Forecasting Seasonal Footwear Demand Using Machine Learning

by
Majd Kharfan
Bachelor of Economics, Accounting, Damascus University, 2011
and
Vicky Wing Kei Chan
Bachelor of Business Administration, Global Supply Chain Management,
The Hong Kong Polytechnic University, 2011

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 2018

Signature of Author...
Majd Kharfan
Department of Supply Chain Management
May 11, 2018

Signature of Author...
Vicky Wing Kei Chan
Department of Supply Chain Management
May 11, 2018

Certified by...
Dr. Tugba Efendigil
Research Scientist, Center for Transportation and Logistics
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

Forecasting Seasonal Footwear Demand Using Machine Learning

by

Majd Kharfan

and

Vicky Wing Kei Chan

Submitted to the Program in Supply Chain Management

on May 11, 2018 in Partial Fulfillment of the

Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

The fashion industry has been facing many challenges when it comes to forecasting demand for new products. The macroeconomic shifts in the industry have contributed to short product lifecycles and the obsolescence of the retail calendar, and consequently an increase in demand variability. This project tackles this problem from a demand forecasting perspective by recommending two frameworks leveraging machine learning techniques that help fashion retailers in forecasting demand for new products. The point-of-sale (POS) data of a leading U.S.-based footwear retailer was analyzed to identify significant predictor variables influencing demand for footwear products. These variables were then used to build two models, a general model and a three-step model, utilizing product, calendar and price attributes for predicting demand. Clustering and classification were used under the three-step model to identify look-alike products. Regression trees, random forests, *k*-nearest neighbors, linear regression and neural networks were used in building the prediction models. The results show that the two forecasting models based on machine learning techniques achieve better forecast accuracy compared to the company's current performance. In addition, the proposed methodology offers visibility into the underlying factors that impact demand, with insights into the importance of the different predictor variables and their influence on forecast accuracy. Finally, the project results demonstrate the value of forecast customization based on product characteristics.

Capstone Advisor: Dr. Tugba Efendigil

Title: Research Scientist, Center for Transportation and Logistics

Acknowledgements

We would like to thank the MIT family of the Supply Chain Management program for giving us the opportunity to challenge ourselves and to broaden our experience. Special thanks go to our advisor, Dr. Tugba Efendigil, for her mentorship and guidance throughout this project. Also, we would like to extend our sincerest gratitude to our sponsoring company. In particular: Shruti, Monica, Stephen, Dan, Luis, Daniele and others who challenged us and were there to answer our questions.

Vicky & Majd

I feel honored and blessed for everything I have learnt and experienced while working on this project at MIT. I want to acknowledge the efforts of my wonderful partner, Vicky. Without her, this work would not have been possible. Also, a huge thanks to my beautiful mom, Nisreen, and my lovely wife, Afnan, for their patience and support throughout this journey.

Majd

I am grateful for the learning opportunities that I have been given while working on this project at MIT. I especially want to thank my amazing project partner, Majd, for his dedication and efforts to this project. I would also like to thank my family and friends who always support me in everything I do.

Vicky

Table of Contents

List of Tables

List of Figures

1. Introduction

This section provides a high-level overview of the state of the retail fashion industry. It discusses how agile supply chain strategies can enable fashion companies to adapt to current trends. Finally, it highlights the essential role of demand forecasting in supporting agile supply chain strategies and the optimization of other business functions.

1.1 Overview of the Retail Fashion Industry

The fashion industry has immensely evolved in the past few decades, especially after the introduction of e-commerce. Consumers' taste has become the major demand driver for fashion products. It is generally influenced by internal and external factors including personalization, omni-channel competition, social media influencers, political movements and others. This continuous change in consumers' behavior has led to shorter product lifecycles and more volatile demand. In addition, consumer expectations have become greater as high quality, guaranteed availability and fast delivery are no longer negotiable.

With all these challenges, fashion companies must develop overarching strategies that are adaptable to the constant changes in the industry. Such strategies should embrace marketing as a demand creation tool and digital capabilities like e-commerce and mobile apps as growth enablers. Innovation and speed to market are other important features a strategy should focus on. These features help companies stay competitive in today's global market where brands like Zara and H&M refresh their assortment every few weeks.

Once such an overall strategy is set, the role of an agile supply chain strategy that focuses on responsiveness, competency, flexibility and quickness comes into play. An agile supply chain will work as an enabler and executor through a number of aligned initiatives that collectively work toward achieving the company's objectives. Examples of such initiatives may include manufacturing lead times, which can be fostered by applying ABC analysis to discover and solve bottlenecks in the process system. Development of postponement strategies through staging materials or semi-finished products at distribution centers (DCs) or factories is also essential to provide flexibility and give the company extra time to see better market signals. Moreover, inventory policies need to be visited to ensure safety stock and order quantity parameters are set based on statistical analysis that considers the trade-offs between cost and level of service.

Having an agile supply chain cannot be accomplished without optimizing demand planning and especially demand forecasting, which will be the area of focus throughout this research project. Demand forecasting can be defined as the art and science of predicting customers' future demand for products. It serves as a major input for planning across different supply chain and business functions, including raw materials planning, supply planning, inventory management, sales and merchandising. Poor forecasting results can lead to stock outs and loss in revenues and market share to competitors, or to excessive inventory, i.e., frozen capital and high obsolescence. Therefore, having good demand forecasting capability is essential to optimize other functions and to support the overall supply chain and company's strategies.

1.2 The Company and Motivation

The sponsoring company for this research is a major footwear manufacturer and retailer based in the US with operations across the globe. The company sells its products through its own inline (full price) and outlet (discounted price) brick-and-mortar retail stores as well as through its online website and wholesale partners.

The United States is the largest market for this company and the scope of this research. Like other fashion retailers, the sponsoring company is at the crossroads of two key macro shifts: the "Buy Now/Wear Now" consumer mentality influenced by social media and the love of personalization, and the economic challenges facing the retail industry in the form of declining mall traffic and the obsolescence of the traditional retail calendar. With that in mind, the company is reworking its strategy to improve its position in the marketplace by becoming closer to consumers and quicker in responding accurately to demand signals. This will consequently bring to the company operational efficiencies in the form of minimized order cancellation rates and healthier levels of inventory in the marketplace, which will be translated into cost savings and additional revenues.

Through carrying out this research project we aim to recommend solutions to the sponsoring company that will improve the demand forecasting capabilities and prediction accuracy. Applying machine learning will maximize the utilization of the point-of-sale (POS) data and help uncover new insights to be used in developing a demand forecasting framework that meets the company's strategic objectives.

2. Literature Review

This section explores the demand forecasting methods and common predictor variables that have been used in industry, compares traditional and machine learning forecasting techniques, and reviews the application of machine learning techniques in different industries. This information sets the stage based on which we built our forecasting models through selecting appropriate predictor variables and using suitable techniques.

2.1 Demand Forecasting Methods

Demand forecasting in the apparel and footwear industry is extremely challenging due to volatile demand, strong seasonality, Stock-keeping-unit (SKU) intensity and for seasonal and fashion items, short lifecycles and lack of historical data (Thomassey, 2010). Consumer demand is the result of the interplay among a number of factors, which ideally should serve as predictor variables in generating demand forecasts. However, in practice sometimes the effect of these factors can be difficult to decouple. For example, price and seasonality are interdependent on each other (Kaya, Yeşil, Dodurka & Sıradağ, 2014). Traditional forecasting methods usually only take into account a single factor or at most a few factors, so part of the variation remains unexplained in the forecasting model when in fact there may be patterns undiscovered. In this research, different machine learning based forecasting techniques will be explored to identify the most suitable approach for the sponsoring company.

2.2 Predictor Variables in Demand Forecasting

The most common type of data used in demand forecasting is POS data, or downstream data, which is widely used in both traditional time-series forecasting and advanced machine learning techniques. For retailers, POS data are usually readily available and relatively accessible, as they are automatically captured at consumers' checkout upon each purchase transaction.

Wholesalers and manufacturers depend on their downstream retail partners for visibility to POS data.

In addition to POS data, there are many other types of data that are being used in industry or proposed in academic research papers in demand forecasting. One important type of data is lost sales. Demand that is not satisfied because of stock-outs is not captured in POS data and results in potential lost sales. In such cases, true demand may be underestimated if sales are treated as being equal to demand (Kaya et al., 2014). Therefore, lost sales need to be taken into account during the forecasting process to reflect true historical demand. Other types of data include price and promotion, consumer loyalty, calendar and holidays, weather, geographic location, competition, item features, fashion trends, store count and mode of distribution, as well as macro-economic trend data such as purchasing power and unemployment rate (Thomassey, 2010). These types of data lead to a large number of decision variables to be explored in improving forecasting accuracy. Some factors are believed to have more impact compared to the others. For example, in building a demand signal repository (DSR) for a fast-moving consumer goods (FMCG) company, Rashad and Spraggon (2013) found year, month, weekday and holidays to be the most significant factors in shaping demand out of the many variables studied.

2.3 Traditional Techniques vs. Machine Learning Techniques

For the past few decades, traditional forecasting methods, including time series (extrapolatory) and regression (explanatory) techniques, have been widely used in demand forecasting. Naïve, moving average, trend, multiple linear regression, Holt-Winters, exponential smoothing and ARIMA are among these traditional techniques. Recently, their performance has been used in research to benchmark against those of advanced machine learning techniques, which have gained attention and popularity in recent years due to the advancement in technology. For example, Carbonneau, Laframboise & Vahidov (2008) performed studies on the application of machine learning techniques such as support vector machine (SVM) and neural networks on demand forecasting and compared the results with traditional methods including naïve, trend, moving average and linear regression.

The emergence of big data, cloud computing and improved computing storage and processing capabilities has led to increased availability and accessibility to large volumes of data, making advanced machine learning techniques a viable option for demand forecasting in the industry.

Traditional and machine learning techniques differ in their capabilities and requirements. Traditional time series and regression techniques normally consider either a single or a few variables such as trend, seasonality and cycle. Machine learning-based techniques are able to process an unlimited number of predictor variables, determining the ones that are significant. The data source for traditional demand forecasting is mainly from demand history, while machine learning-based techniques can make use of limitless data sources. However, this also means that machine learning-based techniques are more reliant on the availability of data. The more data there are, the better the learning will be. In traditional approaches, multiple single-dimension algorithms are used separately for different product styles or categories based on different data constraints. Thus, more manual data manipulation and cleansing work is required and the algorithms are less generalizable. In machine learning, an array of general algorithms is used to fit demand patterns across the entire product portfolio, creating a synchronized and integrated forecast. In terms of technology requirements, machine learning is much more dependent on computing power than traditional methods and may therefore be costlier to implement.

Machine learning and predictive analytics provide an advantage over traditional forecasting methods that use only limited demand factors to create more accurate demand forecasts. Machine learning-based forecasting combines learning algorithms to identify underlying demand drivers and uncover insights (Chase, 2017). Table 1 summarizes the comparison between traditional and machine learning forecasting approaches.

Table 1. Comparison between Traditional and Machine Learning Forecasting Approaches

	Traditional Forecasting	**Machine Learning Forecasting**
Number of predictor variables	Single or a few	Unlimited
Data source	Mainly demand history	Multiple
Algorithms	A number of single-dimension algorithms	An array of integrated algorithms
Manual data manipulation and cleansing need	High	Low
Data requirements	Low	High
Technology requirements	Low	High

2.4 Application of Machine Learning Techniques in Industry

Machine learning techniques that have been applied in demand forecasting in research or practice in the fashion apparel industry include neural networks, support vector machine (SVM), fuzzy inference system (FIS), extreme learning machine (ELM), extended extreme learning machine (EELM), harmony search (HS) algorithm and grey method (GM). In addition, a hybrid combining different techniques tend to perform better than a single method. For example, Wong & Guo (2010) proposed a model combining ELM and HS algorithm. The proposed model performed much better than the traditional ARIMA model and certain other neural networks models in making medium-term forecasts. Choi et al. (2014) also proposed a hybrid model that produced satisfactory forecast accuracy results by utilizing a combination of EELM and GM. Table 2 shows the industries that each technique has been applied to, the preferred input variables and the forecasting horizon.

Table 2. Applications of Machine Learning Techniques in Forecasting in Different Industries

Machine Learning Technique	Industry	Variables	Horizon
Neural Networks (NN)	Apparel Fashion, FMCG, Medical Products (Thomassey, 2010; Vhatkar & Dias, 2016)	POS, Order (Shipment), Product attributes, Consumer attributes	Short term
Fuzzy Inference System (FIS)	Apparel Fashion, Financial Forecast (Stocks), Technology Assessment (Thomassey, 2010; Kaya et al., 2014)	Price, Holidays, Period/ Season, Financial time series, Patent data, Publication data and market research reports	Long term
Support Vector Machine (SVM)	FMCG, Consumer Electronics (Pillo, Latorre, Lucidi, & Procacci, 2016; Lu, 2014)	Promotion, Number of opening hours, Price and number of daily receipts (Forecast), Month, Day of the month, Day of the week, POS	Short term
Harmony Search (HS)	Apparel Fashion (Wong & Guo, 2010)	POS	Medium term
Grey Method (GM)	Apparel Fast Fashion (Choi et al., 2014)	POS	Short term
Decision Trees	Apparel Fast Fashion (Thomassey, 2010)	Prototypes of sales, Descriptive criteria of historical items	Long term
Extreme Learning Machine (ELM)	Apparel Fast Fashion, Apparel Fashion (Wong & Guo, 2010; Choi et al., 2014)	Long term forecasts, Last sales (At least 2 weeks)	Short/ Medium term
k-Means Clustering	Apparel Fast Fashion (Thomassey, 2010)	Historical sales of a range of products	Long term
Multivariate Adaptive Regression Splines (MARS)	Consumer Electronics (Lu, 2014)	POS (Sales amount, Trend, Growth ratios, Volatility)	

3. Methodology

This section explains how we used the data collected to identify significant predictor variables of sales and build the forecasting models. The objective is to find out how the data can be leveraged to improve the demand forecasting capability, especially for seasonal products without sales history. This section is structured as follows: We first describe the types of machine learning methods used in feature selection and forecasting model building, and define the scope and granularity of the data involved. We then move on to describe the process of feature engineering and selection, and finally outline the steps in building two forecasting models: the general model and the three-step model. The flow of the methodology is laid out in Figure 1.

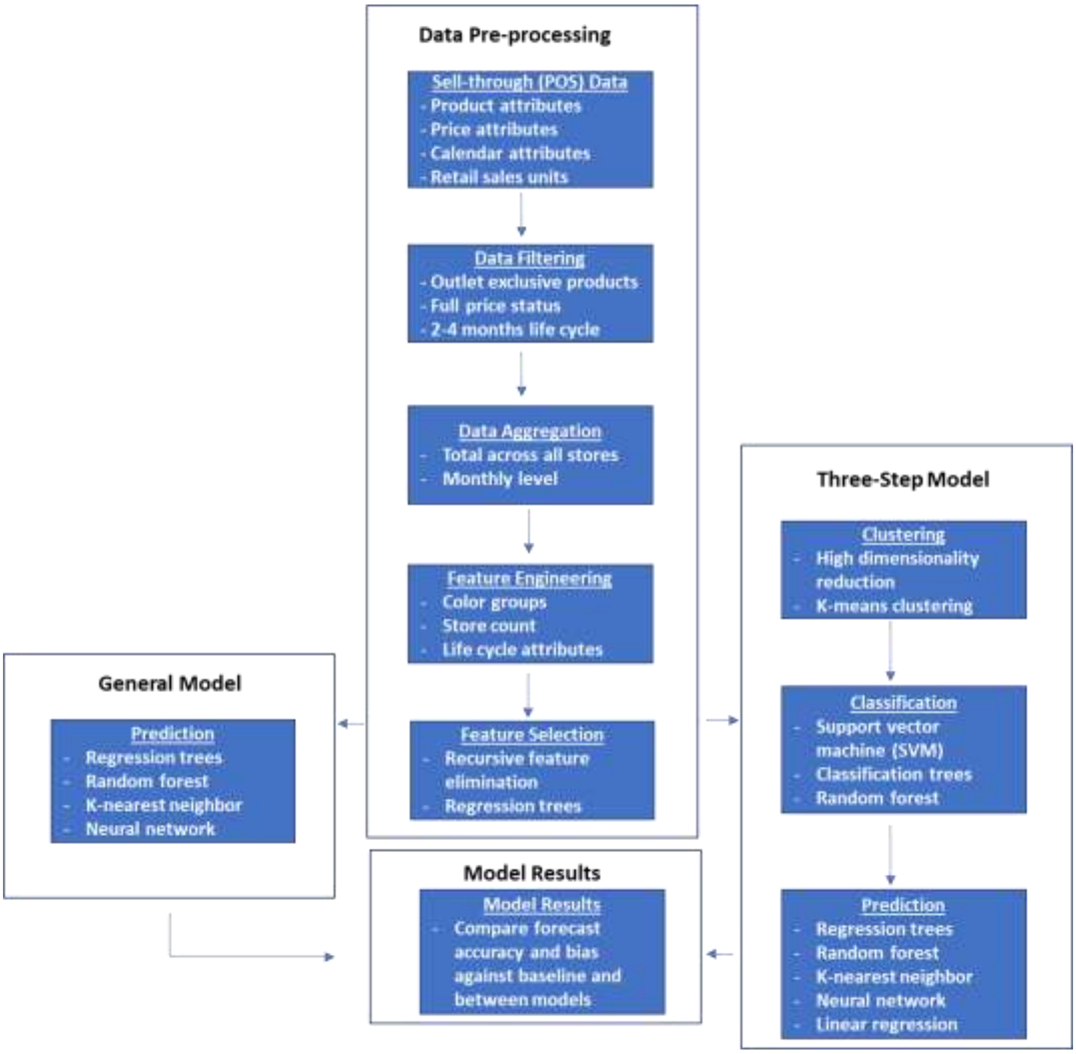

Figure 1. The Proposed Methodology

3.1 Machine Learning Techniques Used

This sub-section describes the types of machine learning techniques used in feature selection and model building.

3.1.1 Supervised Learning Techniques

Supervised learning provides an algorithm with records that have a known output variable. The algorithm “learns” how to predict this value with new records where the output is unknown. The definition of each supervised learning technique used is listed below (Shmueli, Bruce, Yahav, Patel, & Lichtendahl, 2018).

Regression and Classification Trees: Trees separate records into more homogeneous subgroups in terms of the outcome variable by creating splits on predictors, thereby creating prediction or classification rules. These splits create logical rules that are transparent and easily understandable.

Random Forests: Random Forests combine the predictions or classifications from individual trees by drawing random samples from the data and using a random subset of the predictors at each run. The results are obtained either through voting for classification or averaging for prediction.

Neural Networks: Neural networks mimic how human brain works and combine the predictor information in a very flexible way that captures complex non-linear relationships among variables. In neural networks, the user does not need to specify the correct form of relationship. Instead, the network tries to learn about such relationships from the data. A feedforward neural network consists of an input layer with nodes that accept predictor values, hidden layers that receive inputs from previous layers and perform non-linear transformation, and finally an output layer that classifies or predicts the outcome variable.

k-Nearest Neighbor (k-NN): *k*-NN classify or predict a new record by finding "similar" records in the training data. *k*-NN identifies *k* records in the training data that are closest to the new record in terms of predictor variables to derive a classification or prediction for the new record by voting (for classification) or averaging (for prediction).

3.1.2 Unsupervised Learning Techniques

Unsupervised learning attempts to learn patterns in the data rather than predicting an output value. In other words, there is no "correct answer" for the outcome. The definition of each unsupervised learning technique used is described as follows (Shmueli et al., 2018).

k-Means Clustering: *k*-means clustering divides the data into a predetermined number *k* of non-overlapping homogeneous clusters by minimizing a measure of dispersion within the clusters. A common measure of within-cluster dispersion is the sum of distances (or sum of squared Euclidean distances) of records from their cluster centroid.

t-distributed Stochastic Neighbor Embedding (t-SNE): This algorithm is one of the manifold learning techniques. It is used to reduce the dimensionality of the data non-linearly, in a way that helps visualizing the data points on a Euclidean space.

3.2 Scope and Granularity of Data

Two types of data were collected from the company: sell-in (shipment) and sell-through (POS) data. The POS data collected were at the daily style-location level from 115 retail outlet stores and include product attributes, calendar attributes, store attributes, price and promotion attributes as well as the sales units. The total number of records in the sell-through data is 13,295,485, spanning a total of nine and a half seasons from July 2013 to March 2018. The Spring/Summer season consists of January to June while the Fall/Holiday season consists of July to December. Since the focus of this project is to support the decision of how much of each style to order from the manufacturer for the whole season, the data were aggregated to the level at which this decision is made; i.e., across all stores at the monthly level. The list of attributes of the aggregated data is shown in Table 3.

Table 3. List of Attributes from the Aggregated Data by Month at the Style Level

Variable category	Variable	Description
Meta Data	Style	Unique style number of each product
Meta Data	Style Description	Description of the style
Calendar	Year	Fiscal year
Calendar	Month	Fiscal month
Product Attributes	Color	Color code
Product Attributes	Basic Material	Type of upper material
Product Attributes	Gender	Gender or age group description
Product Attributes	Category	Product family
Product Attributes	Sub-Category	Classic vs modern
Product Attributes	Retail Outlet Sub-Department	Basic vs. seasonal
Product Attributes	Cut	Ankle height
Product Attributes	Pillar	Product sub-brand
Product Attributes	Product Class	Product main feature
Price and Promotion	Price Status	Full-price vs mark-down
Price and Promotion	Manufacturer's Suggested Retail Price (MSRP)	Ticket price
Price and Promotion	Average Unit Retail (AUR)	Actual selling price
Sales Units	Retail Sales Units (Target variable)	Retail sales units

The products sold at outlet stores may either be discounted products from regular inline stores or products made exclusively for launching at the outlet stores. In the context of demand forecasting, we were only interested in the latter category. In addition, products with excess inventory after the intended product lifecycle are discounted, and this distorts the demand. Meanwhile, the sponsoring company is specifically interested in studying seasonal products which typically have an intended lifecycle of 2 – 4 months. Therefore, records were removed accordingly so that only records for outlet-exclusive products with full-price status and a product lifecycle of 1 – 4 months were included in our analysis. In this case, product lifecycle was estimated based on the POS data by counting the number of consecutive months with full price sales records for a particular style. The data were pre-processed, filtered and aggregated as described above using Alteryx software package.

3.3 Feature Selection and Engineering

Some features were modified or added in preparation for building the model. There are many unique observations under the attribute color, some of which are very similar. In order to make this attribute more meaningful, colors were aggregated into groups based on similarities. Because it is commonly cited as one of the predictor variables in demand forecast, store count was added as a candidate variable. It refers to the number of stores at which a style was sold, which was estimated using sales record.

Pillar and Category are similar attributes with one-to-one relationship; i.e., they are completely correlated with each other. Therefore, Pillar was dropped as Category already captured the same information. The Retail Outlet Sub-department is the same across all seasonal styles and was therefore dropped as well.

For building the forecasting model, three variables related to product lifecycle were added:

lifecycle, lifecycle month and lifecycle start month. As seasonal styles are launched at different times of the year with short lifecycles, their sales are believed to be dependent on the lifecycle attributes in addition to the calendar attributes; i.e., sales are not only related to which calendar month the sale occurs in, but also to which month the product is launched. Lifecycle refers to the total number of months in the lifecycle of a style. Lifecycle month refers to the number of months since product launch. Lifecycle start month refers to the month the lifecycle started in. The complete list of attributes subsequently being considered in the feature selection process is shown in Table 4.

Table 4. List of Attributes for Feature Selection

Variable Category	Variable	Description
Meta Data	Style	Unique style number of each product
Meta Data	Style Description	Description of the style
Calendar	Year	Fiscal year
Calendar	Month	Fiscal month
Product Attributes	Color Group	Color code
Product Attributes	Basic Material	Type of material
Product Attributes	Gender	Gender or age group description
Product Attributes	Category	Product family
Product Attributes	Sub-Category	Classic vs. modern
Product Attributes	Cut	Ankle height
Product Attributes	Product Class	Product main feature
Price and Promotion	Manufacturer's Suggested Retail Price (MSRP)	Ticket price
Price and Promotion	Average Unit Retail (AUR)	Actual selling price
Lifecycle	Lifecycle	The total number of months in the lifecycle of a style
Lifecycle	Lifecycle Month	The number of months since product launch
Lifecycle	Lifecycle Start Month	The month at which the lifecycle started
Store	Store Count	Number of stores selling a style
Sales Units	Retail Sales Units (Target variable)	Retail sales units

Recursive feature elimination, a backward feature selection method, was used to eliminate features based on their contribution to improving forecast accuracy. A random forests algorithm was used on each iteration to evaluate the model with different subsets of the 14 input variables. A 10-fold cross-validation on the training data was used. Random forests was selected in view of its capability in handling multi-collinearity.

3.4 Dataset Partitioning

For building the general model, the data were partitioned into training and validation sets. The data from the first six seasons (Fall/Holiday 2013 – Spring/Summer 2016) were used as training set for building the model while the data for the next three seasons (Fall/Holiday 2016 – Fall/Holiday 2017) were used as validation set for measuring the predictive performance of the model. The number of styles and records in each data set is listed in Table 5.

Table 5. Overview of Datasets Generated for the General Model

Dataset	Months of sales	Number of Styles	Number of records
Training	36	578	1796
Validation	18	195	560

For the three-step model, we split the database into three sets, a training set, a validation set, and a testing set. For simplicity, the sponsoring company's fiscal year (June – May) was the factor used to split the data. The training set included all the sales records occurred before fiscal year 2017, except for products with sales overlap in both fiscal years 2016 and 2017, which were allocated to the validation set. For example, the records of a style that started selling in April, fiscal year 2016 and continued selling through July, fiscal year 2017 was entirely moved to the validation dataset to prevent data overlap. The validation set covered the sales records in fiscal year 2017 and the overlap from 2016 plus seven months of records from fiscal year 2018.

The testing set included three months of records from fiscal year 2018. Table 6 gives an overview of the three datasets generated for the three-step model.

Table 6. Overview of Datasets Generated for the Three-Step Model

Dataset	Months of sales	Number of Styles	Number of records
Training	35	539	1558
Validation	19	201	591
Testing	3	58	155

3.5 Model Building

3.5.1 General Model

For seasonal styles without sales history, we built a general model utilizing product attributes, calendar attributes, lifecycle attributes, store count and price attributes selected from the feature selection process as described in Section 3.4. We explored using regression trees, random forests, *k*-nearest neighbor (*k*-NN) and neural networks to build the model. In addition, ensemble methods taking the median and average of the outputs from the four individual methods were also considered. The prediction results from each method are compared in Chapter 4.1.

3.5.2 Three-Step Model

The three-step model can be distinguished from the general model that it consists of three separate stages: (*i*) clustering, (*ii*) classification, and (*iii*) prediction. The main objective behind this model is to identify look-alike group of products from the training set. Once these products are identified, their average sales can be used as a proxy to forecast the sales for brand-new products in both the validation and testing sets.

In a similar fashion to the general model, the initial variables used in the three-step model were those that resulted from the feature selection process. However, these variables were mixed differently across the three stages. Additionally, two new variables were created from clustering and then used in classification and prediction. Cluster number in the training set refers to the cluster to which a style belongs. The average sales variable is calculated for a group of products that belong to the same cluster and share similar lifecycle and calendar attributes. A complete list of the variables considered for each stage is presented in Table 7.

3.5.2.1 Clustering

The main objective for the clustering stage was to partition and group all the seasonal styles in clusters based on similarities across eight different attributes. The targeted data were a combination of both the training and the validation sets. The only reason for including the validation set was to later test the classification performance on a dataset (the validation set) that had pre-assigned clusters. The clustering stage included four main sub-steps as illustrated in Figure 2.

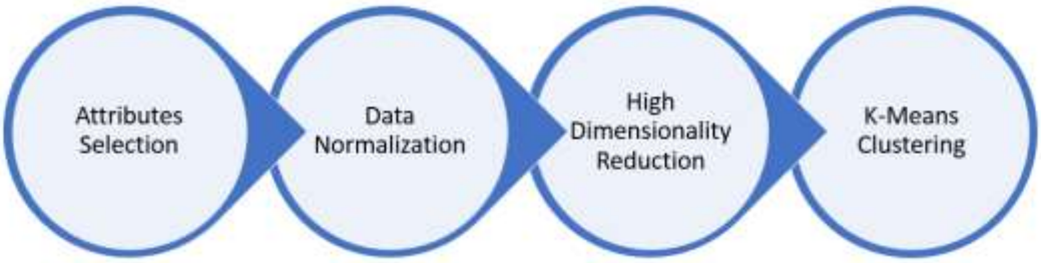

Figure 2. The Four Sub-Steps Followed in Clustering

Attributes selection. It's essential to note that only numerical variables were used for clustering since measuring distances between numerical data points is meaningful, while it is not possible to measure distance between categorical ones. The eight attributes we used were: lifecycle, manufacturer's suggested retail price (MSRP), average unit retail price (AUR) over style lifecycle, average store count over style lifecycle and monthly sales over style lifecycle.

Table 7. List of the Variables Considered by Each Step of the Three-Step Model

Process Name	Variable Name	Variable Category
Clustering	Lifecycle	Lifecycle
	MSRP	Price and Promotion
	Average AUR	Price and Promotion
	Average Store Count	Store
	Retail Sales Units	Sales Units
Classification	Lifecycle	Lifecycle
	MSRP	Price and Promotion
	AUR	Price and Promotion
	Store Count	Store
	Fiscal Year	Calendar
	Fiscal Month	Calendar
	Lifecycle Month	Calendar
	Lifecycle Start Month	Calendar
	Color Group	Product
	Basic Material	Product
	Gender	Product
	Category	Product
	Cut	Product
	Cluster Number	Cluster
Prediction	Lifecycle	Lifecycle
	MSRP	Price and Promotion
	AUR	Price and Promotion
	Store Count	Store
	Fiscal Year	Calendar
	Fiscal Month	Calendar
	Lifecycle Month	Calendar
	Lifecycle Start Month	Calendar
	Color Group	Product
	Basic Material	Product
	Gender	Product
	Category	Product
	Cut	Product
	Cluster Number	Cluster
	Average Sales	Sales Units

Data normalization. To avoid the high level of influence that some variables like sales may have over the others, the eight numerical variables were converted to the same scale by subtracting the average attribute value from each member data point, then dividing it by the standard deviation of the same attribute.

High dimensionality reduction. After normalizing the data, we used the t-SNE algorithm to reduce data dimensionality in preparation for clustering.

k-Means clustering. Once the data were normalized and the data dimensionality were lowered to two components only, we ran *k*-Means clustering algorithm to partition the data records into *k* number of clusters.

3.5.2.2 Classification

By the end of the clustering stage, cluster numbers were assigned to the records of both training and validation sets. Next, the classification stage was initiated to create a link between the styles with pre-assigned cluster from the training set and brand-new styles from the validation and testing sets. The classification drivers were both the categorical attributes and the numerical attributes (except sales). The classification stage had three sub-steps, as illustrated by Figure 3.

Figure 3. The Three Sub-Steps Followed in Classification

Attribute selection. Besides the categorical and numerical variables that were preselected in the feature selection process, the cluster numbers that resulted from the clustering stage were also used in classification. Cluster numbers were treated as a target variable as the objective was to match the records from the validation and testing sets with the clusters from the training set.

Classification. Regression trees, random forests and SVM were the algorithms used for the purpose of classification.

Accuracy evaluation. To evaluate the results of the three classification algorithms, we simply compared the clusters allocated to the validation set against the pre-assigned clusters that resulted from the clustering step.

3.5.2.3 Prediction

As the name indicates, the objective of the prediction stage is to predict the future sales for the brand-new styles in the validation and testing sets. As illustrated in Figure 4, prediction had three sub-steps.

Figure 4. The Three Sub-Steps Followed in Prediction

Attributes Selection. The variables used for prediction were the same ones used in classification. Additionally, a new variable, average sales, was calculated for every record of the training, validation and testing sets.

Prediction. To predict the sales for the products in the validation and testing sets, the following five algorithms were tested and later compared for accuracy: regression trees, random forests, neural networks, *k*-NN and linear regression.

Test & Score. Refer to Section 3.6.

3.6 Performance Measurement

We used two performance metrics for our forecasting models: forecast accuracy and bias. We measured forecast accuracy using Weighted Mean Absolute Percentage Error (WMAPE) and bias using Weighted Mean Percentage Error (WMPE). Absolute forecast error was first calculated for each record at style-month level and then WMAPE was computed at both the style-month and style-lifecycle level for reporting model performance. For seasonal products, since the lifecycle is around 1 – 4 months, normally the entire purchase quantity is confirmed prior to the beginning of the season. Therefore, we are interested in knowing the forecast accuracy for the whole season instead of individual month. Equation 1 was used to calculate the absolute forecast error for each record. We then used Equation 2 to calculate the forecast accuracy at either the monthly or lifecycle level by aggregating the MAPE of each style-month or style-lifecycle weighted by the sales units. We finally used Equation 3 to calculate the forecast bias in a similar fashion.

$$Absolute\ Forecast\ Error = |Forecasted\ Sale - Actual\ Sales| \qquad [1]$$

$$Forecast\ Accuracy\ (WMAPE) = \frac{\sum_{i=1}^{n} |Forecasted\ Sales - Actual\ Sales|}{\sum_{i=1}^{n} Actual\ Sales} \qquad [2]$$

$$Forecast\ Bias\ (WMPE) = \frac{\sum_{i=1}^{n} Forecasted\ Sales - Actual\ Sales}{\sum_{i=1}^{n} Actual\ Sales} \qquad [3]$$

Since the company did not keep records of previous forecasts, the baseline forecast accuracy was estimated using sell-in data, assuming shipment for a season was equal to its demand forecast. It is not feasible to allocate shipments to sales at a monthly level, so the baseline forecast accuracy was estimated on a lifecycle level for each style. If we consider only sales at the full price status, the WMAPE is over 100%. If both full price and markdown sales are considered, the WMAPE is 16%. In this case, both numbers are not directly comparable to our model results. However, this is the best reference we have regarding the company's forecasting performance.

4. Results

This section reports and analyzes the results of feature selection and the two types of forecasting models: general model and three-step model.

4.1 General Model

Based on the results of recursive feature elimination, the model with 12 variables resulted in the lowest error as shown in Figure 5. The list of variables in the order of variable importance is shown in Table 8. These variables were used to build the subsequent forecasting models while the remaining two variables (category and sub-category) were dropped. The top six attributes account for the majority of the variances. Store count, month and lifecycle month are the top three attributes. Among the product attributes, gender, basic material and color group are the top three attributes.

Table 8. List of Attributes Selected for Model Building

Importance Rank	Attribute	Attribute Category
1	Store Count	Store
2	Month	Calendar
3	Lifecycle Month	Lifecycle
4	Gender	Product
5	AUR	Price and Promotion
6	Year	Calendar
7	Basic Material	Product
8	MSRP	Price and Promotion
9	Color Group	Product
10	Lifecycle	Lifecycle
11	Cut Description	Product
12	Product Class Description	Product

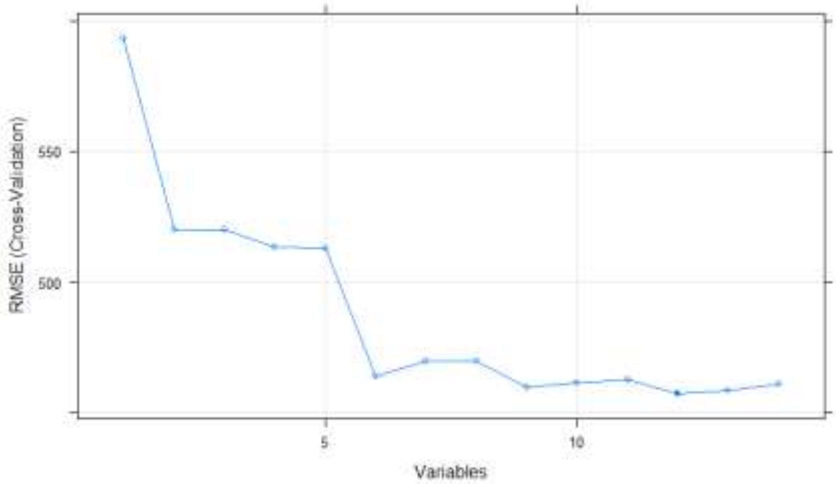

Figure 5. Cross Validation Error by Number of Attributes

The results in terms of forecast accuracy of the four individual models using regression trees, random forests, *k*-NN and neural networks and the two ensemble models using the median and average of individual outputs are shown in Table 9, while the forecast bias are shown in Table 10. MAPE were calculated on both the style-month and style-lifecycle level.

Considering the individual models, random forests gives the best predictive performance on the validation data with the highest accuracy and lowest bias. It achieved 37% WMAPE on the style-lifecycle level and 47% on the style-month level with a negative bias of 2%. Although the regression trees model has a slightly higher WMAPE and also tends to under-forecast, it provides better interpretability and visually gives insights into which predictor variables are more significant. Store count appears at the top of the tree, indicating that it is the most important attribute in predicting demand. Examining the first three layers of the tree, we can see that store count, month and lifecycle month are at the top in terms of feature importance. This is in line with our findings in the feature selection process. *k*-NN also gives reasonably good results in terms of forecast accuracy. However, it is worth noting that *k*-NN will only predict results within the range of the training data, since it is simply searching for the nearest *k* neighbors and predicting sales to be the average of those of the nearest neighbors. Therefore, it may not work as well if the new data are not in the same range as the training data. As for the ensemble methods, taking the median and average of the individual model outputs yields a better forecast accuracy with a WMAPE of 35%. Neural networks show the worst performance in terms of both accuracy and bias, with a 49% WMAPE and a positive bias of 19%, indicating a tendency to over-forecast.

Table 9. Forecast Accuracy of the General Model

	Regression Trees	Random Forests	*k*-NN	Neural Networks	Median	Average
WMAPE (Lifecycle)	38%	37%	39%	49%	35%	35%
WMAPE (Monthly)	49%	47%	50%	56%	45%	45%

Table 10. Forecast Bias of the General Model

	Regression Trees	Random Forests	*k*-NN	Neural Networks	Median	Average
WMPE (Lifecycle)	-12%	-2%	-2%	+19%	-2%	+1%
WMPE (Monthly)	-12%	-2%	-2%	+19%	-2%	+1%

4.2 Three-Step Model

This section discusses the results from each of the stages followed in the three-step model: clustering, classification and prediction.

4.2.1 Clustering and Classification

To determine the right number of clusters that fit our data, we used the silhouette score (silhouette coefficient). The silhouette score is a method of interpretation and validation of consistency within clusters of data. It measures and compare the mean intra-cluster distance to the mean nearest-cluster distance for each data point within a cluster. The silhouette score ranges between -1 (wrong clustering) and +1 (best value) with 0 indicating overlapping clusters. For our dataset, the number of clusters that revealed the highest silhouette score was seven as illustrated in Table 11.

Table 11. The Silhouette Score for Different k Number of Clusters

Number of Clusters	Silhouette Score
2	0.378
3	0.422
4	0.481
5	0.469
6	0.487
7	0.489
8	0.463
9	0.441
10	0.427

In addition to the silhouette score, we took an additional measure to verify which number of clusters work best for our dataset. Using the training and validation sets, we ran a classification exercise for each *k* from the table above, then compared the clusters that resulted from classification against those that resulted from clustering. The number of clusters that revealed the best classification match was five, with an overall accuracy of 93%. Looking up the silhouette score for five versus seven clusters, the difference is minimal. Figure 6 shows the classification results using five clusters and SVM algorithm. The number 590 is the total number of records included in the validation set. The vertical axis represents the number of records that were pre-assigned to each of the five clusters (C1 to C5) based on clustering while the horizontal axis is the number of records allocated to each of the five clusters based on classification.

Figure 6. The Confusion Matrix Resulted from Using Five Clusters and SVM Algorithm

The best performing classification algorithm was SVM. Table 12 compares the classification accuracy based on the three algorithms used.

Table 12. Comparison of the Classification Accuracy by Algorithm

Algorithm Name	Overall Accuracy
SVM	93%
Random Forests	89%
Regression Trees	71%

Additional analyses were performed on the validation set to better understand how clusters were allocated based on lifecycle, sales volume, AUR and store count. The main insights uncovered from these analyses are displayed in Table 13. The analyses showed that lifecycle length was a clear distinguishing driver between the five clusters. Each cluster included one lifecycle length, except Cluster 4, which included styles with mixed lifecycles. However, the styles that were included in Cluster 4 seemed to have relatively smaller sales volume, smaller store count and higher AUR compared to the other clusters. Both clusters C1 and C3 included styles with a three-months lifecycle. However, C3 had smaller sales volume and lower AUR compared to C1.

Table 13. Characteristics of Styles Distribution among Clusters

Cluster	Number of Records	Lifecycle (Months)	Average Monthly Sales (Units)	Average AUR	Store Count
C1	108	3	1298	$38	88
C2	112	2	953	$30	79
C3	84	3	839	$24	83
C4	106	2, 3, 4	462	$44	37
C5	180	4	958	$31	82

4.2.2 Prediction

Running the prediction algorithms on both validation and testing sets resulted in relatively different results in terms of best performing algorithms. However, the forecast accuracy of the ensemble methods for both datasets was much closer. Overall, the testing set had slightly better forecast accuracy and worse forecast bias compared to the validation set. Figure 7 and Figure 8 display the forecast accuracy and forecast bias for each of the two datasets.

Starting with the validation set, both *k*-NN and random forests delivered the highest forecast accuracy on a style-lifecycle level with a WMAPE of 37%. However, *k*-NN had no forecast bias compared to 4% under-forecast bias by random forests. Neural networks were third-best in forecast accuracy on a style-lifecycle level with 39% WMAPE and worst in forecast bias with -27% WMPE (under-forecast). Finally, regression trees had 40% of WMAPE on a style-lifecycle level and 12% over-forecast bias. The two ensemble methods, average and median, delivered the best overall results in forecast accuracy on both a style-month (43%) and a style-lifecycle (34%) levels with a forecast bias around 1% (under-forecast).

For the testing set, regression trees delivered the best performance with a forecast accuracy of 31% (style-lifecycle level) and 2% of forecast bias (over-forecast). Linear regression showed a very close performance with 31% of forecast accuracy (style-lifecycle level) and 3% of forecast bias (over-forecast). The neural networks' performance in the testing set was slightly better compared to the validation set with a forecast accuracy of 33% (style-lifecycle level). However, the neural networks' forecast bias (23% under-forecast) was still relatively high compared to the other four algorithms. Random forests and *k*-NN had the lowest forecast accuracy (around 36%) while they had a forecast bias of 7% and 18% (under-forecast), respectively. Like in the validation set, the ensemble methods also delivered the highest overall forecast accuracy with 30% WMAPE on a style-lifecycle level. However, their forecast bias was relatively high (around 8% under-forecast) compared to the regression trees and linear regression algorithms.

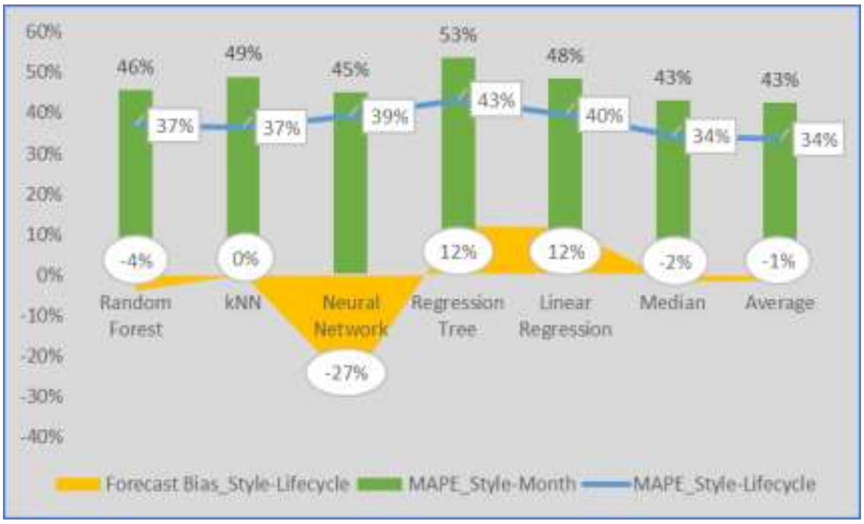

Figure 7. Forecast Accuracy and Bias of the Three-Step Model (Validation Set)

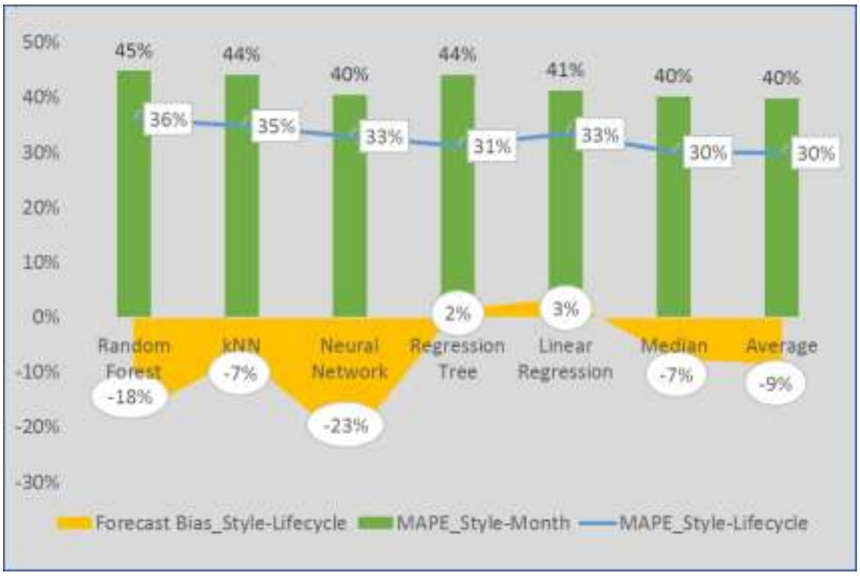

Figure 8. Forecast Accuracy and Bias of the Three-Step Model (Testing Set)

In addition to the analyses above, we clicked down into cluster level to understand how the model performs from one cluster to another. For the validation set, random forests was the best performer in clusters C2, C3 and C5 with a forecast accuracy of 36%, 37% and 34%, respectively (see Table 14). The forecast bias for random forests was -11% (C1), -1% (C3) and +3% (C5). It's essential to note that these three clusters (C1, C3 and C5) share relatively similar store count and average monthly sales. They only differ in the length of lifecycle. *k*-NN was the best performer in cluster C1 with a forecast accuracy of 28% and a forecast bias of 4% (over-forecast). For cluster C4, regression trees performed best with a forecast accuracy of 45% and forecast bias of 30% (under-forecast). The relatively bad performance in cluster C4 could be linked to the complexity of this cluster including multiple lifecycle lengths, low monthly sales volume and high AUR on average.

The testing set included three clusters: C1, C3 and C4. Those were the clusters that resulted from the classification stage. Similar to the validation set, random forests and regression trees delivered the highest forecast accuracy and lowest forecast bias in cluster C3 and cluster C4 (see Table 15). In cluster C1, *k*-NN revealed the best forecast accuracy (28%) and lowest forecast bias (4% over-forecast).

On a cluster level, the performance of the two ensemble methods didn't always outperform the performance of the individual algorithms.

Table 14. Best Performing Algorithm by Cluster (Validation Set)

Cluster	Best-Performing Algorithm	Forecast Accuracy	Forecast Bias	Best-Performing Ensemble	Forecast Accuracy	Forecast Bias
C1	*k*-NN	28%	+4%	Average	26%	-4%
C2	Random Forests	36%	-11%	Median	38%	-10%
C3	Random Forests	37%	-1%	Median	33%	0%
C4	Regression Trees	45%	-30%	Median	51%	-35%
C5	Random Forests	34%	+3%	Average	33%	+14%

Table 15. Best Performing Algorithm by Cluster (Testing Set)

Cluster	Best-Performing Algorithm	Forecast Accuracy	Forecast Bias	Best-Performing Ensemble	Forecast Accuracy	Forecast Bias
C1	Linear Regression	28%	-11%	Median/Average	29%	-22%
C3	Random Forests	32%	+6%	Average	33%	+11%
C4	Regression Tree	39%	0%	Median	39%	0%

5. Discussion

This section discusses implications of the model results, then moves on to describe limitations of our model, and finally outlines some future research opportunities.

5.1 Implications

In evaluating the suitability of the models for the sponsoring company, the ease of implementation was considered in addition to the models' predictive performance. As a starting point, the general model serves as a good framework for an immediate implementation, outperforming the company's current forecasting model in terms of forecast accuracy and bias. Among the different methods tested in the general model, the ensemble methods (median and average) and random forests gave the best predictive performance, thus are the methods that we recommend using when implementing the general model.

The three-step model through the clustering and classification stages offers visibility into the underlying factors that impact demand. With this model, forecasting can be customized to deliver best possible results based on product characteristics such as planned lifecycle, store number and retail price. Regression tree is what we recommend applying on complex clusters with multiple lifecycle lengths. Random forests is the algorithm we recommend using on clusters with mono lifecycle, while *k*-NN and linear regression are what we recommend using on similar clusters but with higher sales volume and AUR.

5.2 Limitations

Due to the limitation in the inventory data available, lost sales were not considered in building our forecast models. Inventory data were provided at the monthly and style level. As a result, we only have one snapshot of the inventory level for each month and at an aggregated level across all sizes of a style. It is therefore not feasible to estimate lost sales, which is needed to reflect true demand not captured in POS data.

Since the company did not keep records of previous forecasts, direct comparison between our model performance and the company's current forecast performance is not possible. Going forward, we recommend that the company keep track of forecast history so forecasting performance can be measured and improvement areas can be identified accordingly.

The intended product lifecycle was estimated based on the POS data by counting the number of consecutive months with full price sales records for a particular style. In practice, the intended product lifecycle will have to be pre-determined in order to be inputted into the forecasting model. There may be some difference between intended and actual product lifecycle.

Store count, which refers to the number of stores that a style is carried in, was estimated using sales record. In the case where inventory is available but there are no sales, store count will be overestimated. In practice, the pre-determined store count should be used as an input since the store count based on actual sales will not be available at the time of forecast.

Due to the complexity of the promotional data provided, price promotions were not used as an explicit attribute, rather they were embedded as a change in the AUR. Price promotions play a major role in driving demand and capturing this component explicitly may help improve the achieved forecast accuracy.

5.3 Future Research

In future research, the company should consider incorporating lost sales, a more accurate measure of intended product lifecycle, as well as store count in building and evaluating the forecast model. The company may also consider extending future research to cover shorter forecast horizons and higher data granularity. While this project focuses on roughly a five-month range forecast for placing orders to manufacturers, there are opportunities to dive deeper into the data at the store and weekly level, for the purpose of store inventory allocation and size curve analysis. In addition, the relationship between price and demand could be studied for price optimization.

6. Conclusion

Our research project proposed a methodology that offers two different forecasting models based on machine learning techniques. These models will enable the company to achieve better forecast accuracy compared to the current performance by considering store count, lifecycle, calendar and product attributes simultaneously.

The data pre-processing phase of the proposed methodology is an important stage that facilitates the formation of the inputs to the models. The feature engineering process helps create new variables that bring additional value to demand interpretation. The feature selection process allows us to gain insights into the importance of the different predictor variables and their influence on forecast accuracy. Another value proposition of this phase is the possibility of using, processing and delivering value out of the categorical variables that have always been considered a challenge when it comes to forecasting demand in the fashion industry.

When it comes to the models, the general model serves as a starting point for easy implementation of the machine learning forecasting framework. The three-step model involving clustering, classification and prediction enables the company further to visualize the relationship between predictor variables and customize the forecasting approaches accordingly.

Finally, the project opens doors for further research that possibly cover store inventory allocation, size curve analysis and price optimization.

References

B., & Thomassey, S. (2016). Intelligent demand forecasting systems for fast fashion. Information Systems for the Fashion and Apparel Industry, 145-161. doi:10.1016/b978-0-08-100571-2.00008-7

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. doi:10.1016/j.ejor.2006.12.004

Chase, C. W., Jr. (2017). Machine Learning is Changing Demand Forecasting. Journal of Business Forecasting,43-45.

Choi, T., Hui, C., Liu, N., Ng, S., & Yu, Y. (2014). Fast fashion sales forecasting with limited data and time. Decision Support Systems, 59, 84-92. doi:10.1016/j.dss.2013.10.008

Lu, C. (2014). Sales forecasting of computer products based on variable selection scheme and support vector regression. Neurocomputing, 128, 491-499. doi:10.1016/j.neucom.2013.08.012

Kaya M., Yeşil E., Dodurka M.F., Sıradağ S. (2014) Fuzzy Forecast Combining for Apparel Demand Forecasting. In: Choi TM., Hui CL., Yu Y. (ads) Intelligent Fashion Forecasting Systems: Models and Applications. Springer, Berlin, Heidelberg

Pillo, G. D., Latorre, V., Lucidi, S., & Procacci, E. (2016). An application of support vector machines to sales forecasting under promotions. 4Or, 14(3), 309-325. doi:10.1007/s10288-016-0316-0

Rashad, A., & Spraggon, S. (2013). Assembling the crystal ball: using demand signal repository to forecast demand (Unpublished master's thesis).

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). Data mining for business analytics concepts, techniques, and applications in R. Hoboken, NJ, USA: Wiley.

Thomassey, S. (2010). Sales forecasts in clothing industry: The key success factor of the supply chain management. International Journal of Production Economics, 128(2), 470-483. doi:10.1016/j.ijpe.2010.07.018

Vhatkar, S., & Dias, J. (2016). Oral-Care Goods Sales Forecasting Using Artificial Neural Network Model. Procedia Computer Science, 79, 238-243. doi:10.1016/j.procs.2016.03.031

Wong, W., & Guo, Z. (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. International Journal of Production Economics, 128(2), 614-624. doi:10.1016/j.ijpe.2010.07.008