Forecast of Short Lifecycle Items in the Fashion Industry

Maria João dos Santos Robles

Master's Dissertation

Supervisor at FEUP: Prof. Paulo Luís Cardoso Osswald

U. PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Master's Degree in Industrial and Management Engineering

2019-07-01

"Success consists of going from failure to failure without loss of enthusiasm."

Winston Churchill

Forecast of Short Lifecycle Items in the Fashion Industry

Abstract

In retail context, quality and organized data can be used to infer future demand patterns. Retailers' awareness about the importance of accurate and reliable sales prediction is becoming more expressed over the years, given the increased competitive advantage that comes alongside with it.

This project was carried out in a company specialized on information systems for retail, which already implements a software that provides solutions related to forecast and replenishment planning. Despite the accurate results obtained, some opportunities were identified to aim for more optimized results in the fashion field. For that reason, an alternative approach to what is currently done was tested in this project.

Historical data from two distinct fashion companies was used to test the prediction accuracy of state-of-the-art forecasting methodologies, and the outcomes of Artificial Intelligence procedures were compared between them.

Several techniques were analyzed, in order to challenge the well-known, faster and easier to understand statistical methods, normally used by firms for the purpose. After a careful cleaning and transformation of the available raw data, Backpropagation Neural Networks technique was tested due to its satisfactory performance achieved over the last decade in multiple studies. Random Forest and Support Vector Machine procedures were also executed given their ability in handling such predictive questions, although there is very few research about the methods.

Generalizing, Random Forest and Backpropagation Neural Networks showed to achieve better evaluation metrics performance than Support Vector Machine for both companies inserted in the fashion industry. However, the latter one showed to be faster than the other two, and accurate during sales periods without many oscillations. The three techniques were capable to learn and infer the demand pattern over the weeks, with special emphasis on Random Forest.

Resumo

Em termos de retalho, a recolha organizada de dados e qualidade inerente dos mesmos pode ser utilizada de forma a inferir padrões de procura futuros. A consciência dos retalhistas em relação à importância de uma previsão de vendas precisa e confiável tem-se pronunciado cada vez mais ao longo dos anos, como consequência da acrescida vantagem competitiva associada à tarefa.

O projeto foi proposto e realizado por uma empresa especializada em sistemas de informação para retalho, a qual implementa um programa que proporciona soluções relacionadas com previsão e planeamento de reabastecimento. Apesar da elevada precisão nos resultados até agora obtidos, foram identificadas oportunidades de resultados mais otimizados no campo da indústria da moda. Desta forma, ao longo do projeto foram testados procedimentos alternativos ao que é utilizado atualmente.

Dados históricos provenientes de duas empresas de moda foram testados de forma a obter as medidas de desempenho do estado-de-arte atual, no que toca a metodologias de previsão, e os resultados finais derivados dos procedimentos de Inteligência Artificial foram comparados entre si.

Várias técnicas foram analisadas com o intuito de desafiar os conhecidos, rápidos e mais facilmente compreendidos métodos estatísticos, normalmente utilizados pelas firmas para o propósito. Após uma cuidadosa limpeza e transformação dos dados disponíveis, a técnica *Backpropagation Neural Networks* foi executada, dado o seu desempenho satisfatório ao longo da década passada. Os procedimentos *Random Forest* e *Support Vector Machine* foram também testados devido à sua habilidade no campo da previsão, apesar de ainda não existir muita investigação sobre os mesmos.

Generalizando, os algoritmos *Random Forest* e *Backpropagation Neural Networks* apresentaram melhores resultados no que toca a medidas de avaliação do que o método *Support Vector Machine* para as duas empresas em estudo. No entanto, o último mostrou ser mais rápido do que os restantes, e obteve resultados precisos durante períodos de vendas com poucas oscilações. Todas as técnicas mostraram ser capazes de aprender e inferir padrões de procura futura ao longo das semanas, com especial ênfase no *Random Forest*.

Acknowledgements

To begin with, I would like to thank my supervisors Mr. Carlos Lima, Mr. Nuno Santos, Mr. Paulo Marques and Mr. Hélder Mota from KSR, SA – Retail Consult for the daily availability and knowledge provided throughout the project.

To Prof. Vera Miguéis from *Faculdade de Engenharia da Universidade do Porto*, for all the afforded know-how and tireless effort.

To Prof. Miguel Gomes and José Luís Borges from *Faculdade de Engenharia da Universidade do Porto*, who gave me some first insights and had an important contribution on the course of the work.

To my supervisor Prof. Paulo Osswald for the continuous guidance and feedback during the dissertation period.

To all the members of Planning and Optimization team from KSR, SA – Retail Consult for their great accessibility and advices.

To my course colleagues Ana Rita Novo and Ana Andrade, who made their internship on KSR, SA – Retail Consult too, for all the patience, opinions and critical judgements.

To my family, especially my mother, Maria de La Salete Robles, father, António Júlio Robles, sister, Ana Raquel Robles, brother, Diogo Pedro Robles, and sister-in-law, Patrícia Jardim for all the support, mainly on more difficult times, for the trust placed on my capabilities during my personal and academic journey and for allowing me the opportunity of following my dreams. To my close friend Adozinda Lopes for never giving up on me and being always there throughout my life. To my cousin and oldest friend Maria Ferreira for all the great adventures, memories and for always believing in my potential.

To all my friends who accompanied my journey on *Faculdade de Engenharia da Universidade do Porto*, André Couto, Catarina Ferreira, Diogo Simões, Filipa Sousa, Inês Sousa, Ricardo Miguel Silva and, in particular, Joana Mota for making this path unforgettable with their never ending friendship.

To my best friends, Ana Isabel Santos, Carolina Vasques, Catarina Moreira and Inês Eiras for the companionship since a very young age and for all the splendid moments shared. A special word of gratitude to Nuno Valente, for always encouraging me to give my best and for being an essential part on the accomplishment of both this project and personal achievements.

List of Contents

Abbreviations

3F – Fast Fashion Forecasting

AI – Artificial Intelligence

AHP – Analytic Hierarchy Process

AR - Autoregression

ARMA – Autoregression Moving Average

ARIMA – Autoregression Integrated Moving Average

AutoES – Automatic Exponential Smoothing

A/W – Autumn/Winter

BIC – Bayesian Information Criterion

BPNN – Backpropagation Neural Networks

DNN – Deep Neural Networks

DT – Decision Tress

EC – Evolutionary Computation

ELM – Extreme Learning Machine

ELME - Extreme Learning Machine Extended

ENN – Evolutionary Neural Networks

GM – Grey Model

HS – Harmony Search

LR – Linear Regression

MA – Moving Average

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

MdAPE – Median Absolute Percentage Error

MSE – Mean Square Error

NN – Neural Networks

nRMSE – Normalized Mean Square Error

OOB – Out-of-bag

PDPF – Panel Data Particle Filter

PF - Particle Filter

PPD – Pure Panel Data

RDF V16 – Retail Demand Forecast Version 16

Relu – Rectified linear unit

RF – Random Forest

RMSE – Root Mean Square Error

RVM – Relevance Vector Machine

SARIMA - Seasonal Autoregression Integrated Moving Average

SES – Simple Exponential Smoothing

SKU – Stock-keeping-unit

SMAPE – Symmetric Mean Absolute Percentage Error

StandardES – Standard Exponential Smoothing

SVM – Support Vector Machine

SVR – Support Vector Regression

S/S – Spring/Summer

Tanh – Hyperbolic tangent function

WMAPE – Weighted Mean Absolute Percentage Error

List of Figures

List of Tables

1. Introduction

1.1. Project Background

The Forecast of Short Lifecycle Items in the Fashion Industry project emerged in KSR, SA – Retail Consult, a multinational firm whose core activity is technology consulting services, providing to their clients' reliable and accurate results within this subject area. The company has a partnership with Oracle Retail, executing the implementation and support of its customer-oriented software, that contributes to a simplified and more accurate overview of a company's internal activities, with little human intervention, and gives guidance for retailers achieving a profitable growth while, for example, decreasing inventory costs and setting the right price of the products (Company 2016).

Oracle Retail acts in six main critical areas: Planning, Supply Chain, Merchandising, Retail Hardware, Omnichannel and Insights and Science (Company 2019). KSR, SA – Retail Consult clients are mainly grocery, fashion, pharmacy, telecommunications and electronics companies, and the company assesses the clients' necessities by implementing one or more of the customized services referred above.

Usually, a firms primary goal is to maximize profits and growth, while decreasing costs. The Planning and Optimization team from KSR, SA – Retail Consult handles this opportunity by implementing forecasting solutions to plan ahead future demand and, therefore, improving warehouse management, transportation costs and reducing excessive or modest stocks.

1.2. Project Specification and Goals

This project aims to find an alternative approach from the currently used ones by companies, focused on performance, accuracy and few user intervention.

The main goal of this paper is to identify new ways of performing forecasts in the fashion field and understanding the power of Artificial Intelligence (AI) techniques, given that, over the past decade, they have suggested increased advance on performances. The work also intends to provide an insight of what are the most used predictive models, their needs, ease of use, speed and input variables. Despite statistical methods being easier to understand and faster, when compared to AI, there could be advantages on performing a more complex and non-linear algorithm, especially in cases where the demand patterns are not easily inferable from the previous year's sales (Makridakis, Spiliotis, and Assimakopoulos 2018).

Every company acts differently, and the way a retailer stores its information over the years differs too. The quality of the data is an important factor to account with in a forecasting task, alongside with the number of observations and available predictors. Due to that, differences between several predictive techniques arise, together with the situations in which each one should be used. In this way, this paper can be consulted for future situations in which the historical data at hand does not fit any of the algorithms executed.

The softwares used for this project were MySQL Workbench to access MySQL database, where the data is stored and processed, and RStudio, an integrated development environment for R, a programming language.

1.3. Methodology and Structure of this Paper

Firstly, the necessary research about the fashion industry and the state-of-the-art prediction methods was performed, and is summarized on chapter 2. A description was provided, ranging from linear statistical procedures to non-linear AI ones, together with a series of the most adequate evaluation metrics for demand forecast purpose.

The currently used techniques at KSR, SA – Retail Consult are referred in chapter 3, as well as the importance of a good forecast and the decisive factors to consider in its choice from a managerial point of view.

Raw data from two distinct fashion retail companies was cleaned, reduced and transformed, so that it could be graphically analysed and used to test the three forecasting algorithms in study: Neural Networks (NN), Random Forest (RF) and Support Vector Machine (SVM). This was the most time consuming task, once it took more or less two months to be completely finished, but also because many alternatives were executed with different data, as explained in chapter 4, which ends with a comparison between the three methods results.

To end up, conclusions about the procedures used on the two companies', the evaluation metrics, inferred patterns and possible future studies are presented in chapter 5.

2. State-of-the-Art Forecasting Techniques

2.1. Sales Prediction on the Fashion Industry

Fashion industry constitutes an uncertain market, especially for new products that are only for sale during one or two months, known by the name of fast fashion (Choi et al. 2014). As in every business, there is a necessity to predict and, consequently, to provide the right amount of product for sale to a client, at the right place and time (Xia and Wong 2014). Fashion demand forecast refers to the process of anticipating future demand, in order to support the operations and supply chain of a firm. Usually, factors that influenced past sales will have a similar weight in future ones, making the use of historical data a good base for prediction. For the particular case of fashion, such data is limited due to the usual small time period in which an item is available to purchase.

The concept of predicting demand is not the same as of predicting sales, since the latter one implies only the on-hand stock, whereas the first one includes the variability that comes from a limited item availability and out-of-stocks. For the same reason, the forecast must be conducted in a real time basis, so that retailers can react in time to make decisions based on the predicted output (Wong and Guo 2010; Yu, Choi, and Hui 2011; Ren, Choi, and Liu 2015). New items do not have historical data, but a prediction can be performed based on their similarities to the previous seasons' available products.

There are numberless forecasting models that can be used to predict future demand, nonetheless the choice of the right one is crucial to achieve accuracy that enables good operational performance, as well as reducing inventory and transportation costs, which in turn will lead to an improvement of the company's profits (Ren, Chan, and Ram 2017). The ever-changing tastes of the consumers and the high volatility of the demand for fashion products increase the complexity of the task (N. Liu et al. 2013). In addition, the fashion market is influenced by factors known as explanatory variables, that increase the randomness and irregularity of the sales. Some of these variables are not possible to control, such as economic factors (purchasing power, unemployment rate), market situation, calendar data (holidays), weather conditions, among others. Item features (colour, size), marketing actions, promotions, fashion trends and price can be better controlled (Ren, Choi, and Liu 2015; Wong and Guo 2010; Loureiro, Miguéis, and da Silva 2018; Choi, Hui, and Yu 2014). In order to deal with all these influencing parameters that introduce non-linearity into the models, the forecasting algorithm should include them as an input, instead of merely considering the time series of historical data (Du, Leung, and Kwong 2015).

2.2. Fashion Industry Concepts

2.2.1. Fashion Products Lifecycle

Almost every product lifecycle curve is constituted by four main phases: introduction, growth, maturity and decline. However, many fashion products, referring to apparel, shoes and accessories, cannot be compared to traditional retail products, such as groceries, due to their different survival time horizon and behaviour alongside the curve, as it can be seen in Figure 1.

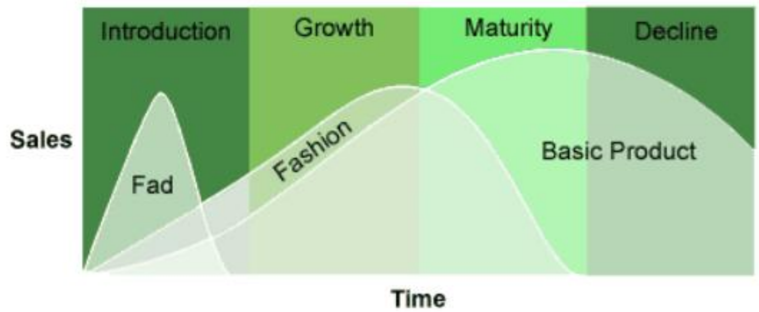


Figure 1 - Product lifecycle curve in (Cornell University 2006).

There are three major categories within short lifecycle fashion products: the basic items, sold during the entire year and usually in some forward periods too; the fashionable ones, the so called "fast fashion", sold during a very short time period; the best-selling items, continually replenished in the course of a season. The second ones, fashionable items, are characterized by a continuous new merchandise offer and reflect the ultimate trends and designs of the market needs and wants, while the latter are usually continually sold year after year, but accompanied by slight adjustments, in order to follow fashion trends and customer tastes. The categories are usually characterized by a strong seasonality throughout the two large fashion seasons: Autumn/Winter (A/W) and Spring/Summer (S/S), but also in smaller periods within each one (Choi, Hui, and Yu 2014; Wong and Guo 2010).

2.2.2. Fashion Products Hierarchy

Given the large number of stock-keeping-units (SKUs) in every fashion collection and the resulting difficulty to analyse all, retailers usually adopt a hierarchical aggregation to classify different products with similar attributes. In addition, this practice allows companies to compile the three different types of products described above, since the upper levels remain commonly unchanged and the fast fashion or best-selling SKUs differ a lot from season to season (Ren, Choi, and Liu 2015; Wong and Guo 2010). The retail industry usually agglomerates items into categories, but the nomenclature can vary according to the client's necessities. An example can be seen in Figure 2.

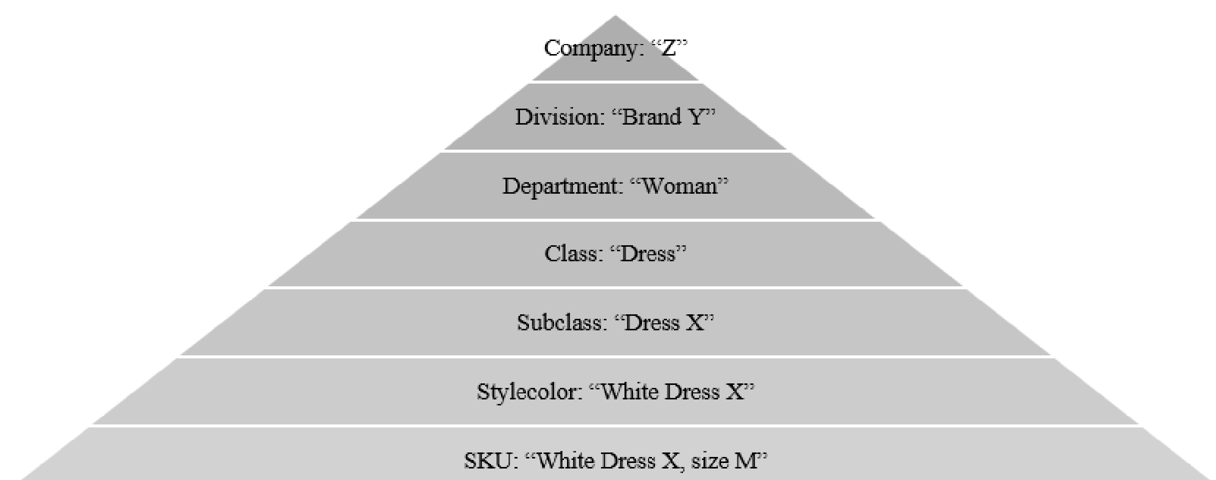


Figure 2 - Product Hierarchy.

2.3. Statistical Forecasting Techniques

There are several studies about demand forecast, which comprise many analytical models developed over the past years. The most widely used techniques for predicting sales are based on time series as an input of the forecast and will be presented in this section.

2.3.1. Time Series

A time series is a series of past sales points through time, and forecasting methods allow to predict the future demand through the analysis of this historical sales data. There are three very important factors that can be present in a time series, and they are represented in Figures 3 to 5: the level, the average value of the series; a trend, which reflects the downward or upward tendency of the series; and seasonality, which refers to the cyclic behaviour of a series (Artley 2018).

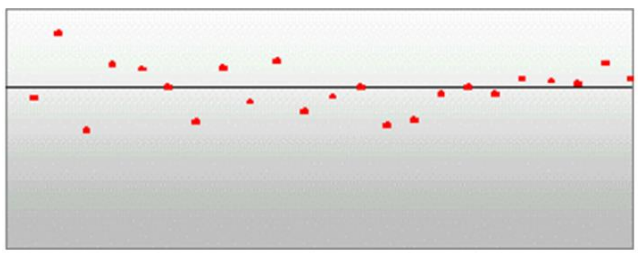


Figure 3 - Level of a time series in (Artley 2018).

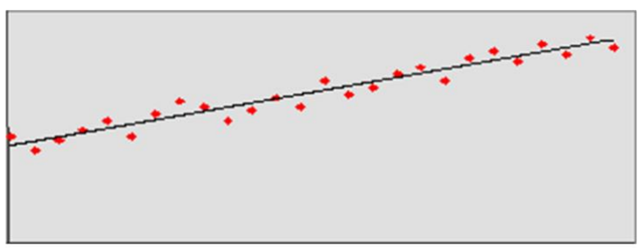


Figure 4 - Trend of a time series in (Artley 2018).

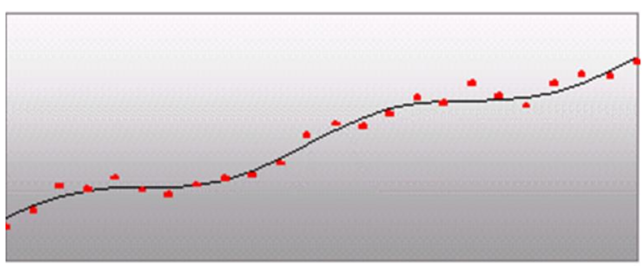


Figure 5 - Seasonality of a time series in (Artley 2018).

2.3.2. Time Series Based Forecasting Techniques

The Autoregression (AR) forecasting model is used to estimate future values by performing a linear regression based on historical values. The regression order defines the number of p past periods that will be used to do it (Agrawal, Adhikari, and Agrawal 2013). AR(p) formula is:

$$y_t = c + \sum_{j=1}^{p} \varphi_j \, y_{t-j} + \varepsilon_t \tag{2.1}$$

Where:

c is a constant,

φ_j is an autoregressive coefficient,

y_{t-j} is the actual value at time $t - j$ and

ε_t is the random error at time t.

The Moving Average (MA) technique performs also a linear regression, but it benefits from the past errors as an explanatory variable to the predictive model (Agrawal, Adhikari, and Agrawal 2013; Artley 2018). The order of the model is defined by q and $\mathrm{MA}(q)$ formula is:

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j} + \varepsilon_t \tag{2.2}$$

Where:

μ is the series mean,

θ_j is a model parameter,

ε_{t-j} is the random error at time $t - j$ and

ε_t is the random error at time t.

Given the characteristics of AR and MA models, they can be combined in order to construct an Autoregression Moving Average (ARMA) model. However, such models can only handle stationary time series, indicating that mean and variance cannot vary in time (Agrawal, Adhikari, and Agrawal 2013). $\mathrm{ARMA}(p, q)$ formula is:

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i \, y_{t-i} + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j} \tag{2.3}$$

The Autoregressive Integrated Moving Average (ARIMA) is similar to ARMA, but can handle non-stationarity time series. The order of the model integrated part is described by d, and being the lag operator $L^i y_t = y_{t-i}$ ("ARIMA Models" n.d.; Agrawal, Adhikari, and Agrawal 2013), $\mathrm{ARIMA}(p, d, q)$ formula is:

$$\left(1 - \sum_{i=1}^{p} \varphi_i \, L^i\right)(1 - L)^d y_t = \left(1 + \sum_{j=1}^{q} \theta_j L^j\right) \varepsilon_t \tag{2.4}$$

Seasonal Autoregressive Integrated Moving Average (SARIMA) covers the seasonal and non-stationarity time series (Agrawal, Adhikari, and Agrawal 2013). The formulation of $\mathrm{SARIMA}(p, d, q) \times (P, D, Q)^s$ is:

$$\Phi_p(L^s)\varphi_p(L)(1 - L)^d(1 - L^s)^D y_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t \tag{2.5}$$

Where:

$\Phi_p(L^s)$ and $\Theta_Q(L^s)$ are the seasonal polynomials.

For AR, MA, ARMA, ARIMA and SARIMA, identical weights are given to every period, which means that when the historical data observations increase, changes in the series are more difficult to identify and, therefore, to predict. An Exponential Smoothing model, as the name suggests, smooths the data by weighting different periods: more recent data is considered to weight more than older one by exponentially decrease it (Artley 2018).

Simple Exponential Smoothing (SES) deals with time series characterized by the level parameter. It is similar to MA, but it uses a smoothing factor, α, so that more recent values of the series are considered to have more weight than older ones in predicting the future demand (Ostertagová and Ostertag 2011; Rob J Hyndman 2014). SES formulas are:

$$\text{Forecast equation: } y_{t+h|t} = l_t, h = 1,2 \ldots \tag{2.6}$$

$$\text{Smoothing level equation: } l_t = \alpha \sum_{j=0}^{t-1} (1-\alpha)^j y_{t-j} + (1-\alpha)^t l_0, (0 \leq \alpha \leq 1) \tag{2.7}$$

Where:

l_t is an estimate of the level parameter,

α is the level weight parameter and

y_{t-j} is the actual value at time $t-j$.

Just like in regression, α and l_0 values are obtained by minimizing the error measure.

Holt's Exponential Smoothing is an algorithm capable of predicting future values of a time series characterized by both level and trend factors. Since it is part of the Exponential Smoothing family, it weights the data according to its proximity to the actual date. In addition to factor α, it also smooths the trend parameter by using $\beta = \alpha \times \beta^*$ (Rob J Hyndman 2014; Gardner and Mckenzie 1985). Holt's formulas are:

$$\text{Forecast equation: } y_{t+h|t} = l_t + hb_t, h = 1,2 \ldots \tag{2.8}$$

$$\text{Smoothing level equation: } l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}), (0 \leq \alpha \leq 1) \tag{2.9}$$

$$\text{Smoothing trend equation: } b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*) b_{t-1}, (0 \leq \beta^* \leq 1) \tag{2.10}$$

Where:

b_t is an estimate of the trend parameter,

y_t is the actual value at time t and

β^* is the trend weight parameter.

For Holt-Winters' Exponential Smoothing model, the time series is characterized by level, trend and seasonal parameters. Besides α and β, it smooths the seasonal component too, γ. If the seasonal variations are approximately constant, the Additive Exponential Smoothing model is preferred. However, when seasonality varies proportionally to the level parameter, the Multiplicative Exponential Smoothing model is the chosen one. The main differences between the two can be seen in Figure 6. A downside of this approach is the difficulty of handling trends within the seasonal indexes (Rob J Hyndman 2014; Gardner and Mckenzie 1985; Artley 2018). Holt-Winters' formulas for Additive and Multiplicative Exponential Smoothing models are, respectively:

$$\text{Forecast equation: } y_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}, h = 1,2 \ldots \tag{2.11}$$

$$\text{Smoothing level equation: } l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}), (0 \leq \alpha \leq 1) \tag{2.12}$$

$$\text{Smoothing trend equation: } b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*) b_{t-1}, (0 \leq \beta^* \leq 1) \tag{2.13}$$

$$\text{Smoothing seasonal equation: } s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma) s_{t-m}, (0 \leq \gamma \leq 1-\alpha) \tag{2.14}$$

Where:

s_t is an estimate of the seasonal parameter,

m is the seasonality period,

k is the integer of $\frac{h-1}{m}$ and

γ is the seasonal weight parameter.

$$\text{Forecast equation: } y_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)}, h = 1,2 \ldots \quad (2.15)$$

$$\text{Smoothing level equation: } l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}), (0 \leq \alpha \leq 1) \quad (2.16)$$

$$\text{Smoothing trend equation: } b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}, (0 \leq \beta^* \leq 1) \quad (2.17)$$

$$\text{Smoothing seasonal equation: } s_t = \gamma \frac{y_t}{l_{t-1} - b_{t-1}} + (1-\gamma)s_{t-m}, (0 \leq \gamma \leq 1-\alpha) \quad (2.18)$$

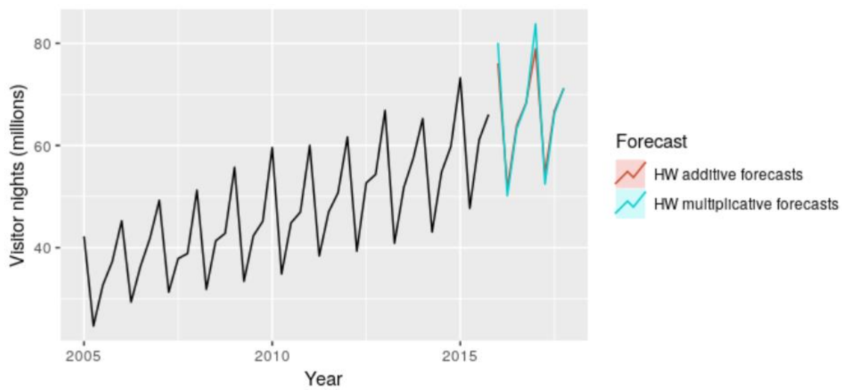


Figure 6 - Difference between Additive and Multiplicative Exponential Smoothing models in (Gardner and Mckenzie 1985).

2.3.3. Advantages and Challenges of Time Series Based Forecasting Techniques

The previously statistical methods described have diverse advantages associated, such as speed, even when dealing with a large dataset, and being easy to understand by the user (Ren, Choi, and Liu 2015). Despite these points in favour and their worldwide popularity, the linearity that categorizes such models carries some limitations when applied to fashion items sales, due to the inherent high complexity and randomness of the historical data available. Moreover, the fashion industry SKUs are constantly replaced by new ones and sales information is characterized by strong non-linear patterns that are not recognized by the algorithms referred above. The influence of other products demand and variations in price, for example, are not taken into account as inputs and such models only accept quantitative data, which are great disadvantages (Yu, Choi, and Hui 2011; Wong and Guo 2010; Loureiro, Miguéis, and da Silva 2018).

Pure Panel Data (PPD) model is a statistical method capable to deal with some of the problems described above. The algorithm has the ability of learning an item's behaviour by analysing other products that can cannibalize or complement it, because it considers the cross-sectional sales data to identify the effect from other items in the one under analysis. Moreover, it also considers the selling price of the fashion components. PPD is a quick and user-friendly method, and as a consequence of the advantages described previously, it expectedly performs better than the methods synthesized above (Ren, Chan, and Ram 2017; Ren, Choi, and Liu 2015). PPD formula is:

$$S_{it} = \alpha_i + \gamma S_{it-1} + \beta P_{it} + \mu_{it}, \quad i = 1, \ldots, I; t = 1, \ldots, T \quad (2.19)$$

Where:

α_i is a constant for item i,

γ is the coefficient for the time series,

S_{it-1} is the linear component of demand in period $it-1$,

β is the coefficient for the cross-sectional data,

P_{it} is the price of item i at time t and

μ_{it} is the associated error of item i at time t (Ren, Choi, and Liu 2015).

There are many other statistical methods that can be used for sales forecasting. However, none of them can handle the non-linearity introduced by fashion industry sales which, together with the advance of computing technologies, boosted the interest on artificial intelligence (AI) models for fashion sales forecasting in the recent years. Their success is mainly due to the learning capability by trial and error, as well as continuous improvement over time (Ren, Choi, and Liu 2015; Loureiro, Miguéis, and da Silva 2018; Ostertagová and Ostertag 2011).

2.4. Artificial Intelligence Forecasting Techniques

Depending on the context and available data, AI methods are believed to be more powerful and versatile than statistical ones, because of their good capacity to handle the high complexity and the non-linearity of the available historical data (Ren, Choi, and Liu 2015).

In order for AI models to perform well, the algorithm must be previously trained by a part of the available dataset, which implies having enough data to be divided into two subsets: for training and testing. Moreover, the execution of this task carries an associated increase of the amount of time needed to run the process, which is the major disadvantage of using AI techniques (Ren, Choi, and Liu 2015).

Many studies emphasize the versatility and good performance of neural network models (NN) when applied in fashion demand forecasting, due to their hidden layers' capacity of capturing non-linear relationships within the available data, and the promising performance on pattern recognition. The most used gradient-descent based learning algorithm for NN is the Backpropagation Neural Network (BPNN), which is a feedforward (information moves in one direction) model that attempts to minimize the error between the desired output and the actual one. However, this algorithm may take hours or even days to complete the forecast because of the large number of iterations steps required on the training task, in order to improve the model accuracy, which represents a limitation for retailers that support their inventory planning and replenishment with the predicted sales. Moreover, many parameters have to be tuned, such as the number of hidden layers, the number of neurons in each hidden layer, the activation function of both input and output layer and the learning rate (Yu, Choi, and Hui 2011; Zhang, Patuwo, and Hu 1998; Xia and Wong 2014; DataSeries 2018; Lahiri and Ghanta 2010; Sun et al. 2008). A more in-depth explanation about NN algorithms is given on subchapter 2.4.3, since it was one of the algorithms studied in this paper.

More recent studies attempted to overcome the disadvantages of the BPNN related to the high running time of the forecast and, consequently, the concept of Extreme Learning Machine (ELM) emerged. This method is a single hidden layer and feedforward NN and has a much faster capacity to learn than the gradient descent based algorithms, making it possible to use it in real time basis applications. Besides, it avoids many difficulties faced by the above method, such as choosing a stopping criteria and learning rate. The training time is shorter and the input weights and hidden layer bias are randomly assigned, while the output weights are determined by the algorithm through a generalized inverse operation of hidden layer output matrices. Although faster, the results from ELM are unstable when compared to BPNN, mainly because of the randomness involved on the definition of the parameters, instead of tuning them, which allows reaching different solutions each time the algorithm runs (Yu, Choi, and Hui 2011; Sun et al. 2008; Wong and Guo 2010).

In order to take advantage of the faster results provided by ELM, some authors proposed an Extended ELM (ELME) algorithm. This model repeatedly performs ELM with the same dataset and calculates the average results, increasing the model accuracy. Yu, Choi & Hui (Yu, Choi, and Hui 2011) assumed colour, size and price as the most significant factors for the predictive task and, consequently, as inputs of the model. They found that, while the number of iterations increase, the time required by the method also increases in an almost perfectly linearly way, assuming that all other parameters remain unchanged. However, there is a point in which this increment leads to a constant forecasting error measure, and increasing the number of iterations adds nothing but a longer duration. The time needed to perform the task also increases with an increment in the number of neurons on the hidden layer. The associated error only decreases until a certain number of neurons, because at some point the model will become too similar to the training set (overfit). The authors also found that more historical observations introduced into the model increase the duration of the task in a really slow path, whereas the forecasting error decreases drastically, indicating that it is preferable to have more data as an input of the method. Another finding was that the increase in the variance of the available dataset does not have significant impact in the time used to run the model due to a normalization performed on the input parameters, although it increases when in the presence of white noise (many different frequencies together). The study proved that ELME achieved slightly better forecasting results than the traditional NN, but the time required to perform it was superior to the latter one, given the increase on ELM number of iterations. The findings referred above are transversal to any NN method.

Sun et al. (2008) also made a study comparing ELME with two BPNN models. The chosen input data to predict jeans demand was month, date, product code, colour, size and price. They found that ELME algorithm achieved almost always better forecasting results and was more stable than the traditional NN ones, and the same happened for ELM in some cases, but the ELME model required far more time to perform the prediction than the other two. Another relevant finding was that NN algorithms perform better than ELM and ELME when there are higher fluctuations in the demand of a product. Nevertheless, Ren, Chan, and Ram (2017) affirm that ELM provides unstable results, similar to ARIMA, a statistical model, in terms of lack of performance. PPD was compared to the two referred models and ended up to be the one with more accurate results.

Another study by Au, Choi, and Yu (2008) deals with the trial and error approach used in the NN methods presented until this point. They proposed a method based on evolutionary computation (EC), capable to quickly locate the high quality areas, even when the search space is very complex and large. Joining this advantage to the NN methods, the concept of Evolutionary NN emerged (ENN). In this specific study a direct binary representation is used for every connection between neurons and, if the value is equal to one, that means that the connection exists, while if it is equal to zero there is no bridge between a neuron and another one from the following layer. Considering a single hidden layer, pruning can occur to a hidden neuron when there is no connection from any input neuron, and can also occur to an input neuron when it is not linked to any other neuron, as respectively shown in ENN pruning in (Au, Choi, and Yu 2008). In this study, the most critical aspect to consider is the number of hidden neurons, because of the time involved and a possible overfitting of the results (the model is too shaped by the training data) when in the presence of a higher number, while a small one may not train well the algorithm. For that reason, a pre-search mechanism was used to find its optimal maximum number: find a number that overfits the results and another one that does not. The authors found that the higher the time consumed by the forecasting task, the better the accuracy associated. Moreover, as the time horizon increases, the more the connections between hidden and output neurons decrease and the ones between input and output layers suffer an increment. The first finding happens because there is more available data to train the model and the second one implies that, in the presence of extra historical data, the relationship between input and output tends to become linear. ENN performs better if the available time series is smaller and has few variance associated. The output was compared to SARIMA results, which proved to be more accurate than ENN when there is a great uncertainty associated to the demand of a product. This finding may be due to the large amount of available data, which allows to use a statistical linear models to perform the predictive task.

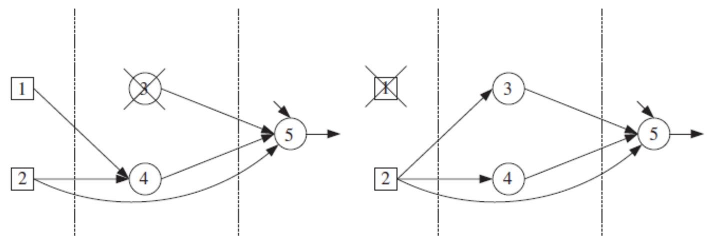


Figure 7 - ENN pruning in (Au, Choi, and Yu 2008).

Deep Neural Network (DNN) is another approach to NN and was studied by Loureiro, Miguéis, and da Silva (2018). It is a feedforward model composed by more than one hidden layer, so that the first layers can analyse more simple features while the latter ones take care of the more complex data. In this specific study, DNN was used to perform the learning process of the algorithm with three hidden layers, all of them containing the same number of hidden neurons. Many techniques were evaluated within the same conditions, in order to find the best predictor of new fashion items. The predictor variables for one season products were family, subfamily, colour, colour type, fashion, segment, store type, price, size and expectation level. With the exception of the linear regression (LR) model, all others achieved satisfactory accuracies. They also found that there is no exceptional method for every case, meaning that the choice of the best one is dependent of the available data characteristics. When in the presence of few historical data, DNN achieves a better output than Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), NN and LR. However, the performance achieved by a single hidden layer NN is not significantly different from the one achieved by DNN. Moreover, the latter one increases the complexity of the task, so it was assumed that there is no great advantage of using it. DT is a weaker method and not as robust as the others because of its simplicity. RF, however, which is a combination of many decision tress, is much simpler than NN and DNN and achieves very similar results to these two. SVM achieves good results too, although with a slightly lower accuracy than RF and NN. A more in-depth explanation about RF and SVM is given on subchapters 2.4.1 and 2.4.2, respectively, given their usage and analysis on the current work.

Some studies deal with the Grey Model (GM) because it uses the first order differential equation, which allows it to have the capacity of conducting a forecast with very few historical data. This is a non-linear method that deals with the uncertainty associated to sales patterns because of its awareness that it is difficult, if not impossible, to classify a system as “white” or “black”. Despite being simple, it is known to be unreliable for time series with high volatility, which is frequent in fashion items (Ren, Chan, and Ram 2017).

2.4.1. Random Forest

RF algorithm is composed by a set of decision trees, each one built under a group of nodes: a root node, the leaf nodes and the splitting nodes. The first one is divided into two different and non-overlapping splitting nodes, after considering every division combination within the input variables from a random sample of predictors. This evaluation is usually supported by the MSE measure and, for regression, the predicted value is calculated by the mean of the response nodes. The input that contributes with the greatest performance and, therefore, smallest MSE, is chosen to perform the split. RF follows a top-down approach without any pruning and the same procedure is applied to the two descendant nodes until reaching a stopping criteria. The average of the last constructed nodes, the leaf ones, dictates the output of the observation in study (Saraswat n.d.; Cutler, Cutler, and Stevens 2012).

The model uses a bootstrap aggregation approach in order to avoid overfitting and, consequently, increase its performance. Different bootstrapped samples are taken from the data to construct all the trees of a RF, however, normally one third of the set is used only to test the algorithm, named the out-of-bag (OOB) observations. This procedure will decrease the variance associated to the learning method and, after constructing each tree, the OOB MSE error is calculated by their average value (Cutler, Cutler, and Stevens 2012; Okun 2011).

2.4.2. Support Vector Machine

SVM algorithm, or more precisely, Support Vector Regression (SVR) (regression is used on this paper) tries to find an optimal hyperplane, named decision boundary, comprised by a margin that maximizes the Euclidean distance by minimizing the error between the observations, the support vectors. Since the output is a real number and there are infinite possible values to predict, a tolerance margin called epsilon is added to the real limit in order for SVR to produce a continuous response that deviates in the maximum or minimum of that value (R-bloggers n.d.; B. Liu 2011; Carvalho 2017).

A kernel is a function that calculates the inner product between two vectors of the training data. Sometimes the provided entrance points are not linearly separable and SVM must use one of its kernel functions to transform it into a higher dimension. In this way, the inner product from a suitable space is calculated and data can be linearly separated (Souza 2010; B. Liu 2011).

This action is called the kernel trick and it is represented as follows:

$$\mathrm{K}(x, y) = \phi(x) * \phi(y) \tag{2.20}$$

Where:

x, y are vectors from the training data and

ϕ is the suitable space.

There are four types of kernel functions, all them with different parameters to be optimized and the representations are the following:

$$\mathrm{Linear} = x * y \tag{2.21}$$

Where:

x, y are vectors from the training data.

$$\text{Polynomial} = (\delta * x * y + \kappa)^d \tag{2.22}$$

Where:

x, y are vectors from the training data,

δ, d are parameters to optimize and

κ is a constant.

$$\text{Radial basis} = e^{(-\sigma(x-y))^2} \tag{2.23}$$

Where:

x, y are vectors from the training data and

σ is a parameter to optimize.

$$\text{Sigmoid} = \tanh(\delta * x * y + \kappa) \tag{2.24}$$

Where:

x, y are vectors from the training data,

δ is a parameter to optimize and

κ is a constant.

For both linear and non-linear separable data there are multiple possible combinations of hyperplanes. In order to choose the best fitting one and, therefore, increase the performance of the model, a penalty cost is applied to the wrongly answered observations having into account the security epsilon (R-bloggers n.d.).

2.4.3. Neural Networks

The NN algorithm procedure is similar to what happens in the human nervous system: many interconnected neurons transmitting processed information between them (Goodman 2016; Günther and Fritsch 2019). NN architecture can be seen in Figure 8.

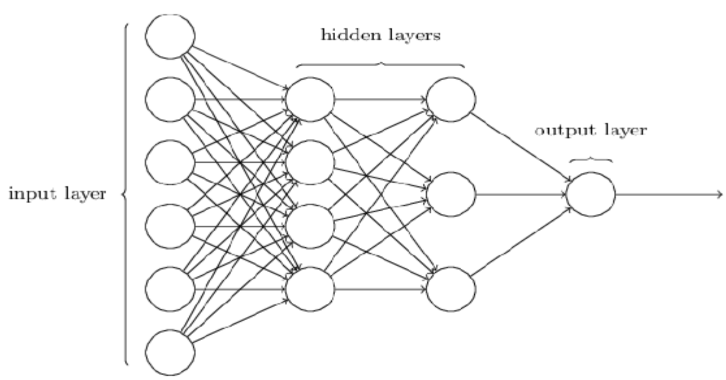


Figure 8 - Artificial Neural Network architecture in (Neapolitan 2018).

There are many distinctive types of this algorithm, but only the one evaluated in this paper is going to be explained in this subchapter: a fully connected feedforward backpropagation NN.

The model is composed by an input, an output and one or more hidden layers, and linked by many feedforward connections. Each connection has an associated weight, and for each layer there is a bias parameter that helps to improve the learning of the algorithm. The number of neurons in the first layer is equal to the number of predictor variables, while the output layer has a single neuron.

The number of neurons in the hidden layer(s) must be optimized in order to achieve the best possible performance. Usually, one hidden layer is enough to execute the algorithm (Alice 2015; Carvalho 2017).

Firstly, the input layer receives the raw data and the initial weights are randomly selected. Then, an activation function between the input and the hidden layer transforms the data received from the first into an output that will be received by the latter one. The same happens between all layers, meaning that each neuron end has an associated activation function. At the output layer, a final answer for an observation is given and the respective comparison to the real value is made. Several iterations named epochs are performed using the same training set and, at the end of each one, the average error is calculated in order to allow, according to the learning rule in use, to update the weights and convergence to the local minima of the error function (Carvalho 2017; SAZLI 2006; AnalyticsVidhya 2017).

A gradient descent learning rule is used to update the weights in each iteration. A positive partial derivate indicates that the weight is contributing to the increase of the error and, consequently, must be decreased. The opposite happens with a negative partial derivate, which means that weights go on the opposite direction of their partial derivate until finding the local minimum. The main goal of NN is to continue converging until reaching the global minimum, however, there are usually regions with local minima within the space and, if the algorithm gets stuck in them, it will never find its way to the global one. The convergence process can be computationally very time expensive (Günther and Fritsch 2019; AnalyticsVidhya 2017).

High learning rates can cause many oscillations within the space that will not lead to the model convergence, while slow ones will perform many unnecessary cycles until finding a local minimum and consequently increase the required time by the algorithm. In this way, finding an optimal learning rate is mandatory to achieve good performance results while minimizing the duration of the task (Alice 2015).

2.5. Hybrid Forecasting Techniques

To overcome some of the AI methods challenges, in recent years the concept of hybrid models has come up. These methods make use of the advantages linked to statistical and AI models together, or linear and non-linear methods (Ren, Choi, and Liu 2015).

Choi et al. (2014) studied a hybrid model called Fast Fashion Forecasting (3F), which combines ELME and GM methods, joining GM advantage of performing the predictive task with few historical data with the adaptability of ELME to non-linear events. Four periods of knitwear sales were analysed, and the input variables were style and colour. It was crucial to find the optimal number of ELME iterations, and the number of hidden neurons was established as two. A first order differential equation is used in GM. The latter one executes the main forecast and ELME takes care of the residual series. The main findings were that, as the running time limit of the model increases, the forecasting accuracy increases too, but at a decreasing rate. Moreover, the results obtained by the 3F model were compared to the ones achieved with GM and a hybrid method composed by GM and NN models. The first comparison showed that 3F performs significantly better when it has enough time to complete the task. The second, instead, showed that a GM-NN hybrid model has lower forecasting error associated, because while using the 3F method, accuracy is slightly compromised by the speed. The relationship between trend, seasonality and noise of the time series and the forecasting accuracy were also studied. When in the presence of a dataset with high variance because of the white noise, the GM, 3F and two static GM-ELME models (fixed number of input and hidden neurons) with different time limits performed all poorly. However, when this component decreases, the accuracies of all of them increase. A decrease in the seasonal variance shows that the forecasting error of the methods also decrease, however it is not possible to differentiate it from the noise. Also, if the slope increases, the results of the algorithms show a better accuracy. The authors finally concluded that the 3F model performs very well when the dataset is characterized by a large trend slope and the seasonal variance is large. Moreover, it has the capacity to do the forecasting of fashion items very quickly and with few historic data available.

Another proposed hybrid model joined ELM and Harmony Search (HS) methods (HS-ELM). The latter is a meta-heuristic technique and searches for optimal input weights and hidden biases, instead of these being randomly generated. Wong & Guo (Wong and Guo 2010) explored the hybrid model and, to start, applied a pre-processing task to remove outliers, interpolate missing values and normalize the data from different periods and cities. Afterwards, the algorithm runs repeatedly for many numbers of hidden neurons and, in the end, the results are analysed by a heuristic fine-tuning process and the final sales forecasting is generated. A comparison was also made between the hybrid model plus the pre-processing task and without it. The results proved that the first hypothesis achieved better performance, essentially because more reliable training samples are constructed, which gives emphasis to the importance of this first step in the predictive process. Moreover, the proposed model was compared to ELME, ENN, ARIMA and AR, and its forecasting results achieved better accuracy than all the others, mainly due to the final part of the procedure, the heuristic fine-tuning process, that eliminates every forecast with a high associated error. However, AR and ENN achieved not so distant results from the HS-ELM model, especially when long term forecasts are conducted. Nonetheless, this model is based on a time series forecast and ignores the explanatory variables such as weather and economics, which are highly present on fashion sales patterns.

Ren, Choi, and Liu (2015) studied a Panel Data Based Particle Filter (PDPF) model. The first model, PPD or Panel Data, allows the study of a product behaviour having into account the other items actions and the second one, Particle Filter (PF) model, is a sequential Monte Carlo methodology and performs well in highly nonlinear environments and consequently predicts uncertain patterns of the sales. Given the characteristics of the two models employed, the inputs of the algorithm were the price of the items, time series trends and cross-sectional sales of correlated products. The analysis showed that both PPD and PDPF achieved stable and better accuracies than ARIMA, PF alone and Relevance Vector Machine (RVM), a Bayesian version of SVM. The last two did not achieve a good performance because they neglect the correlation between items, which emphasizes the importance of considering the interaction between different products in the sales forecasting. Moreover, an analogy was also made between the two Panel Data methods used. Besides PDPF performance being better than PPD almost every time, this last one outperforms the hybrid model when making a prediction for belts because of the presence of a big outlier, which accounts for a high discrepancy, and for bags due to a linear relationship of the data. This study also accentuates the high correlation between changes in price and sales of fashion items and concludes that more quantity of sales data does not improve that much the training of the model.

Thomassey and Fiordaliso (2006) studied a hybrid model based on clustering and decision trees. The choice of these two methods was based on their easy interpretability by the user and good performance. The clustering is an unsupervised method (the output is not known), and the one used in this study was the exclusive k-means. Different results at different runs can arise due to the random beginning of this algorithm, meaning that the centre of each cluster does not represent the optimal value, but a local one. Furthermore, it is crucial to define the desired number of total clusters and in the study it ranged from 2 to the square root of the dataset size, 20. The inputs of the model is the available time series, which is normalized before being introduced, the price of each fashion item and both starting date and duration of the selling. Afterwards, the clustering procedure aggregates the similar sales patterns of the items in families/prototypes and the DT assigns one of them to each item, new or not and performs the forecast. The DT induction algorithm used, which is composed by a root node, leaf nodes and branches, was the C4.5 because of its accuracy and adaptability when dealing with the tree pruning, characteristic that solves the overfitting problem, and the processing of the available numerical and nominal data. There is a DT for each prototype and the criteria to split the nodes relies on the information gain ratio, which analyses what attribute that is not yet in the path contributes most to the information value. The authors adopted a post pruning strategy, meaning that the pruning occurs only after the tree construction by replacing or raising a subtree. Moreover, the learning technique was based on a 10-fold cross validation with a training and a test set. At the end, the number of clusters that minimizes the forecast error is selected. The results had a good accuracy for almost every prediction, and the ones with a higher error were associated to the lack of significant inputs to explain the data. For that reason, it is thought that including factors such as style and textile material would have a great influence on the performance of the algorithm. However, the main disadvantage of this procedure is the many necessary repetitions that increase a lot the computational time required to run the forecast.

Given the very different studies about AI and hybrid forecasting techniques on the literature, Table 26 on Annex B was constructed and synthesizes the main findings of each paper presented above.

2.6. Evaluation Metrics

An error is the quantity by which an estimated value differs from the reality. There are many widely used error measures to evaluate the performance of a predictive model, and they are usually applied to classify the quality of a range of methods in order to choose the best fitting one. However, there is no outperforming and best error measure for all forecasting tasks because of the different prediction goals, data patterns and pre-processing of the available historical data, which is why normally more than one measure is applied to decrease the bias generated by just one (Wong and Guo 2010; Shcherbakov et al. 2013; Du, Leung, and Kwong 2015).

Within the wide variety of existing methods, the second most used ones are the Absolute Forecasting Errors because of their capacity to quantify the difference between predicted and true values and being simple and easy of use. The common factor between the measures in this group is the calculation of the error e_i:

$$e_t = (y_t - f_t^{(m)}) \tag{2.25}$$

Where:

y_t is the real value at the time t and

$f_t^{(m)}$ is the forecasted value by model m at the same time.

The most used measures in the fashion industry forecasting are:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{2.26}$$

Where:

n is the number of observations.

This measure was used by two of the studies referred above, (Du, Leung, and Kwong 2015; Loureiro, Miguéis, and da Silva 2018).

Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i^2) \tag{2.27}$$

Where:

n is the number of observations.

This measure was used by seven of the studies referred above, (Ren, Choi, and Liu 2015; Yu, Choi, and Hui 2011; Xia and Wong 2014; Ren, Chan, and Ram 2017; Loureiro, Miguéis, and da Silva 2018; Sun et al. 2008; Au, Choi, and Yu 2008).

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i^2)} \tag{2.28}$$

Where:

n is the number of observations.

This measure was used by four of the studies referred above, (Wong and Guo 2010; Xia and Wong 2014; Loureiro, Miguéis, and da Silva 2018; Thomassey and Fiordaliso 2006).

The main disadvantage of such methods is the fact that they cannot deal with very high differences in the magnitude of the data and occasional outliers (Shcherbakov et al. 2013). Moreover, they are not good measures to accommodate many zero values (Ren, Choi, and Liu 2015). The Absolute Error can be used if the available data does not suffer large variations in its pattern and if a pre-processing task is performed before the forecast, in order to clean it and remove outliers (Shcherbakov et al. 2013).

There are also many measures based on Percentage Errors that solve some of the drawbacks of the Absolute Errors, and are the most used to measure forecasting performance. Their common factor is the value p_i:

$$p_t = \frac{|e_t|}{y_t} \tag{2.29}$$

Where:

y_t is the real value at the time t and

e_t is the error.

The most used measures of Percentage Errors in the fashion industry forecasting are:

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} 100 * |p_i| \tag{2.30}$$

Where:

n is the number of observations.

This measure was used by six of the studies referred above, (Wong and Guo 2010; Xia and Wong 2014; Choi et al. 2014; Du, Leung, and Kwong 2015; Loureiro, Miguéis, and da Silva 2018; Thomassey and Fiordaliso 2006).

As an alternative to MAPE, there is a measure called Weighted Mean Absolute Percentage Error (WMAPE), which analyses the impact of different frequency sales quantities (Chockalingam 2018):

$$WMAPE = \frac{\sum_{i=1}^{n}|e_i|}{\sum_{i=1}^{n} y_i} \tag{2.31}$$

Where:

n is the number of observations,

y_i is the real value at the time t and

e_i is the error.

This measure was not used in any of the studies referred above.

The Median Absolute Percentage Error (MdAPE):

$$MdAPE = \underset{i=1,n}{median} 100 * |p_i| \tag{2.32}$$

Where:

n is the number of observations and

$median$ an operator.

was used by one of the studies referred above, (Thomassey and Fiordaliso 2006).

The metrics presented above are less sensitive to large errors and outliers than the Absolute Errors, but, excluding WMAPE, they are still quite affected. Furthermore, when the value of y_t is zero, p_i tends to infinite. There is also the problem of non-symmetry because the error differs if the estimated value is bigger or lower than the real one.

The Symmetric Errors have a common factor too, s_t:

$$s_t = \frac{|e_t|}{(y_t + f_t)} \tag{2.33}$$

Where:

y_t is the real value at the time t,

f_t is the forecasted one at the same time and

e_t is the error.

The most used measure of Symmetric Errors in the fashion industry forecasting is:

Symmetric Mean Absolute Percentage Error (SMAPE):

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n} 200 * |s_i| \tag{2.34}$$

Where:

n is the number of observations.

This measure was used by two of the studies referred above, (Ren, Choi, and Liu 2015; Ren, Chan, and Ram 2017).

SMAPE is useful to reduce the bias caused by the values equal to zero, but it has some disadvantages such as the fact that s_t can also be divided by zero if the estimated and the true value

are the same but with opposite signs or if both are equal to zero and this method is still very affected by outliers and different magnitudes.
Shcherbakov et al. uses Normalized Error measures to deal with seasonal time series, and in this method the normalization of the data occurs by cycles:
Normalized Root Mean Square Error (nRMSE):

$$nRMSE = \frac{1}{\bar{y}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i^2)} \tag{2.35}$$

Where:

n is the number of observations and

$\bar{y}$ is the normalization factor which is equal to the maximum value, or to the difference between the maximum and minimum values of the entire time series, or just within an interval of it, which usually corresponds to a season/cycle.

However, this method can be influenced by outliers.

This measure was not used on any of the studies referred above.

3. Current Situation and Further Analysis about Forecasting in the Fashion Industry

3.1. Oracle Retail Demand Forecast

Retail Demand Forecast Version 16 (RDF V16) is one of the customer-oriented solutions offered within Oracle Retail software portfolio, more specifically, by the Supply Chain area, and, as the name implies, performs and delivers a forecasting approach for the demand of a retail item, both on a short and a long-term horizon (Artley 2018).

An accurate prediction is needed to ensure the correct order of an item, as well as the right quantity of it and at the right time. In order to achieve a good forecast, RDF V16 uses a variety of predictive models based on statistical and promotional approaches.

The following three subchapters were based on the Oracle Retail Demand Forecasting online Manual, Version 16 (Artley 2018).

3.1.1. Pre-processing

RDF V16 performs multiple pre-processing tasks prior to the forecast to avoid the possibility of replicating undesired demand patterns, by converting the raw sales into unconstrained demand. Unconstrained demand is the total quantity that would have been sold without purchase limitations, such as stock-outs, strikes and closed stores. Thus, data conversion is made to assert a more typical sales quantity (including the lost sales). In order to properly smooth the data source, up to six pre-processing runs can occur to the same dataset, but usually this value is between three and four.

Oracle Retail considers two main types of inputs to the forecast, the baseline and the causal. The first one requires a four step pre-processing: stock-outs correction, outliers' or/and special events' correction (storm, for example), promotions and short-term price changes correction (peaks of sales) and smoothing of the sales. The second one only requires three steps: stock-outs correction, outliers' or/and special events correction and seasonality correction (deseasonalize).

RDF V16 has the following available options to pre-process the data:

a) Standard Median, an average value for a time length that provides a baseline type of input. A graphical example can be seen in Figure 9, where the black line represents the raw historical data and the red one represents the adjustment performed by Standard Median.

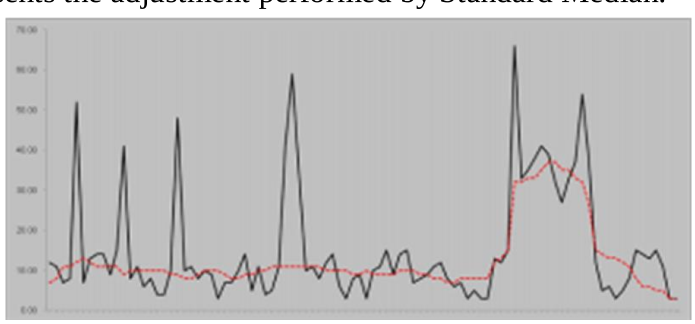


Figure 9 - Pre-processed data by Standard Median vs raw data in (Artley 2018).

b) Retail Median provides also a baseline input for the forecasting task by calculating five Standard Median filters, aiming to contribute with a smoother and more fluid final dataset. A graphical example can be seen in Figure 10, where the black line represents the raw historical data and the red one represents the adjustment performed by Retail Median.

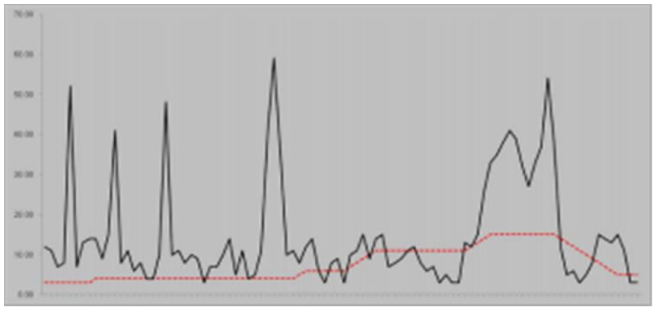


Figure 10 - Pre-processed data by Retail Median vs raw data in (Artley 2018).

c) Standard Exponential Smoothing (StandardES) is used to smooth peaks in demand and cover zero sales situations. The user must give as an input the periods in which such events occurred, and consequently are going to be pre-processed. The past and future exponential weighted averages of a certain time moment are calculated and submitted to a linear interpolation, in order to detect the positive or negative adjustments to the raw history. A graphical example can be seen in Figure 11, where the white line (POS) represents the raw historical data, the red one (OUTAGE) represents the peaks or stock-outs and the blue line (LSOVER) represents the adjustment performed by StandardES.

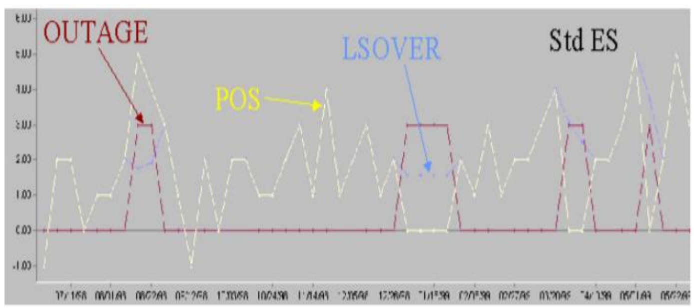


Figure 11 - Pre-processed data by StandardES vs raw data in (Retek Predictive Application Server 2004).

d) Lost Sales Standard Exponential Smoothing (Lost Sales StandardES) is very similar to StandardES, the only difference being the fact that it only makes positive adjustments (stock-outs) and it also smooths the following period after the pre-processed one if the user pretends so. A graphical example can be seen in Figure 12, where the white line (POS) represents the raw historical data, the red one (OUTAGE) represents the peaks or stock-outs (but Lost Sales StandardES only does positive adjustments, meaning that it only deals with the stock-outs) and the blue line (LSOVER) represents the adjustment performed by Lost Sales StandardES.

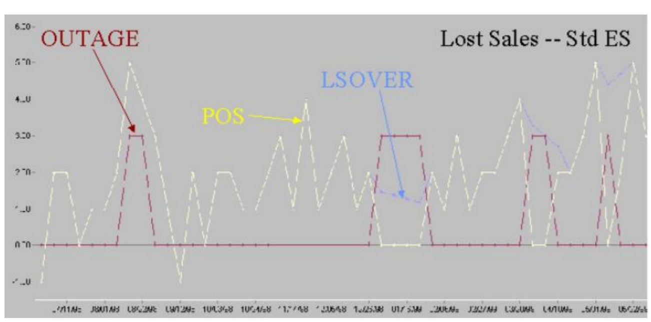


Figure 12 - Pre-processed data by Lost Sales StandardES vs raw data in (Retek Predictive Application Server 2004).

e) Override also deals with peaks and stock-outs, but it requires as input a reference measure, which can be associated to an adjustment percentage. It will only override the periods that are intended to be masked, in case the user pretends it. A graphical example can be seen in Figure 13, where the black line represents the raw historical data, the blue one represents the reference measure, the white dots represent spikes on demand or stock-outs from the history and the red one represents the adjustment performed by Override.

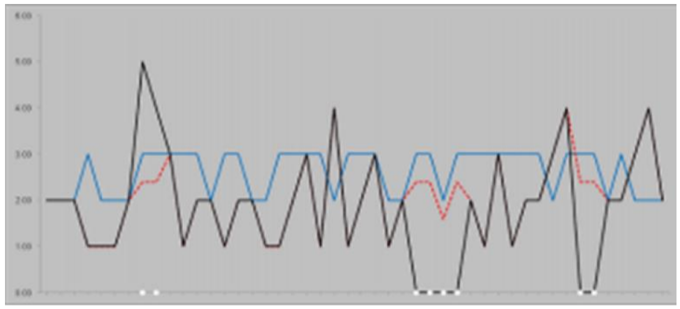


Figure 13 - Pre-processed data by Override vs raw data in (Artley 2018).

f) Increment is very similar to Override, but instead of copying the reference measure, it will increment or decrement the raw data (actual sales plus increment or decrement). The reference measure can also be adjusted by a percentage and only the periods that are intended to be masked will be submitted to a change, in case the user pretends it. A graphical example can be seen in Figure 14, where the black line represents the raw historical data, the blue one represents the reference measure, the white dots represent the peaks or stock-outs and the red one represents the adjustment performed by Increment.

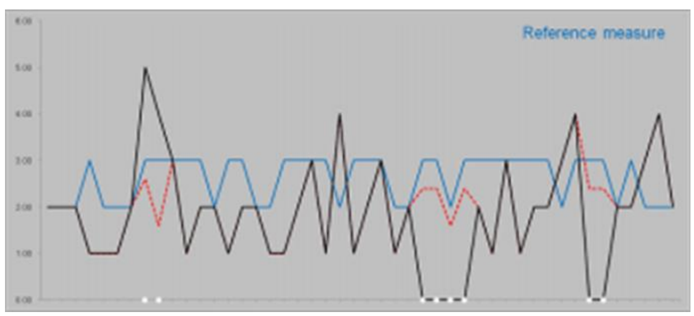


Figure 14 - Pre-processed data by Increment vs raw data in (Artley 2018).

3.1.2. Time Series Based Forecasting Techniques

Table 1 synthesizes the predictive methods used by RDF V16, some of them already explained on the above chapter 2.3.

Table 1 – Time series based forecasting techniques used by RDF V16.

Time Series Parameters	Forecasting Techniques
Level	Moving Average; Simple Exponential Smoothing; Croston's method: used in the presence of intermittent demand. The time series is divided into two: the magnitude series contains all data points different from zero, and the frequency series comprises the intervals between consecutive non-zero data points. After this division, a statistical procedure is executed on the two new time series.
Level and Trend	Damped Holt's Exponential Smoothing: attenuates trend factor over the weeks following the forecast.
Level, Trend and Seasonality	Holt-Winters' Exponential Smoothing; Seasonal Regression: simple linear regression. The slope of the regression line can be used to capture the series trend and the cycle corresponds to its seasonality.
All possible combinations between the three parameters	Profile-Based: firstly, it deseasonalizes historical data. Then, SES is executed and the series is reseasonalized with the seasonal profile. Possibility to predict demand for new items using the profile of a similar SKU; Bayesian: combines historical data with a sales plan developed by the user, and the algorithm is composed by two parameters: shape, the lifecycle of a product, and scale, the total quantity that is going to sell over the time horizon. The latter one can be adjusted while the predicted values start to arise. Possibility to predict demand for new items. Promotional/Causal: requires time series data and the history of previous and future promotions. The methodology consists of a stepwise linear regression, performed in order to estimate what promotional variables are most relevant and their respective effect on the time series. Then, they are applied to the baseline forecast, which is the prediction without promotional/causal effects, and generate a prediction. Nonetheless, sometimes there is lack of information on promotions and/or other events on the historical calendar, but it is possible to see their effects on the time series. Oracle Retail deals with it by eliminating the effects of such events, even though they can have influence on the prediction's accuracy. Some other approaches can be used, such as performing source level forecast (higher level on the product hierarchy), which would force the same effects on different series, or use an average effect of events from a different series.

3.1.3. Choice of the Most Adequate Forecasting Techniques

The algorithms executed in RDF V16 can be applied to any type of retail item, from grocery to fashion, meaning that the user of the algorithm should have some knowledge about the available data in order to choose the best model.

V16 has a module to choose the better performance model to apply on the available data. It makes a comparison between some algorithms using a procedure called Automatic Exponential Smoothing (AutoES), as shown in Figure 15. The Simple Moving Average, Profile-Based, Bayesian and Promotional/Causal methods are not automatically evaluated, so in order to use them the user must select them manually. AutoES evaluates two parameters of the models: goodness-of-fit, which is seen as a reward, and complexity, which is linked to a penalty. The first one is directly related to the quality of the available pre-processed data and the latter is essential to avoid the over-fitting of the model. The Bayesian Information Criterion (BIC) tries to balance the two parameters and the model that minimizes BIC is the one chosen by AutoES. BIC's formula is:

$$BIC = s * n^{k/2n} \tag{3.1}$$

Where:

n represents the number of available historical periods,

s is a measure of goodness-of-fit and

k is a measure of the model complexity.

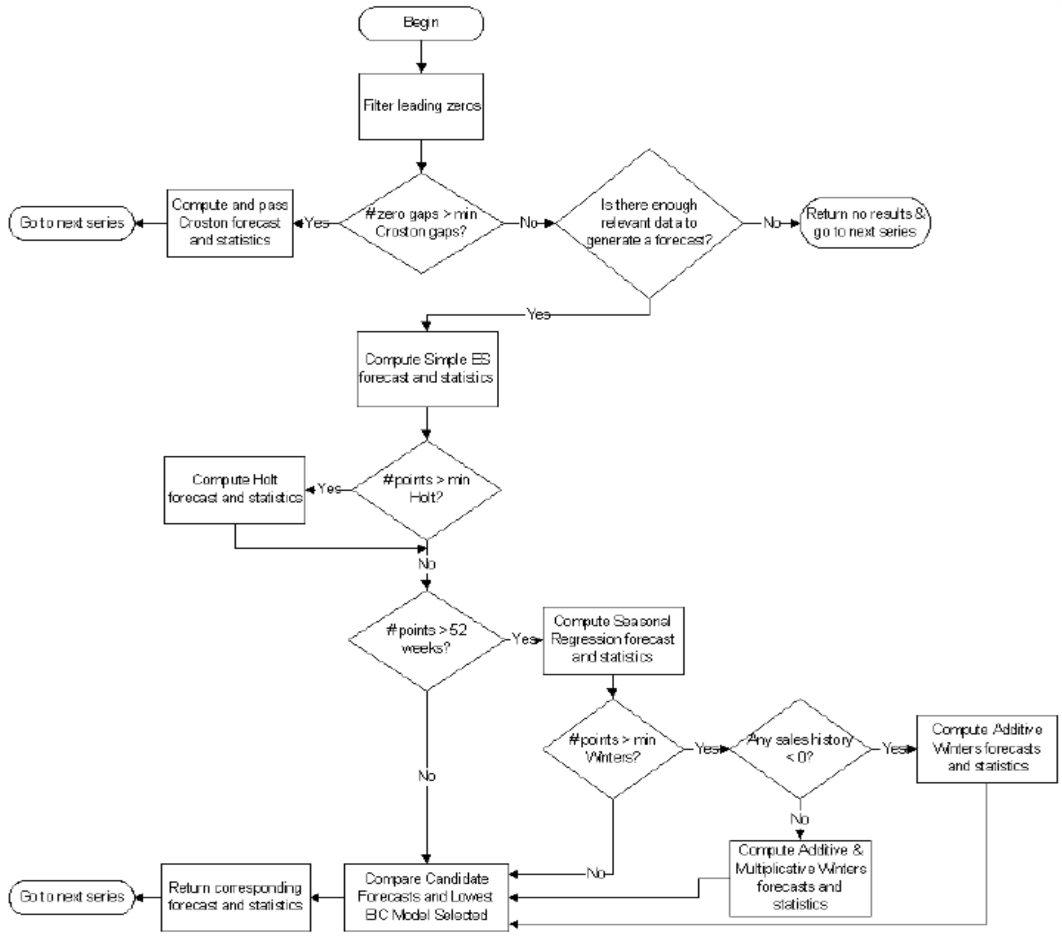


Figure 15 - AutoES behaviour in (Artley 2018).

3.1.4. Challenges and Solutions

Sometimes the available historical information is sparse and noisy, which makes the task of identifying patterns within the data very difficult. This usually happens on the final level of the hierarchy (item/store), and the solution is to aggregate the data until reaching an optimal higher level, called source level, in order to smooth it. By performing an interim forecast it is possible to discover the relationship between the different levels within the chain. The source level is defined by the user and RDF V16 has an option called AutoSource which enables the comparison of many different source levels' accuracy. Afterwards, a forecast is generated for the source and the result is spread down to the final level, having into account the weight of each specific item on the aggregated level. However, every item at the final level is considered to have the same forecast curve shape. The process can be seen in Figure 16.

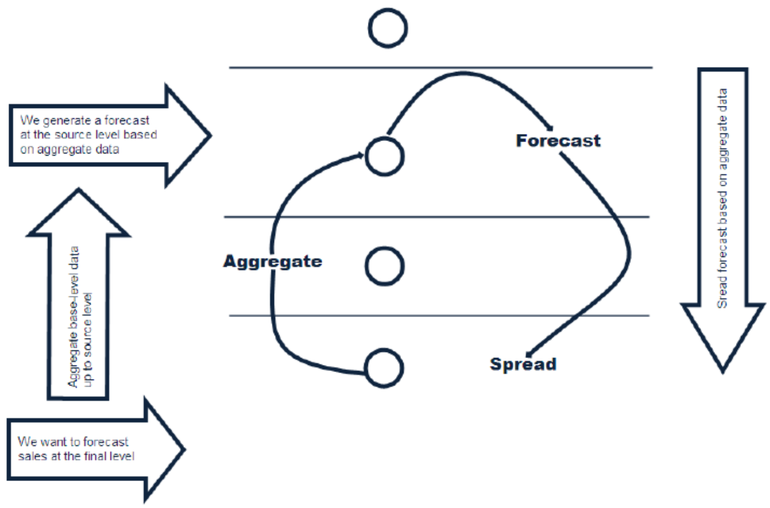


Figure 16 - Source level forecast in (Artley 2018).

Since there is no historical information available for new products and/or locations, the forecast can be based on data of products and/or locations with similar characteristics and history.

RDF V16 can handle the impact of events such as promotions, discounts on price, marketing strategies and others in historical sales that cause unusual behaviour by using the Promotional/Causal forecasting, which allows the program to detect the reasons for such deviations, measure its effects and then use them to predict similar sales on the future. Promotions or events that occur every year, exactly at the same time, are considered seasonality, while floating events that occur at different times of the year are considered causal ones.

In order to manage all different item/store combinations, RDF V16 uses an automated process that provides alerts when some forecasted value is above or below an established threshold range.

All returned items are assumed to be re-sellable and follow a linear function as follows:

$$Returns(t) = Forecast(t - t') * percentage \tag{3.2}$$

Where:

t is the present time,

t' is the period between the purchase and the return,

$Forecast(t - t')$ is the forecasted item in a period and

$percentage$ is the percentage of items returned.

3.2. Importance of Accurate Forecasting Techniques

Predicting the demand of a fashion item is a crucial activity for inventory planning and replenishment. The lack of quality on the forecast output can give rise, for instance, to stock-outs or obsolete products, and such problems consequently bring customer dissatisfaction and an increase in costs. Also, the orders lead time is usually high, meaning that the forecast horizon should cover many weeks of a season, or even an entire season, with a good accuracy. Some companies make one or more reorders of the selling items, however, in the case of fast fashion SKUs that are only sold for a small period of time, reorder lead time can be superior to the items living time in the shelves, making it impossible to correct a forecast error with a replenishment (Nenni et al. 2013).

3.3. Factors that Contribute to the Choice of the Forecasting Techniques

Ren, Chan, and Ram (2017) studied the factors that a retailer values most about a forecast algorithm, which are accuracy, speed, data sufficiency, stability of the results and ease of use. To measure this, the authors created a decision support system based on an analytic hierarchy process (AHP) that has three levels: goal, criteria (same as the factors presented above) and alternative. The main idea is to achieve a specific goal, use the best model, with the range of available criteria and alternatives. After some research, they found that the criteria that has higher influence for a good performance of the model are, by descending order, accuracy, stability, ease of use, data sufficiency and speed.

Given the challenges associated to the time series based forecasting techniques presented above, as well as the recent literature and research on the field, it can be assumed that some predictive methods are capable to outperform or complement statistical ones. This project, Forecast of Short Lifecycle Items in the Fashion Industry, emerged from the recognition of such opportunity.

4. Fashion Industry Forecasting Techniques

4.1. Data Pre-processing

The data pre-processing procedures applied in this paper were defined with the support of the book “Extração de Conhecimento de Dados”, Carvalho (2017).

Raw data is supplied by datasets, which are composed by variables and observations. Figure 17 synthesizes the data, and lack of it, in a common dataset.

Variables / Observations	Product	Quantity	Colour
1	A	*NA*	Black
2	B	1.4	Black
3	C	2.1	*NA*

Figure 17 - Dataset aspect.

Colour is considered a categorical variable because it is associated to specific value of a class, for example, black, green and blue. Quantity is recognized as a numerical attribute. Furthermore, dataset columns can also be characterized as continuous or discrete. The first categorization happens when variables have an infinite number of possible values and an associated measure unit, while the second one can contain both finite and infinite number of them. Binaries, for instance, are considered numerical discrete variables because they can only assume two outcomes: 0 or 1.

A careful data cleaning should be performed before running any AI technique because raw data usually contains redundant, duplicate and missing values (represented by NA in the figure above). Moreover, it is common to find associations between supposedly independent variables among different datasets that are not efficiently aggregated.

Pre-processing is the action of cleaning, integrating and transforming the available historical data into a more reliable set, by approximating the raw values to their real distribution. Usually, it is the longest task and the most important step of the entire AI process, by reason of decreasing inputs complexity and, therefore, facilitating patterns recognition. In addition, a good pre-processing task will cut back on the time required to execute the algorithms, allowing them to work in a more approximate real time basis and aligned with companies’ necessities.

The pre-processing methodology used in this work is represented in Figure 18.

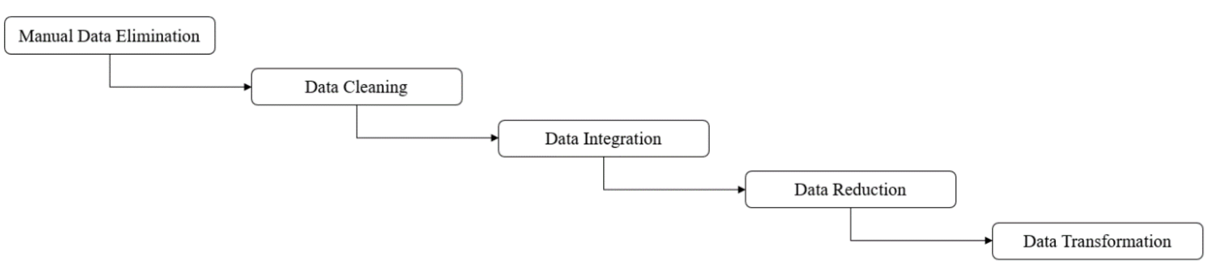


Figure 18 - Pre-processing methodology flowchart.

The first task refers to the manual data elimination, in order to understand the available and most propitious variables to analyse and, consequently, discard irrelevant, unfilled or entirely equally filled columns. It is considered manual because of the task ambiguous character, since it can have different results regarding the person who is performing it.

Subsequently, a cleaning process is executed, comprising the following tasks:

- Handle missing values. Since many AI packages from Rstudio do not accept the presence of such entries in the algorithms, they must be filled with a "No information" label, a possible label among the existent ones, or the entire observation must be deleted;
- Handle noisy, redundant and/or inconsistent values using the same strategy as for NA values. A possible way to decrease noisy and sparse data (errors) is to aggregate the variables at a higher level;
- Handle outliers. According to Hawkins, "an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Before considering an observation as an abnormal value, it is good practice to evaluate possible events that occurred within that period of time that led to such increase or decrease.

The cleaning task requires attention, critical evaluation and much time, since incorrect observations removal can lead to the loss of important information.

Data integration is the process of combining multiple tables through their ID columns into a single or a few ones. Variables often appear in more than one table within the same database, however, occasionally the information linked to them is contradictory or repetitive. When these phenomena happen it is a case of inconsistency or redundancy, respectively.

Data reduction refers to the action of selecting a subset of features for the final dataset, which is going to serve as the predictive model input. For the purpose, new columns can be created with data based on external information, such as websites or catalogues of a given brand. This task should be carefully performed, since increasing the number of attributes consequently increases the complexity and computational time required by AI methods. Notwithstanding, the increment of observations represent more available information to train the algorithm, representing a reduced overfitting probability and increased performance.

Finally, data transformation is performed to transform one or more variables into a common scale. Some AI techniques only accept a specific type of data and, in such cases, transforming the data is an indispensable task. Lastly, transformed data frequently accelerates the predictive methods, especially when in the presence of very dissimilar values. The modification can occur both on numerical and categorical attributes, and there are multiple procedures to perform it.

RF is the only algorithm used in this project that does not require any type of data transformation. SVM and NN do not handle categorical variables, thus both models require data transformation of some attributes into, for example, dummy variables. Also, standardizing the numerical output can lead to a decrease on the model training time and unscaled data can bring complications to the algorithm learning process (Lin 2012; Brownlee 2019).

4.1.1. Company 1

Company 1 is a Mexican fashion retailer with stores widely spread throughout the country. The retailer sells clothes, accessories and shoes for woman, man, children and babies among its different brands, but not all of them sell every listed item.

Raw Data Elimination, Cleaning and Integration

Sales datasets from Company 1 physical stores were provided, as well as tables combining its organization, calendar and products characteristics. The period of time encompassed was from February 2017 until the beginning of March 2019.

Firstly, datasets were imported to MySQL Workbench in order to perform a data analysis and extraction of the relevant information within them. Sparse, noisy and redundant data were eliminated. The remaining documents with at first sight potentially valuable inputs are presented throughout Tables 2 to 7.

Table 2 allows to understand how the company organizes its product hierarchy. SKU is the last available product level, and it is a combination of stylecolour and size. The variable style was discarded on behalf of only differing from the stylecolour by not having the colour encode, and it would, therefore, increase the complexity of the forecasting algorithm with the large number of available levels. The firm sells multiple brands, and it is conceivable to find just one brand or many in the interior of a store. The company column was deleted due to being entirely filled with the same value, Company 1 name. There were no missing values among the columns.

Modifications were made to the attributes brand, department, family, class and subclass. An example of the raw data is presented on the table. Some attributes have an extension of the previous ones, which highly increases the possible input combinations for the predictive model. For this reason, the repetitiveness of the five variables was deleted within the entire dataset. In this way, department only has one feature called “Woman”, instead of (1*number of brands). The new data is represented as follows:

- Brand: Brand X;
- Department: Woman;
- Family: T-Shirt;
- Class: T-Shirt Short Sleeve;
- Subclass: Basic.

Table 2 - Company 1 product hierarchy.

Variable	**Type**	**Key**	**Example**
SKU	Varchar(50)*	Primary*	T-shirt Short Sleeve Basic Stamped White size M
Stylecolour	Varchar(50)		T-shirt Short Sleeve Basic Stamped White
Style	Varchar(50)		T-shirt Short Sleeve Basic Stamped
Subclass	Varchar(50)		T-shirt Short Sleeve Basic
Class	Varchar(50)		T-shirt Short Sleeve
Family	Varchar(50)		T-shirt
Department	Varchar(50)		Woman Brand X
Brand	Varchar(50)		Brand X
Company	Varchar(50)		Company 1

*Varchar(50): varchar represents a character variable with maximum length equal to 50. Primary means that there is only one combination of that specific one or more variables.

Table 3 synthesizes the sales of a given product during an entire week in a specific store. Company 1 considers that week number 1 corresponds always to the first week of February and week 52 to the last week of January. The variable week is encoded, compiling both week of the year and year of sale: “w12_2017”, for example. The quantity sales and retail sales columns had negative values representing the returned items and their value in monetary units, respectively. Since the goal of this project is to predict the demand of a product range, such observations were deleted from the dataset because, if such items were considered in the forecast, clients who are not going to return the product may not have it available at the moment of purchase, because clients who will return it have already bought the stylecolour. Some rows were discordant and consequently deleted because they had positive values for retail sales, for instance, and negative for the subsequent attribute, quantity sales. There were no missing values among the columns, but there were cells with zero value, making it necessary to delete those entire rows.

The attributes price regular, price promotional and price clearance were calculated and added to the dataset by using the formula: $\frac{retail\ sales}{quantity\ sales}$ for each one. Nonetheless, prices among the same stylecolour were sometimes incoherent, with some of them at nearly zero. Because of that, the three calculated variables were not considered as an input of the forecasting models.

Table 3 - Company 1 sales.

Variable	Type*	Key	Explanation
Week	Varchar(50)	Primary	Week, product and store of a purchase.
Stylecolour	Varchar(50)	Primary	
Store	Varchar(30)	Primary	
Retail regular sales	Decimal(5,2)		Sum of retail sales in currency; Regular means non-promotional and non-clearance sales; Promotional means a one or slightly more days period where the product is being promoted; Clearance traditionally means lower prices period in order to clear the existing stock.
Retail promotional sales	Decimal(5,2)		
Retail clearance sales	Decimal(5,2)		
Quantity regular sales	Integer		Sum of quantity sales in units.
Quantity promotional sales	Integer		
Quantity clearance sales	Integer		

*Decimal(5,2) represents a numerical variable with a maximum of 5 units and 2 decimal cases. Integer represents a numerical variable without decimal cases.

Table 4 indicates some characteristics of the 135648 stylecolours available. This number is not an error, since the table comprises products from long before 2017, contrary to the ones pointed out above. Since the initial and final price columns do not have differentiation between the three types of sales (regular, promotional and clearance), both were discarded. The consumer attribute was also ignored because of its 62% of unfilled values, which led to the assumption that such input does not add additional value and may even complicate and slow the predictive model. There were no missing values among the remaining columns.

The column colour was composed by 100 different values and, as it was pointed out before, highly detailed attributes can have a negative contribution to the model performance. For that reason, colour classes were joined in three logical ways, such as follows:

- Navy Blue + Ocean Blue + Baby Blue = Blue;
- Kakhi + Olive + Jade = Green;
- A colour coming up in more than 100000 stylecolours was not combined to another.

Table 4 - Company 1 stylecolour attributes.

Variable	Type	Key	Explanation
Stylecolour	Varchar(50)	Primary	Product code.
Colour	Varchar(20)		Colour code.
Initial price	Decimal(5,2)		Initial Price of that stylecolour in currency.
Final price	Decimal(5,2)		Final Price of that stylecolour in currency.
Consumer	Varchar(20)		Type of consumer that will use that stylecolour.

Table 5 summarizes the quantity sales of a given SKU. There is no separation between regular, promotional and clearance sales. Since there is lack of information about what SKU observation corresponds to each of the sales type, the attribute size was not considered an entry into the predictive task. Besides, it is common that different machines producing an item in one or different plants end up with slightly differing same size products. Moreover, size characteristic varies significantly from season to season among the same brand.

Table 5 - Company 1 size sales.

Variable	**Type**	**Key**	**Explanation**
Week	Varchar(50)	Primary	Week, product, size and store of a purchase.
Stylecolour	Varchar(50)	Primary	
Size	Varchar(30)	Primary	
Store	Varchar(30)	Primary	
Quantity sales	Integer		Sales in units.

Table 6 presents Company 1 seasons. The firm splits the year into two big seasons: S/S and A/W. It is possible to have sub-seasons with smaller horizon within each season. Each brand has its own seasons and sub-seasons, and it is possible to use different season classifications at the same time for woman, man, children, babies, accessories and shoes. There were no missing values among the columns. However, stylecolours presented in this dataset have merely 26% correspondence with the ones presented in Table 3. Given that, the entire table was discarded.

Table 6 - Company 1 seasons.

Variable	**Type**	**Key**
Stylecolour	Varchar(50)	Primary
Season	Varchar(50)	
Sub-season	Varchar(50)	

Table 7 condenses Company 1 store hierarchy. Country attribute was ignored because the retailer acts uniquely within Mexico. There were no missing values among the columns.

Table 7 - Company 1 store hierarchy.

Variable	**Type**	**Key**	**Example**
Store	Varchar(30)	Primary	Store X
Type of store	Varchar(30)		Boutique
Region	Varchar(30)		South Centre
Country	Varchar(30)		Country Y

The following step was understanding which variables presented above could add value to the predictive model and associate the tables in a single dataset. Since there are three different types of sales with very distinctive quantity ranges, a table with week and purchased units was constructed for each in order to perform a more in-depth analysis. The three were imported to RStudio, aiming to graphically analyse one by one and evaluate the meaning of peak values. It was not possible to visualize the three type of sales together, given the great difference between the purchased quantities, for example, regular sales would cover every peak of the promotional sales. Figures from 19 to 21 are relative to regular, promotional and clearance sales, respectively.

Figure 19 incorporates the total quantity of items purchased with regular prices throughout 2017, 2018 and part of 2019. It is possible to observe a pattern within the points, and sales do not suffer many variations for the years in study.

To consider the possibility of outlier values, a study was made in order to find the causes of the main peaks:

- From week 25 to 28 (end of July and start of August) there is a peak of sales with medium amplitude, compared to the remaining ones. This cyclical event is due to the beginning of classes in Mexico and, in fact, the most successful and purchased items during such period were girls' clothes. To corroborate the theory, Statista website conclude that, in Mexico, the month of October is usually the best-selling month in the fashion industry, followed by August, which is in line with increased density of points on the referred periods. (Statista 2019). However, the general increment in October is not reflected by Company 1, at least for regular sales;
- During weeks 42 and 43 (third and fourth week of November) a Mexican event called *Buen Fin* had place in both years 2017 and 2018, which provided 30% of discount in many items of some Company 1 brands during a four days' period. Although regular sales do not represent items in promotion, the increased clients' affluence inside the stores may have led to buying some other products too, in this case, characterized by a regular price;
- The last peak and most notable one happened between weeks 46 and 47 (middle of December) and they were probably a result of the Christmas gifts purchase.

Each stylecolour characteristic with a high purchased quantity was analysed and, in fact, the above events justify the most sold products, with exception of two: 1327 units of girls' swimwear and 1077 units of men swimwear at week 20 of 2017 (third week of June). These two rows were considered outliers, and, therefore deleted of the dataset, since they do not belong to a repeatable situation (is not repeated over the years), which does not mean that they have not actually occurred. Given the justifications for all the remaining abnormal values, the other observations of spikes were considered as an input of the predictive task on a parameter called Events.

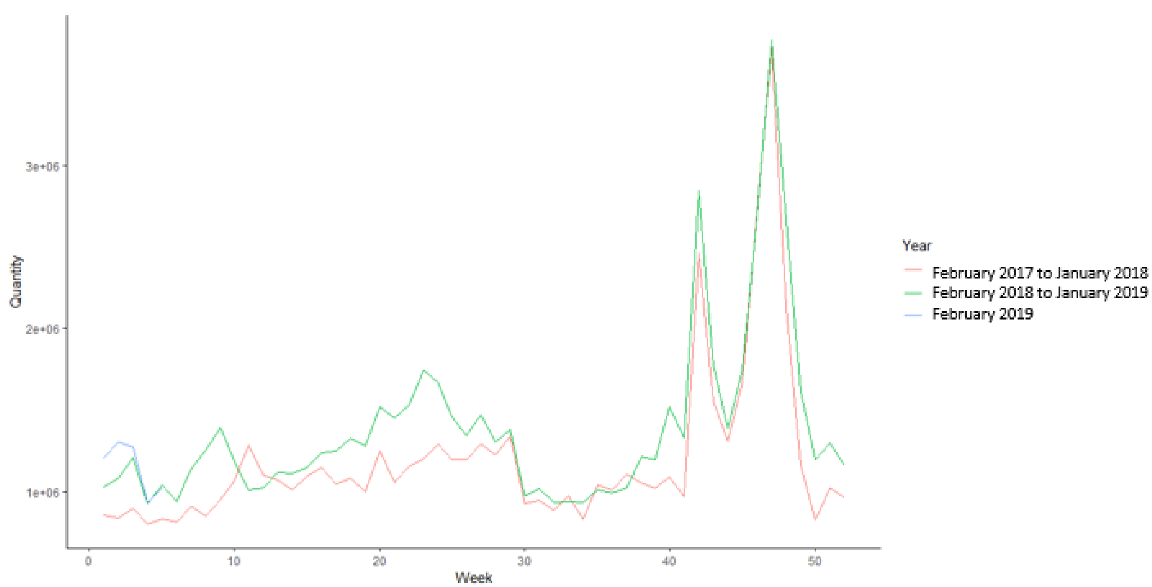


Figure 19 - Company 1 regular total quantity sales.

Figure 20 represents the total items sold on promotion between February 2017 and the start of 2019 and, as it can be observed, the purchased quantities are considerably less than the ones for retail sales. Since there are some weeks without promotions, it is possible to have zero point sales values, as is the case for 2019 year.

The major findings about the spikes in the graph were:

- The increased demand on weeks 14 and 15 (first and second of May) correspond to a special promotion for woman clothes, offered by some Company 1 brands within a small period of days, which in 2017 comprised days from two different weeks (May, 5th to 10th). The event was due to Mother's Day in Mexico, which happens on May, 10th. The main item sold in both years was women's jeans, probably because the most known and profitable brand (Brand X) made a 30% discount on it;
- For week 18 of 2018 (fourth week of May) Brand X announced a Hot Sale promotion with 10% discount on some new products for some days. Nonetheless, it is not noticeable on the graph;
- The sales increment on week 20 (third week of June) was due to Father's Day. Brand X made a 30% discount on men's jeans during four days in both years 2017 and 2018 for this event. In 2017, the Father's Day was on June, 18th and in the following year on June, 17th;
- On weeks 42 and 43 (third and fourth week of November) a Mexican event called *Buen Fin* took place and provided clients with a 30% discount on selected items, as was already mentioned;

Promotional sales were not part of the predictive task because Company 1 usually executes them when there is excessive stock of one or more items. Given that the aim of this project is to prevent situations of out-of-stocks and the inverse, supposedly there will be no need for the retailer to perform so many unexpected discounts on specific items. Moreover, bad predicted quantities can be sold at a lower price on this type of events or clearance sales. The influence of promotional demand is going to be addressed on the predictive task by a column comprising influent events.

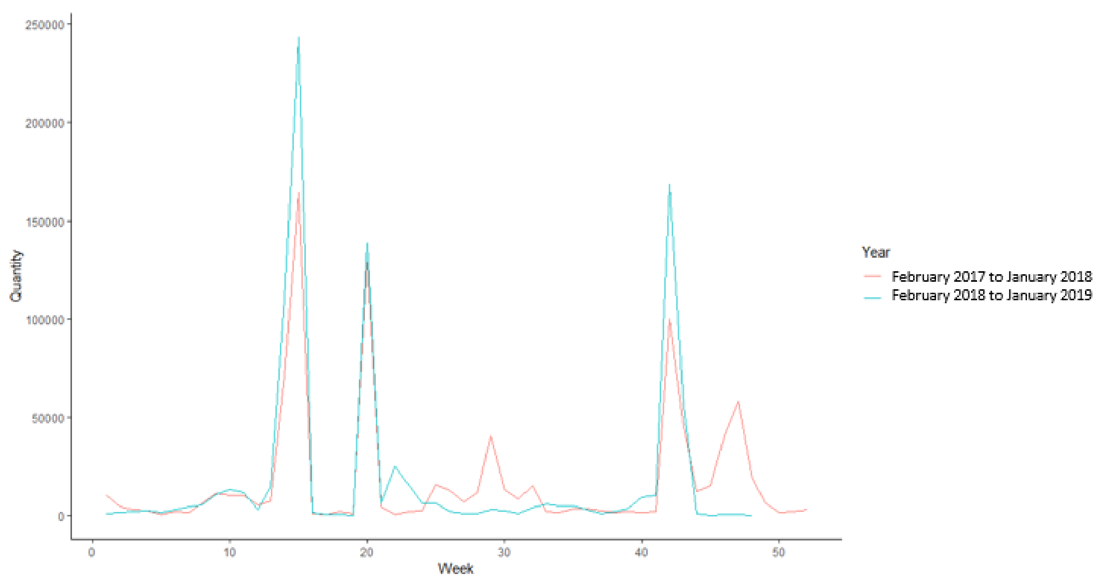


Figure 20 - Company 1 promotional total quantity sales.

Figure 21 refers to the total quantity of clearance sales per week in the period in study. Despite not being immediately noticeable, this dataset has also a pattern over the years. The purchased quantities in clearance are considerably less than the ones in regular sales, as can be noticed on the y-axis scales, but are higher than the promotional ones, probably because the clearance period has a longer duration in which every item of the season has an associated discount, while promotions only happen sporadically and are usually related to a specific range of products.

The periods with higher volume of sales correspond to the start of A/W season lower prices epoch. In Mexico, this period usually starts on the beginning of January, which correspond to weeks 48 to 52, and ends in March, while S/S clearance goes from July to the end of August (Dias Festivos Mexico 2016). The best-sellers products do not follow any pattern, in fact, the stylecolours combinations that were sold are completely random, but they all have a common low price which corroborates the explanation presented above.

Despite the well-defined duration for the clearance period presented, Company 1 sells discounted products during the 52 weeks of a year, probably because of the differences between the weather conditions among Mexico's seven regions, making it possible to have S/S products on discount in one region and A/W products in another (“Mexico’s Seven Climate Regions | Geo-Mexico, the Geography of Mexico” n.d.).

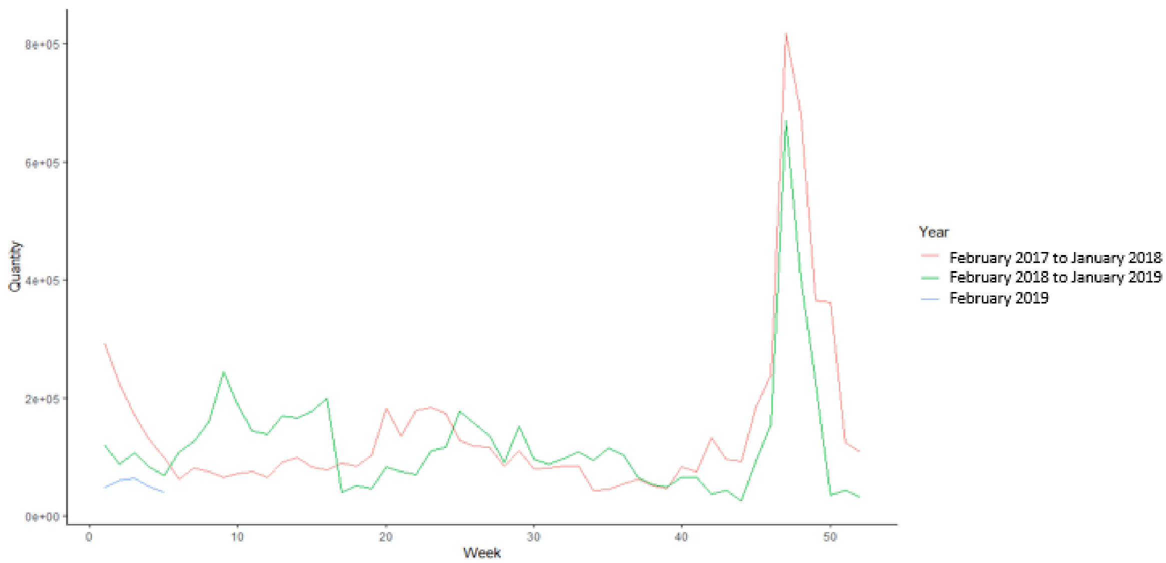


Figure 21 - Company 1 clearance total quantity sales.

Clearance sales were also not considered in the predictive task, given that Company 1 normally orders higher quantities than the ones it expects to sell, in order to have enough stock for these specific periods.

Data Reduction

Analysing promotional and clearance datasets allowed to understand which events bring more clients to the stores and allow the possibility to increase the purchase of regular sales too. Besides, it also gave a more in-depth understanding of the retailer organization, logistics and goals.

Brands usually ask for a prediction of a future season. For that reason, observations of the same period on the past were collected. The chosen season to perform the forecast was A/W, however, since there is lack of information about Company 1 classification of S/S and A/W, an assumption was made based on multiple findings: information collected from the website of Brand X; Mexican seasons calendar; the fact that the retailer calendar starts in February, normally month of S/S launch; differences between the stylecolours encoding, assuming that when Company 1 changes from one epoch to another, the product codes are modified too: A/W was considered to have 30 weeks duration, from the start of August until the end of February, which corresponds to weeks 27 to 52 in one year and 1 to 5 in the following one.

Given the limited time to perform the work, the regular dataset had to be aggregated in terms of region, brand and department. The combination was made after analysing the entire table for all types of products, and accounting that predictive tasks performances are transversal for every attribute referred. The chosen region and brand were due to the fact that both cover around 50% of Company 1 total sales. Besides, the choice of the first one also considered the respective weather conditions, being the most suitable to the chosen periods for A/W season. Department choice was random. The aggregation was the following:

- Region = South Centre;
- Brand = Brand X;
- Department = Woman;

The store variable, with 53 different levels, was excluded from the forecasting algorithm. Such detail level would lead to an increased complexity and possible loss of the model performance. Moreover, during the period in study many stores closed and many others opened.

The total quantity sales for each week within A/W period are displayed in Figure 22. The two main peaks happened during *Buen Fin* and Christmas month, respectively.

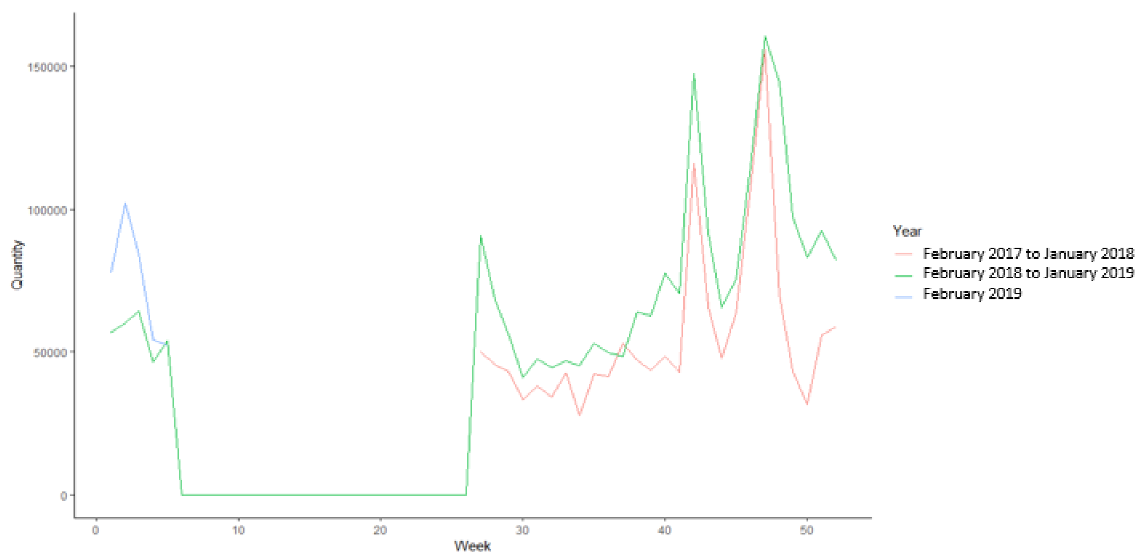


Figure 22 - Company 1 final dataset of total quantity sales.

The dataset introduced on the predictive models is summarized in Table 8.

The week attribute was divided into two, week and year, as it can be seen in Table 8. After creating the new columns, the variable week (Varchar(50)) became repetitive and was not considered in the final dataset.

The stylecolour attribute corresponds to the ID of a fashion item.

Colour is a characteristic of a product.

Subclass, class and family are characteristics of the product hierarchy.

Events attribute was created with information provided from Brand X website, such as special promotions and the beginning of clearance periods. Only events that happen within the A/W season period are presented in the table. The forecasting models do not accept NA values, so weeks without any event were labelled as “NoEvents”.

Lastly, given the arisen difficulties with price variables, a categorical variable price range was created. By looking at the lower and higher prices of similar items to the ones in study on the website of Brand X, it was possible to create five categories. Since the lower expense was 29 in currency and the higher one corresponded to 399 in currency, the calculation (399-29)/5 was made in order to define the ranges of each level.

Quantity is the output variable, the one which purpose is to predict in future seasons.

Table 8 – Company 1 final dataset.

Variable	**Type in mySQL**	**Key in mySQL**	**Type in Rstudio***	**Categories**
Week	Integer	Primary	Factor (31 levels)	[1;5] [27;52]
Year	Integer		Factor (3 levels)	[2017; 2019]
Stylecolour	Varchar(50)	Primary	Character	6527 items
Colour	Varchar(20)		Factor (29 levels)	Beige Blue ... Yellow
Subclass	Varchar(50)		Factor (3 levels)	Basic Basic *Moda* *Moda*
Class	Varchar(50)		Factor (27 levels)	Blouse Long Sleeve Blouse Short Sleeve Coat ... T-shirt Without Sleeve
Family	Varchar(50)		Factor (13 levels)	Blouses Coats, ... T-shirts
Events	Varchar(50)		Factor (4 levels)	Back to School *Buen Fin* Christmas NoEvents
Price range	Varchar(50)		Factor (5 levels)	Very Low Price: [29;103] Low Price:]103;177] Medium Price:]177, 251] High Price:]251, 325] Very High Price]325, 399]
Quantity sales	Integer		Numeric	Minimum: 1 Maximum: 11672 Mean: 51.55 Standard Deviation: 186.28

*Factor is a vector containing values from every possible type (character, numeric, and so on), and the levels are the number of different categories that a factor contains. Character is a character value type. Numeric is a numerical value type.

Data Transformation

A one hot encoding transformation was performed to convert the categorical variables colour, subclass, class, family, events and price range into dummy variables. This procedure remodels the columns by creating the same number of binary variables as there are respective attribute levels. For example, there are 29 different levels for Colour, so 29 new binary variables were constructed. For each column, number 1 refers to the presence of the attribute and number 0 refers to its absence (Potdar, S., and D. 2017).

Moreover, a z-score transformation was performed to the numerical output variable. The algorithms predictions belonged, consequently, to the z-score distribution, and were de-standardized using the same parameters as the entry, so as not to bias the final results.

4.1.2. Company 2

Company 2 data pre-processing structure is similar to the one synthesized above and is explained on Annex A.

4.2. Forecasting Techniques

AI models have the capability of learning the big data patterns by itself after a careful pre-processing task and, consequently, infer the output of future occurrences.

Supervised learning models make use of already labelled data to induce its value and compare it with the real outcome. In this way, the algorithm learns the right answer and tries to adjust to it in order provide the right answer on the following run. These algorithms perform both regression and classification models. The first has the capacity to calculate a numerical output and, therefore, it is the one that is going to be used in this project, and the second one, as the name indicates, refers to the classification of, for example, categorical outputs.

The methods used in AI require two datasets: the training set, from which the models can learn the patterns, and the testing set, which is used to predict the values of its observations. Contrary to the training set, the test dataset only provides the input variables to the algorithm which is in charge of supplying the corresponding output answer. The performance of the predictive task is evaluated by comparing the true values of the test dataset with the predictions of the AI methods.

Sometimes, given the nature of the data and the models' capacity to learn every detail, an overfitting situation occurs. This event happens when the model is too adjusted to the training data and will, therefore, have a bad performance while predicting the testing set. A viable way to check this occurrence is using K Fold Cross Validation, which is a type of data splitting and was the one applied in this project: the training dataset is divided into k sets, $\frac{train}{k}$, in order to construct a new training with the observations of k-1 sets and a validation set with the observations of the remaining set. After that, a similar approach to the one described above with training and testing datasets is performed, but this time with a training and a validation set. In order to do this, k iterations must be performed so that it is possible to crossover the total k sets and, at the end, each data point had a chance to be validated against the other. The performance is evaluated by calculating the error average of the k turns (Payam Refaeilzadeh, Lei Tang 2008). The procedure for k equal to 10 can be seen in Figure 23.

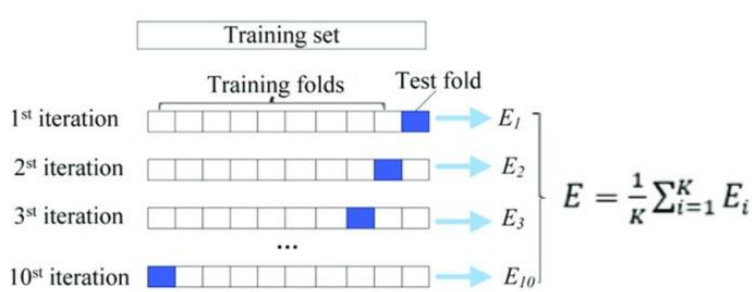


Figure 23 - 10 folds cross validation procedure in (Niu et al. 2018).

Given the available data and the importance given to the accuracy factor in a predictive method, the efficiency of NN regression model is going to be evaluated. Since speed and ease of use is also a critical factor to consider when choosing an algorithm of this kind, RF and SVM regression performances are also going to be analysed and compared between them and with the NN model. The same dataset will be used in the different models. Hybrid models are not going to be considered in this study due to time restrictions.

The procedures referred then were already explored on the above chapter 2. It is important to have in mind that the three were used as regression algorithms, since the predicted attribute was the sales quantity, a continuous numerical variable.

Different models are composed by many different parameters with specific functions. Regarding the type and number of considered entrant variables in the forecasting model, all the potential combination of parameters that can provide good predictors performance should be analysed by executing a parameter tuning. This is the most time consuming task. However, the better the tune, the greater is the probability of improving the quality of the output results on the test set (Feurer and Hutter, n.d.).

The structure of the code used to run the models was based on (Loureiro, Miguéis, and da Silva 2018) article and is summarized on the following Figure 24.

```
Create train dataset = train
Create test dataset = test
Create k folds = folds
Create evaluation metrics = errors
For (j equal to combination of possible parameters) {
        For (k equal to 1 until number of folds) {
                Train the model with k-1 folds
                Calculate predictions for 1 fold
                Save both true and predictive output
        }
        Calculate average errors for the combination of possible parameters
}
Parameters = Choose the best combination of possible parameters to use on the final model
For (i equal to 1 until number of iterations) {
        Train the model with the entire train set
        Calculate predictions for the entire test set
         Save both true and predictive output
}
Calculate average errors for all iterations
Best = Choose the iteration with lower evaluation metrics
Calculate errors for the each week of the best iteration
Calculate average errors for 18 weeks period of the best iteration
```

Figure 24 - Models code structure.

4.2.1. Parameter Tuning

Random Forest Tuning

The number of randomly sampled inputs at each node was the only parameter tuned for both companies, and its optimal value was found by varying it from 1 to 8. At the end, the error measures for the 8 iterations with the training data were compared and 5 was elected by the majority as the number of randomly sampled predictors that maximizes Company 1 RF performance, while 3 was the chosen number for Company 2. The number of trees was established as 2000 for Company 1 and 1000 for Company 2, as a result of the MSE metric stability with such value. Figures 25 and 26 graphically display the variation on the error with the increment of the trees in both cases.

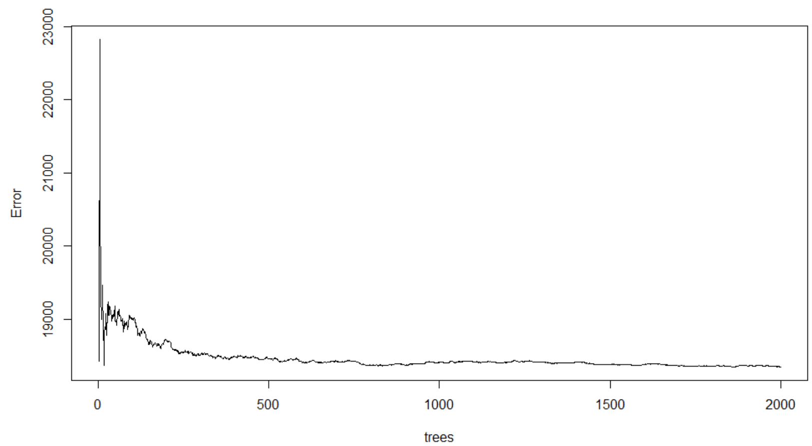


Figure 25 - MSE decrease with the increment of the trees number for Company 1.

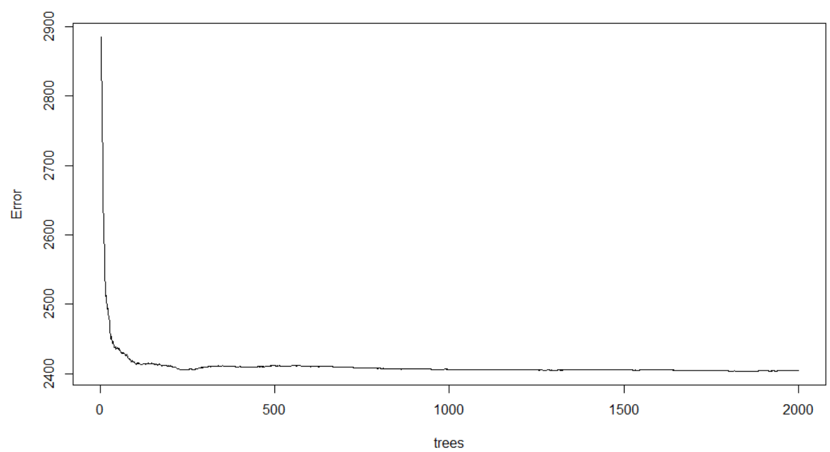


Figure 26 - MSE decrease with the increment of the trees number for Company 2.

Support Vector Machine Tuning

The parameters that were tuned for both SVM algorithms were the four kernel functions and their respective criteria:

- The linear function does not require any criteria tuning;
- For the polynomial function, the degree was tuned between 2 to 5;
- For the radial basis function, the gamma parameter was tuned between 0.5, 1 and 2;
- No tuned was performed for the sigmoid function.

All the parameters referred above perform a total of 9 combinations to tune. Among the 9 runs, the one that achieved the better forecast performance for Company 1 was the linear kernel function, while for Company 2 was the polynomial kernel function with a degree equal to 2. The penalty cost used was equal to 1 and the epsilon equal to 0.1.

Neural Networks Tuning

Activation function between the input layer and the hidden layer, learning rate and number of hidden neurons were the parameters tuned for the NN algorithms performed for the two companies:

- The tested activation functions were the Rectified linear unit (Relu), similar to a linear function but capable to capture the non-linearity within the data, and its output ranges from 0 to infinity, and Hyperbolic tangent function (Tanh), a function with a S-shape with outcome between -1 and 1;
- The learning rates tested were 0.001, 0.01 and 0.1;
- The number of neurons tested was from 100 to 700 with a step equal to 50.

The total number of combinations was equal to 84, and the chosen parameters were tanh activation function with a learning rate equal to 0.1 for both companies, and a number of neurons equal to 400 for Company 1 and equal to 300 for Company 2. There was only 1 hidden layer, and, given the numerical behaviour of the outcome, the activation between this one and the output layer was linear.

4.3. Results and Discussion

The four models were compared one by one for the two companies. The two firms were graphically analysed, and some conclusions were drawn given the proximity of the AI methods results. Besides, the best-seller products were also evaluated separately.

KSR, SA – Retail Consult asked for a comparison between an 18 weeks horizon and an entire season, in order to evaluate the stability of the results while the prediction period gets further away. For that reason, the evaluation metrics were given in two different ways: total error for an entire season and total error for the first 18 weeks of a season: from week 27 to 44 for Company 1 and from week 24 to week 41 for Company 2.

It was not possible to compare the three in study algorithms results with RDF V16 performance, given that it is a very labor intensive procedure and typically requires many months to configure, requiring an extra team working. The available environments were not constructed for Companies 1 and 2 and would not, therefore, be aligned with their characteristics.

The evaluation metrics used were MSE, nRMSE, MAPE and WMAPE. Error characteristics were analysed before choosing them, considering the literature findings in subchapter 2.6. The first one, MSE, was chosen because of its capacity to measure how spread the predicted value is when compared to the real one and, at the same time, how distant it is from the total average. Also, it was the most used evaluation metric on the papers discussed in chapter 2. nRMSE is useful when in the presence of cyclical events, which is the case, particularly in Company 1. MAPE was calculated because it is the second most used measure from chapter 2, however, as already mentioned, it does not deal well with very pronounced extremes. Lastly, WMAPE prevents MAPE issue with peaks by weighting extreme and occasional values versus smaller demand items. The first ones should not have the same importance to the overall dataset as the latter ones.

Given that every error measure has its pros and cons, the four were equally weighted (1/4 for each) in the choice of the best forecasting method for that specific retailer. Each evaluation metric was separately evaluated from the others among the three AI models.

4.3.1. Entry Data

The following Table 9 synthesizes the models entry data.

Table 9 - Companies 1 and 2 training set, testing set and number of predictors.

Company	Train Set	Test Set	Number of Predictors
1	A/W 2017 season: (33053 observations) Minimum: 1 Maximum: 9309 Mean: 52.44 Standard Deviation: 172.05	A/W 2018 season: (47009 observations) Minimum: 1 Maximum: 11672 Mean: 50.91 Standard Deviation: 195.67	Without the output variable, quantity sales, and the ID variable, stylecolour, there are 8 predictors.
2	A/W 2016 and 2017 seasons: (14741 observations) Minimum: 1 Maximum: 677 Mean: 38.81 Standard Deviation: 56.20	A/W 2018 season: (14064 observations) Minimum: 1 Maximum: 663 Mean: 33.10 Standard Deviation: 48.54	Without the output variable, quantity sales, and the ID variable, stylecolour, there are 8 predictors.

The training set should have more observations than the testing set, in order for the predictive algorithm learn the data pattern and infer the results of the test with a good performance. However, for management goals, given the required time horizon (usually one season), the sales prediction of a Company 1 entire season can only be performed with the available A/W 2017 season data, which has less values than the following one.

4.3.2. Company 1

In the following tables, the forecast evaluation metrics results of Company 1 are illustrated for the three performed algorithms in terms of Week/Product/Region level, with a forecast horizon of 31 weeks (an entire A/W season) in Table 10, and 18 weeks' duration in Table 11.

Table 10 - Company 1 evaluation metrics for Week/Product/Region level during an A/W season.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	**19056.75**	**1.20**	872.48	96.00
Support Vector Machine	37294.71	1.70	**576.46**	**91.82**
Neural Networks	23597.36	1.30	967.87	104.34

Table 11 - Company 1 evaluation metrics for Week/Product/Region level during 18 weeks.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	17522.76	**2.00**	954.99	98.81
Support Vector Machine	32453.54	2.70	**579.00**	**91.75**
Neural Networks	**9165.38**	2.30	645.89	95.24

MSE values decrease for the three cases while going from an entire season prediction to 18 weeks. This may be because from weeks 27 to 44, as can be seen in Figure 22, there are no high peaks and the overall quantity sales are closer to the mean value of the total test dataset, when compared to the following weeks. However, nRMSE values slightly increase for the shorter forecasting period, probably as a result of on the course of this 18 weeks, the main peak being the value 6774, approximately half of the bigger one on test set. MAPE and WMAPE do not suffer considerable variations for both periods.

The error measures per week are summarized in Annex B, from Tables 27 to 29, and, as it can be seen, the evaluation metrics do not represent very discrepant values between the beginning and end of a season, which proves the stability of the models. It is possible to observe that weeks with worse performance measures are the ones characterized by high peak values. In fact, such periods are characterized by multiple sales observations, encompassing from very few units to the maximum value of items sold within the test set, which, together with the great standard deviation associated to that specific weeks and increased mean value, could be the reason for the increment on error measures. The worse evaluation metrics were found during *Buen Fin*, the weeks before Christmas and beginning of A/W season period. However, and especially for the two first referred events, WMAPE do not suffer considerable variations, which means that the model is learning the pattern and predicting higher demand outputs.

SVM presents smaller values for MAPE and WMAPE, but higher for MSE and nRMSE in both cases, possibly because this algorithm predicts more accurately lower quantity sales, but does not handle so well the abnormal values within the test set.

The total quantity sales from each three models are represented in Figure 27, as well as the real output. It is observable that every model is capable to learn and represent the pattern over the weeks, being noticeable that NN tended to excessively predict some values, especially the Christmas peak, while SVM almost always modestly predicted the quantities.

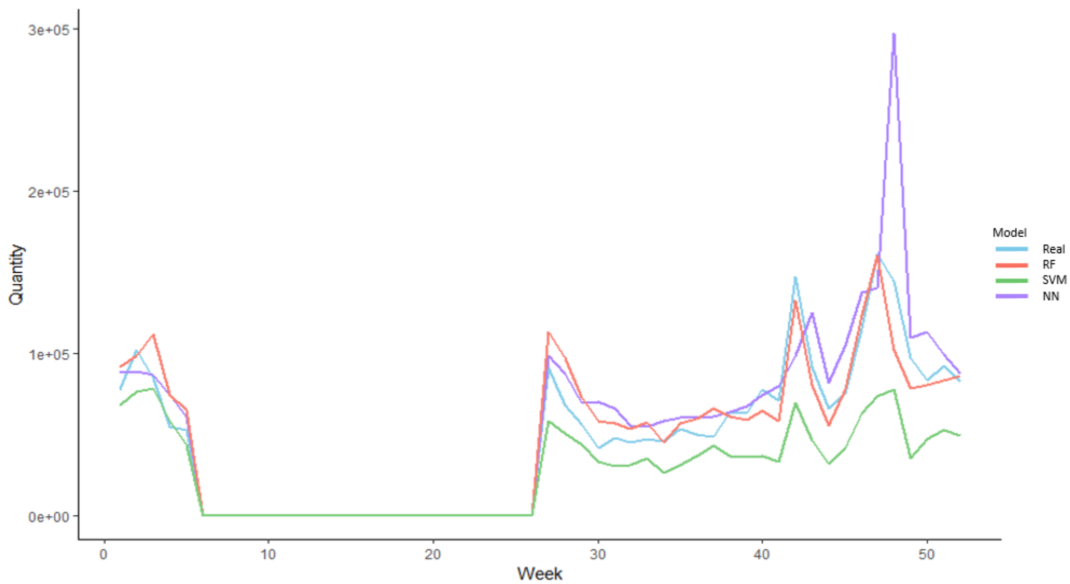
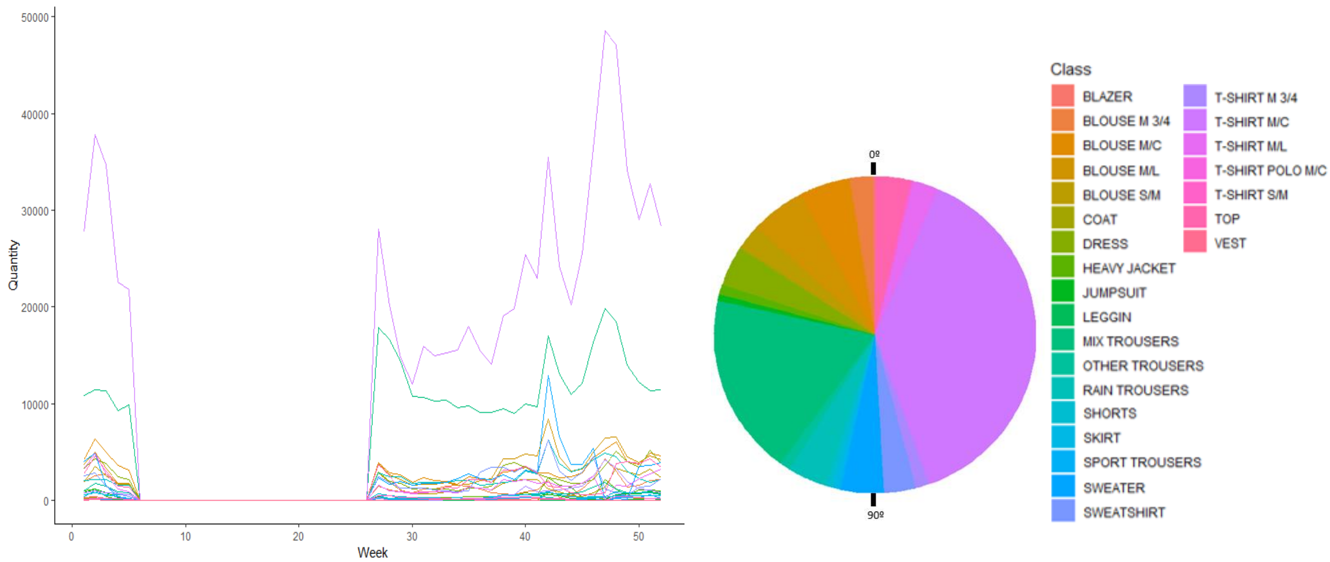


Figure 27 - Company 1 real versus predicted total quantity sales.

The differences between products aggregation by classes on the total purchased quantities are considerably higher than when disaggregated, as can be seen in Figure 28. Short Sleeve T-shirts was Company 1 best-seller product category. Being Mix Trousers an exception, the remaining 23 levels are very similar between them, and highly distant from the two referred above.



(a) (b)

Figure 28 - Company 1 test dataset total quantity sales per class and week (a) and per class (b).

Pareto principle states that 80% of the effects/results are usually obtained by 20% of the causes/activities (Ries et al. 2003). Placing the statement into fashion retail words, it is possible and probable that 20% of the sold items contribute to 80% of a brand success, and those 20% should, therefore, have an increased treatment care. Given that Short Sleeve T-shirts contributes with 32.97% to Company 1 sales, the evaluation metrics for this class were analysed and are addressed in Tables 12 and 13, for an entire season and an 18 weeks' duration (from week 27 to 44), respectively.

Table 12 - Company 1 evaluation metrics for Week/T-shirts Short Sleeve/Region level during an A/W season.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	**5900.04**	**5.30**	944.76	104.51
Support Vector Machine	5905.50	**5.30**	**600.42**	**90.54**
Neural Networks	6319.47	5.50	1020.81	110.17

Table 13 - Company 1 evaluation metrics for Week/T-shirts Short Sleeve/Region level during 18 weeks.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	**6127.27**	**6.60**	1043.00	108.87
Support Vector Machine	6180.43	**6.60**	**607.38**	**91.17**
Neural Networks	6148.78	**6.60**	1037.30	108.42

MSE and nRMSE values increased in the passage between an entire season period and 18 weeks' horizon. Since these two measures give more weight to large discrepancies between the real and predicted output than the other way around (squared measures), a possible explanation can be related to the fact of the short duration period having two high peaks, happening during what it appears to be the most stable sales period of Short Sleeve T-shirts, which could have led to an under prediction of the discrepant quantity sales. MAPE measure increased too, while WMAPE remains approximately constant for both periods.

Comparatively to the entire test set, MSE performance is better, probably due to the lower standard deviation of the class, 78.48, when compared to all together, 195.67, while nRMSE showed an increase that can be related to the maximum quantity sold, 1448, almost ten times smaller. MAPE and WMAPE presented slightly worse evaluation metrics, perhaps due to the increased randomness on the Short Sleeve T-shirts when compared to the overall sales.

The three models do not present so divergent performances as in the case presented above, except for MAPE evaluation metric, in which SVM takes the leadership once again.

4.3.3. Company 2

In the following tables, the evaluation metrics results of Company 2 are synthesized for the three performed algorithms in terms of Week/Product/Country level with a forecast horizon of more or less 42 weeks (an entire A/W season) in Table 14, and 18 weeks' duration in Table 15.

Table 14 - Company 2 evaluation metrics for Week/Product/Country level during an A/W season.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	2281.69	7.20	644.01	97.72
Support Vector Machine	2535.27	7.60	**335.88**	**84.98**
Neural Networks	**2224.41**	**7.10**	463.40	87.97

Table 15 - Company 2 evaluation metrics for Week/Product/Country level during 18 weeks.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	1713.05	12.40	633.50	103.15
Support Vector Machine	1664.09	12.30	**330.95**	**83.11**
Neural Networks	**1506.22**	**11.70**	430.81	87.14

In comparison to Company 1, MSE values are much lower, probably because of the differences between the maxima quantity sales in both test sets. For Company 2, the highest purchased quantity and mean were 663 and 48.54, respectively, while for the other retailer these values were equal to 11672 and 50.91. Considering this, it is possible and likely for the Spanish firm to present diminished mean squared metrics.

Once again, MSE results are better for the 18 weeks' period than for the entire A/W season, possibly as a result of the main peaks absence during weeks 24 to 41, which justifies the decrease on the predicted values variance. nRMSE presents a worse performance for Table 15, probably due to the presence of a peak value equal to 334, approximately half of the higher one on the entire test set. MAPE and WMAPE remain relatively stable for both forecasting periods.

The model appears to be stable both on the beginning and end of a season, and not only for the initial weeks. The evaluation metrics per week are summarized in Annex B, from Tables 30 to 32. The worse performances are related to high quantity sales periods with a great variety of units sold and the opposite. Company 2 launched the new season on week 19 for the first time (it usually happened on week 24), and the demand for A/W stylecolours was very low at start, which probably explains the raised MAPE and WMAPE metrics. The same happens for the end of season, very low units were purchased, and the models were not capable to capture it. During Black Friday and period before Christmas, possibly as a result of the great quantity sales variety, MSE values exhibit their worst performance.

SVM presents the smaller evaluation metrics for MAPE and WMAPE, but higher than NN for MSE and nRMSE for 18 weeks' duration, and worse values than NN and RF for an entire season prediction.

The maximum quantity sales from the three algorithms are presented in Figure 29, as well as the real output. It is observable that RF was the better method at learning the sales pattern over the weeks. Nonetheless, RF does not present any outstanding evaluation metric when compared to the other two models in both Tables 14 and 15, probably because it excessively predicted the demand in almost every week. Although not so accurate, NN and SVM were also capable of doing it. Once again, SVM tended to modestly predict the output.

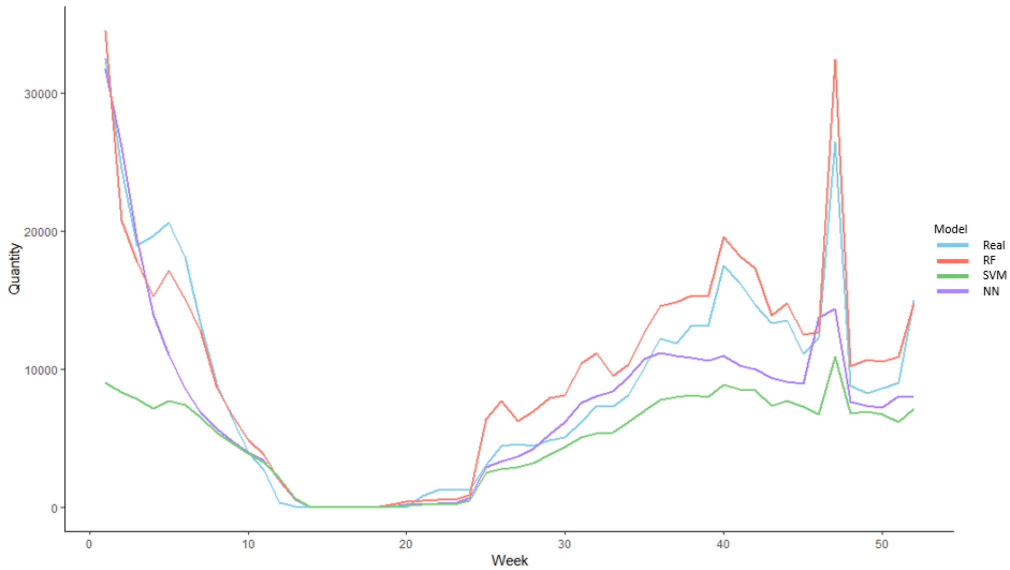
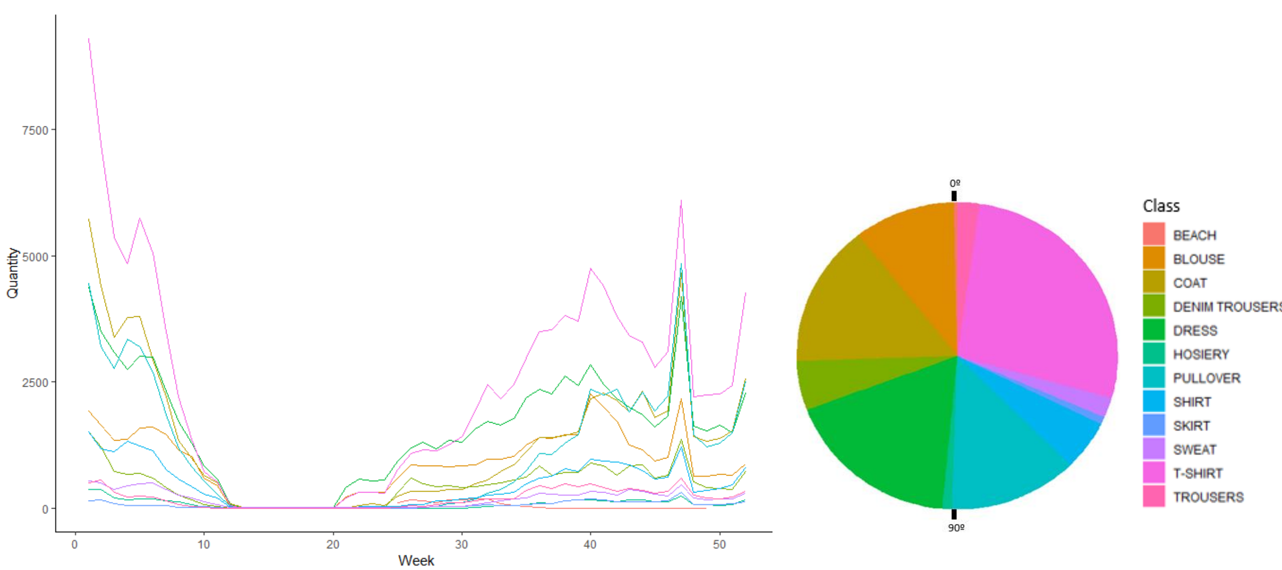


Figure 29 - Company 2 real versus predicted total quantity sales.

The stylecolours junction into classes promote a more significant amplitude difference between the quantity sales of each level. The maximum sold units during the test set period are graphically represented in Figure 30. Company 2 best-seller was T-shirts, contributing with 27.55% of purchases. Despite the pattern dissimilarities over the years, displayed in Annex A, Figure 35, almost all classes of products are represented by the peak sales in the test set.



(a) (b)

Figure 30 - Company 2 test dataset total quantity sales per class and week (a) and per class (b).

Company 2 launched beach products on the season in study, meaning that there were no historical quantity sales available for any of its stylecolours, but only for the similar input variables, such as week, events, price range, and so on. RF and NN could handle the introduction of a new level on the test set without any previous training about it, but SVM could not, forcing an output equal to zero for the category. In Company 2 specific case, and as it can be seen on Figure 30, the item was sold only during a limited period of time, starting on week 25. Moreover, few units were sold for the class, whereby it did not have a great influence on SVM performance.

Having into account Pareto principle, Company 2 T-shirts sales were analysed, and the evaluation metrics are synthesized in Tables 16 and 17, for an entire season and an 18 weeks' duration (from week 24 to 41), respectively.

Table 16 – Company 2 evaluation metrics for Week/T-Shirts/Country level during an A/W season.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	3757.93	9.30	674.87	93.94
Support Vector Machine	4351.42	10.00	**337.57**	**85.48**
Neural Networks	**3723.82**	**9.20**	538.49	88.82

Table 17 – Company 2 evaluation metrics for Week/T-Shirts/Country level during 18 weeks.

Model	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	2548.59	9.30	693.38	97.07
Support Vector Machine	2614.87	10.00	**342.38**	**83.24**
Neural Networks	**2431.18**	**9.20**	490.92	87.73

MSE metrics decrease on the passage from an entire season prediction to 18 weeks', certainly because the main sales peaks do not occur during the latter period. nRMSE, MAPE and WMAPE do not suffer high variations between the two tables.

Comparatively to the entire test set, MSE and nRMSE values suffered an increase, probably due to the increment on T-shirts variance, 62.61, when compared to the overall dataset, 48.45. Moreover, the maximum quantity sales presented above for the test set refers to a T-shirt stylecolour, meaning that the maxima coincide, 663, which, with a higher MSE, forces nRMSE measure to increase too. MAPE measure slightly increased and WMAPE remains approximately constant for both periods.

4.3.4. Best Forecasting Techniques

Table 18 summarizes an evaluation for Week/Product/Region and Week/Product/Country levels during an entire season, based on the factors that contribute most to the choice of a forecasting algorithm, analysed on the above sub-chapter 3.3.

Table 18 - Criteria that contributes most to the choice of the forecasting algorithm.

Criteria (by descending order)	Company 1	Company 2
Accuracy	Random Forest	Neural Networks
Stability	All	All
Ease of Use	Random Forest	Random Forest
Data Sufficiency	All	All
Speed	Support Vector Machine	Support Vector Machine

None of the algorithms had unanimous best evaluation metrics.

The model with better accuracy was chosen by weighting the four evaluation metrics equally between them. Although NN achieved better performance in terms of error measures per product, the overall demand pattern was better graphically represented by RF, as can be observable in Figure

29. Besides SVM presented two of the four best performances, the remaining two showed to have the worse values among the three algorithms.

In terms of user friendliness, RF is the most simple to understand and apply. Besides being the less complex method in study, it is the only one that accepts all types of variables as an input and, therefore, there it is no necessity of data transformation.

The available data was enough to run all the three methods with stable results over time.

Speed was, in fact, the most critical criteria in this project due to the limited time interval to conclude it, for as the number of observations increases, so the duration of the task suffers an increment too. However, more historical data provides a more in-depth learning of the algorithm and, consequently, there is a performance increase. The same happens with the number of predictors, nonetheless it is important to exclude inputs that do not add additional value to the predictive task, or its extent will increase while its performance gets worse. From a managerial point of view, the trade-off between accuracy and computationally time required must be evaluated.

The fastest AI algorithm between the three tested was SVM for both parameter tuning and the predictive task itself. Between RF and NN, the first one was faster in terms of parameter tuning, presumably due to difference on the number of tuned criteria: 1 for RF and 3 for NN, and the latter one was quicker while dealing with the test set. The computational time required by the tasks is summarized in Table 19.

Table 19 – Computational time required by the models tasks.

Model	Company	Task	Task Duration (approximately, in days)
Random Forest	1	Tuning	Four days
		Prediction	Two days
	2	Tuning	One and a half days
		Prediction	3/4 day
Support Vector Machine	1	Tuning	1 day
		Prediction	Half day
	2	Tuning	Half day
		Prediction	1/4 day
Neural Networks	1	Tuning	Five days
		Prediction	One and a half days
	2	Tuning	Three days
		Prediction	One day

5. Conclusions and Future Works

It is not possible to infer an overall better forecasting algorithm, nonetheless, it is observable that the three models are capable to learn the demand patterns over the weeks, and provide satisfactory and accurate total quantity sales results, when compared to the real ones. However, some conclusions may be taken by observing the results for both companies.

To start with, RF approach achieved better evaluation metrics for Company 1, while for Company 2 the prediction is more accurate in terms of error measures when executed by NN. Despite SVM high speed, it does not compete with the other two in terms of accuracy, and ease of use in the case of RF.

Data sufficiency is a critical factor to account, since the increase on the number of observations and valuable input variables would lead to more accurate results. However, the trade-off between time consumption of the models and achieved performance should be measured and weighted by consideration of the retailer's goals. Furthermore, overfitting situations must be prevented with the use of specialized techniques, so that the algorithm does not learn and fit so well to the training data that it is not good enough to predict future values.

The importance of the input variables, summarized in Figures 31 and 32, was assessed by the increase on the MSE error by RF algorithm when a given predictor is withdrawn. However, this measure should be transversal to the three techniques. The two figures are not in the same scale, since the final evaluation metrics also presents very different ranges between the two firms.

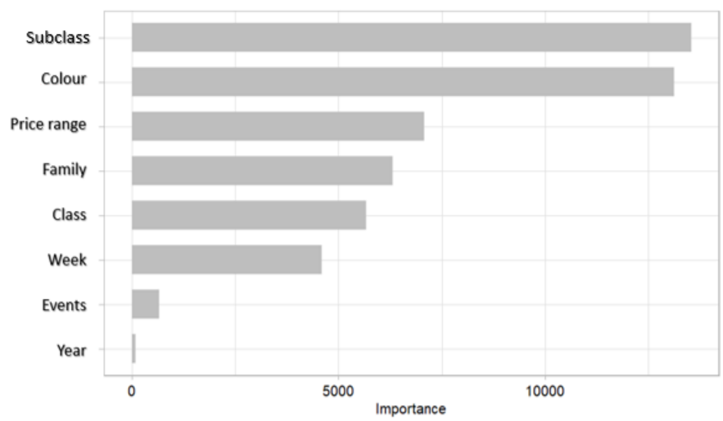


Figure 31 - Company 1 input variables importance.

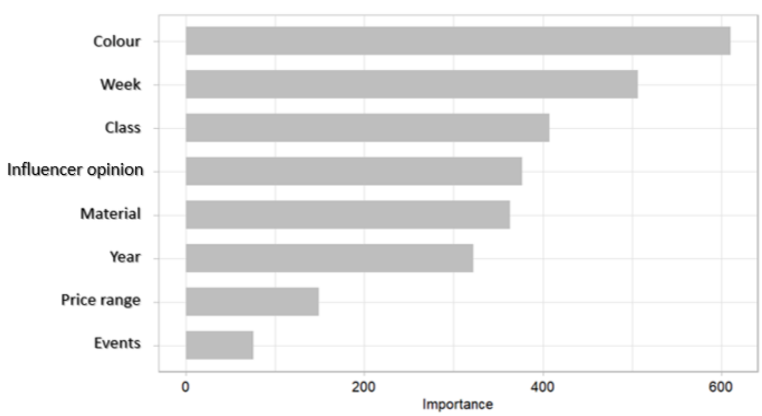


Figure 32 - Company 2 input variables importance.

For Company 1, the most important attributes are the subclass and colour of an item, respectively. Colour was rearranged at the beginning of the pre-processing task in order to be aggregated into less levels and, therefore, not increase the complexity of the prediction. Furthermore, highly detailed inputs do not always lead to a better algorithm performance. Price range is the third most important feature while predicting the retailer future demand, and it was an input previously aggregated into five categories, given the lack of data quality. An accurate collection of this input would probably lead to a better forecast performance, being a numerical entrance on the model. Events and year do not represent a great advantage to the error measure loss, probably as a result of the retailer similar behaviour during the two available A/W seasons. It is important to highlight that the events were collected from Company 1 website, once there was not any available information regarding the topic, neither in the databases or in the internet search. Once again, a rigorous events gathering has potential to positively influence the performance of the methods, since they can be inserted into any week of the year, any year, without having to follow a cyclical presence in every season or period.

The most decisive variable for Company 2 is colour, which was also modified for the purpose of the project. Influencer opinion turned out to have a notorious influence on demand prediction, as discussed in the literature, so perhaps retailers should start collecting human feedback before launching a season. Besides being the third less important feature, and contrarily to Company 1, in terms of error loss, the attribute year highly impacted the algorithm performance, which may be because of the different patterns found on some of the years' training. Price range is the second less important attribute on the prediction of the Spanish brand sales, perhaps due to the fact that the majority of observations present values from very low to medium price, which is possibly due to the gathering between regular, promotional and clearance sales. Having the correct prices within the databases would almost certainly represent an advantage and conduct to an increase of the attribute importance and, therefore, better accuracy. Again, events attribute does not have high impact on the evaluation metrics, presumably due to the lack of information regarding the input and, consequently, low number of different levels. However, it is believed that having as predictor the situations of more affluence within the stores would contribute to the task success.

There is no outstanding predictive method for every brand and possible combination of inputs. In order to choose an algorithm that best fits the test set, a meticulous parameter tuning should be executed, comparing many different criteria aggregation. After tuning the training and running the test dataset, the calculated evaluation metrics must be compared, together with a careful graphical analysis of the predictions, so that the best model can be elected for that specific case. Furthermore, a careful pre-processing task is indispensable to achieve a widely accurate method.

Season provisioning should be the first goal of a brand, from a managerial point of view. Performing a forecast in terms of product sold per week indicates a good proxy of what items and their respective quantities should be produced and stocked on warehouses for the incoming season. Likewise, models can be ran at the season beginning, in order to feed them with historical data regarding that period and, in case of under stock, reorder more quantity. Moreover, a future study could evaluate the monetary impact that a predictive model would have on a retailer expected costs in terms of sales losses, stock-outs and warehousing.

Not considering the week attribute was also a hypothesis tested in this project. Nonetheless, price ranges considerably vary during an entire season, especially for Company 1 since it discriminates its sales between regular, promotional and clearance, so the bias introduced into the model lead to a poor performance of the evaluation metrics. Moreover, many stylecolours are only available for sale during a small period of time, instead of an entire season.

As a future study, the influence of each store could be evaluated as a predictor at a later stage, nearer to the season launch. However, given the uncertainty related to this attribute (with opening and closures of stores), it should be accompanied by information regarding the space characteristics, which would probably provide a more precise forecast. The size of the items could also be addressed in future research.

It would be advantageous for KSR, SA – Retail Consult to perform the same analysis with the same datasets using RDF V16, in order to compare it with the obtained outcomes.

Due to time constraints, it was not possible to test any hybrid model, but since some stylecolours are transversal for more than one season, it would be an interesting approach to analyse.

References

Agrawal, R.K; Ratnadip Adhikari K., Ratnadip Adhikari, and R.K; Ratnadip Adhikari K. Agrawal. 2013. “An Introductory Study on Time Series Modeling and Forecasting Ratnadip Adhikari R. K. Agrawal.” *ArXiv Preprint ArXiv:1302.6613* 1302.6613: 1–68.

Alice, Michy. 2015. “Fitting a Neural Network in R; Neuralnet Package | R-Bloggers.” 2015. https://www.r-bloggers.com/fitting-a-neural-network-in-r-neuralnet-package/.

AnalyticsVidhya. 2017. “Creating & Visualizing Neural Network in R - Analytics Vidhya.” 2017. https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/.

“ARIMA Models.” n.d. Accessed March 28, 2019. http://users.dma.unipi.it/~flandoli/AUTCap4.pdf.

Artley, Melissa. 2018. “Oracle ® Retail Demand Forecasting User Guide for the RPAS Classic Client,” no. November.

Au, Kin Fan, Tsan Ming Choi, and Yong Yu. 2008. “Fashion Retail Forecasting by Evolutionary Neural Networks.” *International Journal of Production Economics* 114 (2): 615–30. https://doi.org/10.1016/j.ijpe.2007.06.013.

Brownlee, Jason. 2019. “How to Improve Neural Network Stability and Modeling Performance With Data Scaling.” 2019. https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/.

Carvalho, João Gama; Ana Carolina Lorena; Katti Faceli; Márcia Oliveira e André Ponce de Leon. 2017. *Extração de Conhecimento de Dados Data Mining*. Edited by Edições Sílabo. 09–2017th ed.

Chockalingam, Mark. 2018. “Forecast Accuracy and Inventory Strategies Forecast Accuracy - Abstract,” 1–42.

Choi, Tsan Ming, Chi Leung Hui, Na Liu, Sau Fun Ng, and Yong Yu. 2014. “Fast Fashion Sales Forecasting with Limited Data and Time.” *Decision Support Systems* 59 (1): 84–92. https://doi.org/10.1016/j.dss.2013.10.008.

Choi, Tsan Ming, Chi Leung Hui, and Yong Yu. 2014. *Intelligent Fashion Forecasting Systems: Models and Applications*. *Intelligent Fashion Forecasting Systems: Models and Applications*. https://doi.org/10.1007/978-3-642-39869-8.

Company, Oracle Retail. 2016. “Oracle Retail Merchandising.”

———. 2019. “Retail Software and Hardware Products | Oracle.” 2019. https://www.oracle.com/industries/retail/products.html.

Cornell University. 2006. “Cutting Edge:Life Cycle.” 2006. https://courses.cit.cornell.edu/cuttingedge/lifeCycle/03.htm.

Cutler, Adele, D Richard Cutler, and John R Stevens. 2012. “Ensemble Machine Learning,” no. February 2014. https://doi.org/10.1007/978-1-4419-9326-7.

DataSeries. 2018. “Back-Propagation Algorithm and Bias | Neural Networks.” 2018. https://medium.com/@neuralnets/back-propagation-algorithm-and-bias-neural-networks-fcf68b5153.

Dias Festivos Mexico. 2016. “Calendario Rebajas 2017 México : Encuentra Las Mejores Ofertas.” 2016. http://www.dias-festivos-mexico.com.mx/calendario-rebajas-2017/.

Du, Wei, Sunney Yung Sun Leung, and Chun Kit Kwong. 2015. “A Multiobjective Optimization-Based Neural Network Model for Short-Term Replenishment Forecasting in Fashion Industry.” *Neurocomputing* 151 (P1): 342–53. https://doi.org/10.1016/j.neucom.2014.09.030.

Feurer, Matthias, and Frank Hutter. n.d. “Hyperparameter Optimization,” 3–37.

Gardner, Everette S., and Ed. Mckenzie. 1985. “Forecasting Trends in Time Series.” *Management Science* 31 (10): 1237–46. https://doi.org/10.1287/mnsc.31.10.1237.

Goodman, Bernadette. 2016. “Oracle® Retail Assortment Planning User Guide.” *Oracle AP User Guide*, no. December.

Günther, Frauke, and Stefan Fritsch. 2019. “Neuralnet: Training of Neural Networks.” *The R Journal* 2 (1): 30. https://doi.org/10.32614/rj-2010-006.

Lahiri, S. K., and K. C. Ghanta. 2010. “Artificial Neural Network Model with Parameter Tuning Assisted by Genetic Algorithm Technique: Study of Critical Velocity of Slurry Flow in Pipeline.” *Asia-Pacific Journal of Chemical Engineering* 5 (5): 763–77. https://doi.org/10.1002/apj.403.

Lin, Chih-jen. 2012. “Large-Scale Linear Support Vector Regression.” *Jmlr* 13: 3323–48.

Liu, Bing. 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*.

Liu, Na, Shuyun Ren, Tsan-ming Choi, Chi-leung Hui, and Sau-fun Ng. 2013. “Sales Forecasting for Fashion Retailing Service Industry : A Review” 2013.

Loureiro, A. L.D., V. L. Miguéis, and Lucas F.M. da Silva. 2018. “Exploring the Use of Deep Neural Networks for Sales Forecasting in Fashion Retail.” *Decision Support Systems* 114 (August): 81–93. https://doi.org/10.1016/j.dss.2018.08.010.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. “Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward.” *PLoS ONE* 13 (3): 1–26. https://doi.org/10.1371/journal.pone.0194889.

“Mexico’s Seven Climate Regions | Geo-Mexico, the Geography of Mexico.” n.d. Accessed June 17, 2019. http://geo-mexico.com/?p=9512.

Neapolitan, Richard E. 2018. “Neural Networks and Deep Learning.” *Artificial Intelligence*, 389–411. https://doi.org/10.1201/b22400-15.

Nenni, Maria Elena, Luca Giustiniano, Luca Pirolo, and Guido Carli. 2013. “International Journal of Engineering Business Management Special Issue on Innovations in Fashion Industry Demand Forecasting in the Fashion Industry: A Review Regular Paper.” https://doi.org/10.5772/56840.

Niu, Mengting, Yanjuan Li, Chunyu Wang, and Ke Han. 2018. “RFAmyloid : A Web Server for Predicting Amyloid Proteins.” https://doi.org/10.3390/ijms19072071.

Okun, Oleg. 2011. “Bagging and Random Forests.” *Feature Selection and Ensemble Methods for Bioinformatics*, 296–313. https://doi.org/10.4018/978-1-60960-557-5.ch018.

Ostertagová, Eva, and Oskar Ostertag. 2011. “The Simple Exponential Smoothing Model.”

Modelling of Mechanical and Mechatronic Systems, no. SEPTEMBER 2011: 380–84. http://www.researchgate.net/publication/256088917_THE_SIMPLE_EXPONENTIAL_SMOOTHING_MODEL.

Payam Refaeilzadeh, Lei Tang, Huan Liu. 2008. “Cross Validation.”

Potdar, Kedar, Taher S., and Chinmay D. 2017. “A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers.” *International Journal of Computer Applications* 175 (4): 7–9. https://doi.org/10.5120/ijca2017915495.

R-bloggers. n.d. “Support Vector Machine Simplified Using R | R-Bloggers.” Accessed June 9, 2019. https://www.r-bloggers.com/support-vector-machine-simplified-using-r/.

Ren, Shuyun, Hau Ling Chan, and Pratibha Ram. 2017. “A Comparative Study on Fashion Demand Forecasting Models with Multiple Sources of Uncertainty.” *Annals of Operations Research* 257 (1–2): 335–55. https://doi.org/10.1007/s10479-016-2204-6.

Ren, Shuyun, Tsan Ming Choi, and Na Liu. 2015. “Fashion Sales Forecasting with a Panel Data-Based Particle-Filter Model.” *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45 (3): 411–21. https://doi.org/10.1109/TSMC.2014.2342194.

Retek Predictive Application Server. 2004. “Retek ® Predictive Application Server ™ 11.0.4 Solution Extension Configuration Guide.” https://docs.oracle.com/cd/B25966_01/RPAS/pdf/1104/rpas-1104-secg.pdf.

Ries, Al, Andrew Campbell, Ashridge Strategic, Management Centre, and Theodore Zeldin. 2003. *The 80/20 Principle: The Secret of Achieving More with Less*. *Long Range Planning*. Vol. 30. https://doi.org/10.1016/s0024-6301(97)80978-8.

Rob J Hyndman. 2014. “Fpp-Notes,” no. September.

Saraswat, Manish. n.d. “Practical Tutorial on Random Forest and Parameter Tuning in R Tutorials & Notes | Machine Learning | HackerEarth.” Accessed June 8, 2019. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/.

SAZLI, Murat Hüsnü. 2006. “A Brief Review of Feed-Forward Neural Networks.” *Communications, Faculty Of Science, University of Ankara*, no. May: 11–17. https://doi.org/10.1501/0003168.

Shcherbakov, Maxim Vladimirovich, Adriaan Brebels, Nataliya Lvovna Shcherbakova, Anton Pavlovich Tyukov, Timur Alexandrovich Janovsky, and Valeriy Anatol evich Kamaev. 2013. “A Survey of Forecast Error Measures.” *World Applied Sciences Journal* 24 (24): 171–76. https://doi.org/10.5829/idosi.wasj.2013.24.itmies.80032.

Souza, César. 2010. “Kernel Functions for Machine Learning Applications – César Souza.” 2010. http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/#kernel_trick.

Statista. 2019. “• Clothing Industry: Sales Value Mexico 2018 | Statistic.” 2019. https://www.statista.com/statistics/735930/sales-value-of-clothing-in-mexico/.

Sun, Zhan Li, Tsan Ming Choi, Kin Fan Au, and Yong Yu. 2008. “Sales Forecasting Using Extreme Learning Machine with Applications in Fashion Retailing.” *Decision Support Systems* 46 (1): 411–19. https://doi.org/10.1016/j.dss.2008.07.009.

Thomassey, Sébastien, and Antonio Fiordaliso. 2006. "A Hybrid Sales Forecasting System Based on Clustering and Decision Trees." *Decision Support Systems* 42 (1): 408–21. https://doi.org/10.1016/j.dss.2005.01.008.

Wong, W. K., and Z. X. Guo. 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): 614–24. https://doi.org/10.1016/j.ijpe.2010.07.008.

Xia, Min, and W. K. Wong. 2014. "A Seasonal Discrete Grey Forecasting Model for Fashion Retailing." *Knowledge-Based Systems* 57: 119–26. https://doi.org/10.1016/j.knosys.2013.12.014.

Yu, Yong, Tsan Ming Choi, and Chi Leung Hui. 2011. "An Intelligent Fast Sales Forecasting Model for Fashion Products." *Expert Systems with Applications* 38 (6): 7373–79. https://doi.org/10.1016/j.eswa.2010.12.089.

Zhang, Guoqiang, B Eddy Patuwo, and Michael Y Hu. 1998. "[10] -Zhang1998 - Forecasting with Artificial Neural Networks: The State of the Art" 14: 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7.

Annex A

Company 2 Data Pre-Processing

Company 2 is a Spanish fashion retailer with worldwide action. The retailer sells products for the categories woman, man, children, babies, beauty, living, sport and miscellaneous products. The observations of the last four departments were immediately deleted, since they represent a very small percentage of the retailer sales.

Raw Data Elimination, Cleaning and Integration

Company 2 sales datasets were provided, along with tables linked to its organization, calendar and products characteristics. The raw data contained information from January 2016 to April 2019.

The procedure of importing, analysing and removing data was similar to Company 1. The remaining information with possible valuable inputs for the predictive task is summarized within Tables 20 to 24.

Table 20 refers to Company 2 product hierarchy. Once again, the SKU variable was not considered. The stylecolour attribute represents the ID of a fashion product and Style only differs from it by not having the colour code. This variable is apparently randomly named and was not considered as an input of the forecasting task, since it comprises many different levels. Given that subclass only adds the season in which a given item is being sold, the variable was modified to a new one, named "Season". Family is a junction between collection and class and was, therefore, deleted. The difference between collection and department is that the last one refers to whom the product is directed for (woman, man, children or babies) and the first one specifies the type of product for sale (accessories, clothes or shoes). Company is always filled with the same value, so it was discarded. There were five missing values among the observations that were filled with a "NoInformation" code.

Table 20 **-** Company 2 product hierarchy.

Variable	**Type**	**Key**	**Example**
SKU	Varchar(50)	Primary	A/W Season Berta White size M
Stylecolour	Varchar(50)		A/W Season Berta White
Style	Varchar(50)		A/W Season Berta
Subclass	Varchar(50)		A/W Season Woman Trousers
Family	Varchar(50)		Woman Trousers
Class	Varchar(50)		Trousers
Collection	Varchar(50)		Clothes
Department	Varchar(50)		Woman
Company	Varchar(50)		Company 2

Table 21 synthesizes the sales of a given stylecolour during an entire week in a specific store. Week number 1 corresponds always to the first week of January, and the attribute is again encoded. The quantity sales and retail sales columns had negative values representing the returned items and their value in monetary units, respectively, and they were deleted as explained for the other retailer. Furthermore, incoherent observations regarding retail sales, quantity sales and zero values were not considered for prediction. There were no NA values among the columns.

Once again, the attribute price was added to the dataset by calculating $\frac{retail\ sales}{quantity\ sales}$, represented in currency. Although not so divergent as for Company 1, prices among the same stylecolour had some inconsistencies. Considering this, and in order to promote an easier comparison between the two companies in study, the calculated price was ignored.

Table 21 - Company 1 sales.

Variable	Type	Key	Explanation
Week	Varchar(50)	Primary	Week, product and store of a specific sale.
Stylecolour	Varchar(50)	Primary	
Store	Varchar(30)	Primary	
Retail sales	Decimal(5,2)		Retail means sales in currency.
Quantity sales	Integer		Quantity means sales in units.

There are 21877 different stylecolours presented on the Table 22. Despite the enormous amount of available attributes, only material, influencer opinion and colour had more than 50% filled values, and, therefore, these were the ones chosen as inputs for the forecasting task. There were some missing values among these three columns, but since they represented a tiny part of the total, a “NoInformation” label was filled in.

The column colour had more than 600 different values, which is not an acceptable number of levels to perform the final task. Once again, colours were joined in the same three logical ways of Company 1 and, beyond that, every unknown designation, for instance “Royal”, was searched in the website of the brand and attributed to a class, a pattern, or replaced by the “NoInformation” label when it was not possible to discover to which group it belonged. The total number of colours was reduced to 28.

Table 22 - Company 2 stylecolour attributes.

Variable	Type	Key	Explanation
Stylecolour	Varchar(50)	Primary	Product code.
Material	Varchar(20)		Material used in the stylecolour.
Shape	Varchar(20)		Shape of the stylecolour.
Channel exclusivity	Varchar(20)		Channels in which the stylecolour is available for sale.
Stamp	Varchar(20)		Stamped figures.
Colour	Varchar(20)		Stylecolour colour.
Characteristic	Varchar(50)		Stylecolour details.
Influencer opinion	Varchar(20)		Influencers’ opinion about the stylecolour.
Retail price	Decimal(40,5)		Stylecolour price.

The SKU sales presented in Table 23 were not part of the predictive task because of the high variance associated to the size of an item, but also to maintain some similarity between the predictions of both companies

Table 23 - Company 2 size sales.

Variable	Type	Key	Explanation
Day	Date	Primary	Day, SKU and store of a specific sale.
SKU	Varchar(50)	Primary	
Store	Varchar(30)	Primary	
Retail sales	Decimal(5,2)		
Quantity sales	Integer		

Although called store, there are many different sales channels included in Table 24. Company 2 acts within 3 different continents and 107 countries. To have a worldwide achievement, the retailer sells its products throughout many channels and type of stores, as follow:

- Type of store: Retail (Company 2 has control on the available quantities for sale).
    - Channels: Department Store, E-commerce, E-Tailer, Franchising, Marketplace, Retail Store, Living Store.

- Type of store: Wholesale (Company 2 has not control on the available quantities for sale).
    - Wholesale Channels: Beauty Store, Department Store, Wholesale, E-Tailer, Flash Sales, Living Store, Travel Store, Multibrand Store.

Wholesale stores were immediately deleted from the dataset, since there is no additional value in predicting quantities that are not managed by the retailer. Regarding Retail stores, the retail store channel was the only one under forecast analysis, given that it refers to Company 2 own physical spaces and is its best-seller. There were no missing values among the columns.

Table 24 - Company 2 store hierarchy.

Variable	**Type**	**Key**	**Example**
Store	Varchar(30)	Primary	Store X
Channel	Varchar(30)		Department Store
Type of store	Varchar(30)		Retail
Country	Varchar(30)		Portugal
Continent	Varchar(30)		Europe

Figure 33 represents the retail stores quantity sales per country. The first bar represents the origin country of Company 2, Spain, with 26.7% of retail sales. The remaining names are not displayed in order to maintain the confidentiality. It is immediately perceivable that the first four countries alone have a major contribution to sales, with 69% of total purchases. The best-seller country, Spain, was chosen for the predictive task, because mixing diverse events that occur among different territories in a unique variable would bias the models.

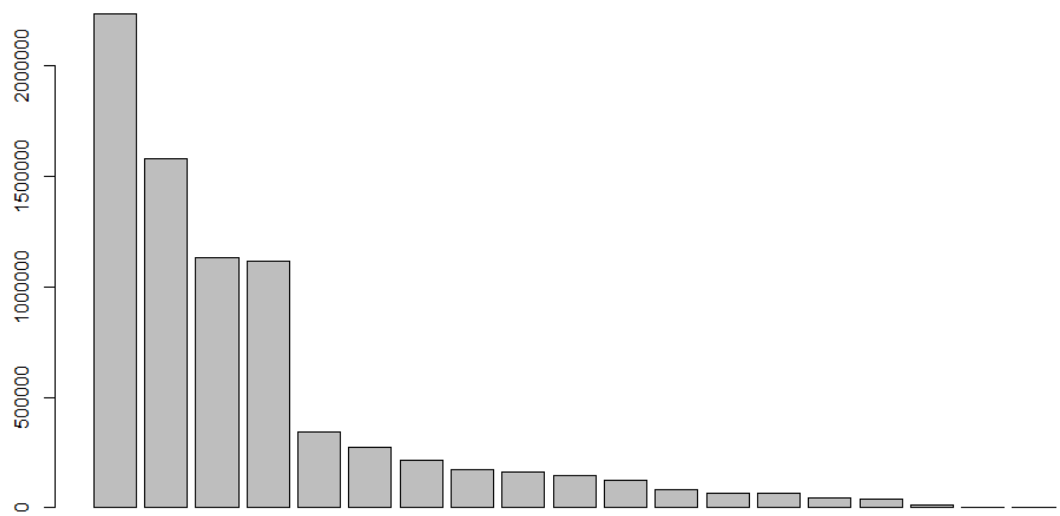


Figure 33 - Company 2 retail stores total quantity sales per countries.

Again, the following step was to understand which of the variables presented in the tables above was more relevant to the predictive model and associate them all in a single dataset. An in-depth analysis was performed to detect possible outliers. Figure 34 represents the Spanish retail stores total quantity sales per week, throughout 2016, 2017, 2018 and four months of 2019. Type of sales is not discriminated.

As it can be seen, Company 2 is also characterized by seasonality over the years, but there are great differences between some of them: 2016 differs a lot from the following periods; 2017 and 2018 have a not concordant beginning. However, it was still possible to find some patterns by also analysing the observations on the dataset.

The main peaks were attributed the following reasons:

- In Spain, Christmas gifts are exchanged only on the 6 of January, which corresponds to the period comprised by week number 1 of every represented year. Mainly in 2017, 2018 and 2019, this fact is in agreement with the respective increased amount of sales on the weeks previous to the event, 51 and 52, and on the week of the event, 1;
- Although not so discrepant when compared to other peaks, right after Christmas the sales were also high in 2018 and 2019, probably due to the beginning of the discount price period. In Spain, the A/W clearance starts on the 7 of January;
- From week 28 to 33 (July month and part of August) there is another high peak, possibly related to the beginning of S/S discounts period, which in Spain starts on the 1 of July;
- The spike on week 47 (last week of November) is assumed to be due to the Black Friday promotion, which happens every year on the same period and with a three-day duration.

Given the lack of information about this topic, it was not possible to analyse the impact of other promotions, such as Father's or Mother's Day. Also, and for the same reason, the great difference between 2016 quantity sales and the following years was not possible account.

Some abnormal values of woman accessories products were found between weeks 35 and 36 of 2016. These numbers, rounding the 200 units of sales for each product, were considered outliers since it was not found a reasonable explanation to their occurrence and it is not a repeatable situation over the years.

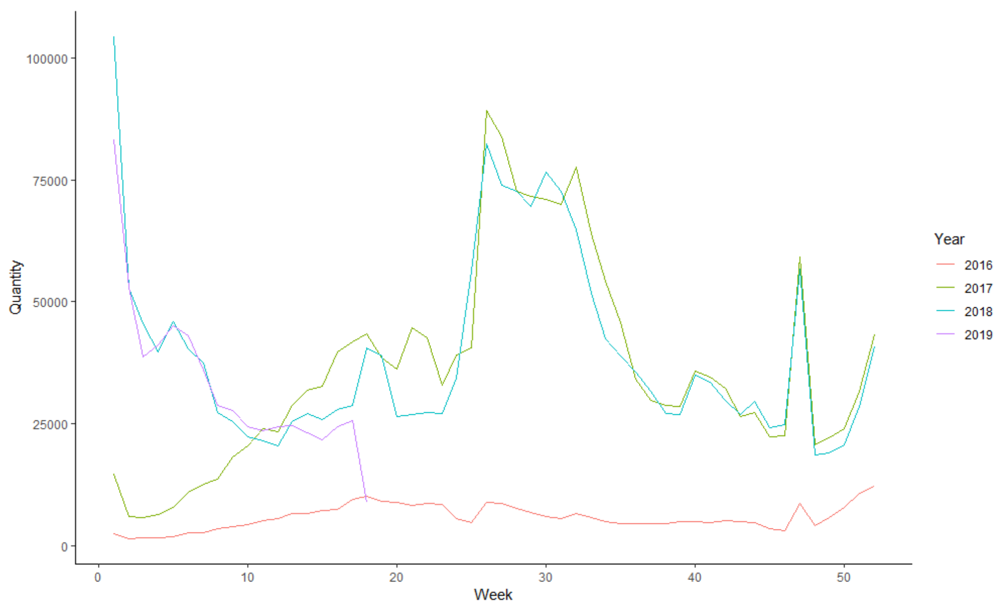


Figure 34 - Company 2 Spanish retail stores total quantity sales.

Data Reduction

Since there is available information about the corresponding season of the stylecolours, there was no need to make any assumption about the topic. A single table with A/W observations collected from Table 21 was constructed. The retailer usually starts A/W season on week 24 (middle of June) and ends it in the following year, on week 13.

Once again, data was aggregated in terms of department and collection. The chosen levels represented around 50% of the entire sales. The aggregation was the following:

- Department: Woman;
- Collection = Clothes.

The store variable, with 78 levels, was excluded from the forecasting. Many stores closed from 2017 to 2018 and few opened, which would probably bias the evaluation metrics.

The total quantity sales per week within A/W period is displayed in Figure 35. The two main peaks happened during clearance period and Black Friday, respectively.

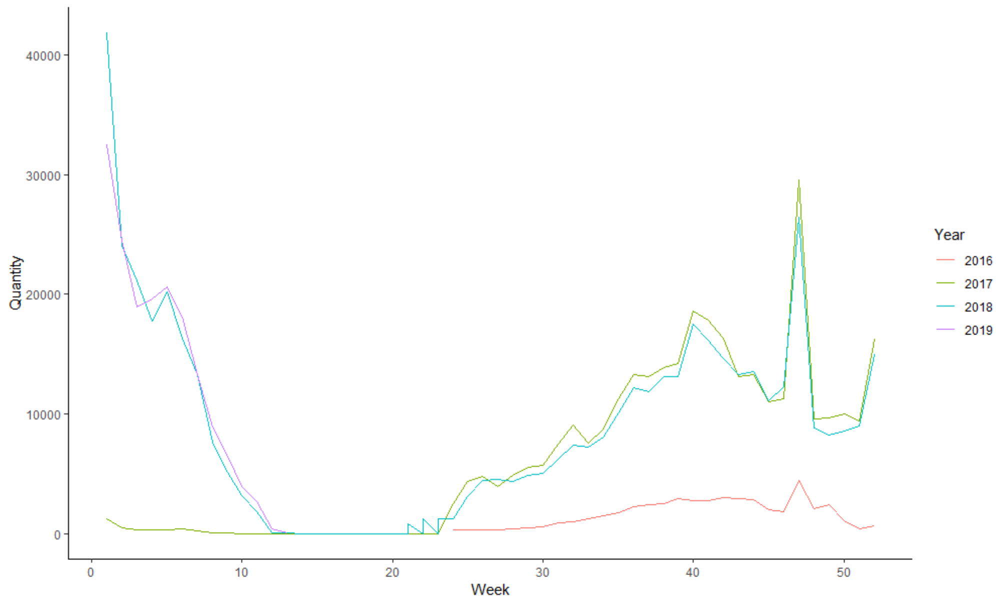


Figure 35 - Company 2 total quantity sales final dataset.

The final regular sales dataset used for the predictive task is presented in Table 25.

Week attribute was divided into two, week and year, and the variable week (Varchar(50)) was not considered. Despite A/W observations occurring usually from weeks 24 to 13 of the following year, a unit from A/W 2016 season was sold at week 15 of 2017. Additionally, on week 19 of 2017 one product of A/W 2017 season was purchased before the season launch. Lastly, A/W 2018 season was launched on week 19, instead of 24.

The stylecolour attribute corresponds to the ID of a fashion item.

Colour and material are characteristics of the product.

Class is a characteristic of the product hierarchy.

The attribute influencer opinion indicates the most frequent influencers' feedback on how the product is going to sell: if he/she thinks that it is going to be a best-seller, the denomination is "Very good", for example. Although this variable does not exist for Company 1, the decision of include it was based on the literature recent findings that underline the importance of human intervention and opinion when executing predictive tasks.

Since there is no reliable information about the products price, the variable price range was created based on the minimum and maximum amount charged for similar items to the ones in study on Company 2 website. The lower expense was 25 in currency and the higher one was 249 in currency. The calculation: $\frac{(249-25)}{5}$ was made in order to define the ranges of the five levels.

Given the fact that season A/W was the chosen one to perform the predictive task, only events that happen within this period are synthesized in Table 25. Weeks without any event occurring were labelled as "NoEvents".

Quantity is the output variable, the one which the aim is to predict in future seasons.

Table 25 – Company 2 final dataset.

Variable	Type in mySQL	Key in mySQL	Type in Rstudio	Categories
Week	Integer	Primary	Factor (48 levels)	[1;13] 15 [19;52]
Year	Integer		Factor (4 levels)	[2016; 2019]
Stylecolour	Varchar(50)	Primary	Character	1191 items
Colour	Varchar(20)		Factor (23 levels)	Beige Black ... Yellow
Class	Varchar(50)		Factor (12 levels)	Beach Blouse ... Trousers
Material	Varchar(50)		Factor (5 levels)	Denim Woven Knit Non Denim Woven
Influencer opinion	Varchar(50)		Factor (5 levels)	Bad Good Medium NoInformation Very good
Events	Varchar(50)		Factor (4 levels)	Black Friday Christmas A/W Clearance NoEvents
Price range	Varchar(50)		Factor (5 levels)	Very Low Price: [25;69.8] Low Price:]69.8;114.6] Medium Price:]114.6, 159.4] High Price:]159.4, 204.2] Very High Price]204.2, 249]
Quantity sales	Integer		Numeric	Minimum: 1 Maximum: 677 Mean: 35.00 Standard Deviation: 52.63

Data Transformation

A one hot encoding transformation was performed to convert the categorical variables colour, class, material, influencer opinion, events and price range into dummy variables.

A z-score transformation was executed to the numerical output variable, the quantity sales. The outcome of the models was de-standardized using the same parameters.

Annex B

Table 26 – Major findings about AI and hybrid forecasting techniques on the literature.

Techniques	Papers	Findings
ELME	Yu, Choi, and Hui (2011); Sun et al. (2008).	The increment on the number of iterations linearly increases the computational time and decreases the error measure until a point where the latter becomes constant; The increment on the number of hidden neurons increases the computational time and decreases the error measure until a point where the model overfits; The increment on the historical data increases the computational time at a slow path and decreases drastically the error measure; The increment on the sales variance does not have a significant impact on the computational time; ELME presents slightly better accuracies than NN, however, it is slower; NN performance is better than ELME when in the presence of high fluctuations on demand.
ENN	Au, Choi, and Yu (2008)	The accuracy increases with the increment on the computational time; The increment on the historical data tends to linearly approximate the relationship between the input and output neurons; ENN presents satisfactory results when dealing with few historical data; ENN achieves better results when the dataset is characterized by low variance.
DNN	Loureiro, Miguéis, and da Silva (2018)	DNN is very complex; DNN achieves slightly better accuracies than NN; DNN presents satisfactory results when dealing with few historical data; The paper highlights and proves the importance of human judgement on the predictive task.
GM	Ren, Chan, and Ram (2017)	GM presents satisfactory results when dealing with few historical data; GM is unreliable when dealing with high variance sales patterns.
ELME+GM	Choi et al. (2014)	ELME+GM presents satisfactory results when dealing with few historical data; The accuracy increases at a decreasing rate with the increment of the computational time; In the presence of white noise, ELME+GM, GM and NN+GM perform poorly; NN+GM achieves better accuracies than ELME+GM, since in the latter the performance is compromised by speed; ELME+GM achieves satisfactory results when dealing with high trend slopes and high seasonal variance.
ELME+HS	Wong and Guo (2010)	ELME+HS achieves better performances when the data is previously pre-processed than when it is not; ELME+HS achieves slightly better results than ENN, and considerably better results than ELME alone.
PDPF	Ren, Choi, and Liu (2015)	The increment on the historical data does not improve much the learning process of the model; The paper highlights that price fluctuations have a great influence on the sales pattern.
Clustering+DT	Thomassey and Fiordaliso (2006)	The paper highlights that the lack of significant predictors to explain the data patterns leads to an increased error measure.

Table 27 – Company 1 RF evaluation metrics per week for Week/Product/Region level during an A/W season.

Model	Year	Week	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	2018	27	15206.97	3.50	1019.86	110.86
		28	10775.75	4.00	1197.64	120.81
		29	12556.90	4.40	1342.41	114.15
		30	6166.23	4.60	1047.76	117.77
		31	8887.63	4.80	1082.36	107.50
		32	7497.10	5.20	893.70	103.63
		33	8982.74	5.30	867.57	102.22
		34	8696.01	5.50	621.17	94.10
		35	10114.28	5.20	871.05	98.08
		36	9267.58	5.80	785.10	102.81
		37	8857.99	5.80	845.10	110.76
		38	9629.29	5.80	779.52	92.19
		39	6873.36	5.20	675.71	86.90
		40	10383.19	4.50	587.77	80.30
		41	10315.66	5.00	557.66	78.90
		42	118075.56	5.10	1765.91	97.01
		43	44331.65	4.00	896.68	87.62
		44	16112.53	4.60	770.03	84.27
		45	17966.12	3.90	1133.03	93.75
		46	51286.09	4.00	1685.78	98.92
		47	206825.32	3.90	3864.88	94.83
		48	67615.09	3.20	689.65	80.81
		49	19724.04	3.40	560.28	83.57
		50	5330.85	3.10	439.93	83.95
		51	5505.87	3.00	480.77	81.82
		52	4110.60	3.40	538.89	91.61
	2019	1	4152.17	3.70	609.77	104.63
		2	4324.16	3.90	527.10	93.68
		3	4322.22	3.20	768.83	113.60
		4	3116.42	3.50	771.84	114.22
		5	4728.35	4.50	954.16	107.62

Table 28 – Company 1 SVM evaluation metrics per week for Week/Product/Region level during an A/W season.

Model	Year	Week	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Support Vector Machine	2018	27	24010.92	4.40	593.01	94.40
		28	18135.32	5.20	661.66	99.45
		29	22680.56	5.90	916.40	104.90
		30	11400.15	6.20	660.26	103.10
		31	15923.55	6.40	637.79	96.23
		32	13822.90	7.00	618.81	95.15
		33	16933.46	7.20	574.59	93.47
		34	15125.57	7.30	389.37	88.32
		35	18817.40	7.10	519.82	90.34
		36	15806.35	7.60	525.24	92.35
		37	15383.78	7.70	633.62	100.28
		38	15576.45	7.30	474.83	85.30
		39	12415.93	7.00	435.99	84.02
		40	19238.74	6.10	330.31	80.41
		41	18818.74	6.70	323.52	80.11
		42	228048.22	7.10	1006.68	96.15
		43	80874.28	5.40	540.85	88.05
		44	32150.26	6.40	464.79	86.60
		45	41403.92	5.90	703.95	92.15
		46	109068.93	5.80	1115.88	93.88
		47	471294.91	5.90	2266.40	99.84
		48	125350.51	4.30	663.04	88.63
		49	40217.33	4.80	288.62	85.47
		50	10454.69	4.40	264.96	82.02
		51	11241.49	4.30	322.03	82.58
		52	7219.70	4.50	304.23	83.52
	2019	1	6392.48	4.60	475.11	93.38
		2	6348.56	4.70	453.90	88.02
		3	7140.30	4.10	586.71	100.52
		4	5169.38	4.60	712.20	109.67
		5	8009.99	5.90	752.78	101.67

Table 29 – Company 1 NN evaluation metrics per week for Week/Product/Region level during an A/W season.

Model	Year	Week	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Neural Networks	2018	27	4272.63	8.20	875.53	104.72
		28	3813.24	7.90	1001.88	113.97
		29	4701.28	8.80	1209.01	114.37
		30	5371.43	10.10	1270.92	146.15
		31	7268.29	11.30	1301.37	128.94
		32	5641.19	10.40	919.84	113.05
		33	6874.33	11.30	799.00	107.11
		34	6845.15	11.40	850.96	115.00
		35	6882.39	11.10	948.63	108.72
		36	7506.69	12.00	826.13	110.06
		37	7022.44	11.70	783.79	113.88
		38	7600.72	12.00	795.72	94.83
		39	6726.39	11.50	765.26	95.37
		40	7563.46	11.40	654.10	87.82
		41	7219.47	11.30	748.35	95.62
		42	26214.38	19.70	1250.09	99.90
		43	25426.47	19.60	1395.39	130.29
		44	22312.96	19.10	1123.86	118.64
		45	31573.35	22.50	1490.11	138.42
		46	45310.69	26.20	1910.21	131.00
		47	55004.56	28.40	3529.29	117.93
		48	40401.86	24.40	2512.58	197.47
		49	10500.81	12.80	793.63	108.59
		50	7283.11	11.10	608.53	109.62
		51	6868.21	10.70	608.36	93.42
		52	7320.34	11.50	527.78	93.47
	2019	1	3094.77	7.60	600.79	101.53
		2	3494.17	8.10	491.05	92.94
		3	2701.50	6.90	559.12	99.63
		4	2457.12	7.00	729.10	114.26
		5	3410.14	8.40	773.04	101.26

Table 30 - Company 2 RF evaluation metrics per week for Week/Product/Region level during an A/W season.

Model	Year	Week	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Random Forest	2018	19	2022.00	2248.30	3020.00	2625.00
		20	1366.91	1232.40	1784.09	1414.81
		21	1717.00	33.70	86.16	56.33
		22	4317.67	41.60	82.89	63.03
		23	4327.60	49.80	48.70	59.02
		24	2769.70	31.50	939.55	89.33
		25	2006.40	27.50	940.84	147.46
		26	2733.07	18.50	937.63	132.62
		27	1996.56	13.40	650.05	106.92
		28	1646.96	15.70	612.52	116.83
		29	1479.37	15.00	689.84	121.61
		30	1203.86	16.40	702.62	122.57
		31	1518.09	14.50	723.70	127.98
		32	1590.93	15.20	606.76	115.87
		33	1109.86	17.20	634.04	104.71
		34	1081.38	17.00	522.87	102.77
		35	1291.49	18.20	639.73	102.02
		36	1623.04	20.30	625.07	96.78
		37	1587.66	16.20	574.29	100.00
		38	1701.39	18.40	527.11	92.97
		39	1673.88	17.50	542.65	94.46
		40	2527.85	16.40	602.76	92.18
		41	2458.84	20.00	700.21	94.30
		42	2156.66	14.60	623.58	98.85
		43	1837.77	14.80	591.80	94.25
		44	1846.10	15.90	597.68	95.95
		45	1239.27	14.70	592.07	98.66
		46	1186.74	13.90	543.82	91.57
		47	4933.83	19.10	1055.10	101.13
		48	790.11	12.40	546.59	95.59
		49	767.4165	15.50	611.15	102.82
		50	846.2875	12.50	613.76	99.12
		51	772.0758	19.20	617.74	99.26
		52	1758.5189	17.80	635.14	85.57
	2019	1	9985.90	16.60	1160.43	94.18
		2	5320.01	11.00	685.92	86.92
		3	3595.50	14.70	770.65	95.88
		4	5730.32	13.40	666.57	95.65
		5	4699.44	15.30	669.37	91.74
		6	3311.23	18.30	642.21	89.17
		7	2124.50	15.20	536.25	91.84
		8	1181.41	13.50	461.28	90.26
		9	816.85	12.50	441.66	95.17
		10	395.35	16.90	425.77	105.98
		11	234.60	22.90	432.54	115.10
		12	128.34	103.00	744.51	395.88
		13	157.57	627.60	1143.48	1061.54

Table 31 - Company 2 SVM evaluation metrics per week for Week/Product/Region level during an A/W season.

Model	Year	Week	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Support Vector Machine	2018	19	246.80	785.50	1176.67	975.00
		20	161.27	423.30	659.88	503.70
		21	3062.50	45.00	77.23	77.97
		22	6664.46	51.70	81.65	83.95
		23	6674.40	61.90	79.17	83.01
		24	3300.37	34.40	567.86	89.95
		25	1153.39	20.80	407.11	87.89
		26	2148.66	16.40	353.71	85.15
		27	2118.12	13.80	321.65	84.50
		28	1411.61	14.60	304.81	83.46
		29	1316.14	14.10	369.30	89.60
		30	1078.77	15.60	411.31	95.56
		31	1312.14	13.50	369.16	91.48
		32	1448.58	14.50	314.86	86.19
		33	1066.98	16.80	363.64	85.59
		34	1057.63	16.80	328.37	85.36
		35	1267.48	18.00	370.85	83.50
		36	1630.31	20.40	328.27	80.98
		37	1533.20	15.90	295.24	81.16
		38	1764.44	18.80	289.96	78.79
		39	1712.91	17.70	283.35	79.52
		40	2743.19	17.10	298.90	80.03
		41	2494.28	20.10	334.11	80.93
		42	2158.61	14.60	318.93	82.55
		43	1950.25	15.30	316.67	82.57
		44	1911.20	16.10	323.25	83.40
		45	1232.39	14.60	353.35	85.55
		46	1313.02	14.70	297.92	82.60
		47	5667.96	20.50	347.74	84.91
		48	786.30	12.40	363.13	85.49
		49	705.98	14.80	391.87	86.32
		50	820.93	12.30	377.43	86.34
		51	810.44	19.60	330.55	86.27
		52	2165.52	19.80	320.19	83.04
	2019	1	13394.77	19.30	320.42	87.63
		2	6413.86	12.10	274.64	84.37
		3	4088.06	15.60	323.99	85.45
		4	6427.13	14.10	301.82	87.67
		5	5450.65	16.40	306.32	85.53
		6	3912.59	19.90	332.04	84.56
		7	2382.57	16.10	275.53	82.63
		8	1280.18	14.00	281.53	81.09
		9	863.05	12.90	293.34	83.78
		10	388.56	16.70	334.08	95.82
		11	233.90	22.80	364.30	107.01
		12	180.06	122.00	884.67	464.16
		13	186.78	683.30	1246.31	1157.31

Table 32 - Company 2 NN evaluation metrics per week for Week/Product/Region level during an A/W season.

Model	Year	Week	MSE	nRMSE(%)	MAPE(%)	WMAPE(%)
Neural Networks	2018	19	382.20	977.50	1326.67	1137.50
		20	276.72	554.50	729.55	629.62
		21	2724.71	42.40	67.39	72.42
		22	6159.73	49.70	78.59	80.21
		23	6126.20	59.30	74.62	78.61
		24	2768.03	31.50	673.10	83.20
		25	958.75	19.00	396.32	85.81
		26	1890.72	15.40	384.33	83.34
		27	1846.79	12.90	410.78	85.27
		28	1217.27	13.50	346.40	87.33
		29	1170.98	13.30	474.37	96.28
		30	992.09	14.90	543.06	103.94
		31	1240.78	13.10	548.44	105.00
		32	1347.26	14.00	452.79	97.17
		33	1015.28	16.40	554.10	97.24
		34	1017.94	16.40	467.83	97.47
		35	1161.24	17.20	541.79	92.28
		36	1480.57	19.40	462.16	85.66
		37	1396.80	15.20	387.17	85.32
		38	1552.89	17.60	351.95	79.90
		39	1527.64	16.70	343.61	80.72
		40	2471.86	16.20	324.52	78.85
		41	2271.53	19.20	374.45	79.30
		42	1970.63	13.90	350.34	81.78
		43	1837.99	14.80	390.94	85.94
		44	1822.09	15.80	359.17	85.42
		45	1171.24	14.30	389.16	88.66
		46	1165.99	13.80	603.70	94.69
		47	5241.20	19.70	469.42	85.39
		48	763.70	12.20	413.04	87.34
		49	693.67	14.70	427.69	88.51
		50	797.68	12.10	412.86	87.86
		51	721.32	18.50	456.40	87.74
		52	2060.94	19.30	353.80	82.14
	2019	1	9801.10	16.50	1113.73	91.56
		2	4935.09	10.60	852.18	90.76
		3	3422.88	14.30	803.24	96.35
		4	5716.78	13.30	594.68	92.44
		5	5023.79	15.80	409.22	85.30
		6	3744.84	19.40	343.02	83.07
		7	2334.65	15.90	275.78	81.60
		8	1257.59	13.90	291.00	80.41
		9	854.55	12.80	296.81	84.66
		10	391.61	16.80	310.65	94.82
		11	230.29	22.60	348.98	106.60
		12	145.62	109.70	769.98	404.63
		13	117.30	541.50	990.57	923.07