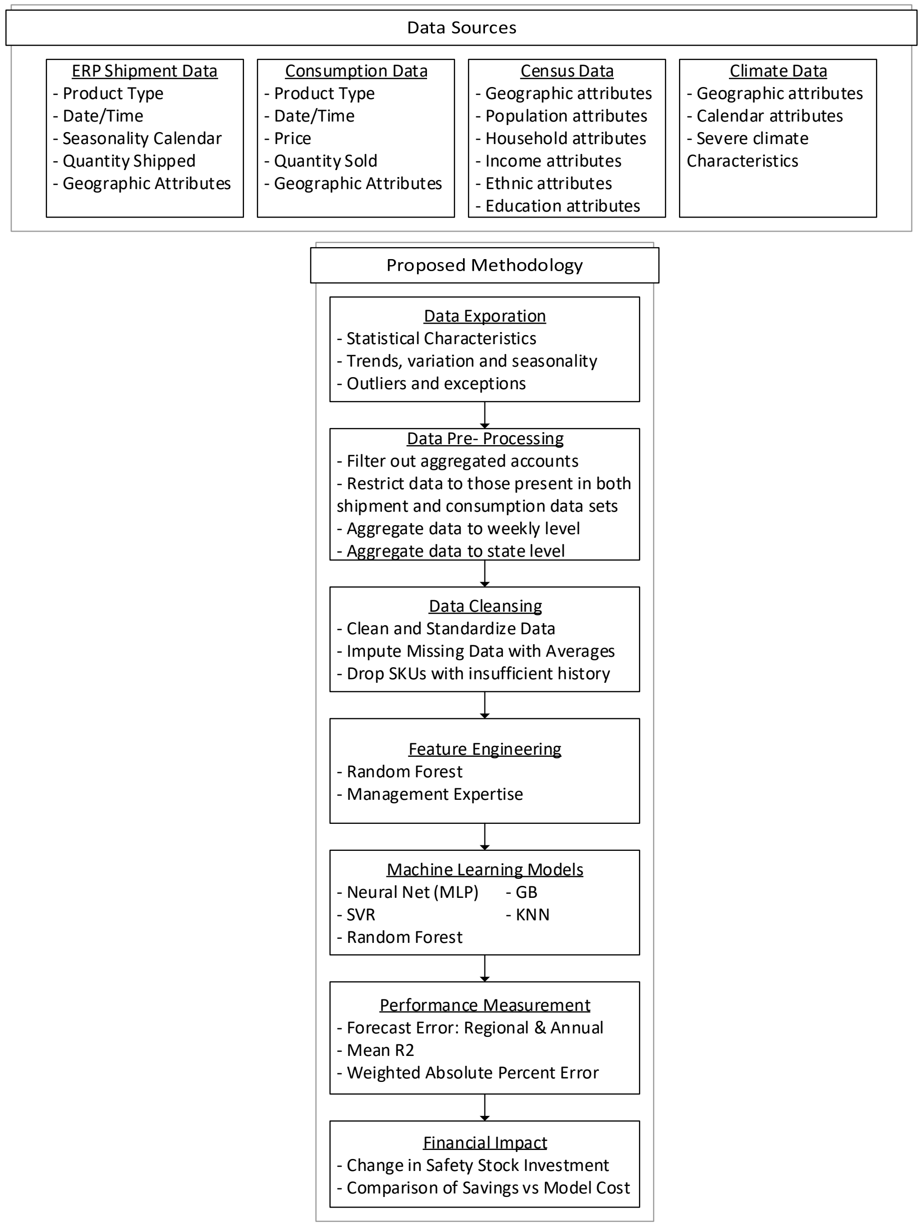


Fig. 3.1 *Data Sources and Process Map. Data sources, and path of data for forecast generation and comparison*

Every model underwent a combinatorial evaluation for each of the two hyper-parameters selected, as well as each feature attribute set identified, resulting in search across not just the two dimensions of the hyperparameters but the third dimension of the feature attribute set as well. The cube search process is detailed in Fig 3.2

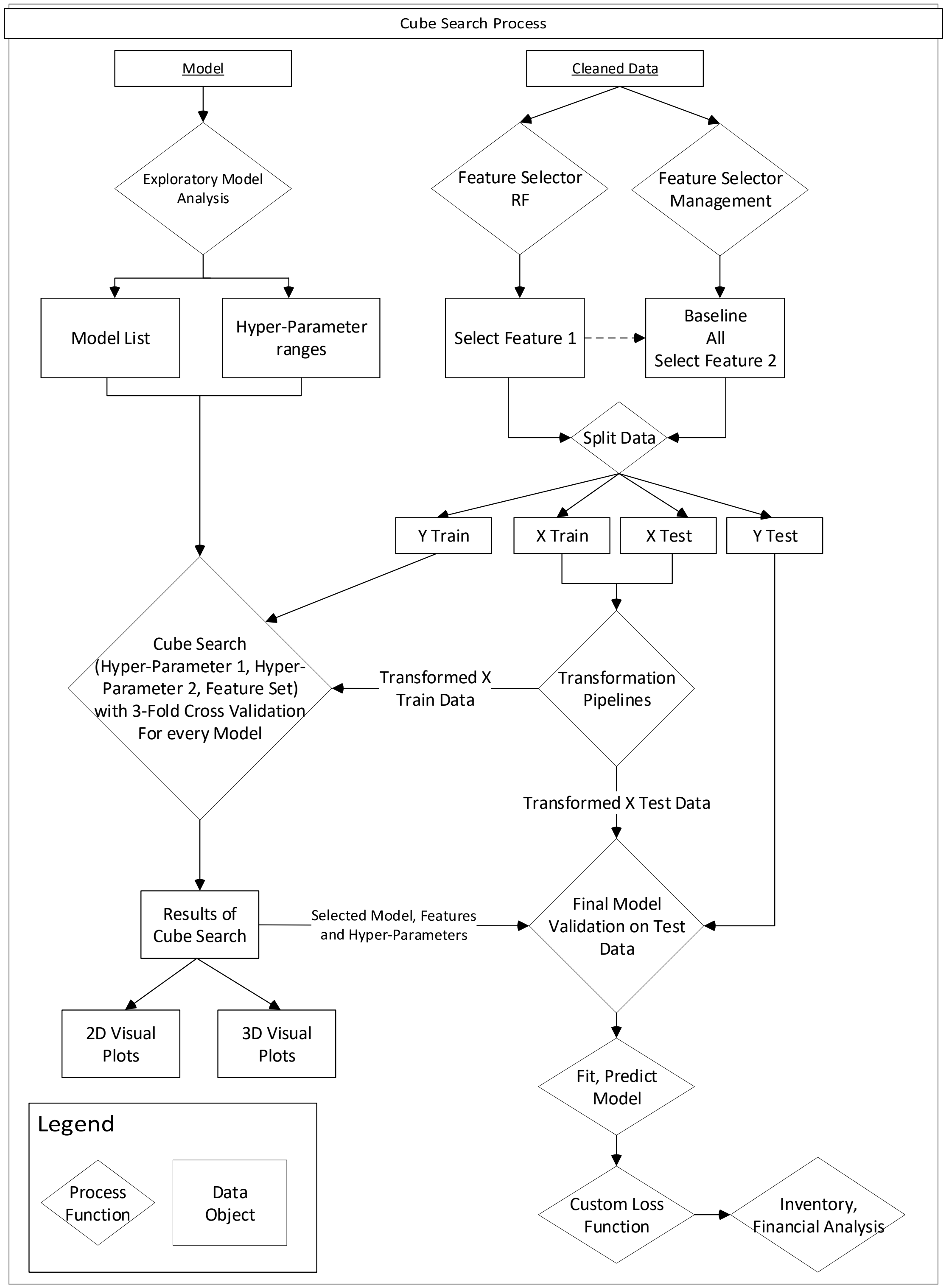


Fig 3.2 *Cube Search Process Flow. The data flow in the Cube Search methodology as data moves from pre-processing through model selection to final validation and analysis*

3.1 Data Collection

3.1.1 Shipment Data from Internal ERP

Shipment data from King's Hawaiian ERP system, MAS500, were collected and housed in an internal SQL data table. These data reflected the units sold to retail customers, not end consumers. Shipment data have the attributes identified in Table 3.1.

Table 3.1 *Shipment Data Attributes*

Data Source	Attribute	Description
Shipment Data	SKU	Unique Product Number
Shipment Data	Replacement SKU	Closest Active SKU to Historical SKU
Shipment Data	Quantity	Units, or Pounds, Shipped
Shipment Data	Customer	Name of Account Receiving Product
Shipment Data	Ship To	Receiving Address of Customer
Shipment Data	3PL	Unique Warehouse Designator
Shipment Data	Ship Date	Day, Week, Month, Year
Shipment Data	Seasonal ID	Identifier Based on Historic Seasonality

Historical shipment data are the primary data source for the current King's Hawaiian statistical forecast. Six years of historical data (2013 to 2018) were available in the data warehouse. The shipment data comprise 372,074 records, from 58 unique active SKUs from 45 states. The seasonal ID was assigned based on historical shipment demand related to national holidays that had differential volume of shipments. The duration of elevated shipments, lull periods afterward, as well as the offset from the holiday, were also identified. The data set was restricted to have SKUs that are no longer sold, or have been re-launched, to align to current SKU IDs (Replacement SKU) to preserve the validity of historical records.

3.1.2 Statistical Forecast Data from Internal Model

The current statistical forecast methodology in use at King's Hawaiian is a Holt-Winters model. The forecast was built from historical shipment data going back to 2012, when available. The data set that was fed into the model to align historical SKUs to their current closest analog was first scrubbed. Attributes for the statistical forecast are shown in Table 3.2.

Table 3.2 *Statistical Forecast Attributes*

Data Source	Attribute	Description
Statistical Forecast	SKU	Closest Active SKU to Historical SKU
Statistical Forecast	Forecasted Demand	Units/LBs Forecasted to Ship
Statistical Forecast	Actual Demand	Units/LBs Shipment Actuals
Statistical Forecast	Regional Demand	Historic Regional Demand by Percentage
Statistical Forecast	WAPE	Weighted Absolute Percent Error
Statistical Forecast	Ship Date	Week, Month, Year
Statistical Forecast	Seasonal ID	Identifier Based on Historic Seasonality

The statistical forecast is updated as part of the monthly Sales & Operations Planning (S&OP) process. The current King's Hawaiian forecast utilizes a current period lockout, limiting adjustments to forward periods only. The King's Hawaiian demand forecast, which covers the total shipment demand, regardless of location, is defined as the network level forecast. This network level forecast is disaggregated to the regional level by applying historic demand distribution percentages, by month, to each 3PL within the network. Historic demand percentages are calculated from the historical sales using the default (expected) 3PL as the shipping point. Actual demand is evaluated using the same default shipping 3PL to correct for temporary distribution deviations due to capacity and obsolescence. The regional forecast generated by this process is used by King's Hawaiian to distribute product throughout the network.

3.1.3 Safety Stock from Internal Model

Safety stock inventory is currently calculated in an Excel model. For SKUs with sufficient history, the regional demand forecast is evaluated against actuals. The root mean squared error (RMSE) of the weekly forecast error is calculated for each period, for each SKU, and then multiplied by the standard deviation equivalent to cover 99.5% of demand. The value of the required safety stock is calculated at the average value for each SKU. Figure 3.3 visualizes the forecasted weekly demand and safety stock against the actual 2017 demand at a specific 3PL (Nordic Cold Storage, abbreviated NA).

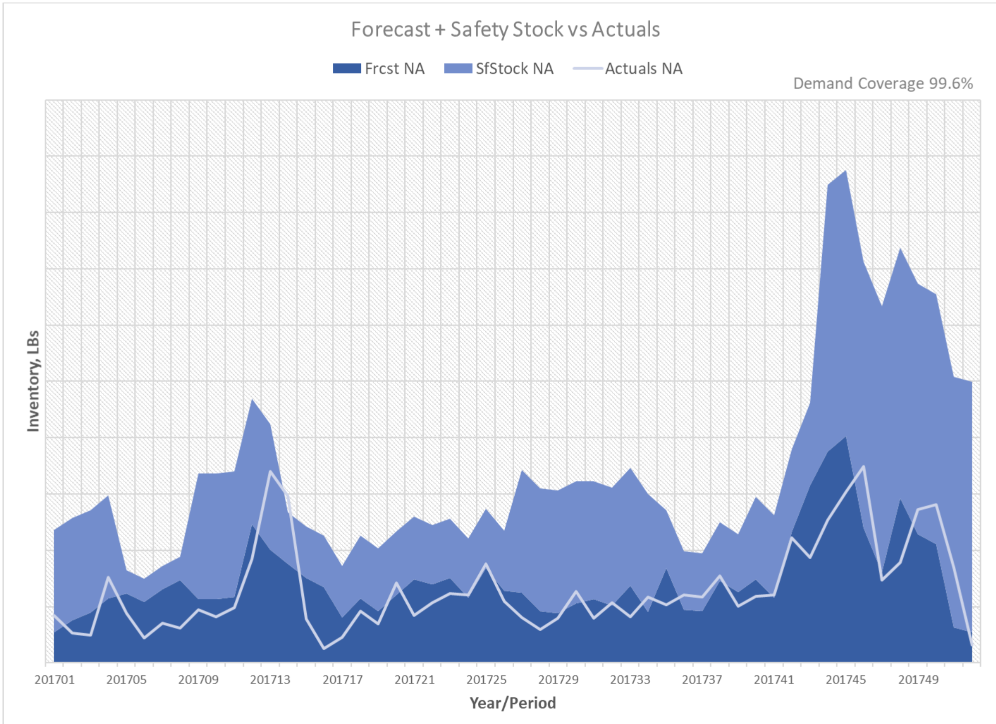


Fig 3.3 *Visualization of Safety Stock. Inventory carried as a result of forecasted demand (dark blue area) and safety stock (light blue area) with actual demand (line)*

3.1.4 Consumption Data from External Vendor

Consumption reflects the sell-through of the product to the end customer, also known as point-of-sale or scan data. Consumption data are currently purchased from an external data supplier for most major customers. Data can be captured by account, or by geographic region, but not both simultaneously. For the purposes of this research, geographic data were used that did not include any customer information. Consumption data are pre-processed through proprietary third-party algorithms and are scaled up to reflect the impact of the total market. Historic consumption was available for the past six years (since 2013). This range of consumption data yields 876,000 records across all 50 states and two US territories. Attributes of the available consumption data are presented in Table 3.3.

Table 3.3 *Consumption Data Attributes*

Data Source	Attribute	Description
Consumption Data	Unit	Fundamental Selling Unit
Consumption Data	Quantity	Units Consumed
Consumption Data	Avg. Price	Average Price per Unit
Consumption Data	Date	Week, Month, Year of Sale
Consumption Data	Geography	State Level Resolution
Consumption Data	ACV ($MM)	All Commodity Volume per Million Dollars, A Measure of Distribution
Consumption Data	Points Distribution	A Measure of Number of Stores Selling
Consumption Data	Share of Isle	Measure of Physical Space Allocated

Data do not align to the specific SKU but instead report on the fundamental selling unit, resulting in 19 unique product classifications. While these selling units tie to the King's Hawaiian SKU in most cases, it does not uniquely identify the sales that originate from product sold in specialty merchandised corrugated displays. Instead, the data reflect the consumption of the base level selling unit.

3.1.5 Population Data from External Database

Data were collected from the United States Census using the American Fact Finder tool (United States Census Bureau, American Fact Finder, 2019). Data on characteristics such as population, education, ethnicity, households, and income were collected from 2013 to 2017. Census data for these time ranges are estimates, not actual counts. Each data set covers all 50 states and US territories. The number of attributes within the data sets range from a low of 21 to a high of 769 unique fields.

3.1.6 Climate Data from External Database

Data were collected from the National Oceanic and Atmospheric Administration's database (National Oceanic and Atmospheric Administration, National Centers for Environmental Information, 2019) on severe weather events. These data include all significant weather and climatological events that occurred in each state from October 2018 going back to 2013.

3.2 Data Exploration

The sections below discuss the data pre-processing and initial characteristics of the source data sets.

3.2.1 Summary Statistics

3.2.1.1 Shipments

The shipments data reflect a comprehensive view of King's Hawaiian demand, extending beyond the scope of this research, which is focused on the core bread SKUs. Therefore, it was restricted to reflect only those products that fit this criterion. The elimination of 37 unique SKUs leaves 21 SKUs, which will be further reduced based on the availability and consistency of data.

Shipment statistics were evaluated at the resolution of total pounds sold, regardless of region, for each week beginning in 2013 and through period 11 of 2018. Summary statistics are presented for core bread shipments in Table 3.4. Shipment data used in the forecast included all of 2018.

Table 3.4 *Summary Statistics for Shipment Actuals*

Year	2013	2014	2015	2016	2017	2018 (partial)
Count (Weeks):	52	52	52	53	52	47
Average: % of Annual	1.92%	1.92%	1.92%	1.89%	1.92%	2.13%
Minimum: % of Annual	0.86%	0.48%	0.27%	0.07%	0.67%	1.21%
Maximum: % of Annual	4.53%	5.24%	5.09%	4.98%	5.19%	5.25%
Median: % of Annual	1.65%	1.69%	1.69%	1.64%	1.57%	1.90%
Standard deviation: % of Annual	0.85%	0.90%	0.91%	0.90%	0.91%	0.82%
First quartile: % of Annual	1.35%	1.34%	1.39%	1.40%	1.48%	1.62%
Third quartile: % of Annual	2.19%	2.06%	1.99%	1.97%	2.18%	2.24%
Skewness:	1.48	1.88	1.85	1.57	1.9	2.12
Excess Kurtosis:	1.54	4.05	3.66	2.76	3.34	4.83
CV	0.51	0.53	0.54	0.55	0.58	0.43

The increase in total pounds sold year over year indicates the presence of a positive trend (growth). The range between the minimum and maximum volumes year to year highlights the variable nature of shipments.

3.2.1.2 Consumption

The consumption data of King's Hawaiian bread are reported separately from that of other King's Hawaiian food categories, and therefore are already restricted to the relevant products and aligned with the filtered shipment data set. Consumption statistics were evaluated as the total number of units sold, regardless of region, for each week beginning in 2013 through period 9 of 2018. Table 3.5 presents the summary statistics for consumption. Consumption data used in the forecast included all of 2018.

Table 3.5 *Summary Statistics for Consumption Actuals*

Year	2013	2014	2015	2016	2017	2018 (partial)
Count (Weeks):	52	52	52	52	53	39
Average: % of Annual	1.92%	1.92%	1.92%	1.92%	1.89%	2.56%
Minimum: % of Annual	1.52%	1.36%	1.52%	1.43%	1.45%	2.17%
Maximum: % of Annual	4.48%	4.91%	4.87%	5.34%	5.30%	4.67%
Median: % of Annual	1.72%	1.74%	1.71%	1.69%	1.66%	2.45%
Standard deviation: % of Annual	0.58%	0.70%	0.65%	0.75%	0.73%	0.43%
First quartile: % of Annual	1.63%	1.61%	1.65%	1.64%	1.58%	2.36%
Third quartile: % of Annual	1.92%	1.93%	1.82%	1.81%	1.85%	2.61%
Skewness:	2.77	2.56	3.19	3.47	3.39	3.3
Excess Kurtosis:	7.46	6.67	9.9	11.78	11.3	13.43
CV	0.34	0.4	0.38	0.44	0.44	0.17

As expected, due to the correlation between unit sales and shipments, a similar positive trend was detected in the annual consumption volume. Also similar is the large range between the minimum and maximum amount purchased. Unlike shipments, the minimum volume of units sold each year does not vary as significantly as the minimum pounds shipped. Additionally, the coefficient of variation for the consumption is lower than that of shipments, suggesting shipments are subject to the bullwhip effect.

3.2.2 Seasonality

It is well understood within King's Hawaiian that their demand is highly seasonal. The primary season is centered on the US holiday Thanksgiving, with the Christmas holiday a close second in terms of the magnitude of sales. Easter also drives high sales, as do other key events such as championship football games, and federal holidays such as Labor Day. Seasonality is mapped at King's Hawaiian using a calendar that assigns a label, or seasonal identifier, to each week of the year. A heat map (Fig 3.4) helps identify the duration of the seasons, and whether there is a lag between the change in shipments and the known seasonal event. The seasonal labels uniquely identify the timeframe, lag, and duration for key recurring seasonal events that historically affected (both positively or negatively) shipments.

Heat Map of Shipments

	Year Desc					
Week, Ship..	2013	2014	2015	2016	2017	2018
01	9e+05	6e+05	4e+05	1e+05	2e+06	2e+06
02	1e+06	1e+06	2e+06	2e+06	2e+06	2e+06
03	2e+06	2e+06	2e+06	2e+06	2e+06	2e+06
04	1e+06	2e+06	2e+06	2e+06	3e+06	3e+06
05	2e+06	2e+06	2e+06	2e+06	2e+06	3e+06
06	1e+06	1e+06	2e+06	2e+06	2e+06	3e+06
07	1e+06	2e+06	2e+06	2e+06	2e+06	2e+06
08	1e+06	1e+06	2e+06	2e+06	2e+06	2e+06
09	2e+06	1e+06	2e+06	2e+06	3e+06	3e+06
10	2e+06	2e+06	2e+06	4e+06	2e+06	4e+06
11	3e+06	1e+06	3e+06	4e+06	3e+06	5e+06
12	3e+06	2e+06	4e+06	4e+06	4e+06	4e+06
13	2e+06	3e+06	3e+06	3e+06	4e+06	3e+06
14	2e+06	3e+06	3e+06	2e+06	4e+06	2e+06
15	1e+06	2e+06	2e+06	2e+06	3e+06	2e+06
16	2e+06	3e+06	2e+06	2e+06	2e+06	2e+06
17	1e+06	2e+06	2e+06	2e+06	3e+06	2e+06
18	1e+06	2e+06	2e+06	3e+06	2e+06	3e+06
19	2e+06	2e+06	3e+06	3e+06	2e+06	3e+06
20	2e+06	2e+06	3e+06	3e+06	3e+06	3e+06
21	3e+06	3e+06	3e+06	3e+06	3e+06	3e+06
22	2e+06	2e+06	2e+06	3e+06	2e+06	3e+06
23	2e+06	2e+06	3e+06	2e+06	3e+06	2e+06
24	2e+06	2e+06	3e+06	3e+06	3e+06	3e+06
25	2e+06	2e+06	2e+06	3e+06	4e+06	4e+06
26	2e+06	3e+06	3e+06	3e+06	3e+06	3e+06
27	1e+06	2e+06	2e+06	3e+06	2e+06	2e+06
28	2e+06	2e+06	4e+06	2e+06	3e+06	3e+06
29	2e+06	2e+06	3e+06	2e+06	2e+06	3e+06
30	1e+06	2e+06	2e+06	2e+06	2e+06	2e+06
31	1e+06	2e+06	2e+06	2e+06	2e+06	2e+06
32	2e+06	2e+06	2e+06	2e+06	3e+06	3e+06
33	1e+06	2e+06	2e+06	3e+06	2e+06	3e+06
34	2e+06	2e+06	2e+06	2e+06	3e+06	3e+06
35	2e+06	2e+06	2e+06	3e+06	3e+06	3e+06
36	2e+06	2e+06	2e+06	3e+06	3e+06	2e+06
37	2e+06	2e+06	2e+06	2e+06	2e+06	2e+06
38	2e+06	2e+06	2e+06	2e+06	3e+06	3e+06
39	1e+06	2e+06	3e+06	2e+06	3e+06	3e+06
40	2e+06	2e+06	2e+06	3e+06	3e+06	3e+06
41	2e+06	2e+06	2e+06	3e+06	3e+06	3e+06
42	2e+06	2e+06	2e+06	2e+06	4e+06	4e+06
43	3e+06	3e+06	3e+06	3e+06	5e+06	5e+06
44	5e+06	4e+06	5e+06	5e+06	7e+06	5e+06
45	5e+06	6e+06	7e+06	8e+06	9e+06	7e+06
46	4e+06	6e+06	7e+06	7e+06	7e+06	8e+06
47	5e+06	5e+06	6e+06	5e+06	4e+06	5e+06
48	2e+06	2e+06	3e+06	6e+06	6e+06	4e+06
49	4e+06	4e+06	4e+06	6e+06	7e+06	7e+06
50	3e+06	4e+06	5e+06	5e+06	5e+06	7e+06
51	3e+06	3e+06	4e+06	5e+06	4e+06	4e+06
52	1e+06	2e+06	2e+06	3e+06	1e+06	2e+06
53				2e+06		

Fig 3.4 *Heat Map of Total Shipments. Darker colors indicate higher volume of shipments and reveal annual seasonal patterns*

3.2.3 Severe Weather Data

Climate data in the form of severe weather events were gathered from NOAA using their Storm Events Database. Records of storm events were collected from 2013 through October 2018. Initial records included geographic areas, bodies of water and US territories that were excluded from the analysis which reduced the number of regions from 68 down to the 50 states. Storm events were recorded at the resolution of daily events, which were then aggregated to the weekly level. Storm events cataloged range from dense fog to hurricanes, with a total of 56 unique categories. After excluding marine specific events and less severe categories, 17 categories of severe weather events remained for consideration. To quantify the severe weather events, the counts of unique events in each category were aggregated for each state, each week. To normalize the impact of typical weather patterns, the weather events were assigned a positive binary value only if the sum of the count of the specific weather events exceeded one standard deviation above the mean for the month for each state.

3.2.4 Census Data

Data were collected for all US States, Washington D.C., and Puerto Rico from 2013 to 2017. The data for these date ranges are estimates, unlike the physical counts obtained every ten years. 2018 data were not available. The inability to collect census data for the current year highlights a challenge of using publicly available data sets. To make use of the census data as a predictive indicator of demand, 2017 data were used as a proxy for 2018. This was judged this to be a safe assumption as the year-over-year variation within each category was not large, and typically fell within the census published margin of error. The imputation of the annual census data allows for the inclusion of 2018 data and is a key assumption made to align data sets with complementary data.

Since the approach was to use a random forest model to determine the appropriate combination of attributes for forecasting demand, the census data were restricted to only the primary attributes, ignoring attribute combinations and subsets. The census data required enhanced scrubbing to align timeframes. Not all attributes were consistently estimated each year, or else had slight differences in descriptions. To ensure a complete data set, attributes were restricted to those that were present across the entire timeframe considered and matched descriptions. Attribute descriptions were then abbreviated and, in some cases, truncated to fit within database parameters. Table 3.6 shows the starting attributes and the result of multiple rounds of transformations.

Table 3.6 *Census Data Attribute Summary*

Classification	Original Attributes	Refined Attributes
Population	357	27
Education	769	15
Ethnicity	21	8
Household	201	22
Income	169	14

- Population data were collected from the US Census data set designated DP05. The population data set included attributes that were subsets or combinations of attributes. The data set was restricted to select only total population and age range percentages, by state.
- Education data were collected from US Census data set S1501. 2013-2014 data presented education ranges as percentages of the population within an age range. 2015-2017 data showed the count of the population within that age and education range as opposed to the percentage. To normalize the data, all attributes were converted to percentages of the population for that category age range.
- Ethnicity data were selected from the US Census data set B02001. Due to the potential combination of ethnic backgrounds, attributes were restricted to either single defined ethnic groups or a combination. To normalize the data by state, attributes were converted to percentages of the state population. Ethnicity data were unavailable for 2017 and 2018, so 2016 data were used as a proxy for both 2017 and 2018.
- Household data were selected from the US Census data set S1101. The data were filtered into primary categories and restricted subset combinations of attributes. Data were presented as

percentages of homes that fell within the primary category, or the mean percentage of homes that fell within the category.

- Income data were selected from the US Census data set S1902. The attribute median income, in US dollars, was selected as the primary measure.

3.3 Data Pre-Processing

Integrating multiple data sources required significant levels of pre-processing and secondary transformations to ensure alignment and relevancy. This section discusses the process of aligning multiple data sources.

3.3.1 Aggregation

To capture the unique perspectives of the external data sources that can be leveraged by machine learning forecast models, an appropriate level of resolution was required. Shipment data are available down to the city, state level. Data are based on the customer ship to location, which is typically a large DC that feeds multiple different stores in different cities. Consumption is similarly discrete, reflecting sell-through data at the city level. However, like shipments, these data are often the aggregated consumption for the wider metropolitan area. King's Hawaiian's 3PL network generally aligns to geographic regions such as North East, and Midwest, with each 3PL supplying product for multiple states.

For demand forecasting, the level of aggregation selected was the state resolution. Aggregation to the state level allows for the use of both internal shipment data and external data sets such as consumption, with a higher degree of confidence that they are reflective of the broader area. Additional data sources such as socio-economic census data and weather data must be aggregated to the same resolution.

King's Hawaiian's current demand forecasting methodology is at the weekly level. To align the proposed machine learning forecast as much as possible, the data were aggregated to the resolution of a week. King's Hawaiian shipment data and weather data were the two data sets that had a temporal resolution down to the day and had to be aggregated to weekly. Shipments were aggregated by summing the daily sales within that week. For climate data, the counts of severe weather events one standard deviation above the mean that occurred each week were used. Consumption data as currently purchased is already set at the weekly resolution and therefore will require no aggregation. Census data are annual and were therefore disaggregated to the weekly level. As a simplifying assumption, weekly census data was assumed to mirror that of the annual data for each year.

3.3.2 De-Seasonalizing

Demand was de-seasonalized to improve the predictive performance of the machine learning models. Machine learning models can be overly influenced by strong seasonal patterns, ignoring other important patterns and trends due to the magnitude of variation caused by seasonal factors. For regression models in general, de-seasonalizing the data is expected to improve the fit and predictive power of the model.

3.3.2.1 Shipments

Visualization of the actual shipment demand (Fig. 3.5) confirmed the presence of a strong seasonal element.

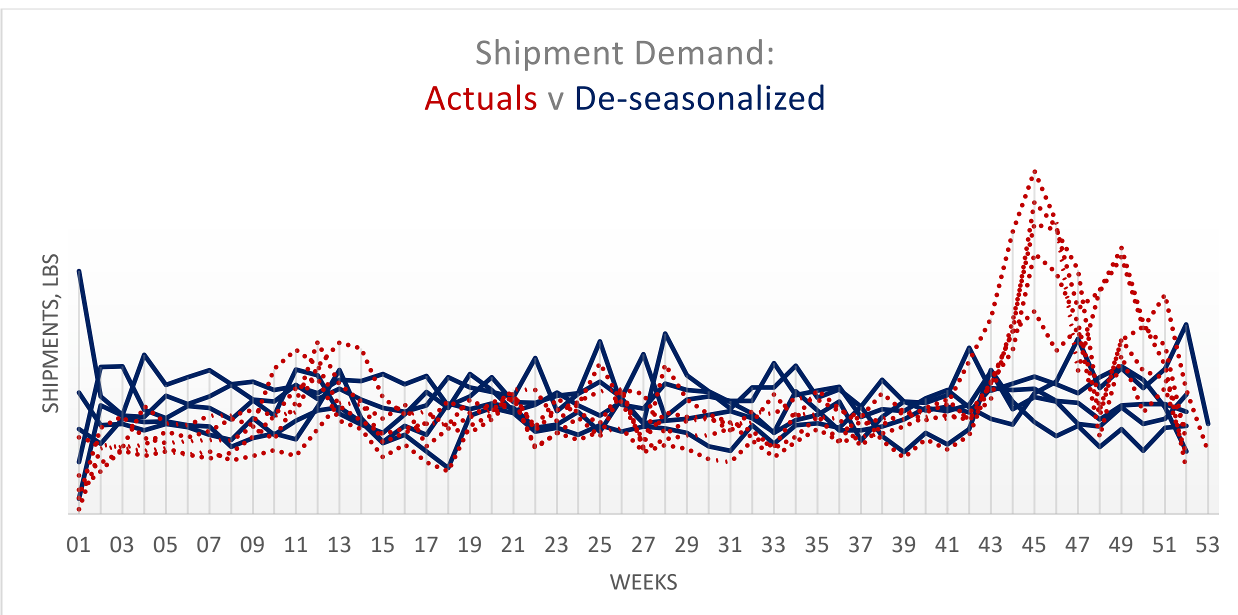


Fig 3.5 *Shipments, De-seasonalized. Weekly total brand shipment data shown in both seasonal (red dotted lines) and de-seasonalized (solid blue) for multiple historic years*

Magnitudes were calculated by taking the quotient of actual shipments and the annual mean on a per SKU, per year basis. The mean seasonal magnitude was then calculated for each seasonal identifier. Shipments were de-seasonalized using the company assigned seasonal identifiers and the mean SKU specific magnitudes.

Outliers such as the first week of 2017 can generate variation as significant as normal seasonality, even after the shipments have been normalized. This event, while believed to be replenishment from an unexpectedly large sell-through effort due to a single large retailers' aggressive merchandising, highlights the volatility in shipments and the outsized impact a single retailer can have on demand.

3.3.2.2 Consumption

Visualization of the consumption demand (Figure 3.6) aligned with, and demonstrated an even more distinct seasonality than, the shipment data.

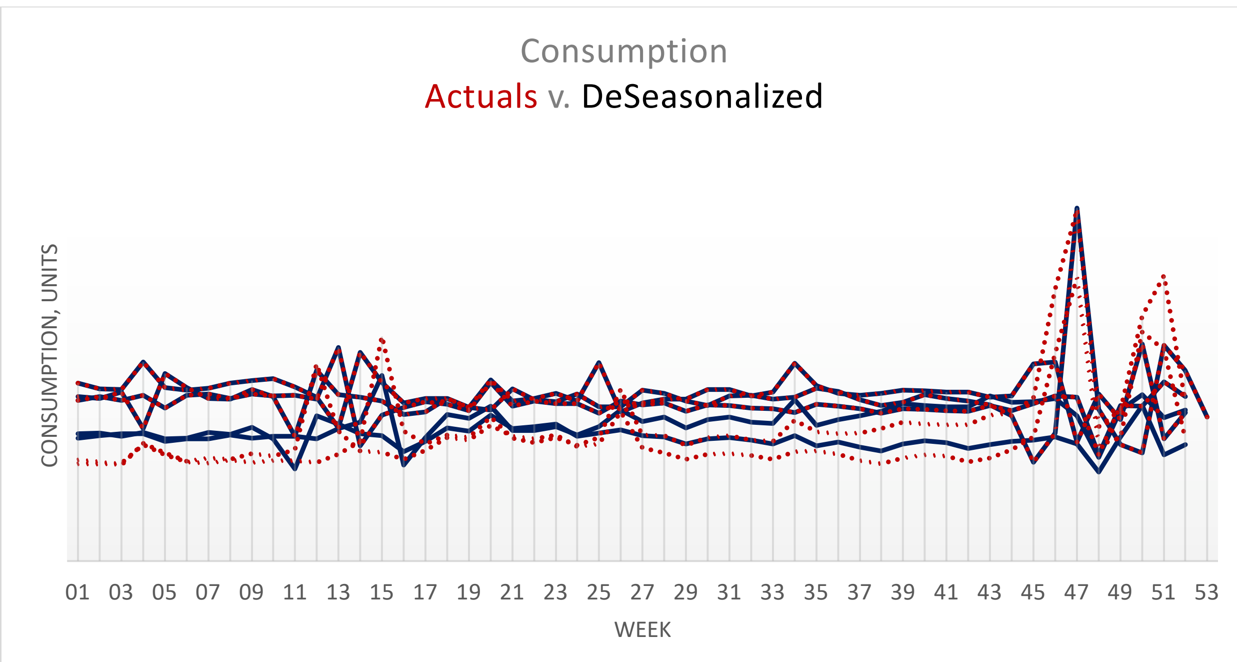


Fig. 3.6 *Consumption, De-seasonalized. Weekly total brand consumption data shown in both seasonal (red dotted lines) and de-seasonalized (solid blue) for multiple historic years*

Consumption was de-seasonalized using the company assigned seasonal identifiers. Seasonal magnitudes were calculated by taking the quotient of actual consumption and the annual mean on a total unit basis. The mean magnitudes were then calculated across 2013-2017 for each seasonal identifier.
Outliers such as week 46 consumption demand, which was likely caused by an aggressive merchandising effort at a large retailer, demonstrates how sensitive consumption demand signals can be to specific actions at the retail level.

3.3.3 Aligning by State

The data were aggregated to the state level because this resolution allowed for the best alignment of the data sets with the highest degree of confidence. While nearly all of the data sets cover all 50 states, the shipment data available from King's Hawaiian show that from 2013 through 2018, they had not shipped to all 50 states. Since the target of the forecast is shipments, external data sets were aligned to the resolution of the shipment data.

3.3.4 External Attribute Incorporation

External data were incorporated into the analysis by joining data sets on two different attributes. The first attribute was regionality, set to the resolution of state. The second attribute was a temporal variable with data joined by year and week number. Additionally, consumption data were joined on the product attribute. External data sets were joined in SQL to improve the speed of analysis, as well as allow for analysis on the characteristics of the individual data sets.

3.3.5 Data Partitioning

The data were partitioned into two sections for training and testing. The models were trained on data from 2013 through 2017. The models were then tested against actuals from 2018. Each model also underwent a three-fold cross-validation to improve understanding of the accuracy of the model and reduce the likelihood of overfitting.

The King's Hawaiian statistical model was trained on data beginning in 2012 and undergoes a monthly update as part of the ongoing S&OP process. It was this S&OP updated 2018 statistical forecast that was evaluated against the machine learning forecast.

3.4 Data Scrubbing

Data were preprocessed and joined into a unified data set in SQL. After data were pre-processed in SQL, it underwent a scrubbing process in Python prior to input into the models. Data were randomly sampled and visualized for each of the data scrubbing steps, and summary statistics generated, to ensure the data cleansing process was effective and data integrity was not compromised.

3.4.1 Temporal Alignment

To ensure comparable availability of data for demand forecasting in practical use, shipment data (the target of the demand forecast) was offset by one week to all other associated data. This lag of one fiscal week ensured that the predicted value for the shipment demand was based on consumption and weather events from the preceding week and that the model would not have visibility to the current week's shipment performance for use in its predictions.

3.4.2 Product Specific

SKU specific shipment data was not consistently available for every product for the complete timeframe being evaluated. To align data and account for the updated products and limited time promotional offerings, all products were classified based on their fundamental selling unit. Since demand was being evaluated in pounds of product, this eliminated any issue of variable secondary or tertiary packaging configurations.

Not all products had historical data back through 2013, such as newer products for club store channels. Due to the nature of machine learning algorithms, gaps or holes in data can significantly skew the results. Since newer products did not always have a comparable analog in the historical data or could not be accurately segregated from the historical data to prevent double counting the demand, any SKU without a complete historical data set was dropped from the analysis. This reduced the product count down to 10 final SKUs.

The shipment demand for the remaining 10 SKUs was visualized through Tableau to confirm the consistency of shipments at the national level. SKUs that had gaps identified were investigated as to the underlying cause. If the demand was found to be missing as a result of non-market driven forces, such as manufacturing disruption causing inventory stock outs, then those records were also dropped from the analysis.

3.4.3 Regionality

Regional data labels were standardized to two characters and any out of scope region was dropped from the data set. While census, weather and consumption data were available for every state, shipment data were not. This is because not every state has King's Hawaiian product sold directly to an entity in each state, as certain customers and regions are supported by out of state distribution centers or distributors.

3.4.4 Weather

Weather data were not available for the final two months of the 2018 test data set. This timeframe was included in the evaluation absent this data as it represents a real limitation of using publicly available data for forecasting.

Weather data were also engineered into three new binary attributes; summer, winter, and high damage. If the occurrence for any of the underlying weather types were identified as having occurred per the selection criteria described in section 3.2.3, then the new attributes indicated the occurrence.

3.4.5 Census

Census attributes that evaluated the same underlying data in different subsets were restricted to a single set that was mutually exclusive and collectively exhaustive. Categories that exhaustively covered a characteristic were broken down into two or more complementary attributes. Census data is often reported as a percentage of the population for the geography; however, the unit of measure was not consistently represented in the data. All census data that were reported in percentages were normalized to range from zero to one. Attributes categories that had data represented as both a percentage as well as the numerical counts of populations were restricted to the percentages only when possible.

3.4.6 Data Cleansing

The data set was scrubbed of any outliers or duplicates. Duplicated records were dropped after data aggregation had been performed. Records that contained multiple nulls or negative values for any of the attributes were removed. A small number of missing records with missing data were imputed with the mean of the attribute for the class within that feature. Negative values for shipments were extremely limited, and an artifact of an inventory reconciliation mechanism for accounting purposes.

3.5 Data Normalization

To ensure optimum results from the machine learning algorithms and to minimize the impact of dissimilar data scales, all data were normalized prior to forecasting. Categorical values were converted to binary representations via one-hot encoding to improve the effectiveness of the predictions. Census data reported as percentages were normalized to ranges between zero and one. Weather data and the engineered weather category attributes were converted to binaries as part of pre-processing. Non-range bound data such as shipment and consumption data, as well as population totals, were standardized by removing the mean from each value and dividing by the standard deviation as part of the transformation pipeline processing. These transformations were calculated from the training set only, with the values being applied to the test set for final verification.

3.6 Feature Selection and Engineering

One important advantage of using machine learning models versus traditional methods in demand forecasting is the ability to perform analysis using not just the internal demand data but also incorporate data from different, external sources. By doing so, demand planners can answer a broad range of questions (e.g. what is the impact of weather events on demand?). However, answering those questions requires including more features in the model and as more features are added the model becomes more complicated, potentially overfitting and running slower. Therefore, feature selection is an important step in building models. For this project, multiple feature sets were evaluated, using different evaluation criteria.

The Feature Select 1 set employed a random forest model, a type of ensemble machine learning technique that is frequently used in feature selection, to identify attributes that have an assigned importance rating above a set threshold of 0.1%. Random forests are described in section 3.7, along with other machine learning models, as the same technique will also be used in developing forecasting models.

Feature Select 2 set was also created using the consumption and census features from Feature Select 1, but replaced storm events with the engineered weather feature attributes discussed in section 3.4.4.

3.7 Machine Learning Models

A large number of machine learning models can be applied in demand forecasting. Five models were selected that were considered to be most suitable for the task based on the literature review. Based on the research, random forests, artificial neural networks, support vector machines, gradient boosting, and k-nearest neighbors have a high potential in decreasing demand forecasting error. A three-dimensional cube search was performed of those models by varying different hyper-parameters (e.g. different number of layers in artificial neural network models) and using different features (e.g. seasonal vs de-seasonalized target data) to compare the performance of the different models.

3.7.1 Random Forests (RF)

Random forests are a type of ensemble methods where multiple decision trees are used for classification or regression. The goal is to decrease the variance by combining trees built using random samples of the data and random subsets of features. Decision trees are built by separating the data at nodes using an algorithm that determines the best split. Multiple algorithms can be used to construct a decision tree. Random forest is an ensemble machine learning approach since it combines multiple different trees when evaluating the best fit.

Random forests are powerful enough to model complex nonlinear relationships. On the other hand, their outputs are not easily interpretable as they come from a combination of multiple trees.

3.7.2 Artificial Neural Networks (MLP)

Artificial neural networks are inspired by how neurons and synapses in the brain work and they can be used to model complex relationships. Neural networks consist of nodes that are used to calculate the weights of the features in the model. The inputs to each node are outputs of the nodes in the previous layer and their associated weights. The output of a node is calculated by evaluating an activation function using the weighted average of the previous layers' outputs. The nodes are organized in layers and as the number of layers increases, the neural network is considered to become a deep neural network. Depending on the number of nodes, the layer structure and the algorithms used, neural networks can take many different forms. The most commonly used type of neural networks called multilayer perceptron (MLP) were used.

Multilayer perceptron models are feed forward neural networks where the information flows in one direction, i.e. from the input layer to the output layers. An arbitrary number of hidden layers can exist between the input and output layers. A multilayer perceptron with one hidden layer can approximate any continuous function. The network can be trained using a backpropagation algorithm.

3.7.3 Support Vector Machines (SVR)

Support vector machines (SVM) are used in tasks like classification and regression. They employ a decision boundary called a hyperplane. The approach is to maximize the minimum margin, i.e. the distance between the hyperplane and the nearest data point.

Support vector machines can model non-linear relationships using a kernel method, which maps the data points that have a non-linear boundary to a higher dimensional space where it is easier to separate them. Polynomial and Gaussian are two of common kernels. SVMs can be slow to run and the outputs can be difficult to interpret.

3.7.4 Gradient Boosting (GB)

The goal of boosting methods is achieving better results by combining weaker models. There are many boosting models like adaboost and gradient boosting. They can be used for both classification and regression tasks.

Gradient boosting is an ensemble method where predictors are added sequentially. In each stage, the new predictor is fit into the residual errors. The new predictors are regression trees in the gradient boosting regression model that was used from the scikit-learn library.

3.7.5 K-Nearest Neighbors (KNN)

K-nearest neighbors is a model that can be used for both classification and regression. Its intuitiveness and ease of implementation made KNN a popular tool among practitioners, especially for classification tasks. In classification, the class of an observation is determined by a vote of the k number of neighbors closest to the observation based on a distance metric. Similarly, in regression tasks, the value of the target is determined by taking an average of the k closest observations.

KNN models are generally fast to train as they do not need to take all the data into consideration, they evaluate the value of the targets based on only k observations. The optimal value of k is generally determined by running the model with different k values and comparing the results.

3.8 Performance Measurement

3.8.1 Forecast Error

King's Hawaiian's current statistical Holt-Winters forecast performance is calculated at two different levels. The first level is the total network forecast or the performance for the SKU across the sum of all selling regions. The forecast error is calculated as the absolute percent error, weighted by forecasted demand for each SKU (WAPE). While the forecast is generated at a week level, the current forecast error is calculated monthly for inventory policy purposes (Fig 3.7).

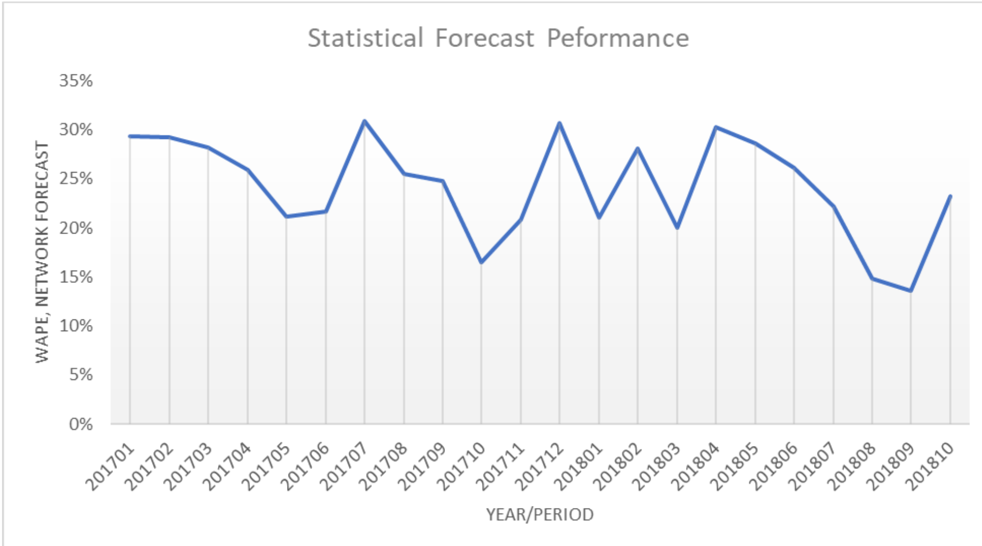


Fig 3.7 *Current Network Level Forecast Performance. Forecast error (WAPE) by period for all core bread SKUs at the network level*

The performance of this regional forecast is evaluated similarly to the network forecast (Fig 3.8), using the weighted absolute percent error (WAPE).

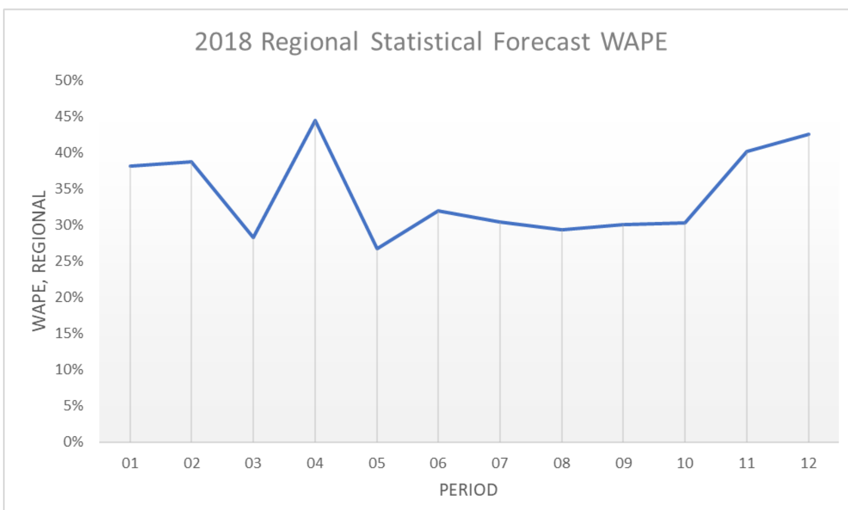


Fig 3.8 *Current Statistical Forecast Regional Forecast Performance. Weighted absolute percent error by period by region when aggregated to King's Hawaiian distribution regions*

The regional forecast has achieved a mean monthly forecast error of ~34% in 2018. This is a significant change in the accuracy of the regional forecast, which improved dramatically from 2017 to 2018 as a result of the implementation of an improved methodology for the calculation of the regional distribution at the start of 2018. It was, therefore, appropriate to evaluate the machine learning model's performance against King's Hawaiian's recent 2018 regional forecast.

3.8.2 Financial Impacts

A primary objective of this research was to determine if the cost savings from a reduced forecast error justify the expense of running a machine learning demand forecast. To evaluate this, the cost savings resulting from the decrease in safety stock due to a reduced forecast error were analyzed. The costs were calculated as the value of the inventory multiplied by the change in the aggregate monthly units of inventory required as safety stock compared to the current practice.

A simple framework is discussed to evaluate the expected savings from likely changes to the forecast error for companies with similar demand characteristics. The financial assessment compares the expected savings from reduced safety stock only, using the average inventory value, and customer service levels specific to the company. These potential savings were evaluated against the incremental costs identified with a machine learning forecast due to the data and personnel required to run it. For the purposes of this financial analysis, additional factors such as interest, anniversary costs, and operational efficiencies were not considered in the evaluation of cost savings, however, changes to forecast accuracy would certainly have an impact on these and other areas of business.

4. RESULTS

This section presents the results of the multiple machine learning models that were run on the cleaned and transformed data. Results are presented in the order of the data flow through the model. A comparison of the performance of all models ran through the cube search process indicated that a KNN model was the best fit for predicting shipments when running on a de-seasonalized feature set. The KNN model selected scored the highest mean R^2 value on cross-validation during the cube search while maintaining a low Train-Test variance. A KNN model with the hyper-parameters defined by the output of the grid search was trained and then validated on the out of sample 2018 test data. The machine learning model achieved a lower total WAPE than the current King's Hawaiian Statistical forecast model, which translates to financial savings due to lower safety stock inventory levels required to cover demand variance.

4.1 Feature Selection

Four unique feature sets were selected for evaluation as part of the cube search. Each of these sets was evaluated using either seasonal or de-seasonalized demand as the target variable, for a total of eight unique feature sets for evaluation. The feature set ultimately selected was based on the model-feature set combination that achieved the highest performance during the cube search process.

4.1.1 All Features

The 'All Features' category included every attribute of the final processed data set, excluding the target variable shipment demand. Models were run with two variants of the all features category, one with seasonal and the other with de-seasonalized demand.

4.1.2 Baseline

The baseline feature set included only geographic and calendar data, without any of the additional consumption, census and weather attributes. The baseline feature set was run with two variants, one with seasonal and the other with de-seasonalized demand.

4.1.3 Feature Select 1

A random forest regressor with a depth of 68 was run on all features. The aggregated weather attributes and calendar data were not considered for this feature selection evaluation. All categorical values (e.g. SKU, Seasonality) that had been converted to binary attributes were preserved and were not considered during feature selection. An importance threshold of 0.1% was selected and features that scored below this threshold were excluded from the feature set. The cumulative importance of the features is presented graphically in Figure 4.1, with the threshold indicating the cumulative importance captured in the features selected.

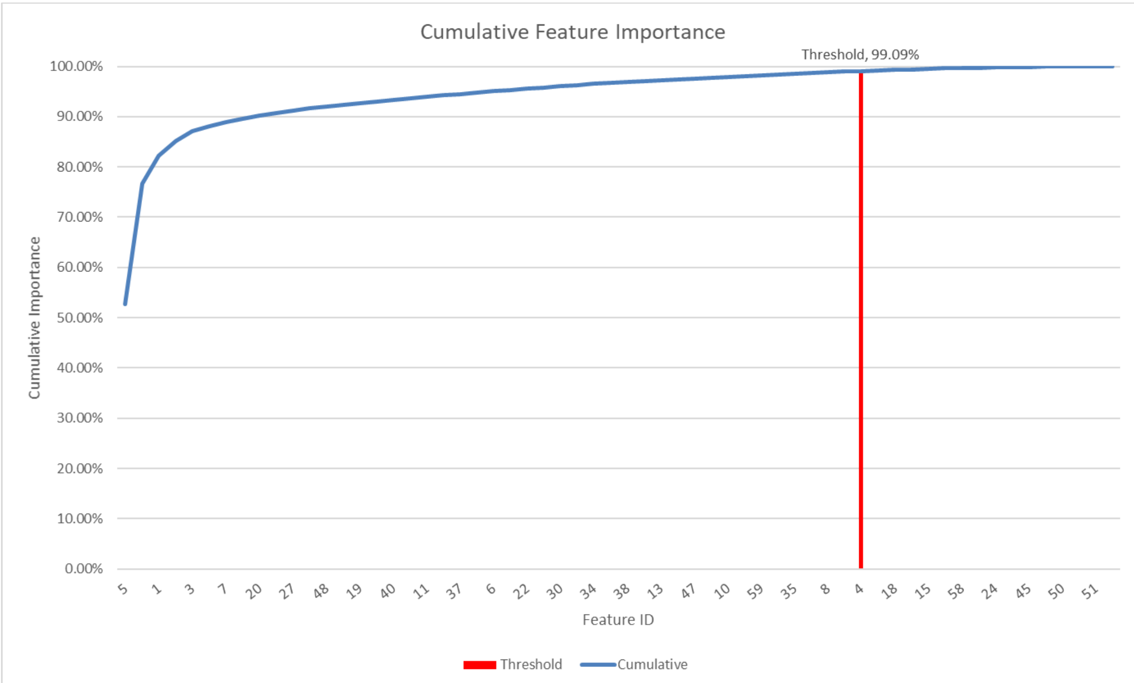


Fig 4.1 *Feature Importance Curve. Number of features plotted against the cumulative weighted value for each feature according to the random forest classifier, with the threshold indicating cumulative importance captured in features selected*

4.1.4 Feature Select 2

The Feature Select 2 set was built from the features selected by the random forest classification done for Feature Set 1; however, for this set, all discrete weather events were dropped, and the engineered weather categories used. This feature set was run on both seasonal and de-seasonalized demand.

4.2 Model Selection

Support vector regression (SVR), multilayer perceptron (MLP), random forest (RF), gradient boosting (GB), and k-nearest neighbors regression (KNN) models were tested with the cleaned data to select the ultimate model for demand forecasting. Each model underwent a cube search across a range of hyper-parameters and feature sets to determine viable settings and input data. Models were also evaluated using a cross-validation test to determine the forecast accuracy. Computation run times were captured for each model as well to understand the practical limitations of each algorithm. The model with the lowest forecast error was selected for final evaluation on the untrained test data set from 2018.

4.2.1 Cube Search Process Results

The cube search successfully utilized a pipeline process to split and normalize the data and include only the features specified for each feature set evaluated. The Python code was set up to be dynamic, allowing for the swift addition of different machine learning models for testing, or different combinations of features.

Overall the cube search was used to evaluate 386 different models to determine the best combination of model algorithm, hyper-parameters, and feature set. The length of time required to train and score varied widely between model types. A comparison of the log time in seconds for each model type is presented in Fig 4.2.

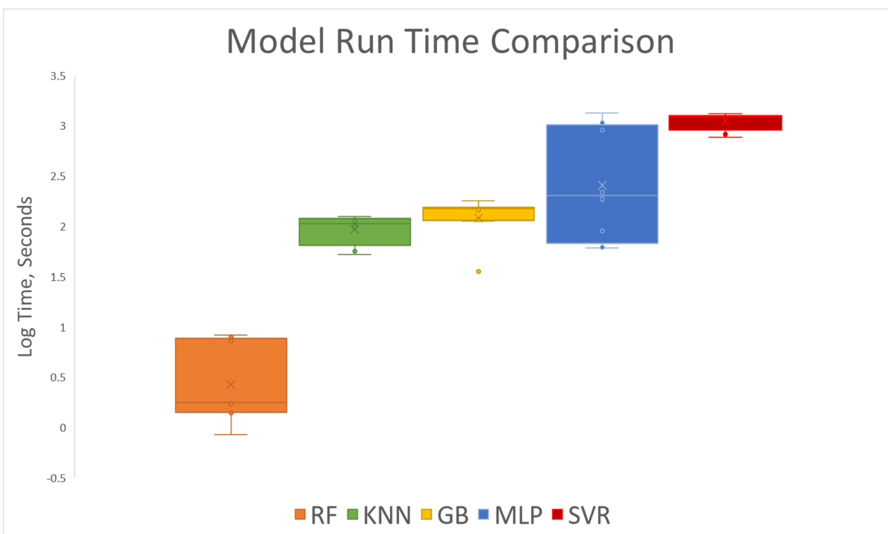


Fig 4.2 *Comparison of Model Run Time. Log time in seconds for each feature set run during cube search*

Run time calculations were calculated in the Python code to capture computational time for each iteration of the cube search.

4.2.2 Cube Search Performance Results

Each of the five models underwent a systematic cube search across every combination of a set of two different hyper-parameters unique to each algorithm and the eight feature sets. The performance, as measured by the mean R^2 value, for each model is discussed below. All models were run in Spyder 3.3.1 running Python 3.7 on a Windows 7 64-bit virtual machine with an Intel Xeon E5-2697A processor running at 2.59 GHz with 32GB of RAM.

4.2.2.1 Support Vector Regression Results

SVR performance results were consistently the lowest of all the models evaluated. The mean run time per model was 112 minutes, making it the most time-consuming model to run. Figure 4.3 shows the combinatorial impact of the hyper-parameters (where “C” is the penalty for the error and “epsilon” is the margin of tolerance for errors) evaluated and the resultant performance for the feature set that produced the highest mean R^2 of the eight feature sets evaluated.

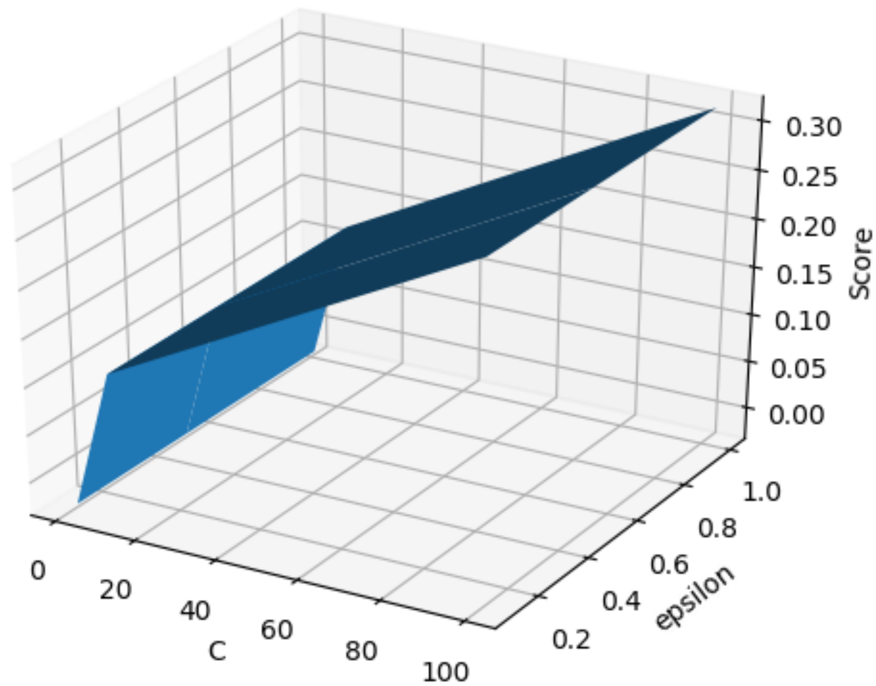


Fig 4.3 *SVR Grid Search Results. Graphical representation of the mean R^2 for each set of hyper-parameters evaluated; C and epsilon on the de-seasonalized Feature Select 2 set*

A total of 72 models were run using the SVR algorithm. For all feature sets tested, the de-seasonalized versions achieved the highest mean R^2 values. Figures 4.4 and 4.5 show the individual hyper-parameter performance, in R^2, between the training and testing data sets during cross-validation.

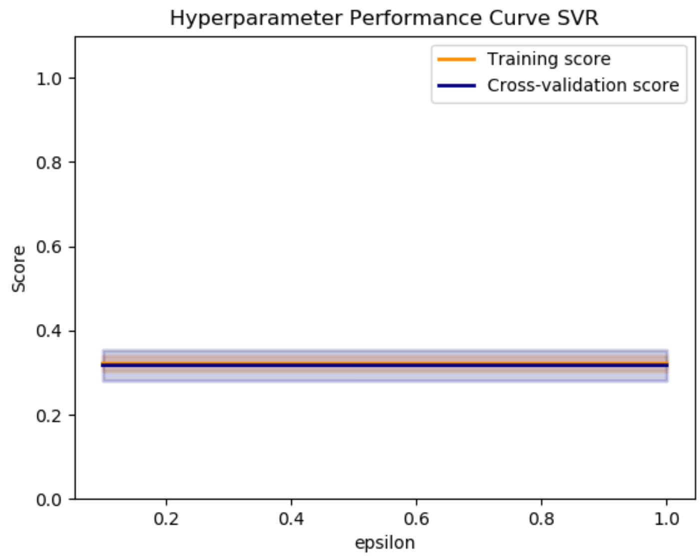


Fig 4.4 *SVR Epsilon Hyper-parameter Performance Curve. Comparison of Train-Test mean R^2 for varying values of epsilon on the de-seasonalized Feature Set 2*

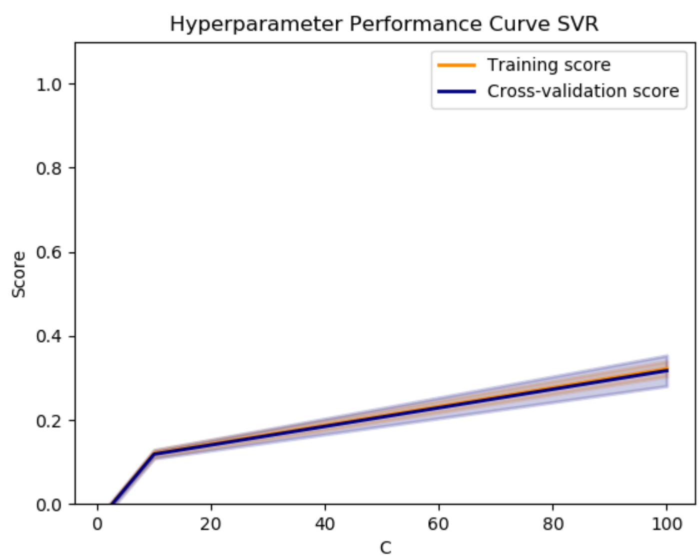


Fig 4.5 *SVR C Hyper-parameter Performance Curve. Comparison of Train-Test mean R^2 for varying values of C on the de-seasonalized Feature Set 2*

The results of the cube search indicated that an SVR model with a C value of 100 and an epsilon of 0.1 when running on the de-seasonalized Feature Select 2 set produced a mean R^2 of 0.317, the best of any of the SVR combinations evaluated.

4.2.2.2 Multilayer Perceptron Results

MLP results, on average, scored above SVR but below all other models tested. The mean run time per model was 39 minutes, making it the second most time-consuming model to run, but still nearly three times faster than SVR. Figures 4.6 show the comparison of the Training and Test data during cross-validation on mean R^2 of the number of layers evaluated for the feature set that produced the highest mean R^2 and lowest Train-Test variance of the eight feature sets evaluated.

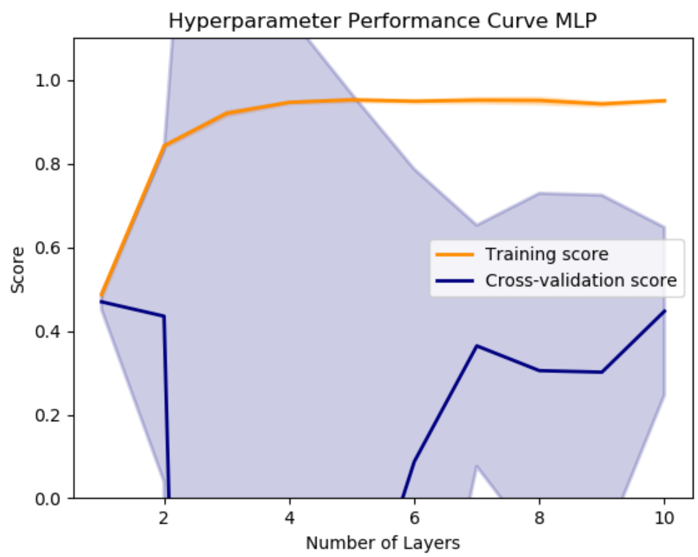


Fig 4.6 *MLP Layers Hyper-parameter Cross Performance Results. Comparison of the mean R^2 for the Train and Test data for a varying number of layers (fixed at 150 neurons per layer) on the seasonal Feature Select 1 set*

A total of 74 models were run using the MLP algorithm. The results of the cube search show that an MLP model with hyper-parameters of 1 layer and 150 neurons per layer, when running on the seasonal Select Feature 1 set, produced a mean R^2 of 0.470. While this was not the highest R^2 that the MLP models generated, this model performed best when evaluating both the R^2 and the Train-Test variance. The model with purely the highest mean R^2 of 0.566 was achieved on the de-seasonalized Feature Select 1 set with 8 hidden layers of 150 neurons each; however, this model demonstrated one of the highest variances between Training and Test data of any of the models tested, suggesting it was highly overfit.

4.2.2.3 Random Forest Results

The RF model results scored, on average, better than both SVR or MLP. The mean run time per model was 14 seconds, making it the fastest model to run, and substantially faster than both SVR or MLP. Figure 4.7 shows the combinatorial impact of the hyper-parameters (where "max depth" limits the depth and fit of the tree and "max features" defines the limit of features considered for each split) evaluated and the resultant performance for the feature set that produced the highest mean R^2 and lowest Train-Test variance of the eight feature sets evaluated.

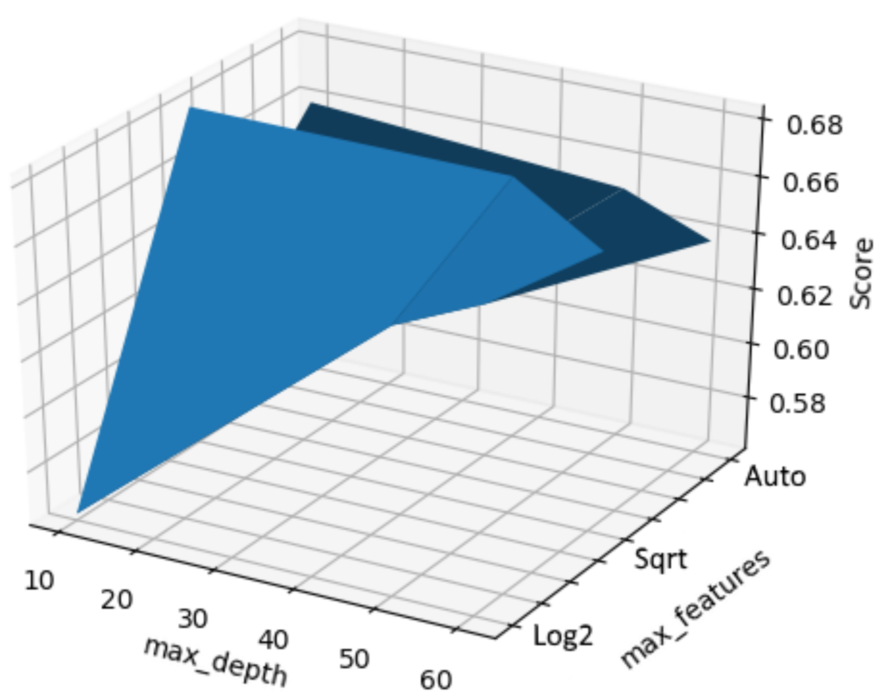


Fig 4.7 *RF Grid Search Results. Graphical representation of the mean R^2 for each set of hyper-parameters evaluated; maximum features and maximum feature depth on the de-seasonalized Feature Select 1 set*

A total of 72 models were run using the RF algorithm. The result of the cube search indicates that RF running on the de-seasonalized Feature Select 1 set and a feature depth of 10 (see Fig 4.8) and the maximum number of features equal to the square root of all features (see Fig 4.9), produced a mean R^2 of 0.682 and the lowest Train-Test variance of 0.09, half that of any of the combinations tested. This low Train-Test variance indicated that this model was the least likely to be overfit while having a mean R^2 value just 0.01 below the highest mean R^2 value result for any of the MLP models tested.

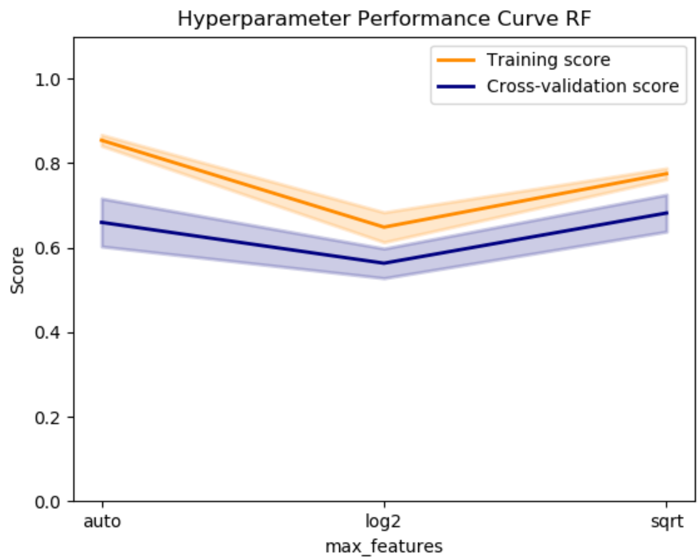


Fig 4.8 *RF Max Features Hyper-parameter Performance Curve. Comparison of Train-Test mean R^2 for varying values of the maximum number of features on the de-seasonalized Feature Set 1*

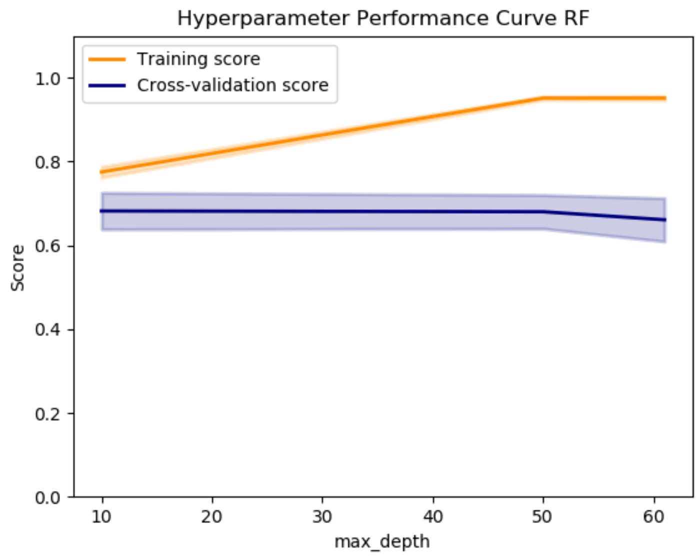


Fig 4.9 *RF Max Depth Hyper-parameter Performance Curve. Comparison of Train-Test mean R^2 for varying values of the maximum depth of the tree on the de-seasonalized Feature Set 1.*

4.2.2.4 Gradient Boosting Results

The gradient boosting model delivered consistently high mean R^2 values, outscoring other models tested on many of the feature sets. The mean run time per model was 7.5 minutes and it achieved the second highest mean R^2 value of 0.696, below only KNN. Figure 4.10 shows the combinatorial impact of the hyper-parameters (where “n estimators” sets the number of boosting stages and “min sample split” defines the samples required to split a node) evaluated and the resultant performance for the feature set that produced the highest mean R^2 and lowest Train-Test variance of the eight feature sets evaluated.

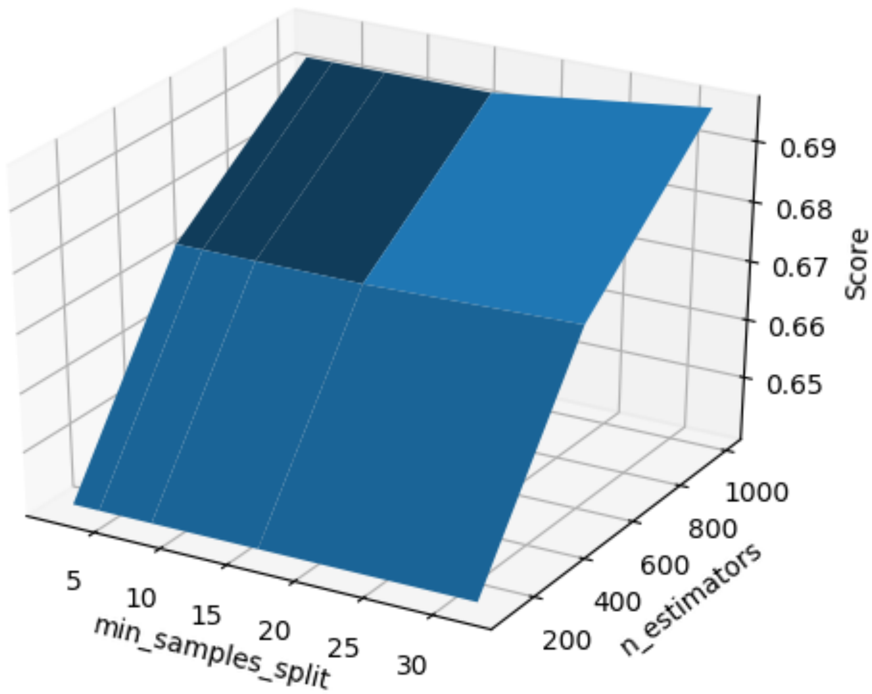


Fig 4.10 *GB Grid Search Results. Graphical representation of the mean R^2 for each set of hyper-parameters evaluated; minimum sample split and number of estimators on the de-seasonalized Baseline Feature set*

A total of 120 models were run using the GB algorithm. The result of the cube search indicates that GB running on the de-seasonalized baseline feature set with a minimum sample split of 32 (see Fig 4.11), and 1000 estimators (see Fig 4.12) produced the highest mean R^2 of any of the combinations tested. The model achieved the lowest Train-Test variance and also performed nearly as well on the mean R^2 when running on the de-seasonalized All Feature set with the minimum sample split was 32 and the number of estimators set to 100.

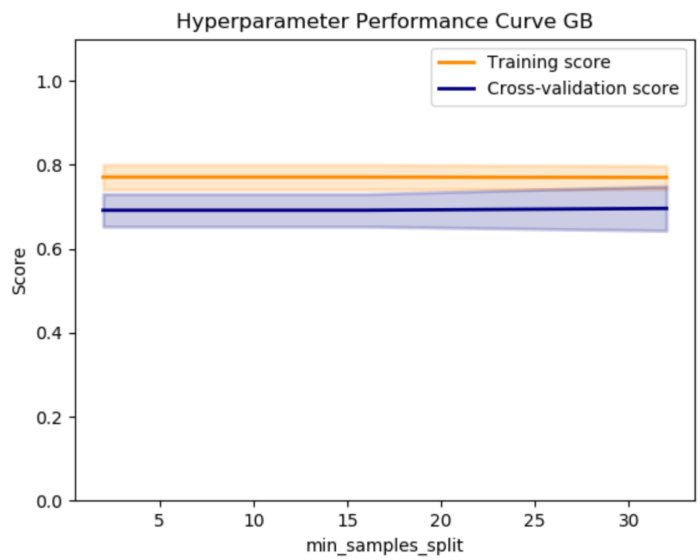


Fig 4.11 *GB Min Sample Split Hyper-parameter Performance Curve. Comparison of Train-Test mean* R^2 *for varying values of the minimum sample split on the de-seasonalized Baseline Feature set*

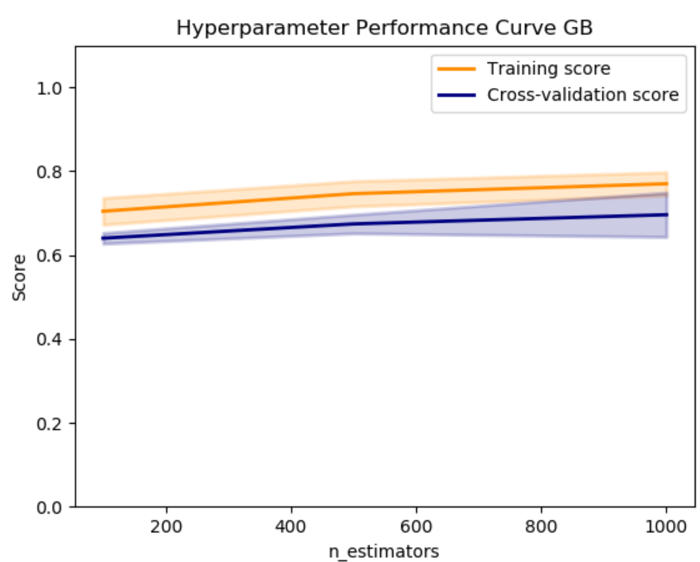
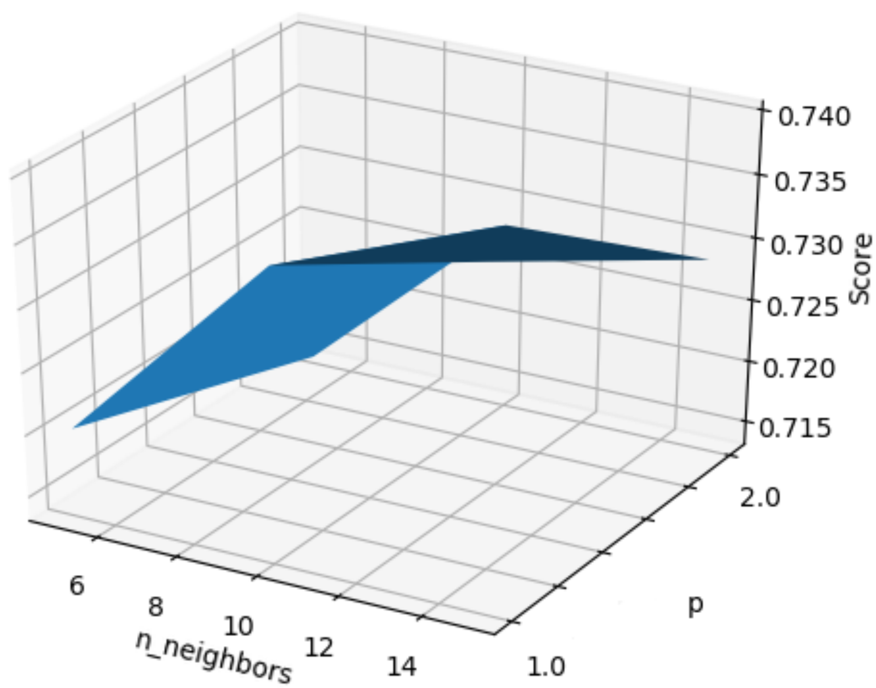


Fig 4.12 *GB N-Estimator Hyper-parameter Performance Curve. Comparison of Train-Test mean* R^2 *for varying values of the number of estimators on the de-seasonalized Baseline Feature set*

4.2.2.5 K-Nearest Neighbors Regressor Results

The KNN model achieved the highest mean R^2 of all models evaluated. The mean run time per model was 13 minutes, making it the third fastest model to run, behind RF and GB. Figure 4.13 shows the combinatorial impact of the hyper-parameters (where "p" is the Minkowski distance parameter and "n neighbors" determines the number of neighbors evaluated for each observation) evaluated and the resultant performance for the feature set that produced the highest mean R^2 and lowest Train-Test variance of the eight feature sets evaluated.



KNN 3D Error Results for Best Feature Set

Fig 4.13 *KNN Grid Search Results. Graphical representation of the mean R^2 for each set of hyper-parameters evaluated; number of neighbors and distance calculation on the de-seasonalized Feature Select 1 set*

A total of 48 models were run using the KNN algorithm. The result of the cube search indicates that KNN running on de-seasonalized Feature Select 1 set with the number of neighbors set to 15 (see Fig 4.14) and a Manhattan distance calculation (see Fig 4.15) produced the highest mean R^2 of any combinations tested. This model also achieved a very low Train-Test variance from cross-validation, indicating it is likely not overfit.

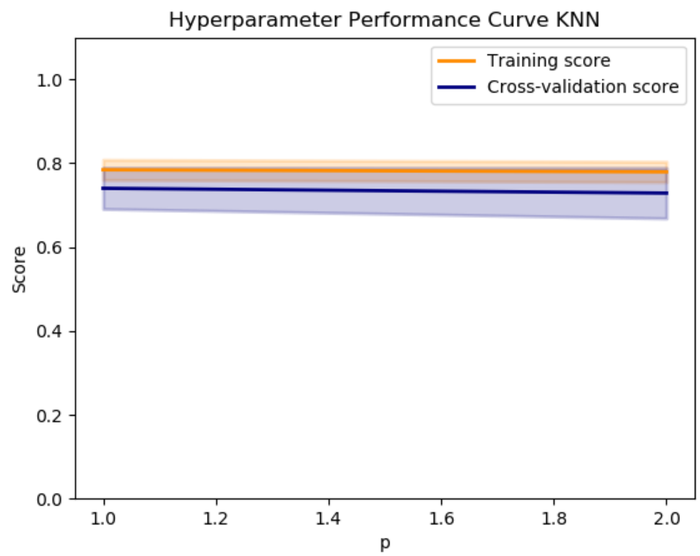


Fig 4.14 *KNN P Hyper-parameter Performance Curve. Comparison of Train-Test mean* R^2 *for the two different values of the p- parameter which determines the distance calculation on the de-seasonalized Feature Select 1 set*

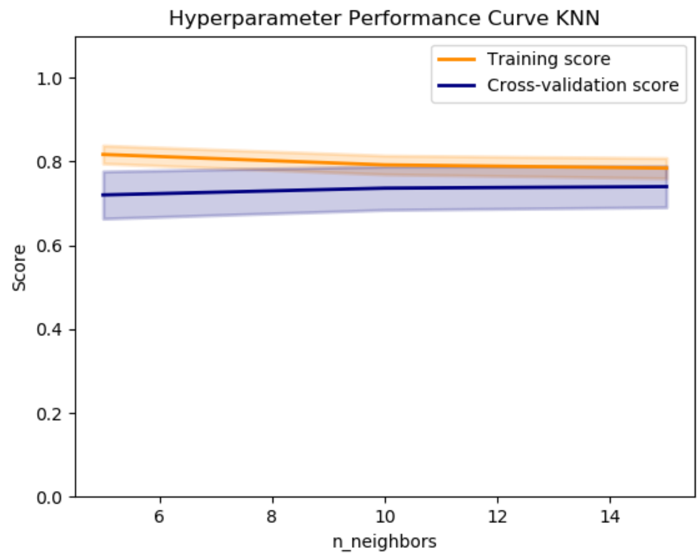


Fig 4.15 *KNN N-Neighbors Hyper-parameter Performance Curve. Comparison of Train-Test mean* R^2 *for varying values of the number of neighbors on the de-seasonalized Feature Select 1 set*

4.2.2.6 Model Performance Comparison

Support vector regression (SVR), multilayer perceptron (MLP), random forest (RF), gradient boosting (GB), and k-nearest neighbors (KNN) models were first compared using the hyper-parameters that yielded the highest mean R^2 for every feature set evaluated. Fig 4.16 visualizes the comparison of the mean R^2 for the best of each model for each feature set.

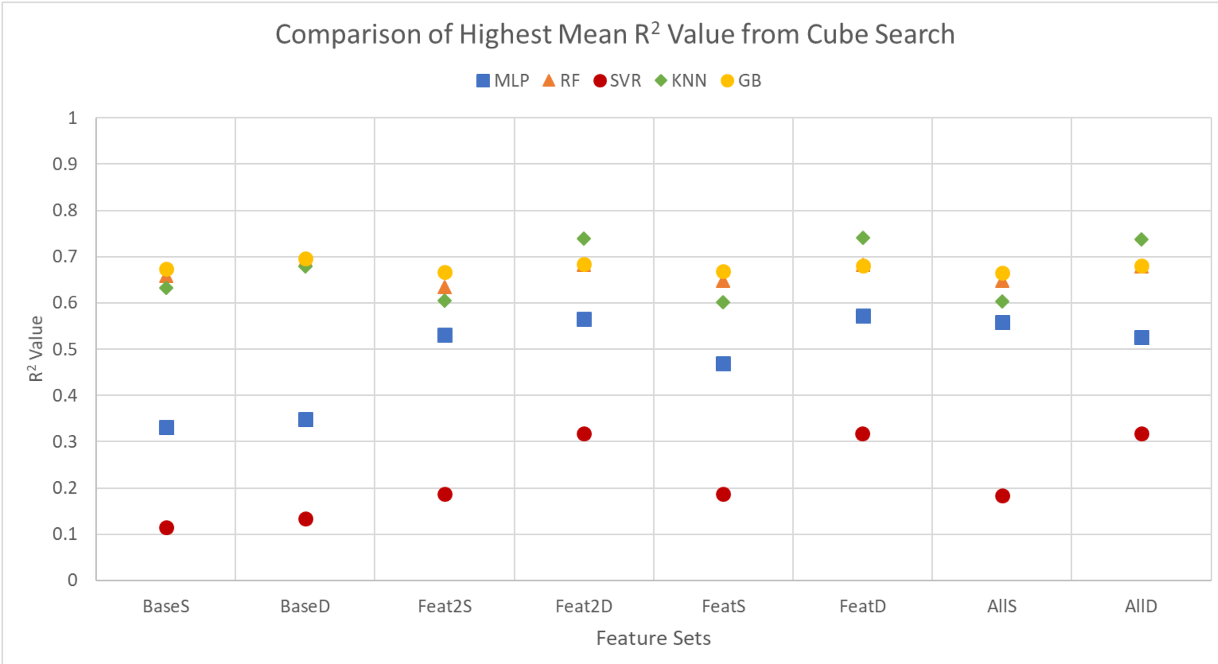


Fig 4.16 *Comparison of Error from Cube Search. This plot compares the highest mean R^2 value from each model for each feature set and each hyper-parameter evaluated in the cube search*

Based on the comparison, KNN consistently presented a higher R^2 value than all other models when additional features were added above baseline and the demand was de-seasonalized. In general, the highest mean R^2 values tended to come from models run on de-seasonalized feature sets. The addition of features tended to improve mean R^2 for most models above the Baseline feature set. MLP showed the largest increase in performance with the addition of features above Baseline. SVR was universally lower than any other model tested, while also requiring the longest run time.

The Train-Test variance was then evaluated for all models with the highest mean R^2 values to help resolve the differences between models with similar mean R^2 values as well as to help determine the likelihood of the model being overfit. Fig 4.17 below shows the comparison of the absolute variance between mean R^2 values of the Training and Test data sets during the cross-validation phase of the cube search for all features sets from the best models.

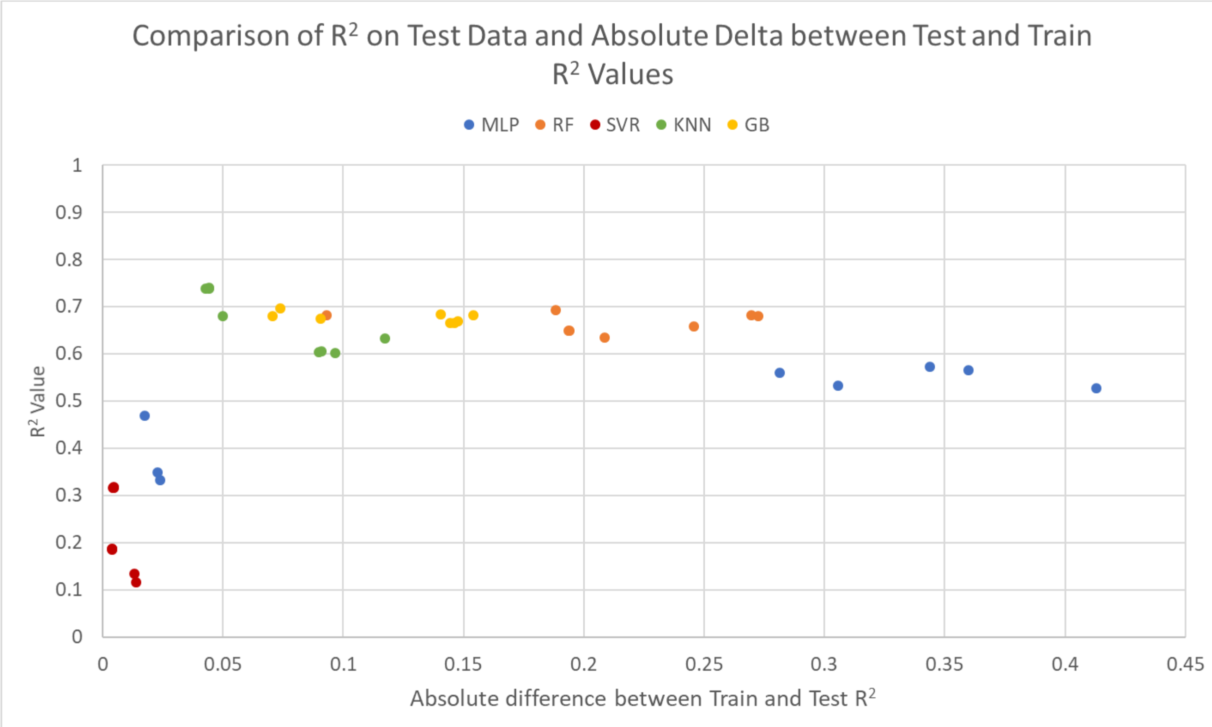


Fig 4.17 *Comparison of Train-Test Variance. The absolute difference between Test and Train mean R^2 values generated during cross-validation vs mean Test R^2 values. Lower deltas between test and train are considered less overfit, and thus indicate a better model*

4.3 Selected Model

The results of the cube search determined the model selected. The model and feature set that achieved a combination of the highest mean R^2 and lowest Train-Test variance of all models and combinations tested was determined to be the best model. This model with its corresponding hyper-parameters and feature set was then run on the unseen 2018 test data to generate the final forecast performance data that was compared to the current King's Hawaiian statistical forecast. The evaluation of the performance of this model was calculated using a custom loss function (WAPE) that matches the current King's Hawaiian's forecast error metric.

4.3.1 Model, Features, and Hyperparameters

The results showed that KNN achieved the highest mean R^2 while maintaining a low Train-Test variance, of any of the models tested. While KNN scored the highest mean R^2 value for de-seasonalized feature sets, it was outperformed by GB for seasonal feature sets (but still scoring higher than the remaining models). KNN performed better on de-seasonalized feature sets compared to seasonal (see Fig 4.18), and when additional features were included in the model.

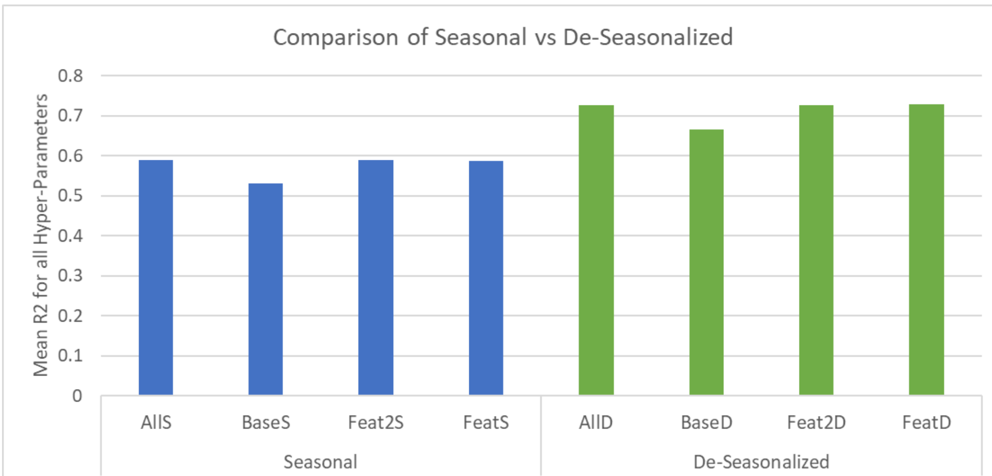


Fig 4.18 *KNN Seasonal vs De-Seasonalized Performance. Comparison* of mean *R^2 values for all KNN hyper-parameters tested on seasonal vs de-seasonalized feature sets*

The model achieved the best results when running on the de-seasonalized Feature Select 1 set. From the hyper-parameter ranges tested the settings that achieved the best results were when the number of neighbors was set to 15, and the distance was calculated with the Manhattan distance metric. The best KNN model achieved a mean R^2 of 0.74 on cross-validation during the cube search. The three-fold cross-validation showed a range of R^2 values from 0.304 to 0.800. The train time ranged from a minimum of 0.79 seconds to a maximum of 4.5 seconds, while the score time ranged from a minimum of 47.9 seconds to 153.2 seconds.

4.4 Forecast Accuracy

The selected model was used to predict the previously unseen data for 2018 as validation of model accuracy. The results of the selected model are discussed in section 4.4.1, along with some additional evaluation metrics. The ultimate goal of this research was to determine if machine learning can reduce the forecast error compared to a standard statistical model. A comparison of the performance of both models on the same 2018 data is included, along with the resultant financial impact as a result of changes to safety stock levels due to forecast error.

4.4.1 Error Measurement

The selected model was validated against the unseen 2018 test data set and used both R^2 as well as WAPE to determine its performance. Using the same R^2 evaluation metric as the three-fold cross-validation performed during the cube search, the selected model achieved an R^2 of 0.67. The results of the KNN model on the 2018 test data align with the mean R^2 values observed during cross-validation.

An additional error metric was used to evaluate the selected model to align it to the forecast error metric in practice at King's Hawaiian. Due to the relative disparities in the volumes sold of each product, and the highly seasonal patterns seen throughout the year, King's Hawaiian uses a weighted absolute percent error (WAPE) metric to address the total performance of the forecast. Since King's Hawaiian does not currently forecast at the resolution of state, and occasionally makes use of supplemental 3PLs in the same region, the forecast and actuals were aggregated to five primary distribution regions (see Fig 4.19). These distribution regions are aligned with the primary 3PLs in the King's Hawaiian network for evaluation purposes.

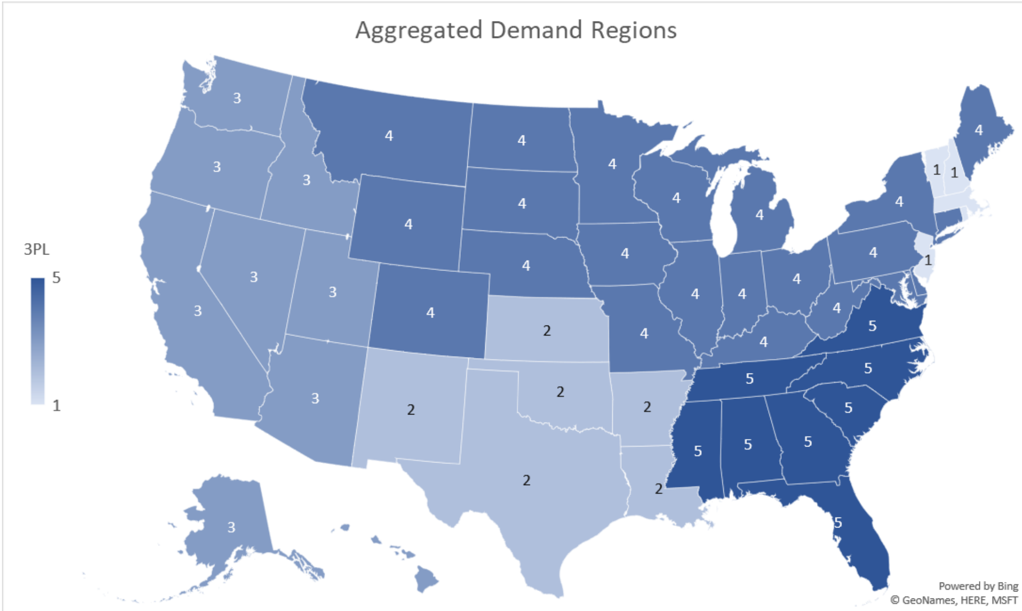


Fig 4.19 *Geographic Distribution Regions. The states included in each aggregated region for forecast evaluation and comparison*

The actual and forecasted shipments were first aggregated to these geographic distribution regions while still maintaining product and weekly resolution. The WAPE was calculated as the absolute percent error (|(Actuals – Forecast)|/Actuals) multiplied by the ratio of the specific demand to the annual demand. The specific demand is calculated at the same level of the forecast error (e.g. product and week specific for each aggregated geographic region). The final validation of the selected model resulted in an annual WAPE of 0.305 or 30.5%.

4.4.2 Error Comparison to Statistical

The current King's Hawaiian statistical model utilizes a Holt-Winters process that is trained on data from 2012. The model uses the same aggregate product categories for limited time offer products and product refreshes as used in this analysis. King's Hawaiian updates its model monthly as part of its S&OP process. Currently, the statistical forecast is run at the total national level only and uses a distribution matrix built from historic demand to disaggregate the forecast down to the resolution of the regional 3PL network. Actual shipments and King's Hawaiian regional forecast were aggregated from the resolution of

3PLs to the five distribution regions discussed in section 4.4.1 to achieve an equivalent comparison between the selected KNN model and the current King's Hawaiian regional forecast error.

The performance of the 2018 King's Hawaiian statistical model evaluated on the same product mix, geographic resolution, and weekly timeframe was a WAPE of 0.344 or 34.4% annually. The selected KNN model run on the de-seasonalized Feature Select 1 set achieved a forecast error (WAPE) 3.9% lower than the current King's Hawaiian model.

4.5 Inventory and Financial Impact

King's Hawaiian's regional forecast is at the weekly resolution; however, for calculating statistical safety stock they use the root of the mean of the weekly forecast error squared (e.g. root mean squared error or RMSE) for each region for each month. To evaluate the impact on safety stock, the RMSE of the KNN model was calculated at a monthly resolution to match the current safety stock methodology. As discussed in section 4.4.1, the geographic resolution was aggregated to distribution regions. The RMSE is then scaled by the corresponding standard deviation value associated with King's Hawaiian's 99.5% customer service level.

Using the forecast accuracy of King's Hawaiian's current statistical methodology, the average periodic pounds of the safety stock for the sum of all regional geographies and products selected equates to 6.5% of the total annual demand for the 2018 timeframe. By contrast, the same value when using the KNN model's forecast error is 5.8% (Fig 4.20).

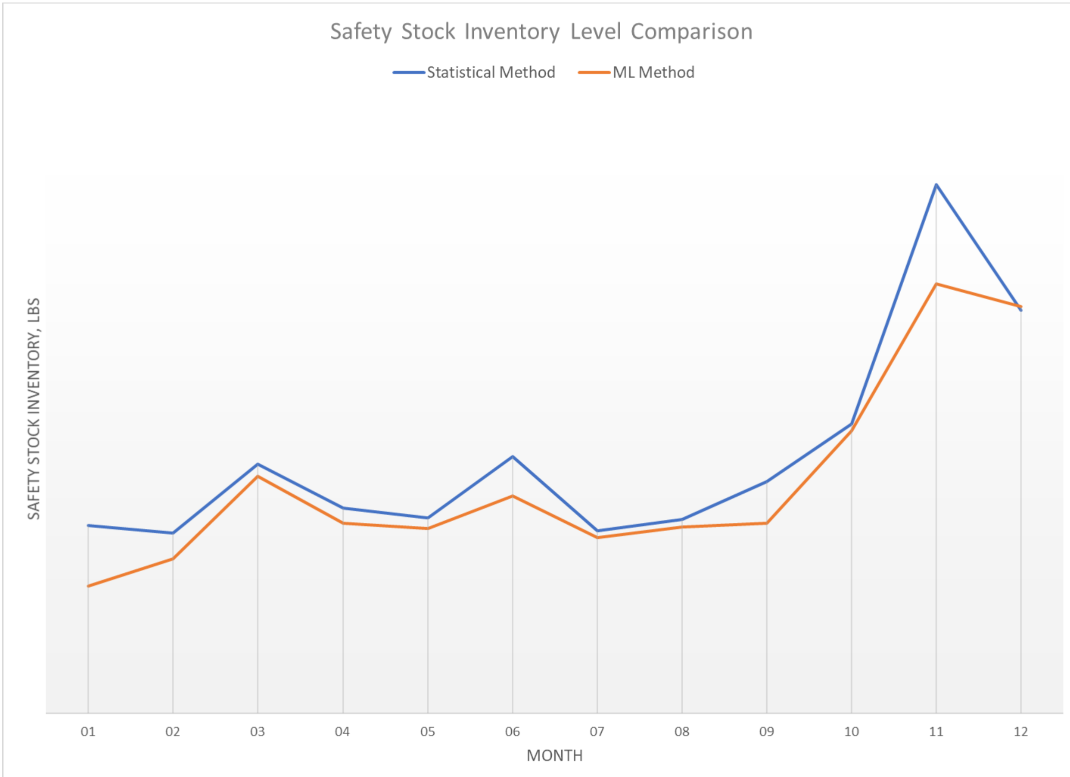


Fig 4.20 *Comparison of Safety Stock Levels for Statistical vs Machine Learning. Pounds of inventory required for safety stock as calculated from the RMSE of both King's Hawaiian statistical model and the final KNN machine learning model*

The machine learning forecast error driven equation represents a decrease in the total level of safety stock of 10% as a direct result from the increase in forecast accuracy of 3.9%. The change in safety stock equates to a proportional financial change of the same magnitude and can be converted to any financially relevant metric on a value per unit forecasted basis. While other associated positive financial impacts could be expected from decreases in the forecast error, they have not been considered in this analysis. Increased costs are incurred by running a machine learning model over a statistical model that can be built in Excel. The features determined to be the most important during feature selection were those attributes that are purchased from a third party regarding detailed consumption data. Additional data was included from free sources but require specialized personnel to process and integrate the data set. A CPG firm can expect to spend at least half a million dollars a year to support advanced analytics such as the selected

KNN model. Approximately 30-40% of this incremental cost will come from the collection of data, and 10-20% from the incremental compensation for employees with higher technical specialization to initialize and maintain the models. The remainder of the cost comes from the hosting and licensing fees associated with commercially available demand forecasting systems.

For a machine learning model to make financial sense for a company, the value it brings should exceed the incremental costs of running the model. The KNN model's lower forecast error results in a projected reduction in the value of safety stock carried of over $900k. The savings exceed the annual incremental costs of $500k in data, software and personnel support required to maintain the KNN model, justifying the pursuit of the more advanced forecasting methodology.

5. CONCLUSION

We believe the results of this research help frame the impact that more advanced analytical techniques, ones that integrate additional relevant attributes, can have at King's Hawaiian or similarly seasonal CPG businesses. The model selection process resolved models that were a good predictor of the business and those that were not, demonstrating that machine learning models are not universally better than traditional statistical methods. The variable feature sets showed that forecast accuracy tended to improve with the addition of relevant attributes. The improved demand forecast accuracy achieved with the KNN model had a beneficial impact on inventory, resulting in decreased safety stock. The inventory benefits demonstrate that the incremental costs associated with the more complex analytical technique, and the data required to run them, can be overcome by savings in other areas of the supply chain.

5.1 Feature Selection

The results of the cube search demonstrated the importance of considering additional features when forecasting demand. The varying degrees of performance from the same model on differing feature sets highlight the importance of feature selection when demand modeling, as well as what the business can glean from this process. The statistical feature selection using RF gives insight into the relative impact different regional characteristics have on demand. Such insights can be used to improve regional service levels or increase the impact of targeted promotions and marketing. Understanding the impact of a severe weather event can help planners react accordingly, improving operational readiness and visibility to financial risk.

The additional features beyond King's Hawaiian internal shipment data were collected from a variety of sources, both at-cost and free from government sources. It is not surprising that the feature selection assigned the highest importance to pricing and distribution attributes that came from the at-cost consumption data. However, the free-to-use census data accounted for 10% of the importance among the top 15 features. This demonstrates the insight that can be gleaned from available open data sets, and the value of incorporating these data in the demand forecasting models.

The feature selection element of the cube search can be expanded to evaluate additional attributes in the future that may provide deeper insight into customer behavior. Regional occurrences such as sporting events or local holidays can be included to better capture highly regionalized demand variation. Larger economic factors such as GDP or employment rates could be included to help capture fundamental economic trends that impact consumer behavior. The impact of product price changes could be evaluated for specific regions. While the nature of how these factors influence the customer may not always be clear, they can be efficiently evaluated to determine their impact on the predictive power of the model.

5.2 Model Selection

K-nearest neighbors regression achieved the best performance among the five models tested. It was also relatively fast to train and predict. Therefore, the model is a good candidate to be implemented as part of the S&OP. One cautionary point is that features should be normalized for KNN; otherwise, features with different magnitudes can result in poor model performance. Thus, proper feature engineering is critical. gradient boosting and random forest also showed promising results, with performance scores close to KNN. Conversely, the neural network and support vector machine models tested consistently performed worse than the other models.

Further investigation into the hyper-parameters of GB and RF models, and at a more granular level than the hyper-parameter tuning evaluated in this research, could result in models performing on par with the K-nearest neighbors models. For neural network models, further refinement on models with varying number of neurons per layer is suggested.

Another consideration for future modeling improvement would be the development of separate models for different segments of the data. For example, the evaluation and development of a different model for each SKU or region, instead of a single model for the entirety of the data.

5.3 Financial Impact

The financial impact due to changes in safety stock as a result of a change in forecast error is explained by the calculation (where the standard deviations of coverage is set by the customer service level target):

weighted average value of inventory x % change in forecast error x standard deviations of coverage

= $ change in safety stock investment

The formula shows the relationship that forecast error, and therefore demand uncertainty, have on statistically calculated safety stock. As demonstrated by this research, not every machine learning model will result in a higher performance forecast; therefore, each application requires a thorough evaluation to determine the best fit model, hyper-parameters, and feature data.

A company with similar characteristics to King's Hawaiian, a fast-moving seasonal CPG, should expect a similar degree of change when moving from a comparable statistical model to a machine learning model. Associated costs for a machine learning model should be roughly in line with those discussed in section 4.5 depending on the size of the company; however, the threshold of adoption of the model will depend on the individual company's weighted cost of inventory and other associated savings.

While the only financial impact evaluated was the change to safety stock, to build a more complete picture of the impact to the business, there are many more direct and indirect costs a company should include in their final analysis of a model. Some potential savings to consider include the cost of holding inventory, insurance on inventory, increases in operational capacity, reductions in obsolescence, reduction in transportation costs associated with regional deployment, and increases in the ability to meet customer service levels.

5.4 Further Investigation

We believe there is much more research to be done to understand the potential value machine learning models have to offer demand forecasting. Our research was limited to five machine learning model classes; however, there are many that may prove more predictive than current statistical models. Additionally, the cube search was restricted to two hyper-parameters over narrow ranges and eight feature sets. Further insights could be gained by evaluating the impact of different hyper-parameters over broader and more granular ranges and the inclusion of additional features for evaluation. As the costs for advanced analytical models continue to decline, and data continues to become easier to acquire and cheaper to store, the value of a machine learning approach to demand forecasting will only increase.
For King's Hawaiian, the implementation of such a model, even in a limited capacity for select SKUs, is financially justified. The next step in King's Hawaiian's adoption of more advanced analytic techniques is to build the processes around data collection and cleaning to support the continued use of the model created. The KNN model should be evaluated in parallel with their existing demand forecast for a trial period to build confidence in the model and help train the S&OP team on viable insights that can be collected from the model. While the current model was built to predict customer shipments, the model could easily be modified to predict consumption at the consumer level as well. Future versions of the model could expand down to the resolution of customer to improve understanding and help shape strategy with key customer accounts. King's Hawaiian should also explore additional features that it believes might be valuable in predicting demand for inclusion in future models and as part of the feature selection process. This feature selection process could prove valuable not only in improving the predictive accuracy of the forecast but in gaining insight into previously unrecognized drivers of demand.

REFERENCES

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5-6), 594-621.

Amirkolaii, K. N., Baboli, A., Shahzad, M. K., & Tonadre, R. (2017). Demand Forecasting for Irregular Demands in Business Aircraft Spare Parts Supply Chains by using Artificial Intelligence (AI). IFAC-PapersOnLine, 50(1), 15221-15226.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3), 1140-1154.

Chui, M., Manyika, J., Miremadi, M., Henke, C. N., Chung, L. R., & Valley, S. (2018). Notes from the AI frontier: Insights from hundreds of use cases.

Demir, L., & Akkaş, S. (2018). A comparison of sales forecasting methods for a feed company: A case study. Pamukkale University Journal of Engineering Sciences, 24(4).

Efendigil, T., Önüt, S., & Kahraman, C. (2009). A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. Expert Systems with Applications, 36(3), 6697-6707.

Gaur, M., Goel, S., & Jain, E. (2015). Comparison between Nearest Neighbours and Bayesian Network for demand forecasting in supply chain management. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1433-1436). IEEE.

Guanghui, W. (2012). Demand forecasting of supply chain based on support vector regression method. Procedia Engineering, 29, 280-284.

Hribar R., Potočnik, P., Šilc, J., & Papa, G. (2019). A comparison of models for forecasting the residential natural gas demand of an urban area. Energy, 167, 511-522.

Johansson, C., Bergkvist, M., Geysen, D., De Somer, O., Lavesson, N., & Vanhoudt, D. (2017). Operational demand forecasting in district heating systems using ensembles of online machine learning algorithms. Energy Procedia, 116, 208-216.

Kandananond, K. (2012). A comparison of various forecasting methods for autocorrelated time series. International Journal of Engineering Business Management, 4, 4.

Kharfan, M., & Chan, V. W. K. (2018). Forecasting Seasonal Footwear Demand Using Machine Learning.

Liu, T., & Fricke, L. M. (2018). Demand Forecasting of the Bike-sharing Service in Beijing.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS one, 13(3), e0194889.

Mupparaju, K., Soni, A., Gujela, P., & Lanham, M. A. (2018). A Comparative Study of Machine Learning Frameworks for Demand Forecasting. In 2018 Midwest Decision Sciences Institute Conference (Indianapolis, IN).

National Oceanic and Atmospheric Administration, National Centers for Environmental Information. (2019). Storm Events Database. Retrieved from https://www.ncdc.noaa.gov/stormevents/

Pezente, O. (2018). Predictive demand models in the food and agriculture sectors: an analysis of the current models and results of a novel approach using machine learning techniques with retail scanner data (Doctoral dissertation, Massachusetts Institute of Technology).

Saloux, E., & Candanedo, J. A. (2018). Forecasting District Heating Demand using Machine Learning Algorithms. Energy Procedia, 149, 59-68.

Slimani, I., El Farissi, I., & Achchab, S. (2015, December). Artificial neural networks for demand forecasting: application using Moroccan supermarket data. In 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA) (pp. 266-271). IEEE.

Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 46(1), 411-419.

Taghizadeh, E. (2017). Utilizing artificial neural networks to predict demand for weather-sensitive products at retail stores. arXiv preprint arXiv:1711.08325.

United States Census Bureau. (2019). American Fact Finder. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Vargas, C. A., & Cortés, M. E. (2017). Automobile spare-parts forecasting: A comparative study of time series methods. International Journal of Automotive & Mechanical Engineering, 14(1).

Vhatkar, S., & Dias, J. (2016). Oral-care goods sales forecasting using artificial neural network model. Procedia Computer Science, 79, 238-243.

Yu, X., Qi, Z., & Zhao, Y. (2013). Support vector regression for newspaper/magazine sales forecasting. Procedia Computer Science, 17, 1055-1062.